How is your feedback perceived? An experimental study of anticipated and delayed conversational feedback

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

How is Your Feedback Perceived? An Experimental Study of Anticipated and Delayed Conversational Feedback

Auriane Boudin,[1,2,3] Stéphane Rauzy,[1,3] Roxane Bertrand,[1,3] Magalie Ochs,[2,3] and Philippe Blache[1,3]
[1)]*Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France*
[2)]*Aix Marseille Univ, CNRS, LIS, Marseille, France*
[3)]*Institute of Language Communication and the Brain, Aix-en-Provence, 13100, France*[a)]
auriane.boudin@univ-amu.fr, stephane.rauzy@univ-amu.fr, roxane.bertrand@univ-amu.fr, magalie.ochs@univ-amu.fr, philippe.blache@univ-amu.fr

Abstract: This article presents a novel experiment examining the impact of feedback timing on its perception. Dialog sequences, featuring a main speaker's utterance followed by a listener's feedback, were extracted from spontaneous conversations. The original feedback instances were manipulated to be produced earlier, up to 1.5 seconds in advance, or to be delayed, up to 2 seconds later. Participants evaluated the feedback acceptability and the engagement level of the listener. The findings reveal that 76% of the time feedback remains acceptable regardless of the delay. However, engagement decreases after a one-second delay, while no consistent effect is observed for feedback anticipation.

[Editor: —] https://doi.org/

Date: 4 September 2024

1. Introduction

During conversations, listeners produce vocal, visual and multimodal responses or reactions known as feedback, which serve as explicit markers of attention, interest and understanding (Allwood *et al.*, 1992; Bunt, 2012; Schegloff, 1982) and guide conversational flow (Bertrand, 2021; Gandolfi *et al.*, 2023).

Various cues within the main speaker's speech, such as intonation patterns, pauses, or eye gaze, may serve as triggers for listener's feedback. Building upon studies of transition-relevance places (Sacks *et al.*, 1974), which denote moments when it is relevant for a speaker to take a turn, (Heldner *et al.*, 2013; Howes and Eshghi, 2017) proposed investigating the potential space for feedback realization, termed *feedback relevance spaces* (or *backchannel relevance spaces*).

However, there is no strict temporal alignment among the main speaker's cues, the conditions for feedback production are gradually met, making the boundary of the potential feedback position blurred. While studies have shown that the gap between turn-taking variations is around 250 ms (Stivers *et al.*, 2009), there has been no investigation into the optimal timing of feedback. Our goal is to determine the temporal limits beyond which feedback is no longer acceptable.

Otherwise, the acceptability of the feedback is not the only factor to consider when evaluating the quality of listening. Feedback also serves as a means to demonstrate engagement (Dermouche and Pelachaud, 2019; Ishii *et al.*, 2013; Leite *et al.*, 2015; Sidner *et al.*, 2004). Engagement is characterized as the perceived connection between speakers (Sidner and Dzikovska, 2002). According to (Pellet-Rostaing *et al.*, 2023), engagement is defined as a *"state of attentional and emotional investment in contributing to the conversation by processing partner's multimodal behaviors and grounding new information"*.

We propose to evaluate, for the first time, the optimal window for feedback production, distinguishing between generic feedback (i.e., reactions that show understanding) and specific feedback (i.e., reactions that involve some form of evaluation or display a certain attitude towards the main speaker's discourse) (Bavelas *et al.*, 2000). Our study will explore the impact of timing on feedback acceptability, as well as on the perceived level of engagement of the listener. In this study, we extract from spontaneous conversations original sequences featuring a main speaker production and listener subsequent feedback. We manipulate these sequences by anticipating and delaying feedback. Participants were asked to assess the acceptability of feedback and the level of engagement of the listener. We found that participants generally consider original feedback as the most acceptable response for both generic and specific types. Notably, increased anticipation or delay in feedback leads to a decline in acceptability rates. Additionally, specific feedback demonstrates higher engagement compared to generic feedback, while delays in feedback negatively impact listener engagement. These findings contribute valuable insights into our understanding of feedback timing and its consequences on listeners' engagement.

[a)]Author to whom correspondence should be addressed.

2. Context and Hypotheses

Research has extensively explored the functions, types, and forms of feedback (Bavelas *et al.*, 2000; Schegloff, 1982; Stivers, 2008; Tolins and Fox Tree, 2014). Simultaneously, numerous studies have investigated cues in the main speaker's signals that precede feedback, also known as feedback inviting-features (Allwood *et al.*, 2007; Bertrand *et al.*, 2007; Brusco *et al.*, 2020; Ferre and Renaudier, 2017; Gravano and Hirschberg, 2011; Koiso *et al.*, 1998).

Feedback-inviting features encompass various modalities, including prosodic features (e.g., a rising intonation followed by a pause), mimo-gestural features (e.g., gaze, nodding), and morpho-syntactic features (e.g., determinant-adverb-noun trigram) (Brusco *et al.*, 2020; Gravano and Hirschberg, 2011; Poppe *et al.*, 2010). These features have been leveraged in computational models designed to predict vocal, visual and multimodal feedback in human-human and human-machine interactions (Cathcart *et al.*, 2003; de Kok *et al.*, 2010; Morency *et al.*, 2010; Mueller *et al.*, 2015; Ozkan and Morency, 2010, 2013; Ruede *et al.*, 2019; Truong *et al.*, 2010; Ward and Tsukahara, 2000).

These models typically predict, at small intervals (usually 40 ms or 50 ms), whether feedback should be produced based on preceding main speaker features extracted within a given window (e.g., 2 s). Evaluation of continuous feedback predictive models often involves comparing model predictions with observed feedback in corpora. One common approach is to assess whether the prediction falls within a brief window around the observed feedback onset, typically $\pm$ 500 ms, as proposed in the seminal work by Ward and Tsukahara (2000). This evaluation window (also called *margin of error*) has been reused and adapted in various studies (Poppe *et al.*, 2010; Ruede *et al.*, 2019; Truong *et al.*, 2010). For a comprehensive review, see (de Kok and Heylen, 2012).

However, it is important to note that, as far as we know, the validity of this 500 ms error window has never been experimentally confirmed, neither by Ward and Tsukahara (see quote on p. 1192 of Ward and Tsukahara (2000), "*The decision to tolerate misalignments of up to 500 milliseconds was based on informal judgments of 'how much earlier or later a back-channel could appear and still sound appropriate' in various contexts.*") nor by the subsequent studies mentioned above.

The problem raised is that the choice of evaluation window can significantly influence the assessment of model performance. A wider evaluation window may capture more predicted feedback instances, consequently inflating the number of correct predictions and, consequently, the overall performance score (e.g., F-score) (Boudin *et al.*, 2024).

Moreover, most of these models have focused on a limited set of feedback types (e.g., nods or vocalizations). In (Boudin *et al.*, 2024), we proposed a feedback predictive model of feedback position by considering two main types of feedback in order to be as comprehensive as possible.

Following (Bavelas *et al.*, 2000) and then (Bertrand and Espesser, 2017; Stivers, 2008; Tolins and Fox Tree, 2014) we distinguished between *generic* and *specific* feedback. **Generic** feedback expresses understanding. It plays a role in encouraging the main speaker to continue his/her speech. It is conveyed by different components such as nods, vocalizations "*mhm, yeah, ok*" and/or smile. In contrast, **specific** feedback is dealing with the semantic and pragmatic context of the main's speaker discourse, providing a form of assessment and displaying various attitudes (e.g. happiness, surprise, etc.). Different feedback components can be used such as eyebrow movements, laughter, lexicalization, etc.

Specific feedback is highly context-dependent, involving the evaluation of the semantic and pragmatic content of the main speaker, as opposed to generic feedback, which may simply demonstrate an update of the common ground or show understanding and can fit into a multitude of contexts (Tolins and Fox Tree, 2014).

In this study, we introduce an original behavioral experiment aimed at gaining a deeper understanding of the variability in feedback production timing. To achieve this goal, short sequences from the Cheese! (Priego-Valverde *et al.*, 2020) and PACO (Amoyal *et al.*, 2020) corpora have been extracted to create our material of utterance-feedback. Through video editing, the original feedback, both generic and specific, was artificially anticipated (up to 1 500 ms) or delayed (up to 2 000 ms) by steps of 500 ms. Participants evaluated the response produced by the feedback-producer.

We test four hypotheses. The first one is that feedback can be delayed or anticipated by more than 500 ms and remain acceptable. The second hypothesis is that the maximum acceptable delay for generic feedback is longer for generic feedback than for specific feedback. The third hypothesis is that the perceived engagement of the listener gradually decreases with delay until the feedback is ultimately rejected. For example, feedback with a delay of 1 000 ms may still be considered acceptable in the conversation, but the listener's perceived level of engagement decreases significantly. Delayed feedback can imply disinterest or distraction, giving the impression of reduced engagement from the listener. The fourth and final hypothesis posits that when feedback is anticipated, the listener will be perceived as equally engaged as with the original feedback. Indeed, we believe that feedback can be anticipated and produced with a short reaction time in relation to the feedback target without being misperceived thanks to predictive mechanisms (Gandolfi *et al.*, 2023; Pickering and Garrod, 2021), demonstrating a significant investment in interaction and a strong collaboration.

3. Method

3.1 Participants

One hundred and twenty-eight participants have been involved in the experiment (mean age = 24, sd = 4.6, min = 18, max = 49). One hundred eight participants identified themselves as a woman and 20 participants as a man. All participants

Main Speaker Speech	**Feedback**
Non moi j'avais fait un master de linguistique un master recherche *No I'd done a master's in linguistics a master in research*	Gen: "Ah d'accord" + Nod "Oh ok"
Ca me fait 20€ par mois *It saves me €20 a month*	Gen: "Ouais" + Nod *"Yeah"*
Hum à Paris y'a un truc qui s'appelle la Cité de la musique *Hum In Paris there's a thing called the Cité de la musique*	Gen: "Ouais" + Nod *"Yeah"*
J'attends de finir l'année pour partir à l'armée *I'm waiting until the end of the year to go to the army*	Spe: "Allez" ↗ + Eyebrows ↗ + Smile *"Really"*
Ca fait 6h par jour si tu t'inscris à tous les créneaux donc c'est pas mal quoi *That's 6 hours a day if you subscribe to all the slots* *so it's not bad at all*	Spe: "Ah ouais c'est cool hein" + Nod + Eyebrows ↗ *"Oh yeah it's cool huh"*
Ah c'est marrant parce que moi c'était l'inverse au début je voulais faire le CRPE *Ah it's funny because for me it was the opposite* *at first I wanted to do the CRPE*	Spe: "Ah bon" ↗ + Laughter + Eyebrows ↗ + Smile *"Oh really"*

Table 1. Examples of utterance-feedback sequences. The first three lines show examples of generic (Gen) feedback, while the next three lines display examples of specific (Spe) feedback.The ↗ symbol indicates a rising intonation or rising eyebrows.

reported being native speakers of French. All were recruited from different students Facebook groups in different regions of France (Strasbourg, Bordeaux, Lyon, Toulouse, Aix-en-Provence and Montpellier) and through the mailing lists of *Laboratoire Parole et Langage*. The experiment was conducted online via the *FindingFive* platform and participants received a compensation of 7€ on PayPal. One participant was excluded due to response times exceeding 30 minutes.

3.2 Material

For this experiment, conversation excerpts from the Cheese! (Priego-Valverde *et al.*, 2020) and PACO (Amoyal *et al.*, 2020) corpora were used to construct the stimuli. These corpora involved participants seated face-to-face in a soundproof room, engaging in free conversation for 15 minutes. Each participant was recorded by a front-facing camera. We used 10 dyads, selecting sequences consisting of an utterance from one interlocutor followed by feedback from the other. We used Sony Vegas Pro software to artificially anticipate or delay the feedback from its original production. We test eight temporal steps (separated by 500 ms steps): three feedback anticipation steps (-1 500 ms, -1 000 ms, -500 ms), four feedback delay steps (+500 ms, +1 000 ms, +1 500 ms, +2 000 ms) and the original time of production. We test the feedback delayed up to 2 000 ms seconds and the feedback anticipated up to 1 500 ms. We have chosen not to go beyond 1 500 ms of anticipation, as typically, beyond this threshold, the feedback is either produced simultaneously with or before the main speaker utterance.

In order to test both generic and specific feedback, we select 32 feedback per type. Our final set of stimuli is composed of 512 video clips (64 original sequences, each manipulated in every temporal condition) with an average duration of 5.66 sec (sd = 1.85, min = 1, max = 12). Among specific feedback, we exclusively retained the most prevalent type observed in our dataset: *positive-new* feedback, which responds to a positive stance expressed by the main speaker and pertains to newly introduced information. The selection of utterances consistently ensured syntactic saturation. To streamline our experimental design, we opted to avoid testing various combinations of verbal, gestural and multimodal feedback, which could introduce unnecessary complexity. It is anticipated that perceived engagement may vary between unimodal verbal, gestural and multimodal feedback. Furthermore, multimodal feedback is prevalent in our dataset, constituting 68.55% of feedback instances among the 26 annotated participants (Boudin *et al.*, 2024). Therefore, only multimodal feedback instances have been selected for both generic and specific types. Examples of utterance-feedback sequences are provided in Table 1.

To avoid speaker effect and dyad effect, we created 6 or 7 stimuli per dyad. We balanced speakers' roles (main speaker vs. listener) and the types of feedback (generic vs. specific) within each dyad. Each speaker provided both types of feedback and took on the main speaker role at least once. The main speaker always appears on the left of the screen and the listener on the right of the screen. In few cases, when feedback is anticipated or delayed, it is possible for non-feedback-related gestural or verbal components (e.g., the listener's previous turn-taking) to be visible in the video. Through video editing techniques, we ensure that these extraneous components are removed from the final stimuli. We accomplish this by replacing them with sequences where the listener remains still and silent (either duplicate a video frame multiple times or insert a sequence of the same duration without any gestures or speech).

3.3 Experimental Design

Eight experimental lists were elaborated so that a participant evaluated all sequences and all temporal conditions (-1 500 ms, -1 000 ms, -500 ms, 0 ms, 500 ms, 1 500 ms, 2 000 ms) but a participant could not see the same sequence twice in different temporal conditions. Each list comprises 64 stimuli, divided into two blocks of 32 each. One block consists of 16 generic and 16 specific stimuli, with each type presented twice across all temporal conditions. In summary, each participant evaluates a total of 64 items, including 32 distinct generic feedback and 32 distinct specific feedback instances. A participant evaluates each temporal condition 8 times, including 4 times for each type of feedback.

3.4 Procedure

Participants were first informed of their rights and signed a consent form. They were given a personal link and password to access the experiment on *FindingFive* (FindingFiveTeam, 2023) from their home computer. Each participant was informed that the purpose of the study is to better understand spontaneous conversation. They were instructed that they would be watching short video clips of conversation between two interlocutors, where the person on the left was speaking while the person on the right was listening. They were asked to focus on the person on the right of the screen and answer two questions for each video (illustrated in figure 1): 1/ *Does the reaction of the participant on the right of the screen seem strange to you?* - Yes: the reaction seems strange, inappropriate or unnatural ; No: The reaction seems normal and appropriate. 2/ *Does the participant on the right of the screen seem involved/interested by the conversation?* - 1: not at all involved/interested ; 2: not very involved/interested ; 3: somewhat involved/interested ; 4: interested/involved ; 5: very involved/interested.

They are asked to respond as quickly and accurately as possible. After reading the instructions, participants begin the experiment with a training block containing 11 trials not used in the blocks. The stimuli are separated by 1 second of white screen. The first question appears on the screen 300 ms after the video ends and the second question 300 ms after the participant answers the first question. The experiment is divided into two blocks each containing 32 trials. Blocks are separated by a maximum break of 2 minutes. The order of the blocks remains consistent, while the presentation of stimuli within each block is randomized. At the end of the blocks, we ask them to make comments on the experience if desired and to answer to two questions to find out if they perceived the editing of the videos. The first question is "*Did you feel that some of the videos were buggy?*" and "*Most of the videos you just saw were edited. Did you realize that?*". The average total duration of the experiment was 18.72 minutes (sd = 3.72, min = 12.81, max = 30.13).

Fig. 1. A Snapshot of a Trial.

3.5 Data Preprocessing

Trials duration and reaction time were automatically recorded by *FindingFive*. After manually reviewing all responses and reaction time, trials with abnormal duration were removed (greater than 110 000 ms). In a second step, all trials whose duration was more than 2.5 σ compared to the logarithmic mean reaction time were removed. Thus, 2% of the data were deleted. From the responses to the first question, participants who always respond in the same way (always "no" or "yes" answer) were removed. Seven participants were removed. In the same vein, participants showing a too small variability in their responses were discarded. In practice, participants with a standard deviation of responses too small (at the 2.5 σ level) compared to the mean standard deviation over the participants were excluded. The criterion concerns only one participant of the analysis. Finally, participants who noticed the video editing were removed, corresponding to 49 participants. A total of 57 participants was finally removed for the following results.

4. Results

4.1 Question 1: Feedback Acceptability

All the following analyses were performed with Rstudio (R version 4.2.2) (RStudio Team, 2020). Figure 2 shows the average proportion of "yes" responses to the question "*Does the reaction of the participant on the right of the screen seem strange to you?*" with generic feedback in yellow and specific feedback in blue. These average proportions and associated 1σ error bars are obtained from the distribution of the individual proportion of each participant for a given time delay and feedback type. The original feedback timing (0 ms) obtained a proportion of "weird" responses of 9.27% for specific feedback and 10.92% for generic feedback. The proportion of feedback rated as "weird" increases as the feedback is anticipated or delayed. However, even for the minimum and maximum timing, the proportion of feedback rated as "weird"

Fixed effects	Estimate	SE	Z-value	P-value	P-level
type at 0 ms	0.24610	0.29518	0.834	0.404429	
- 1 500 ms	-1.15416	0.25260	-4.569	4.90e-06	***
- 1 000 ms	-0.76916	0.25712	-2.991	0.002777	**
- 500 ms	0.03817	0.28364	0.135	0.892962	
+ 500 ms	-0.33270	0.26886	-1.237	0.215924	
+ 1 000 ms	-0.91186	0.25468	-3.580	0.000343	***
+ 1 500 ms	-0.84594	0.25265	-3.348	0.000813	***
+ 2 000 ms	-1.23980	0.24900	-4.979	6.39e-07	***
Interaction terms					
- 1 500:type	-0.04394	0.36805	-0.119	0.904966	
- 1 000:type	-0.18809	0.37420	-0.503	0.615207	
- 500:type	-0.82825	0.39691	-2.087	0.036913	*
+ 500:type	-0.27516	0.38952	-0.706	0.479940	
+ 1 000:type	0.09963	0.37583	0.265	0.790944	
+ 1 500:type	-0.46310	0.36645	-1.264	0.206326	
+ 2 000:type	0.06239	0.36607	0.170	0.864680	

Table 2. Estimate, Standard Error (SE), z-value and p-value obtain by the general linear mixed-effects model ran to test the impact of feedback timing and feedback type on the feedback acceptability rates. The significance level (p-level) are defined as follow: '***' indicates a p-value inferior to 0.001, '**' indicates a p-value inferior to 0.01 and '*' indicates a p-value inferior to 0.05.

never exceeds 30%. The general trends drawn in Figure 2 need to be confirmed by assessing the statistical significance of these findings. In order to assess the increase in proportion of "weird" responses as the timing condition moves apart from 0, we analysed the responses to the first question by applying a general linear mixed-effects model (*glmer* function from R *lme4* package (Bates *et al.*, 2015)) using the binomial family and the bobyqa optimizer. The original feedback timing was defined as the reference level and all other timing as the contrast levels. The variable *type* was treated as a categorical predictor in the model (using dummy coding with *generic* type as reference level). The type of feedback, and its interaction with timing was defined as fixed effects. The model also incorporates participants as random effects. Results of the model are presented in table 2.

The model revealed a significant effect of the feedback timing conditions beyond ± 500 ms on the perceived feedback acceptability. Additionally, we found an interaction effect between type and timing at -500 ms.

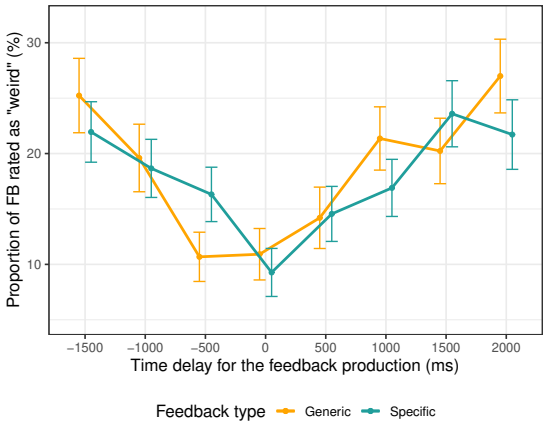

Fig. 2. Average Proportion of Feedback Evaluated as *"weird"* (Q1: Feedback Acceptability) for Generic and Specific Feedback.

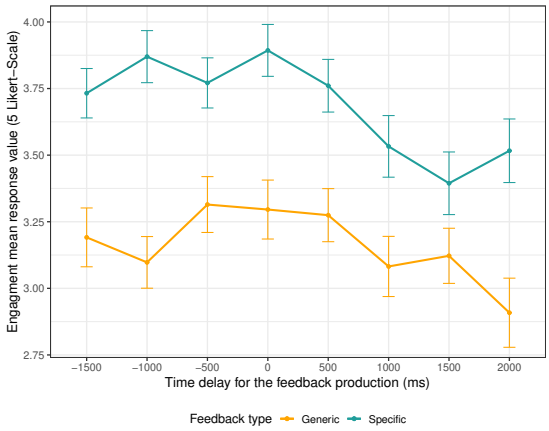

Fig. 3. Average Response Score for the Listener Engagement (Q2), Based on Timing Conditions for Generic and Specific Feedback. The Scale Ranges from 1 (Low Engagement) to 5 (High Engagement).

4.2 Question 2: Listener Engagement

Figure 3 presents the average response score to the question "*Does the participant on the right of the screen seem involved/interested by the conversation?*" depending on the timing condition and the type of feedback. The score varies from 1 corresponding to the response "not involved/interested at all" to 5 for the "very involved/interested" response.

Similar to the previous section, we ran a mixed model with feedback type and timing as predictor variables. Given that the response variable exhibits a Gaussian distribution over time, we opted for a linear mixed-effects model using the *lmer* R function. The results of the model are presented in Table 3. As a first result, the model reveals an effect of feedback

Fixed effects	Estimate	SE	t-value	P-value	P-level
type at 0 ms	0.59782	0.08805	6.789	1.28e-11	***
- 1 500 ms	-0.09810	0.08933	-1.098	0.27219	
- 1 000 ms	-0.20102	0.08797	-2.285	0.02235	*
- 500 ms	0.02010	0.08757	0.230	0.81849	
+ 500 ms	-0.02182	0.08757	-0.249	0.80328	
+ 1 000 ms	-0.21736	0.08806	-2.468	0.01361	*
+ 1 500 ms	-0.17995	0.08662	-2.077	0.03782	*
+ 2 000 ms	-0.35954	0.08813	-4.079	4.59e-05	***
Interaction terms					
- 1 500:type	-0.05266	0.12583	-0.419	0.67560	
- 1 000:type	0.17878	0.12436	1.438	0.15059	
- 500:type	-0.14228	0.12441	-1.144	0.25281	
+ 500:type	-0.10531	0.12424	-0.848	0.39668	
+ 1 000:type	0.15441	0.12469	-1.238	0.21563	
+ 1 500:type	-0.32113	0.12352	-2.600	0.00936	**
+ 2 000:type	-0.01447	0.12504	-0.116	0.90789	

Table 3. Estimate, Standard Error (SE), t-value and p-value obtain by the linear mixed-effects model ran to test the impact of feedback timing and feedback type on the perceived level of engagement. The significance level (p-level) are defined as follow: '***' indicates a p-value inferior to 0.001, '**' indicates a p-value inferior to 0.01 and '*' indicates a p-value inferior to 0.05.

type on perceived level of engagement. Concerning the relationship between engagement and timing, listeners' perceived engagement begins to be affected from 1 000 ms to 2 000 ms of delay. In terms of anticipated feedback, the perceived engagement of listeners is not significantly affected, except for an anticipation of -1 000 ms. Regarding interaction effects, a significant effect was found between feedback anticipated by 1 500 ms and type.

These findings suggest that both the timing and type of feedback production significantly influence the perceived engagement of the listeners.

5. Discussion and Conclusion

In this study, our objective was to investigate the optimum window, which has never been experimentally validated, for the occurrence of a conversational feedback. For this purpose, we designed an online behavioral experiment in which the time taken for a feedback to appear is manipulated. Participants were asked to evaluate the level of feedback acceptability (Q1) and the level of engagement of the listener (Q2). Participants were unaware of the timing manipulation nor the precise purpose of this experiment.

5.1 Question 1: Feedback Acceptability

Original generic feedback was judged acceptable 89.08% of the time, while original specific feedback was judged acceptable 90.73% of the time. The findings suggest that feedback timing between -500 ms and +500 ms is not perceived by participants. However, we found that the acceptability rate decreased significantly when feedback was anticipated or delayed by more than one second. For a maximum feedback anticipation of -1.5 s, generic feedback is judged acceptable 74.76% of the time and specific feedback 78.05% of the time. For a maximum feedback delay of +2 s, generic feedback is judged acceptable 73.01% of the time and specific feedback 78.29% of the time. Therefore, the unacceptability of these feedback production delays is not so clear-cut, as there is still a low rejection rate, even in the most extreme cases. These results tend to support our first hypothesis that feedback can be anticipated and delayed by more than 500 ms without becoming unacceptable. This also seems to validate the notion that feedback should be apprehended within a time window rather than at a specific point in time. However, it is essential to note that this temporal apprehension is not arbitrary, as this window of occurrence depends on necessary conditions (feedback inviting-features). Finally, the analysis of responses to the feedback acceptability question with respect to timing conditions does not reveal consistent differences between the two types of feedback. However, interaction effects were observed: the acceptability at -500 ms and perceived engagement at +1 500 ms varied significantly between the two types of feedback. Despite these findings, the present study does not allow us to validate our second hypothesis, which posited that the window of acceptability for feedback realization is larger for generic feedback than for specific feedback.

5.2 Question 2: Listener Engagement

The results of the second question about listener engagement show slightly different outcomes. Specifically, the model identifies a significant effect in the level of engagement between generic and specific feedback, with specific feedback eliciting higher engagement. The original generic feedback obtained an average engagement score of 3.30 but decreased to 2.91 with a delay of +2 s. In contrast, original specific feedback obtained an average score of 3.89, reaching its lowest point at a delay of +1.5 s with a score of 3.39 (with no significant difference observed compared to the 2-second delay). This finding is unsurprising given that specific feedback typically includes more salient components such as laughter, eyebrow

movement, and larger intonational span. However, it offers valuable insights into how listener engagement is expressed. Additionally, the perceived level of listeners' engagement significantly decreases as feedback is delayed by more than one second, supporting our third hypothesis. As a third observation, we noticed that listeners engagement is not significantly affected by anticipated feedback, except for the timing of -1.0 s. However, this is not a consistent effect, as the more extreme anticipation of -1.5 s does not show a significant effect. This finding provides support for our fourth hypothesis, which states that anticipated feedback does not impact perceived engagement, at least not consistently.

5.3 Concluding remarks

The first contribution of this work lies in the design of an online experiment with a third-party analysis, necessary because it is not possible to ask a person to anticipate or delay naturally his/her feedback production. This method is relevant to study the impact of different conversational behaviors that are not consciously manageable in spontaneous and natural conversations. Nevertheless, given the baseline error rate of 10% identified in Section 4.1, it might be worth considering using an equal number of original sequences and manipulated sequences for subsequent studies.

The experiment presented in this paper serves a dual purpose. Firstly, it seeks to deepen our understanding of how the timing of feedback delivery impacts its acceptability in conversation and listener level of engagement. Secondly, we aim to validate the window of evaluation (margin of error) used to assess the performance of feedback predictive models. In existing literature, it has been claimed that an acceptable delay for generating feedback typically falls within approximately 500 ms relative to the onset time of the original feedback produced by a listener (Ward and Tsukahara, 2000). However, various studies have employed different windows. For example, (Ruede *et al.*, 2019) used a window of 1 s after the feedback onset based on the assumption that anticipated feedback may not be acceptable, whereas delayed feedback is acceptable with up to a 1 000 ms delay. Mueller *et al.* (2015) used a window of $\pm$ 200 ms. Nevertheless, these windows are based on arbitrary choice. With the insights gained from these results, our goal is to propose an objective metric that provides a more nuanced evaluation of predictions, considering the temporal distance of the prediction from the feedback onset. This approach aims to refine the assessment beyond binary classifications of good or bad predictions.

Our results suggest that participants treat the issues of acceptability and engagement as distinct concepts in conversation. In essence, evaluating acceptability requires semantic interpretation and inference from context, whereas engagement is influenced directly by the type of feedback and its position. Participants indicated that it was easier to answer the second question. We plan to conduct two follow-up experiments. The first will reproduce this procedure by not manipulating the timing of the feedback, but only its type and content. The second experiment will evaluate the cumulative effect of timing on participant responses. In this experiment, instead of displaying only one feedback instance at a time, participants will be presented with longer sequences containing several feedback instances. This approach will provide more context to the participants.

Finally, it's important to note that this experiment exclusively focuses on manipulation of multimodal feedback. However, investigating the individual roles of visual and vocal modalities is essential for gaining a deeper understanding of their respective contributions to feedback perception. Furthermore, the exclusion of a significant number of participants who noticed the video manipulation could call for caution in interpreting our results. However, we would like to nuance this point by specifying that none of the participants reported perceiving any delays or anticipations in the feedback (which was our main variable of interest). Instead, some participants who responded 'yes' to the question related to video editing believed that the editing was done by compiling individuals who never interacted together, possibly resulting from the most anticipated or delayed feedback. A potential approach to mitigate this problem could be to test only audio feedback or to adopt a cumulative experimental design, as mentioned above.

Acknowledgments

This work, carried out within the Institute Convergence ILCB (ANR-16-CONV-0002) and Laboratoire Parole et Langage (UMR 7309), has benefited from support from the French government. Our warmest thanks go to Sophie Dufour for her help with design and analysis, and to Amandine Michelas for her valuable advice. AB warmly thanks Morgane Peirolo and Lydia Dorokhova for their help with the FindingFive platform. We would also like to thank the reviewers for their extremely valuable suggestions, which have helped to considerably improve the reporting of our results.

Author Declarations

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References and links

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (**2007**). "The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena," Language Resources and Evaluation **41**, 273–287, doi: 10.1007/s10579-007-9061-5.

Allwood, J., Nivre, J., and Ahlsén, E. (**1992**). "On the semantics and pragmatics of linguistic feedback," Journal of Semantics **9**(1), 1–26, doi: 10.1093/jos/9.1.1.

Amoyal, M., Priego-Valverde, B., and Rauzy, S. (**2020**). "Paco: a corpus to analyze the impact of common ground in spontaneous face-to-face interaction," in *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 628–633, https://aclanthology.org/2020.lrec-1.79.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models usinglme4," J. Stat. Softw. **67**(1), http://dx.doi.org/10.18637/jss.v067.i01, doi: 10.18637/jss.v067.i01.

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

Bertrand, R. (**2021**). "Linguistique interactionnelle : Du corpus à l'expérimentation," Mémoire d'habilitation à diriger des recherches, Aix-Marseille Université.

Bertrand, R., and Espesser, R. (**2017**). "Co-narration in french conversation storytelling: A quantitative insight," Journal of Pragmatics **111**, 33–53, doi: https://doi.org/10.1016/j.pragma.2017.02.001.

Bertrand, R., Ferré, G., Blache, P., Espesser, R., and Rauzy, S. (**2007**). "Backchannels revisited from a multimodal perspective," in *Auditory-visual Speech Processing*, Hilvarenbeek, p. paper P09.

Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., and Blache, P. (**2024**). "A multimodal model for predicting feedback position and type during conversation," Speech Communication 103066, doi: https://doi.org/10.1016/j.specom.2024.103066.

Brusco, P., Vidal, J., Štefan Beňuš, and Gravano, A. (**2020**). "A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool," Speech Communication **125**, 24–40, doi: https://doi.org/10.1016/j.specom.2020.09.004.

Bunt, H. (**2012**). "The semantics of feedback," in *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012), Paris, France*, edited by S. Brown-Schmidt, J. Ginzburg, and S. Larsson, University Paris-Diderot, Paris Sorbonne-Cite, pp. 118–127.

Cathcart, N., Carletta, J., and Klein, E. (**2003**). "A shallow model of backchannel continuers in spoken dialogue," in *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, Association for Computational Linguistics, pp. 51–58, doi: 10.3115/1067807.1067816, 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) 2003 ; Conference date: 12-04-2003 Through 17-04-2003.

de Kok, I., and Heylen, D. (**2012**). "A survey on evaluation metrics for backchannel prediction models," in *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, University of Texas, pp. 15–18, interdisciplinary Workshop on Feedback Behaviors in Dialog, Stevenson, Washington, USA ; Conference date: 07-09-2012.

de Kok, I., Ozkan, D., Heylen, D., and 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, ICMI-MLMI '10*, ICMI-MLMI '10, Association for Computing Machinery, New York, NY, USA, doi: 10.1145/1891903.1891908.

Dermouche, S., and Pelachaud, C. (**2019**). "Engagement modeling in dyadic interaction," in *2019 International Conference on Multimodal Interaction*, ICMI '19, Association for Computing Machinery, New York, NY, USA, p. 440–445, doi: 10.1145/3340555.3353765.

Ferre, G., and Renaudier, S. (**2017**). "Unimodal and bimodal backchannels in conversational english," in *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, SEMDIAL, Saarbrücken, Germany, http://semdial.org/anthology/Z17-Ferre_semdial_0006.pdf.

FindingFiveTeam (**2023**). *FindingFive: An online platform for creating, running, and managing your experiments*, https://www.findingfive.com/.

Gandolfi, G., Pickering, M. J., and Garrod, S. (**2023**). "Mechanisms of alignment: shared control, social cognition and metacognition," Philosophical Transactions of the Royal Society B: Biological Sciences **378**(1870), 20210362, doi: 10.1098/rstb.2021.0362.

Gravano, A., and Hirschberg, J. (**2011**). "Turn-taking cues in task-oriented dialogue," Comput. Speech Lang. **25**(3), 601–634, doi: 10.1016/j.csl.2010.10.003.

Heldner, M., Hjalmarsson, A., and Edlund, J. (**2013**). "Backchannel relevance spaces," in *Nordic Prosody : Proceedings of XIth Conference, Tartu 2012*, pp. 137–146.

Howes, C., and Eshghi, A. (**2017**). "Feedback relevance spaces: The organisation of increments in conversation," in *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*, edited by C. Gardent and C. Retoré, https://aclanthology.org/W17-6913.

Ishii, R., Nakano, Y. I., and Nishida, T. (**2013**). "Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze," ACM Trans. Interact. Intell. Syst. **3**(2), doi: 10.1145/2499474.2499480.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (**1998**). "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," Language and Speech **41**(3-4), 295–321, doi: 10.1177/002383099804100404 pMID: 10746360.

Leite, I., McCoy, M., Ullman, D., Salomons, N., and Scassellati, B. (**2015**). "Comparing models of disengagement in individual and group interactions," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, Association for Computing Machinery, New York, NY, USA, p. 99–105, doi: 10.1145/2696454.2696466.

Morency, L.-P., Kok, I., and Gratch, J. (**2010**). "A probabilistic multimodal approach for predicting listener backchannels," Autonomous Agents and Multi-Agent Systems **20**(1), 70–84, doi: 10.1007/s10458-009-9092-y.

Mueller, M., Leuschner, D., Briem, L., Schmidt, M., Kilgour, K., Stueker, S., and Waibel, A. (**2015**). "Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques," in *Human-Computer Interaction: Interaction Technologies*, edited by M. Kurosu, Springer International Publishing, Cham, pp. 329–340, doi: 10.1007/978-3-319-20916-6_31.

Ozkan, D., and Morency, L.-P. (**2010**). "Concensus of self-features for nonverbal behavior analysis," in *Proceedings of the First International Conference on Human Behavior Understanding*, HBU'10, Springer-Verlag, Berlin, Heidelberg, p. 75–86, doi: 10.1007/978-3-642-14715-9_8.

Ozkan, D., and Morency, L.-P. (**2013**). "Latent mixture of discriminative experts," IEEE Transactions on Multimedia **15**(2), 326–338, doi: 10.1109/TMM.2012.2229263.

Pellet-Rostaing, A., Bertrand, R., Boudin, A., Rauzy, S., and Blache, P. (**2023**). "A multimodal approach for modeling engagement in conversation," Frontiers in Computer Science **5**, doi: 10.3389/fcomp.2023.1062342.

Pickering, M. J., and Garrod, S. (**2021**). *Understanding Dialogue: Language Use and Social Interaction* (Cambridge University Press).

Poppe, R., Truong, K. P., Reidsma, D., and Heylen, D. (**2010**). "Backchannel strategies for artificial listeners," in *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, IVA'10, Springer-Verlag, Berlin, Heidelberg, p. 146–158.

Priego-Valverde, B., Bigi, B., and Amoyal, M. (**2020**). ""cheese!": a corpus of face-to-face french interactions. a case study for analyzing smiling and conversational humor," in *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 467–475, https://www.aclweb.org/anthology/2020.lrec-1.59.

RStudio Team (**2020**). *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, http://www.rstudio.com/.

Ruede, R., Müller, M., Stüker, S., and Waibel, l. (**2019**). *Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor*, 247–258 (Springer International Publishing, Cham), doi: 10.1007/978-3-319-92108-2_25.

Sacks, H., Schegloff, E. A., and Jefferson, G. (**1974**). "A simplest systematics for the organization of turn-taking for conversation," Language **50**(4), 696–735, doi: 10.2307/412243.

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

Sidner, C., and Dzikovska, M. (**2002**). "Human-robot interaction: engagement between humans and robots for hosting activities," in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, IEEE, pp. 123–128, doi: 10.1109/ICMI.2002.1166980.

Sidner, C. L., Kidd, C. D., Lee, C., and Lesh, N. (**2004**). "Where to look: a study of human-robot engagement," in *Proceedings of the 9th International Conference on Intelligent User Interfaces*, IUI '04, Association for Computing Machinery, New York, NY, USA, p. 78–84, doi: 10.1145/964442.964458.

Stivers, T. (**2008**). "Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation," Research on Language and Social Interaction **41**(1), 31–57, doi: 10.1080/08351810701691123.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (**2009**). "Universals and cultural variation in turn-taking in conversation," Proceedings of the National Academy of Sciences **106**(26), 10587–10592, doi: 10.1073/pnas.0903616106.

Tolins, J., and Fox Tree, J. E. (**2014**). "Addressee backchannels steer narrative development," Journal of Pragmatics **70**, 152–164, doi: https://doi.org/10.1016/j.pragma.2014.06.006.

Truong, K., Poppe, R., and 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, http://www.interspeech2010.jpn.org/, 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010 ; Conference date: 26-09-2010 Through 30-09-2010.

Ward, N., and Tsukahara, W. (**2000**). "Prosodic features which cue back-channel responses in english and japanese," Journal of Pragmatics **32**(8), 1177–1207, doi: https://doi.org/10.1016/S0378-2166(99)00109-5.