A Multimodal Model for Predicting Conversational Feedbacks

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

A Multimodal Model for Predicting Conversational Feedbacks

Auriane Boudin Roxane Bertrand Stéphane Rauzy
Magalie Ochs Philippe Blache

LPL-LIS, CNRS Marseille, France
Author version

Abstract

We propose in this paper a statistical model in the perspective of predicting listener's feedbacks in a conversation. The first contribution of the paper is a study of the prediction of all feedbacks, including those in overlap with the speaker with a good accuracy. Existing model are good at predicting feedbacks during a pause, but reach a very low success level for all feedbacks. We give in this paper a first step towards this complex problem. The second contribution is a model predicting precisely the type of the feedback (generic vs. specific) as well as other specific features (valence expectation) useful in particular for generating feedbacks in dialogue systems. This work relies on an original corpus.

1 Introduction

Conversational interactions are characterized by different phenomena showing participant's engagement. Among them, the most important are undoubtedly conversational feedbacks [33] which usually consist in brief signals produced by the interlocutor during the main speaker's speech and can be verbal (e.g. *yes*), vocal (e.g. *mhm*), and/or gestural (head movements, smiles). All linguistic interaction theories underline their crucial role during an interaction [16, 19]. We know in particular that they are mandatory to update the shared knowledge (*common ground*) and promote the alignment between participants which is necessary for mutual comprehension and success of the interaction [4, 24]. Understanding their behaviors and the conditions of their realization in a natural context is then of deep importance both for theoretical reasons, but also for applications in the perspective of human-machine interaction.

Feedbacks predictive cues have been studied in different modalities: prosodic [18, 25, 38], syntactic [9], gestural [17] semantic and pragmatic [2]. However, to date, no global model proposing a multimodal account of feedback realization and a prediction of feedback types exists. One of the reasons is that there exist only few corpora providing such information. The vast majority of existing resources focus on audio only, which explains the fact that most of the works rely on prosody, taking other modalities into account only to a certain extent. Corpora providing both video and audio are limited and generally do not bear annotation of the gestures involved in the description of these phenomena. Moreover, the processing of

these multimodal data is still an open question. Overall, a multimodal model of feedbacks in natural interaction, describing and predicting their realization, has yet to be established.

We present in this paper an overview of the different feedback models and propose a new model involving a large set of multimodal features. Moreover, we propose to join time and type prediction, making it possible to explore the prediction of feedback occurrences concerning not only the site of realization for the feedbacks, but also their type. This work is based on an original corpus annotated at different levels. We propose a modeling approach rendering possible a clear interpretation of the model. The model and its evaluation are finally discussed.

2 Related works

In this section, we propose an overview of the main feedback modeling works by focusing on three main information: the features they rely on, the time span of the observation window (in other words the segmentation) and the results.

Most of the works study feedbacks as a specific case of turn-taking. Among them, a seminal study has been proposed in [18] for predicting three different situations: turn changes (smooth switches) vs. turn retentions (holds) and backchannels (a feedback subtype). The signal is segmented into inter-pausal units (IPU) and at each pause longer than 50ms, a set of features is extracted from the preceding IPU. Seven features are used in turn-yielding prediction: a falling or high-rising intonation at the end of the IPU; a reduced lengthening of IPU-final words; a lower intensity level; a lower pitch level; a point of textual completion; a higher value of three voice quality features: jitter, shimmer, and NHR; and a longer IPU duration. Six features are identified more precisely for backchannels: a rising intonation at the end of the IPU; a higher intensity level; a higher pitch level; a final POS bigram (Det-N, Adj-N, N-N); a lower value of noise-to-harmonics ratio (NHR); and a longer IPU duration. They fit multiple logistic regression to assess the relative importance of the different cues. This study show that the likelihood of the occurrence of a feedback is 30% when all 6 cues are present (the percentage of turn-taking being 65% with the 7 relevant cues). A cross-lingual study based on 3 languages has been proposed in [12].

This approach is adapted by [21] for detecting "*response relevant places*" (a response being feedbacks, short answers, clarification questions, etc. organized into 8 classes). The task called Response Location Detection consists in classifying into 3 main classes: hold (a response would be inappropriate), expected response, optional response. They completed the model of [18] by taking into account three different types of features: prosodic (pitch, intensity), contextual (turn and IPU length, last system dialogue act, pause duration) and lexico-syntactic (word form, POS, semantic classes). The observations are also based on an IPU segmentation (2,272 IPUs). Different learning algorithms and feature combinations have been applied. The model based on lexico-syntactic features obtains a 84.42% score of accuracy for the voted perceptron model, only slightly improved when adding prosodic features (84.64% accuracy with naive bayes).

In his work, [34] proposes a continuous model of turn-taking, predicting turn-shifts and their types (short backchannel or a longer utterance). It uses a set acoustic features (voice activity, pitch, intensity,spectral stability), completed by POS. Feature vectors are extracted from the signal each 50ms. The model learned by a RNN reaches an F-Score of 0.762, using all features for the turn-shift prediction.

When focusing on feedback prediction independently from turn-taking, we find several rule-based approaches. In their reference work, [38] propose a famous rule based on prosodic cues (pitch region, its level, duration and localization and previous feedback realization). This rule has been completed by other algorithms in [25] aiming at determining the placement of feedbacks. These rules are based on prosodic and gaze features. They are implemented in a virtual listener. Note that this work also provide interesting information about human listener behavior and has been refined by a corpus study in [37]. In their work, [23] also propose to use several lexical and prosodic cues as well as the indication of speaker's gaze. Speaker features are sampled at a rate of 30Hz. Different probabilistic sequential model are trained for different encoding templates. The results show an F-score of 0.2562 (outperforming the Ward & Tsukahara method). On their side, [13] present an approach for predicting a specific type of feedbacks (backchannel continuers) on the basis of pause duration and POS. Their results show an F-measure of 35%.

To summarize, the different works usually focus on a subset of predictive cues, usually coming from a unique modality. None of them tries to give a global picture involving a multimodal set of features. Second, these works aim at predicting the site or the time for the feedback realization, but not the feedback type. Finally, as for the modeling aspects, no clear indication are usually given on the relative importance of the cues and more generally on the model's interpretation. The great variability and the scores obtained by the different techniques show the difficulty of the task.

3 Feedbacks: typology, cues

We distinguish between two feedback types: *generic* (displaying attention to the speaker) and *specific* (expressing responses to the content of the speaker's production) [4, 36, 6]. Generic feedbacks are often vocal item such as *mh* or gestural signals like nods. They express interest and understanding while specific feedbacks (e.g. brief verbal utterances, a particular tone of voice, etc.) correspond to an evaluation or comment. Some feedbacks are prototypical of one type (e.g. *mhm* is typically generic whereas *oh no* is specific), some others not. Feedbacks can be expressed in different modalities: verbal, visual or multimodal. Besides verbal feedbacks, which are in most of the cases brief lexical expressions, visual feedbacks can be realized in different ways: head movement (*nod, jerk, shake, tilt, turn, waggle*), facial expressions (*smile, laugh*) or eyebrow (*frowning, raising*). Bimodal feedbacks (involving both visual and verbal productions) are also very frequent and play an important role [17]. In some cases, bimodality can reinforce the function: for example, bimodal BCs show a stronger agreement than unimodal ones. Table 1 give some examples of the main feedback types and their modalities associated to different possible subtypes (note that our work focuses on French, but can be directly generalized to other languages).

Typology: As shown in the table, specific feedbacks are described according to two levels of analysis: polarity (positive or negative) [1, 15] and their expected/unexpected character [27]. This question of feedback subtype is crucial [33, 4] and could explain the various and sometimes controversial cues found in the literature (*yes* for example can be either generic or specific).

Feedback cues: Based on the literature as well as our experience, we propose to explore the role of a large set of features, from three modalities: speech (prosodic features, [18,

Type	Modality	Feedback
Generic	Verbal	`oui, mh`
	Visual	`nod, smile`
	Bimodal	`nod+yeah, smile+ok`
Specific / agreement	Verbal	`oui, d'accord, ok`
	Visual	`nod, smile`
	Bimodal	`nod+yeah, nod+ooh`
Specific / disagreement	Verbal	`non`
	Visual	`shake`
	Bimodal	`shake+no, shake+mh`

Type	Modality	Feedback
Specific surprise	Verbal	`ah bon`
	Visual	`raising, frowning, tilt`
	Bimodal	`raising+no`
Specific / fear	Verbal	`non, oh non`
	Visual	`frowning, raising, shake`
	Bimodal	`shake+no, frowning+non`

Table 1: Feedback Types

Speech	Short silent pause lasting at least 200ms and maximum 1200ms Long silent pause lasting at least 1200ms Bigrams and trigrams of tones
Verbal	POS bigrams and trigams Discourse markers Positive, Negative, Concrete words
Visual	Laughs, Smiles Nods

Table 2: Predictive features

5, 20, 34, 22, 12, 38]), verbal (lexico-syntactic features [21, 10, 18]) and visual (gestures, expressions, attitudes, [1, 17, 22, 25]). Besides widely used features (e.g. POS, pauses, etc.), we also propose to involve less studied features.

Before entering into feature description, it is necessary to specify the frame unit into which the predictive features will be analyzed. In most of the cases, inter-pausal units (IPUs) are chosen, the features of the end of the main speaker's IPU are the input variable of the model. The problem in this case is that we miss many feedback produced in overlap with the speaker (representing 40% of feedbacks in our corpus). It is then necessary to chose another segmentation. One solution consists in segmenting the input by means of a rolling window. In some works, an arbitrary frame size of 30-50ms is chosen [23, 34]. However, this type of segmentation entails a huge problem of imbalanced classes when trying to predict feedback vs. no-feedback, including during speakers speech. We propose instead to segment on the basis of *events*, at each word, or during no-speech segments at nod or laugh. Such events form the right boundary of a predicting window of 2 seconds. This duration is arbitrary, but usually correspond to segments larger enough to contain complete units (in terms of syntax and pragmatic contents). Table 2 present the complete set of features involved in our model.

At the prosodic level, besides pauses, we examine the intonation pattern given by a sequence of tone. Some specific tone n-grams before a pause followed by a feedback could correspond to a final intonation pattern which is often correlated with the introduction of a new information [14] and carries an important part of the interactional meaning [26]. Several studies have shown that the final intonation contour could be a good predict or for feedback occurrence. Tones also represent an intermediate level between low-level acoustic features such as pitch and phonological interpretations. Taking into account tone n-grams makes it possible to compare the influence of tones taken separately or by sequence.

Concerning lexico-syntactic features, we propose to include POS and semantic information about word polarity (positive, negative) and aspect (concreteness). These information can be associated to specific listener's reaction concerning a certain level of emotion, but also the use of a discourse referent associated to concrete words. On their side, discourse markers are the sign of discourse organization, often associated to transition between discourse units, that can be associated to reactions.

Finally, introducing visual features complete the multimodal description. Nods, laughs and smiles, that can by themselves constituting feedbacks, are also to be taken into account in the prediction, not only because of mimicry (a laugh can trigger a laugh feedback) , but also due to the importance of their communicative function [7].

4 The dataset

This work focuses on French. Our dataset is built upon an existing corpus of natural conversations that we have completed with the different annotations. The corpus, called Cheese-Paco [28, 3], contains 7 hours of audio-video recording. The participants, in dyads, were installed face-to-face. They first had to read a short story before having a free conversation. Cheese-Paco is composed of 27 interactions, each lasting an average of 17 minutes. A manual transcription has been done, including different information such as noise, laugh, pause, elision, unexpected events. This transcription has been automatically aligned onto the signal thanks to the SPPAS system [8] that returns the list of phonemes, syllables and IPUS. The MarsaTag analyzer [31] has been applied to extract the lemmas and POS. Moreover, smiles and nods were annotated semi-automatically using the HMAD toolkit [30, 29]. In this work, we used a subset of data composed of 4 dyads for a total of about one hour. The corpus contains 769 feedbacks, 2,739 IPUs and 15,215 words.

Feedback annotation: Feedbacks have been manually annotated by 3 annotators. The annotation guide considers the distinction between specific and generic feedbacks to which we added the two subtypes (valence +/- positive and +/- expected). In order to facilitate the annotation, a pre-processing has been done for identifying automatically the possible feedbacks on the basis of different signals: laughs, smiles shorter than 200ms, repetitions, interjections. Annotators had to check whether these suggestions was correct and when necessary add feedbacks non identified during pre-processing. The second annotation step consists in identifying the feedback type. Five categories are possible: one for generic feedbacks and four for specific feedback sub-types (+/- positive, +/- expected). Finally, annotators were asked to determine the feedback boundaries.

Lexico-syntactic features: Besides POS, we annotated lexical-semantic information (concreteness, valence) on the basis of word lists given in [11]. We also identified discourse markers. Concerning POS, we kept only bigrams and trigrams with a frequency higher than 40.

Prosodic features: We used the automatic pitch modeling tool MOMEL-INTSINT [32] which consists in two steps. The first consists in modeling the f0 based on a sequence of transitions between successive points on the curve (anchor points). The procedure of 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 used by the authors allows to compare speakers with different pitch ranges (typically males versus females) In a second step, the anchor points obtained from MOMEL are automatically coded by an alphabet of tonal symbols T(op), B(ottom), M(id) referring to absolute values and *Higher, Lower, Same, Upstepped, Downstepped* (referring relative values) and give rise to an intonation pattern represented by the key/midpoint and the span of the speaker's pitch range. We also encode, as it is the case in different studies, the length of the pause (with a threshold of 1,2 seconds between short and long pauses).

Gestures, expressions: Nods has been annotated semi-automatically. The automatic step relies on the HMAD system [30] which returns the time interval of the nod. A manual correction is then done, identifying when necessary missing nods and correcting time boundaries. The annotation of smiles follow the same procedure, and distinguishes 2 levels of smile (noted S3 and S4 in the following) [29].

5 Data analysis

We give in this section a brief overview of the main statistical characteristics of our dataset. These statistics are in line with the literature [25]: feedback are frequent and their production is a consistent phenomenon. Our data revealed a frequency of 9.3 feedback per minute (see Table 3). Specific feedback are slightly more frequent than generic. Regarding the sub-type of specific feedbacks, there is more often positive than negative feedback. The most frequent is the positive-unexpected feedback, the less is negative-unexpected.

Feedback type		Frequency per minute
Specific	positive-expected	1.62
	postive-unexpected	2.06
	negative-expected	0.84
	negative-unexpected	0.34
	total specific	**4.85**
Generic		**4.44**
Total feedback		**9.3**

Table 3: Feedback frequency per minute and per type

The most frequent feedback realizations are verbalization, laughs [6] and nods [35, 1]. We find interesting to see how nods, laughs, verbalization and smiles are used according to the generic/specific function of the feedback. Figure 1 reveal the sum of feedback produced with at least one of these items (for a total of 769 feedbacks). Even if all feedbacks are mainly produced with verbal items (interjections, repetitions and other short lexical elements), this tendency is stronger for specific feedbacks. This can be explained by the fact that specific feedbacks need more details and context-dependent interventions. Verbal items for generic feedbacks are mostly interjections and plays the role of continuers. As expected, nods are significantly more used for generic than specific feedback [4]. Conversely, laughs and smiles are widely used as specific feedbacks (and are very rare for generic ones). Finally, generic feedbacks are mainly produced with nods and/or verbalization, whereas specific feedbacks are primarily produced with verbalization/laugh and/or smile. We also noticed that generic

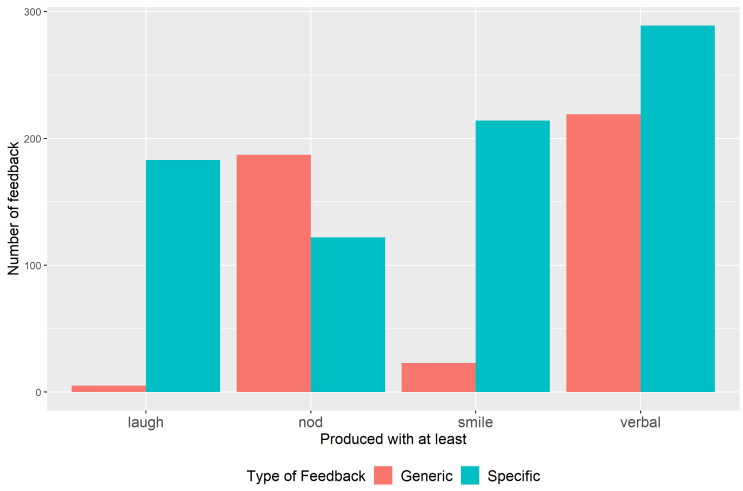

Figure 1: Description of feedback production

feedbacks are produced in overlapping 34% of the time and specific feedback are produced in overlapping 41% of the time.

6 The model

We applied a statistical processing, logistic regression, that fits with our dataset specificities. At the difference with most of other predicting models, our goal is to predict both the position and the type of feedback to be produced during a conversation. We use for that a two-stage approach based on two models: one predicting the realization of a feedback, the second predicting its type (generic vs. specific). In this approach, *logit* is then used as a hierarchical classification technique. One interest of using *logit* also lies in the fact that it returns a probability for the classification. This is interesting in the perspective of implementing the model in a human-machine communication system, offering the possibility to introduce a fine variability in the feedback production. We also chose logistic regression in order to take into consideration different questions about the dataset (the size is rather small and learning techniques are prone to overfitting) as well as dimensionality and interpretability of the model.

The probability to produce a given type of feedback (or the probability that the feedback is produced at a given time location) is modeled by the *logit* equation:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = a_0 \; + \; a_1\, x_1(t) \; + ... + \; a_j\, x_j(t) \; + ... \tag{1}$$

where $x_j(t)$ are the predictors which depend on the time location t and which 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 on the training sample. An inspection of the result allows to decide which predictor contributes significantly to the prediction. The model is

finally formed with the subset of relevant predictors. A probability p is thus attributed to any combined values of the predictors. To convert this probability into a binary response, we apply the following classifier equation:

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

where $p_{\text{threshold}}$ is for example the averaged probability to produce a feedback. This threshold value is estimated on the training sample. The model thus provides us with a binary prediction which depends on the predictor values in input.

The performance of the model is afterwards evaluated by using standard metrics based on the confusion matrix of the predictions versus the observations. A cross-validation strategy is applied in order to quantify the potential presence of overfitting problems. This evaluation procedure will be also applied in order to measure the relative contribution of a given subset of predictors (e.g. grouped by modalities).

Predicting the occurrence of a feedback The problem of predicting the time location of the feedbacks is difficult since it requires to include in the training table events where feedback does not occur. In our strategy, we considered all time locations corresponding to the end of an event (e.g. end of a word, of a pause, of a laugh or smile, etc.) as an entry of the table and we encoded the value of the binary dependent variable as 0 (i.e. absence of feedback). In addition we inserted in the table the list of observed feedbacks (i.e. with the binary dependent variable at 1). This strategy comes to the standard binary classifier model. In our case, the difficulty lies in the imbalanced character of the binary distribution (i.e. the no feedback events are massively dominant).

Note that dynamical models based on HMM, CRF, LSTM RNN have been proposed for predicting occurrences of feedbacks or related situations (see for example [23, 34]). These models integrate the information produced by the main speaker as well as information on the current state of the listener. However, they reach very low F-score results unless they restrict the prediction of feedbacks during pauses only.

Predicting the type of feedback For the task of predicting the type of feedback the training sample is a table containing for each observed feedback:

- The binary dependent variable (encoded as 1 if the feedback has the desired type an 0 otherwise)
- The predictors of binary type (e.g. absence or presence of a given POS trigram, absence of presence of a pause, ...) or continuous type (e.g. number of tokens).

The *logit* model takes this training table as input, the relevant predictors are identified and the final model is adopted which allows to compute for any combination of predictor values the predicted type of feedback in output through equation 2.

7 Results

We propose in the following to discuss the results obtained for different feature combinations and clustered by modality. For each combination of coefficients, the accuracy and the Cohen's kappa scores have been obtained by running a Monte Carlo cross-validation (on 100 trials with a ratio 80%-20% for the training versus the evaluation sample). Quoted errors are the 1σ standard deviation on the estimates of these parameters.

Feedback occurrence: Concerning the prediction of feedback occurrence, let's note first that our modeling leads to a good accuracy whatever the feature combination. Also note that in their work, [18] mention that the likelihood of occurrence of a feedback is 30% when involving all their predicting cues. As underlined previously, the main difficulty in this task is that classes are imbalanced and most of the input samples concern no-feedbacks, explaining the rather low kappa value. But also note that this value is significantly higher than random.

Table 4 shows the emergence of 12 features: pauses, laughs, smiles (S3 and S4), 4 tone bigrams, 3 POS bigrams and number of tokens. Interestingly, each modality contribute to the prediction. Expression and visual features (laughs and smile) are unsurprisingly correlated with feedback realization, confirming the literature. In terms of prosody, the presence of pauses also confirm existing results. More precisely, we observe 4 tonal bigrams: *DH, SS, SD* and *MD*. The DH pattern could correspond to a rise intonation contour (to be compared with the L-H% contour in [18]). On its side, the MD bigram corresponds to a fall. This pattern can be find in several works, even though with different consequences: [18] indicate that this intonation contour is likely to result in turn change or a continuation by the same speaker (with no feedback) where [38] indicate that a region of low pitch lasting at least 110 milliseconds is considered as a feedback-inviting cue. The *SD* and *SS* bigrams could also refer to such feedback-inviting-cues. In terms of lexico-syntactic cues, the most important feature is the number of tokens. This information is in fact complex: it can be related to speech rate, but also to the syntactic structure (grammatical words are small, their number increase when the syntactic structure is more complex). In both cases, this information could be in relation with "completude": more tokens could be associated to a more complete unit.

Feat. combination	Formula	Accuracy	Kappa
All	S4 + S3 + Pause + Laugh + S4 + N-Det + S3 + Det-N + DH + SS + Dem-V + SD + MD + TokenNb	0.78 ±0.008	0.099 ±0.01
Visual	Laugh + S4 + S3	0.81 ±0.005	0.039 ±0.012
Prosody	Pause DH + SS + SD + MD	0.82 ±0.013	0.114 ±0.019
Lexico-synt	TokenNb + N-Det + Det-N + Dem-V	0.64 ±0.109	0.007 ±0.009

Table 4: Prediction of feedback occurrences

Generic/specific prediction: Table 5 shows the accuracy for predicting the different feedback subtypes. In this case, the classes are more balanced, leading to a better kappa. We obtain, in spite of the task difficulty, interesting results.

In this table, we present the different models for each subtype : generic vs. specific, and for the specific feedbacks positive vs. negative and expected vs. unexpected. In the +/- generic classification, the main features are visual (laughs, smiles and nods), which is in line with the literature. As for prosody, three patterns occur: *DHL*, *DH*, *TL*. The T tone refers to a maximum in the speaker's pitch range and DH refers to a rise. We know that high values of f0 are often considered more salient (focus, emphatic style) which could be more associated with specific feedbacks (conveying for example stance, emotion, etc.). When comparing the different combinations, the all features gives the best results. The POS bigrams tend to show the opening of a new structure, which could play in favor of generic

Feat. combination	Type	Formula	Accuracy	Kappa
All	generic	Laugh + Smile + Nod + DHL + Adv-Clit + Det-N + N-Prep + DH + TL + Dem-V	0.62 ±0.0369	0.255 ±0.067
	positive	Clit-V-Det + Pause(¡ 1s2) + Smile + Pd.Vm. + Disc.Mark + SD + MS + Aux-V	0.61 ±0.046	0.093 ±0.081
	expected	V-Adv + LH + LU + N-Adj + Det-N + LUD + PositiveToken	0.51 ±0.048	0.037 ±0.098
Visual	generic	Laugh + Smile + Nod	0.59 ±0.035	0.229 ±0.054
	positive	Laugh	0.41 ±0.046	0.007 ±0.053
	expected	x	x	x
Prosody	generic	DHL + DH + TL	0.50 ±0.034	0.069 ±0.049
	positive	Pause(¡ 1s2) + SD + MS	0.67 ±0.16	0.054 ±0.1
	expected	LH + LU + LUD	0.48 ±0.048	0.048 ±0.073
Lexico-synt	generic	Adv-Clit + Det-N + N-Prep + Dem-V	0.57 ±0.036	0.134 ±0.073
	positive	Clit-V-Det + Dem-V + Disc.Mark + Aux-V	0.57 0.068	0.047 ±0.087
	expected	V-Adv + N-Adj + Det-N + PositiveToken	0.51 ±0.05	0.045 ±0.094

Table 5: Prediction of feedbacks subtypes

feedbacks (continuers).

Concerning the prosody, a short pause just before the feedback is the most salient cue for positive feedbacks. It is difficult to interpret this result, unless considering that positive responses are preferred and preferentially occur in a short delay. Note that surprisingly, the feature *PositiveToken* does not play a role in this task. Results concerning +/- expected information are less reliable. Prosody features show two tonal bigrams, LH and LU, corresponding to a rise that could be associated with a new information (that could be for example associated with a surprise feedback).

8 Conclusion

We have presented in this paper different models addressing for the first time both the prediction of feedbacks and of their types. Our approach shows that a multimodal combination of predictive features can lead to a good accuracy level and represent, to the best of our knowledge, a state of the art for this double classification task. This statistical modeling constitutes the first step for future systematic studies based on different machine learning techniques. In terms of application, this model has been implemented in an automatic dialogue system and is currently under evaluation.

References

[1] Allwood, J., Cerrato, L.: A study of gestural feedback expressions. In: First Nordic Symposium on Multimodal Communication. pp. 7–22 (2003)

[2] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation **41**(3), 273–287 (2007)

[3] Amoyal, M., Priego-Valverde, B., Rauzy, S.: PACO : A corpus to analyze the impact of common ground in spontaneous face-to-face interaction. In: LREC procs (2020)

[4] Bavelas, J., Cates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology **79**(6) (2000)

[5] Beňuš, Š., Gravano, A., Hirschberg, J.: Pragmatic aspects of temporal accommodation in turn-taking. Journal of Pragmatics **43**(12), 3001–3027 (2011)

[6] Bertrand, R., Espesser, R.: Co-narration in french conversation storytelling: A quantitative insight. Journal of Pragmatics **111** (2017)

[7] Bertrand, R., Ferré, G., Blache, P., Espesser, R., Rauzy, S.: Backchannels revisited from a multimodal perspective. In: Auditory-visual Speech Processing (2017)

[8] Bigi, B.: Sppas: a tool for the phonetic segmentations of speech. In: The eighth international conference on Language Resources and Evaluation. pp. 1748–1755 (2012)

[9] Blache, P., Abderrahmane, M., Rauzy, S., Ochs, M., Oufaida, H.: Two-level classification for dialogue act recognition in task-oriented dialogues. In: COLING'20 (2020)

[10] Blache, P., Abderrahmane, M., Rauzy, S., Bertrand, R.: An integrated model for predicting backchannel feedbacks. In: IVA procs. pp. 1–3 (2020)

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

[12] Brusco, P., Vidal, J., Beňuš, Š., Gravano, A.: A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. Speech Communication **125**, 24–40 (2020)

[13] Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: European ACL. pp. 51–58. Citeseer (2003)

[14] Chafe, W.: Discourse, consciousness and time. University of Chicago Press, Chicago (1994)

[15] Chovil, N.: Discourse-oriented facial displays in conversation. Research on Language & Social Interaction **25**(1-4), 163–194 (1991)

[16] Clark, H.: Using language. Cambridge University Press (1996)

[17] Ferré, G., Renaudier, S.: Unimodal and bimodal backchannels in conversational english. In: SEMDIAL procs. pp. 20–30 (2017)

[18] Gravano, A., Hirschberg, J.: Turn-taking cues in task-oriented dialogue. Computer Speech and Language **25**(3) (2011)

[19] Horton, W.: Theories and Approaches to the Study of Conversation and Interactive Discourse (01 2017)

[20] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y.: 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 (1998)

[21] Meena, R., Skantze, G., Gustafson, J.: Data-driven models for timing feedback responses in a map task dialogue system. Computer Speech and Language **28**(4) (2014)

[22] Morency, L.P., De Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: International Workshop on Intelligent Virtual Agents. pp. 176–190. Springer (2008)

[23] Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems **20**(1), 70–84 (2010)

[24] Pickering, M.J., Garrod, S.: Understanding dialogue: Language use and social interaction. Cambridge University Press (2021)

[25] Poppe, R., Truong, K.P., Reidsma, D., Heylen, D.: Backchannel strategies for artificial listeners. In: IVA-2010 (2010)

[26] Portes, C., Bertrand, R.: Some cues about the interactional value of the continuation contour in french. In: IDP05 procs. pp. 1–14 (2005)

[27] Prévot, L., Gorisch, J., Bertrand, R.: A cup of cofee - a large collection of feedback utterances provided with communicative function annotations. In: LREC-2016 (2016)

[28] Priego-Valverde, B., Bigi, B., Amoyal, M.: "cheese!": a corpus of face-to-face french interactions. a case study for analyzing smiling and conversational humor. In: LREC. pp. 467–475 (2020)

[29] Rauzy, S., Amoyal, M.: Smad : A tool for automatically annotating the smile intensity along a videorecord. In: HRC2020 (2020)

[30] Rauzy, S., Goujon, A.: Automatic annotation of facial actions from a video record: The case of eyebrows raising and frowning. In: WACAI 2018. Porquerolles, France (2018)

[31] Rauzy, S., Montcheuil, G., Blache, P.: Marsatag, a tagger for french written texts and speech transcriptions. In: Second Asia Pacific Corpus Linguistics Conference (2014)

[32] Rossi, M., Di Cristo, A., Hirst, D., Martin, P., Nishinuma, Y.: L'intonation: de l'acoustique à la sémantique. (1981)

[33] Schegloff, E.: Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences. In: Tannen, D. (ed.) Analyzing discourse: Text and talk. Georgetown University Press (1982)

[34] Skantze, G.: Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In: SIGdial. pp. 220–230 (2017)

[35] Stivers, T.: Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. Research on language and social interaction **41**(1), 31–57 (2008)

[36] Tolins, J., Tree, J.F.: Addressee backchannels steer narrative development. Journal of Pragmatics **70** (2014)

[37] Truong, K.P., Poppe, R., Kok, I.d., Heylen, D.: A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In: Interspeech (2011)

[38] Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in english and japanese. Journal of Pragmatics **32**(8) (2000)