Interdisciplinary Corpus-based Approach for Exploring Multimodal Conversational Feedback

Auriane Boudin

Interdisciplinary Corpus-based Approach for Exploring Multimodal Conversational Feedback

AURIANE BOUDIN*, Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France and Aix Marseille Univ, CNRS, LIS, Marseille, France

During spontaneous conversation, interlocutors have three possible actions: speak, be silent or produce feedback. In order to better understand the mechanisms that render spontaneous interactions successful, this PhD research focuses on conversational feedback. It is the reactions/responses produced by an interlocutor in a listening position. Feedback is a phenomenon of deep importance for the quality of the interaction. It allows interlocutors to share relevant information about understanding, establishment/upgrading of the common ground, engagement and shared representations. The objective of the PhD is to propose a multimodal model of conversational feedback. The methodological approach is interdisciplinary, combining a corpus analysis, based on machine learning enhanced by a linguistic interpretation. The resulting model will be evaluated through its integration in an Embodied Conversational Agent (ECA) with perspective studies.

Additional Key Words and Phrases: Feedback; Multimodality; Linguistic interaction; Statistical model; Corpus study

ACM Reference Format:
Auriane Boudin. 2022. Interdisciplinary Corpus-based Approach for Exploring Multimodal Conversational Feedback. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3557029

1 INTRODUCTION

During spontaneous conversation, interlocutors collaborate in order to reach the communicative goal and make the interaction successful. The role between the main speaker (who holds the floor) and the listener (who is actively listening) is dynamically changing. In listening position, the interlocutors provide information to the main speaker about his/her understanding but can also provide an evaluation of the semantic and pragmatic content produces by the main speaker. These reactions are called feedback (or backchannel) and are of deep importance to help the main speaker in the elaboration of his/her discourse and to guide the conversational directory [Bertrand 2021]. Feedback is also crucial for updating the shared knowledge (common ground) [Clark 1996; Horton 2017] and promoting alignment between interlocutors [Bavelas et al. 2000; Pickering and Garrod 2021, 2013]. Without any reactions from the listener, achieving a successful interaction is almost impossible.

If feedback is essential in human/human interaction, it is also of deep importance to obtain natural conversation between a human and a machine. Indeed, a virtual agent for instance, have to produce feedback at the right time and of the appropriate type. Moreover, depending on the goal of the interaction, the virtual agent has to balance it feedback production to provide the appropriate listening behavior [Glas and Pelachaud 2015; Poppe et al. 2011, 2010].

The aim of this PhD research is to better characterize human/human spontaneous conversations at the linguistic, cognitive and neurophysiological level. Thanks to an inter-disciplinary approach, a theoretical and computational

*PhD student under the supervision of Philippe Blache (Laboratoire Parole et Langage) and Magalie Ochs (Laboratoire d'Informatique et des Systèmes)

Manuscript submitted to ACM

model of feedback will be proposed, by studying it at three different levels: the feedback perception (by the main speaker), the feedback production (by the listener) and the role of feedback in the production/perception loop (the whole interaction). The elaboration of a theoretical and computational model of feedback is twofold. First, the theoretical model will allowed to compute an efficient model that automatically predicts the feedback position and type. Secondly, implementing the computational model in a dialog system will allow to use a virtual agent as an experimental platform to validate the theoretical description of feedback.

The contributions of this thesis is to propose a model of great utility in interactional linguistic and in human/machine interaction field but also to explore new methods of scientific experimentation.

2 FEEDBACK DEFINITION

2.1 Background

The notion of feedback was introduced a long time ago [Schegloff 1982] in the perspective of underlining the collaboration between interlocutors during interaction. Feedback is an unimodal and/or a multimodal reaction produces by an interlocutor in a listening position (referred to as a listener). It can have different forms: nods, smile, laughter, short verbalization, facial expressions, eyebrow movements, hand gestures etc. According to [Bavelas et al. 2000], feedback can be a generic response or a specific response. Generic feedback simply shows understanding and interest by mostly using nods and/or short vocalizations (e.g. *"yeah", "mhmh", "okay"*). On their side, specific feedback is a context-dependent response, that react to the semantic and pragmatic content produce by the main speaker. Their form of realization can thereby be more or less complex, and can include longer verbalization (e.g. *"oh my god", "oh wow", "really"*) and/or particular element (wince, exclamation, rising tone, etc.). Specific feedback can show different attitudes (surprise, amusement, enthusiasm, etc.) and can have an evaluative function. Several studies [Bertrand and Espesser 2017; Stivers 2008; Tolins and Fox Tree 2014] have confirmed the relevance of such a typology.

2.2 Proposed definition

The first step of this thesis was to propose a new definition of feedback that fits with the linguistic and machine learning aspects.

The proposed definition is based on two types (generic and specific) and two sub-types for the specific feedback (positive/negative; given/new). This enabled us to take into account the multi-functionality of feedback with only five categories (generic, positive-new, positive-given, negative-new, negative-given).

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 elements. Generic feedback has more restricted functions than specific feedback. Since generic feedback is homogeneous in form (limited to a defined list of realizations [Prévot et al. 2016]) and in function (mostly grounding and understanding), they do not need a more detailed typology. Conversely, specific feedback is more context-dependent and related to semantic interpretation. It can be composed of various visual and/or vocal elements (marked intonation, longer lexicalization, laughter, eyebrow movements, smiles, head movements or facial expressions). Specific feedback can convey different attitudinal/emotional values that we represent by a finer classification using two levels of sub-classes.

Positive vs. Negative: Specific feedback is classify regarding to the polarity of the main speaker discourse. Did the feedback react to something positive or to something negative? Did the speaker talk about an experience or an event that he/she was evaluating negatively by expressing a criticism, sadness or hunger, etc? Or on the contrary, was he/she

evaluating positively with joy or humor, etc? This positive versus negative aspect of feedback represents the polarity of the semantic and pragmatic content produced by the main speaker, to which the listener responds. It reflects the polarity alignment between the interlocutors.[1]

Given vs. New: Specific feedback is also classify regarding to the common ground (i.e. the shared knowledge between the interlocutors). The *given* type refers to an information already present in the common ground, the listener is reacting to an information that he/she already knows. The *new* type refers to a new instance in the common ground of the listener, here the feedback reacts to the introduction of a new information or to an element that modifies an information already known.

Both speaker and listener information define the feedback category. Each category should be associated with stereotypical characteristics composing the feedback form (i.e. the elements used by the listener to produce the feedback). There is not a one to one mapping between the feedback category and the feedback form. Nonetheless, the assumption is that some combinations of elements are more frequently used in one category of feedback (e.g. raised eyebrows and smile for a *positive-new* feedback ; raised eyebrows and neutral face for a *negative-new* feedback). By using this typology, we can imagine to propose a directory of possibles feedback form within each category. This will allow to introduce more variability in the behavior of ECA during feedback production.

3 RELATED WORKS

3.1 Feedback inviting-features

Several works based on corpus analysis have shown that the production of feedback can be triggered by different multimodal cues from the speaker's speech. The different features are summarized in 1.

- Prosodic features [Brusco et al. 2020; Gravano and Hirschberg 2011; Koiso et al. 1998; Poppe et al. 2010; Terrell and Mutlu 2012; Ward 1996; Ward and Tsukahara 2000]
- Morpho-syntactic features [Bertrand et al. 2007; Gravano and Hirschberg 2011; Ward and Tsukahara 2000]
- Mimo-gestural features [Allwood and Cerrato 2003; Ferré and Renaudier 2017; Poppe et al. 2010; Stivers 2008; Terrell and Mutlu 2012]

3.2 Feedback predictive models

Different computational models have been proposed to automatically predict feedback. Nowadays, they are mainly based on a machine learning approach from a conversational corpus. In most cases, the models focus either on verbal feedback [Cathcart et al. 2003; Okato et al. 1996; Ward and Tsukahara 2000] or gestural feedback [Morency et al. 2010; Ozkan and Morency 2010]. Very few works have considered both gestural and verbal feedback prediction [De Kok et al. 2010; Fujie et al. 2004]. Moreover, existing models usually only focus on the most general type of feedback such as *mh, yeah* or nods. Only a few works have recently started to study more complex feedback [Jang et al. 2021; Kawahara et al. 2016; Ortega et al. 2020].

We can find in the literature two types of method for predicting feedback. The first concerns the temporal prediction and consists in identifying whether or not a feedback may occur, for each timestamp (for example every 50ms). The second method consists in studying what happens at specific positions such as pauses. In this case, the task consists in predicting whether the next event after the pause will be a feedback, a turn change or a turn hold. The works are

[1]Note that we did not consider negative feedback as an inappropriate feedback or a feedback rejected by the main speaker, as other works have done in the past

Table 1. *Summary of feedback and non-feedback predictive cues listed in the state of the art per type of features: Prosodic, Lexico-syntactic and Mimo-gestural. Abbreviations: Inter-Pausal-Unit (IPU), Determinant (Det), Common Noun (NN), Adjective (Adj).*

Cue	Feedback	No Feedback
Prosodic	Flat-fall ; rise-fall ; high-rise ; low-rise	Flat intonation
	High peak of energy	Low peak of energy
	Final vocalic lengthening	Short duration of the final phonemes
	Low pitch period ; High pitch	
	Long IPU duration	
	Speech rate	
	Low noise-to-harmonics-ratio	
Lexico-syntaxic	POS bigram: Det-NN ; Adj-NN, NN-NN	Det
	Connectives close	Interjections
	Disfluencies	Conjunctions
	Final particles	Speech markers
Mimo-gestural	Speaker looks at the interlocutor	Speaker does not look at the interlocutor
	Headnods	
	Smile	

Table 2. Summary of the literature on feedback prediction with objective evaluation. *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**, **Latent Mixture of Discriminative Experts (LSTM)**. *Feedback* columns refers to the feedback studied, first letter indicates the type predicted: only **Generic (G)**, or also **Specific (G/S)**; second letter refers to modality: **Verbal (V)** and/or **Gestural (G)**. *Features* column refers to the type of feature: **Prosodic (P)**, **Morpho-syntactic (M)**, **Gestural/Visual (G)**, **Auto-regressive (A)**. Margin of error (*MoE*) column indicates the window used to evaluate feedback from the onset of the ground truth (- indicates missing information). *Scores* column contain the metrics and associated scores: **f-score (F)**, **Precision (P)**, **Recall (R)**.

Paper	Method	Feedback	Features	MoE	Scores
[Ward and Tsukahara 2000]	RB	G-V	P	± 500ms	P = 0.18 R = 0.48
[Cathcart et al. 2003]	RB	G-V	P M	-	F = 0.35 P = 0.29 R = 0.43
[Truong et al. 2010]	RB	G/S-V	P	± 500ms	F ≈ 0.14 P ≈ 0.22 R ≈ 0.11
[Ozkan and Morency 2010]	CRF	G-G	P M G	FB interval	F = 0.32 P = 0.24 R = 0.49
[Morency et al. 2010]	CRF HMM	G-G	P M G	FB interval	F = 0.26 P = 0.19 R = 0.41
[De Kok et al. 2010]	CRF	G-V/G	P M G	-	F = 0.26 P = 0.27 R = 0.26
[Ozkan and Morency 2012]	LMDE	G-G	P M G	± 500ms	F = 0.30 P = 0.20 R = 0.64
[Mueller et al. 2015]	DNN	G-V	P	± 200ms	F = 0.11
[Ruede et al. 2019]	LSTM	G/S-V	P M A	+ 1s	F = 0.39 P = 0.31 R = 0.51

summarized in Table 2 and 3. They focus on methodological differences in feedback definition, feature selection and evaluation methods.

4 TECHNICAL & THEORETICAL CHALLENGES

In this section, I aim at highlighting the different reasons explaining the task complexity of feedback prediction.

- **Feedback definition**: the first difficulty is the variability of feedback definition from one study to another. By their multimodal and functional characteristics, feedback can include a large set of behaviors. Between the strictest and the broadest definition, problematic becomes completely different.

Table 3. *Summary of the literature on feedback classification in an offline fashion. Prediction columns refers to the classification task:* ***Feedback (FB)***, ***Turn Taking (TT)***, ***Turn Taking Willingness (TTW)***, ***Waiting (W)***, ***Feedback Form (FF)***, ***Feedback Type (FT)***. *Location column describe the site where classifications are done. Features column refers to the type of features used:* ***Prosodic (P)***, ***Morpho-syntactic (M)***, ***Gestural (G)***, ***Contextual (C)***, ***Lexical (L)***, ***Sentiment (S)***, ***Acoustic (A)***. *Scores column present the metrics used and scores:* ***Accuracy (A)***, ***F-score (F)***.

Paper	Algorithm	Prediction	Location	Features	Scores
[Kitaoka et al. 2006]	C4.5	FB/TT/W	Pauses	P M	F(FB) = 27 F(TT) = 54 F(W) = 60
[Meena et al. 2014]	J48	Hold/Response	IPUs end	P M C	A = 84
[Kawahara et al. 2016]	Logit	FF	Boundaries end	P M	F = 0.66
[Ishii et al. 2021]	Adam	FB/TT/TMW	IPUs end	A G L	F = 0.85
[Jang et al. 2021]	LSTM KoBert	FT	Feedback interval	P L S	F = 0.76

- **The data** raise two major points to consider. First, data used can differ by the type of interaction and by the task realized by the participants during data collection. Previous studies mainly used audio-only corpus, map task, doctor/patient dialog or short narration. In order to investigate feedback in the most complete way and to propose a generic model of feedback, audio-visual corpora of spontaneous conversation are necessary. The second point is the imbalance character of these data. For continuous prediction, feedback instances are much less present in interaction than no-feedback instances. Specifically, during spontaneous conversation, it is complicated to predict who is currently speaking. It involves difficulties for training and evaluating the models. By consequences, data of substantial size are required. This implies to lead automatic, semi-automatic and manual annotations.
- **Features**: the state-of-the art has identify feedback predictive cues in different modalities (prosodic, morpho-syntactic, lexical, acoustic, gestural). The problem of dimensionality in machine learning techniques, confronts to make choices in the information give as input to the model. Moreover, the time elapsed between the start of a feature and the time required to trigger a feedback, is not yet very well established. A features selection work is then of deep importance.
- **Models**: related works list many techniques to predict feedback, from very manual (rules-based approach) to fully automatic, whether for features extraction or prediction. The results are hardly comparable given that different data, features, evaluation, margin of error are used in each study. It is thus complicated to settle which model is the best.
- **Evaluation**: this is the consequence of all the problems mentioned above. As explained, feedback can cover several behaviors. Interlocutors are able to grab feedback opportunity, according to the signal produced by the main speaker. There is a lot of variability in feedback production. A listener can decide to be more or less expressive and to pick a few or a lot of feedback opportunities. Listener have also different possibilities to realize the feedback (modality of production, intensity, lengthening, etc.). This variability exists between and within subjects. That renders objective evaluation difficult and provide scores that are not relevant enough to say if a predictive model is accurate or not. These problems are raised in a large quantity of works, referred as the *expressiveness problem* in [Morency et al. 2010].

5 HYPOTHESIS

5.1 Feedback production

- The first hypothesis is that the new feedback classification proposed (generic, positive-new, positive-given, negative-new, negative-given) will allow to cover most of the listening behaviors. This classification is large enough to use machine learning techniques and enough fine-grained to segregate the main strategies use by the listeners.
- The second hypothesis is that within each of the five feedback categories, we can find some stereotypical pattern in the feedback form. The feedback form is defined by the elements that composed the feedback: verbalization, prosody, energy, facial expressions (e.g. eyebrows, smile), gesture, posture, etc. If this hypothesis is confirmed, the feedback classification proposed will be of great importance to improve feedback generation in dialog system.
- The third hypothesis is that the signal of the main speaker provides relevant information (referred to as features) that allow the prediction of feedback (for the listener and for machine learning). Feedback can be triggered either by prominent features or by the combination of less prominent features. The probability to obtain a feedback depends on the weight and on the number of features produced by the main speaker in a given window. For each generic and specific types and sub-types, different sets of feedback-inviting features should be significant.

5.2 Feedback perception

- The fourth hypothesis is that more a feedback form is complex, more the engagement of the listener is perceived as high. In other words, if a lot of elements composed the feedback and/or the feedback is realized with high variation and intensity (prosody and gesture) more the listener will seem involved in the interaction.
- The fifth hypothesis is related to the acceptable delay (or reaction time) to produce a feedback. We observe in the literature, a high variability in the score of the models depending on the window of evaluation used. Most studies use a margin of error of 500ms when evaluating the models. Nonetheless, I think a longer delay than 500ms may be correctly perceived and thus acceptable (up to 1000-1500ms), as long as the feedback type is respected.

5.3 Production-Perception

- The sixth hypothesis is that the absence of feedback or it decreases in quantity, will negatively impact the main speaker production and his/her engagement in the interaction. This can be measured at different multimodal levels: discourse complexity, lexical richness, speech rate, F0 variation, utterance duration, quantity and amplitude of gesture.
- The seventh hypothesis is that an alignment between a speaker and a listener can be found on the frequency band oscillation between a speaker and a listener that are fully engaged in the interaction, where a misalignment can be measured when the interlocutors are disengaged.

6 RESEARCH PLAN

6.1 First year: State-of-the-art and building a predictive model to explore feedback characteristics

- State-of-the art in linguistic and machine learning.
- New feedback classification based on the original distinction between generic and specific from [Bavelas et al. 2000].

- Annotation of feedback according to the generic/specific dichotomy and the two sub-types for specific feedback (positive/negative ; given/new).
- Computation of a hierarchical model to predict the feedback position and type with objective evaluation using a logistic regression.
- Interpretation of the significant features selected by the model.
- Analysis of the element that composed the feedback according to the feedback type.

6.2 Second year: Improve the predictive model and identifying the best parameters of the model by an experimental approach

- Improve the prediction of feedback by adding more features and by testing other machine learning algorithms (CRF, SVM, etc.).
- Perceptual experiment: editing videos of utterance/feedback sequences with delay from -1500ms to 2000ms by steps of 500ms in order to find the maximal time of anticipation and of delay of feedback for the objective and subjective evaluation of the model. The feedback selected as stimuli have been correctly predicted by the first model, in a window of 2 seconds around the original feedback onset.
- Behavioral experiment: dyadic interaction with audio-video and dual EEG recording where a participant is telling a story in front of a) an attentive listener or b) a distract listener (adaptation of the experimental protocol proposed by [Bavelas et al. 2000]).
- Analysis of the perceptual experiment results.

6.3 Third year: Integrating the model in an dialog system and embodied agent to evaluate it in human/machine interaction

- Implementation of the predictive model in a dialog system and its evaluation with a virtual agent.
- Replication of the behavioral study by replacing the listener by a virtual agent. For the distracted condition, the agent will produce only generic feedback and with less occurrences.
- Analysis of the behavioral results.

7 RESULTS

The results of the first year of the PhD led to three publications: two conference papers in international conferences [Boudin et al. 2022a, 2021], and one journal paper in an international journal currently under review [Boudin et al. 2022b].

In [Boudin et al. 2021] we present our fine-grained feedback classification and a hierarchical model that predicts the feedback position and the feedback type.

In [Boudin et al. 2022b], we present also a hierarchical model that predicts feedback position and type with major improvements of the model at several points. A larger dataset is used and an important work of features engineering have been realized. An extensive state-of-the-art allowed to present the main complexity factors of feedback prediction. Finally, we propose a linguistic interpretation of the features selected by the models which is of deep importance for further works in the field.

In [Boudin et al. 2022a], we present a more focused study where we have looked at the smile, smile intensity and laughter used during feedback production. The alignment between the speaker and the listener have been computed and a predictive model of the smile and smile intensity during feedback production lead to encouraging results.

REFERENCES

Jens Allwood and Loredana Cerrato. 2003. A study of gestural feedback expressions. In *First nordic symposium on multimodal communication*. Copenhagen, 7–22.

Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology* 79, 6 (2000), 941.

Roxane Bertrand. 2021. *Linguistique Interactionnelle: du Corpus à l'Expérimentation.* Ph. D. Dissertation. Aix Marseille Université.

Roxane Bertrand and Robert Espesser. 2017. Co-narration in French conversation storytelling: A quantitative insight. *Journal of Pragmatics* 111 (2017), 33–53.

Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, and Stéphane Rauzy. 2007. Backchannels revisited from a multimodal perspective. In *Auditory-visual Speech Processing*. 1–5.

Auriane Boudin, Roxane Bertrand, Magalie Ochs, Philippe Blache, and Stéphane Rauzy. 2022a. Are you Smiling When I am Speaking?. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*. 6.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, and Philippe Blache. 2022b. A Multimodal Model for Predicting Feedback Position and Type During Conversation. (2022), 37 pages. Under review.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A Multimodal Model for Predicting Conversational Feedbacks. In *International Conference on Text, Speech, and Dialogue*. Springer, 537–549.

Pablo Brusco, Jazmín Vidal, Štefan Beňuš, and Agustín Gravano. 2020. A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. *Speech Communication* 125 (2020), 24–40.

Nicola Cathcart, Jean Carletta, and Ewan Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *European ACL*. Citeseer, 51–58.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Iwan De Kok, Derya Ozkan, Dirk Heylen, and Louis-Philippe Morency. 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*. 1–8.

Gaëlle Ferré and Suzanne Renaudier. 2017. Unimodal and bimodal backchannels in conversational english. In *SEMDIAL 2017*. 27–37.

Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 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*. Citeseer, 379–384.

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

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25, 3 (2011), 601–634.

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

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling?. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 131–138.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. BPM_MT: Enhanced Backchannel Prediction Model using Multi-Task Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3447–3452.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems.. In *Interspeech*. 2890–2894.

Norihide Kitaoka, Masashi Takeuchi, Ryota Nishimura, and Seiichi Nakagawa. 2006. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *Information and Media Technologies* 1, 1 (2006), 296–304.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 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 (1998), 295–321.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech & Language* 28, 4 (2014), 903–922.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems* 20, 1 (2010), 70–84.

Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques. In *International conference on human-computer interaction*. Springer, 329–340.

Yohei Okato, Keiji Kato, M Kamamoto, and Syuichi Itahashi. 1996. Insertion of interjectory response based on prosodic information. In *Proceedings of IVTTA'96. Workshop on Interactive Voice Technology for Telecommunications Applications*. IEEE, 85–88.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 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)*. IEEE, 8064–8068.

Derya Ozkan and Louis-Philippe Morency. 2010. Concensus of self-features for nonverbal behavior analysis. In *International Workshop on Human Behavior Understanding*. Springer, 75–86.

Derya Ozkan and Louis-Philippe Morency. 2012. Latent mixture of discriminative experts. *IEEE transactions on multimedia* 15, 2 (2012), 326–338.

Martin Pickering and Simon Garrod. 2021. *Understanding Dialogue*. Cambridge University Press.

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

Ronald Poppe, Khiet P Truong, and Dirk Heylen. 2011. Backchannels: Quantity, type and timing matters. In *International workshop on intelligent virtual agents*. Springer, 228–239.

Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *International Conference on Intelligent Virtual Agents*. Springer, 146–158.

Laurent Prévot, Jan Gorisch, and Roxane Bertrand. 2016. A cup of cofee: A large collection of feedback utterances provided with communicative function annotations. (2016).

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced Social Interaction with Agents*. Springer, 247–258.

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

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

Allison Terrell and Bilge Mutlu. 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*. 280–289.

Jackson Tolins and Jean E Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics* 70 (2014), 152–164.

Khiet P Truong, Ronald Poppe, and Dirk Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. In *Eleventh Annual Conference of the International Speech Communication Association*. Citeseer.

Nigel Ward. 1996. Using prosodic clues to decide when to produce back-channel utterances. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, Vol. 3. IEEE, 1728–1731.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics* 32, 8 (2000), 1177–1207.