Verbal and nonverbal feedback signals in response to increasing levels of miscommunication

Maëva Garnier[1], *Eric Le Ferrand*[1,2], *Fabien Ringeval*[2]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[2]Grenoble INP, LIG, Univ. Grenoble Alpes, Inria, CNRS, Grenoble, France

maeva.garnier@gipsa-lab.fr,eric.le-ferrand@univ-grenoble-alpes.fr
,fabien.ringeval@imag.fr

Abstract

This study aims to explore in detail how listeners respond to communication disruptions in a task-oriented dialogue. We conducted an experiment with participants playing a map task with a partner via a video conferencing system that showed seemingly random breakdowns. In fact, the breakdowns were scripted to induce increasing levels of miscommunication. After an initial interactive session, a second non-interactive session was recorded with one-sided communication from the task leader. Among the fifty or so verbal and nonverbal feedback signals observed, twelve were produced by more than half of the participants. A detailed analysis of their use in different situations, their timing and their co-occurrence, supported that they may have different functions: some appear to be personal reactions of uncertainty, misunderstanding, or inability to complete the task, whereas others were clear repair initiators or turn-taking signals deliberately addressed to the interlocutor.
Index Terms: conversational dynamics, backchannels, feedback, repair, multimodality, verbal and nonverbal communication

1. Introduction

When listeners have difficulty hearing or understanding their interlocutor, they typically use a set of conversational strategies [1] to request a repetition [2], correction [3], or clarification of what was said [4, 5]. These repair initiators can take different linguistic forms, ranging from short non-lexical sounds (”hmm?”) to simple words (”sorry?”) or sentences (”what did you say?”) [6, 7]. They can also vary in their level of specificity, from open repairs that do not specify what or where the problem is (”what?”), to restricted repairs (”when?”) that specify which part of the utterance needs to be clarified, and restricted suggestions that indicate even more specifically what element was misunderstood, and offer suggestions for it (”did you say tomorrow?”) [8, 9]. Finally, like other types of backchannels, these repair initiators are multimodal, and verbal expressions have been shown to be accompanied by prosodic patterns [10], movements of the body or head, such as leaning forward or turning towards [11, 12], and facial expressions [13]. Previous studies have shown that the frequency and form of these repair initiators depend on the communication context. Thus, repair is more frequent in task-oriented conversations, compared to spontaneous interactions [14], and contexts requiring precision involve more specific forms of repair [9]. Recent work has also shown that listeners use specific feedback to indicate the nature of their impairments, in particular to distinguish between hearing impairments (e. g. , lifted eyebrows) and comprehension impairments (e. g. , facial freeze) [13].

The purpose of this study is to further explore the multimodality of these repair initiators, specifically facial expressions, head movements, and gaze, which listeners may use to indicate that they are experiencing hearing or comprehension difficulties in task-oriented dialogues. Since these visible manifestations may be shared with other backchannels or communicative attitudes [15], we sought to disentangle these different aspects by exploring: 1- whether certain nonverbal cues are particularly synchronized with the instant of communication disruption, and 2- whether the magnitude of certain cues correlates with an increasing level of miscommunication, or whether they are used specifically to indicate a certain level of miscommunication.

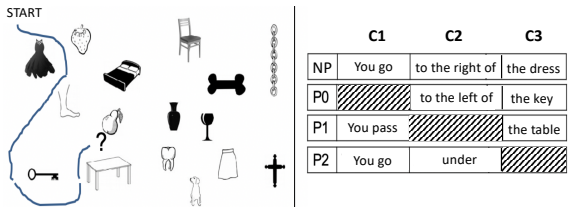

Figure 1: *Example of an interaction map (left) with the instructions the leader could give (right), and the perturbations (P0, P1, and P2) that could occur either on the first (C1), second (C2), or third (C3) component of the utterance, leading to increasing levels of miscommunication.*

2. Material and Method

2.1. Participants and experimental setup

Fifteen participants (seven females and eight males, ages 20-60), were recorded while playing a map task with a partner. Both players were in two separate rooms and communicated through a special video conferencing system that allowed them to look at each other as if in a natural face-to-face interaction situation. Sound was recorded with a headset microphone. Participants played the role of *follower* in the map task, while one of the experimenters, posing as another participant, always played the role of *leader*. The different maps of the game were displayed and filled in by participants, using a graphics tablet, whose screen was duplicated to also be displayed to the leader. An electronic device equipped with a push button was inserted at the output of the leader’s auvio-visual stream, to simulate random disruptions of the transmission by interrupting the signals sent to the participant when the button was pressed, producing a black screen and audio silence.

Although the interaction seemed spontaneous, the game was in fact entirely scripted. First of all, the instructions for the map task, given by the leader, were written in advance, and

were all built on the same model, made of three components (cf. Figure 1):

- **C1**: You + non-specific verb of movement ("go", "pass", "should go")
- **C2**: Preposition of place ("to the left of", "to the right of", "above", "below")
- **C3**: Monosyllabic word designating an object of the map ("the chair", "the key", ...)

Transmission disruptions were triggered by the leader at very specific times in order to mask a component (either C1, C2, or C3) of the leading instructions (cf. Figure 1). The whole scenario was designed so that the disruptions appeared to occur randomly, but in the end, we obtained an equal number (N=12) of instructions that fell into one of the following categories:

- **NP**: instruction without any perturbation.
- **P0**: instruction with a missing onset, which was expected to cause hearing difficulty, but did not impact overall comprehension.
- **P1**: instruction with a missing place preposition, which was expected to have a moderate impact on comprehension, without completely preventing the tracing of a path.
- **P2**: instruction with a missing target, which was expected to have a severe impact on comprehension and, in the absence of clarification, prevent completion of the task.

After a first session of the map task in remote interaction, a second session was recorded, in which the participant again received audiovisual instructions from the map task leader. However, in return, his/her own audio-visual signals could no longer be transmitted to the leader, making the communication one-sided and non-interactive.

2.2. Data labeling and analysis

The audio and video signals of both the map task leader and the participants were recorded synchronously. The leader's instructions were manually annotated, using Praat software [16]. All participants' verbal and nonverbal expressions were labelled manually and freely, using ELAN software [17]. This also enabled us to verify that all the undisrupted instructions (NP) were indeed well understood. We measured the frequency of each category of feedback signal across the eight conditions – i. e. , {NP, P0, P1, P2} x {interactive, non-interactive situations} –, their average onset time after the start of each leader instruction, and, for facial expressions, the degree of Facial Action Units (FAU) activation that was automatically estimated using OpenFace software [18]. Statistical analyses were performed using R software, considering mixed models with Interaction and Disruption position as fixed factors, and a random effect on the participant.

3. Results and discussion

3.1. Categorization of the observed signals

Fifty multimodal feedback signals were identified from the annotations, including sixteen acoustic or verbal signals that have been grouped into the six following categories:

- **Confirmation marks**: non-lexical sounds (e. g. , "hmm"), small affirmative words (e. g. , "yeah", "ok", "all good", ...), or an affirmative repetition of the previous utterance.
- **Hesitation marks**: non-lexical sounds (e. g. , "uuh") and small phrases (e. g. , "Dunno ...")
- **Open repair initiators**: small interrogative words (e. g. , "what?", "sorry?"), sentences to tell a problem (e. g. , "There are breakdowns"), a hearing or comprehension difficulty (e. g. , "I didn't hear you"), or request repetition of the instruction (e. g. , "Can you repeat that?")
- **Restricted repair initiators**: small interrogative words ("where?"), repetitions of part of the instruction, with an invitation to complete a misunderstood part (e.g. "I need to go to the right of ... ?"), or sentences to indicate what was not understood (e.g. "I didn't hear the end of your sentence")
- **Restricted suggestions**: repetition of the utterance with a focused, interrogative suggestion in place of the disrupted constituent (e.g. "Should I go to the left of the dress ?")
- **Brief signal** whose meaning is less clear, nor if they are really addressed to the interlocutor: non-lexical sounds (e. g. , laughter, loud breath, sigh), and swearing.

The thirty-four remaining signals are all nonverbal and have been classified into four categories, based on the facial region involved and Ekman's system of facial actions coding [19]:

- **Head**: vertical (up-down) and horizontal (left-right) nodding, rocking (left-right tilt), sudden forward or back movement, or rotation to one side.
- **Upper face**: Making eye contact, squaring, squinting or voluntary closing of the eyes, looking up or sideways, frowning of one or both eyebrows, raising of one or both outer eyebrows (as in a frightened or surprised face), or raising of the inner eyebrows (as in a sad or sorry face).
- **Lower face**: spontaneous or artificial smile (lip corners raised), lip spreading on one or both sides (straight lip corners), "inverted" smile (lip corners depressed + chin raised), lip compression in the middle or on one side, "pout" (compression + slight forward movement) in the middle or on one side, lip protrusion (closed), frowning of the chin, dropping of the jaw with closed or open lips, lowering of the nasolabial region, raising of the upper lip (with open lips).
- **Other communicative behaviors**: finger on the lips or chin, raising the shoulders, touching one's ear.

3.2. Frequency of the feedback signals according to the type of interaction, and the level of miscommunication

Twelve feedback signals (six verbal, and six nonverbal) were observed in more than half of the participants (cf. Figure 2). On one hand, the six most common verbal signals were observed almost exclusively in the interactive condition (cf. top of Figure 3). Two nonverbal signals: "Horizontal nod" and "Smile", followed the same pattern and were also observed with a sig-

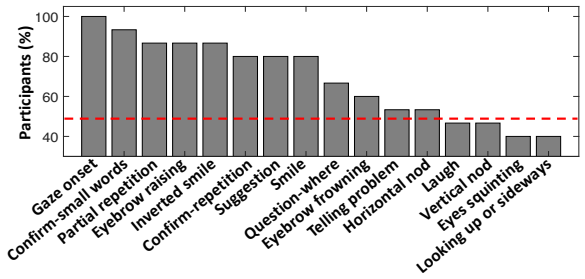

Figure 2: *Verbal and nonverbal feedback signals, ranked by decreasing percentage of participants who showed them.*

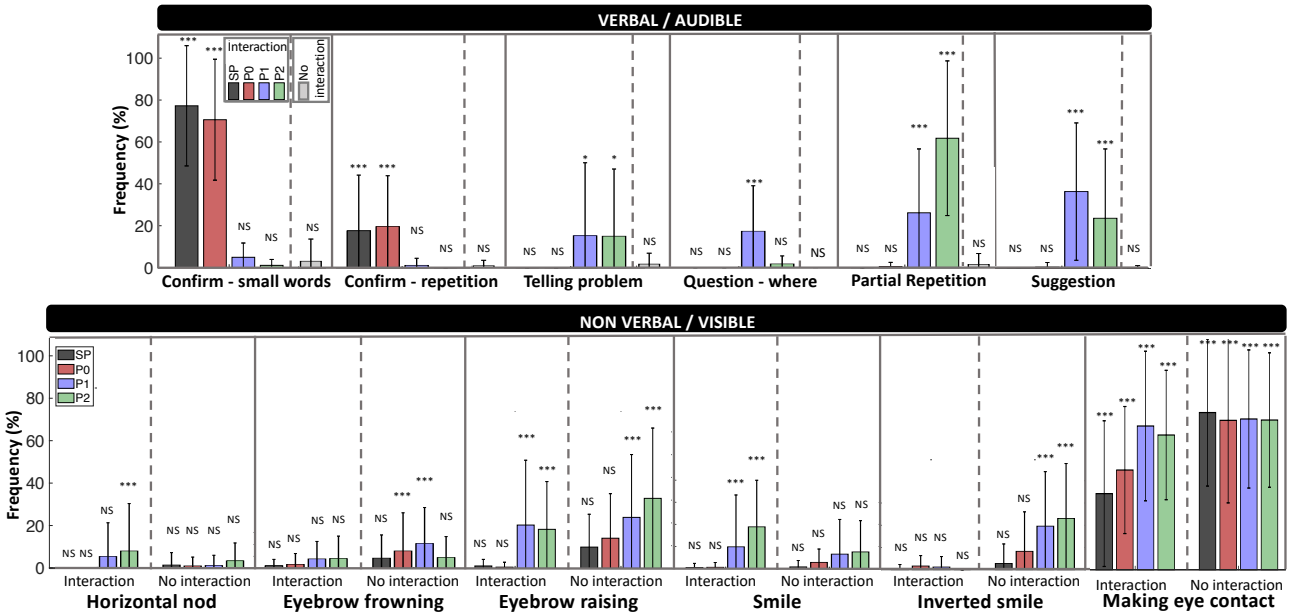

Figure 3: *Percentage of instructions for which the different verbal and nonverbal feedback signals were observed, in either interactive or non-interactive conditions, in an undisrupted situation (NP) or for increasing levels of miscommunication (P0, P1, P2).*

nificant frequency in the interactive condition only (cf. Figure 3 (bottom)). We can therefore consider these eight behaviors as communicative signals deliberately addressed to the interlocutor.

Among them, we observed, as expected, that small confirmation words (e. g. , "ok", "yeah", . . .) and affirmative repetitions of the instruction were produced only following undisrupted instructions (NP) or disrupted instructions that can still be understood (P0). We can therefore consider them as marks of understanding and task completion. Conversely, we can also consider as a mark of incomprehension the absence of these confirmation signs after moderately and severly disrupted instructions (P1, P2).

In contrast, the other four verbal signals "Telling a problem", "Question-where", "Partial repetition" and "Suggestion", and the two nonverbal signals "Horizontal nod" and "Smile" were observed with significant frequency in the two perturbed conditions P1 and P2 only, in which participants lacked information to complete the task and needed misunderstood instructions to be repaired (cf. bottom of Figure 3). These six signals can therefore be considered as repair initiators deliberately addressed to the interlocutor. The fact that none of these signals were also observed in P0 (in addition to P1 and P2), shows that individuals do not (or rarely) express their hearing difficulty until it actually affects their comprehension. Interestingly, the verbal signal "Partial repetition" and the nonverbal signal "Smile" had their frequency significantly increased by the degree of disruption (P2 vs. P1: respectively $+35.1 \pm 7.5\%$, $p < .0001$; $+9.1 \pm 4.1\%$, $p = 0.028$). We also observed that the two nonverbal signals "Horizontal nod" and "Smile" were rarely observed alone, but almost always with verbal signals (93% and 81% of the time, respectively), suggesting that they may be accompanying signals, rather than being repair initiors *per se*.

On the other hand, two nonverbal signals: "Eyebrow raising" and "Making eye contact" were observed in both interactive and non-interactive conditions, although with a significantly higher frequency without interaction (cf. Figure 3). This means that these signals are not deliberate communication signals addressed to the interlocutor. Instead, they may rather reflect the listener's cognitive and emotional state [20, 21], such as surprise, hesitation, disturbance, reflection. Thus, "Eyebrow raising", was observed 46% of the time without a co-occuring verbal signal, with a significant frequency for incomprehensible utterances only (P1, P2). "Making eye contact", however, occured for both disrupted and undisrupted instructions (cf. Figure 3), with a comparable frequency of occurence in the non-interactive situation, but with an increased frequency for incomprehensible utterances (P1, P2), compared to undisrupted and mildly disrupted utterances (SP, P0) in the interactive situation.

Finally, two nonverbal signals : "Eyebrow frowning" and "Inverted smile" were observed with a significant frequency in the non-interactive situation only (cf. bottom of Figure 3), rarely with a co-occuring verbal signal, and for moderate and severe degrees of miscommunication (P1 and P2). It turns out that the task was difficult or impossible to perform in these two specific conditions, due to the fact that no repair could be provided by the leader for the misunderstood instructions in the non-interactive situation. It is therefore possible that these two signals reflect the participant's sense of inability to complete the task.

3.3. Timing of the feedback signals

As expected, participants' verbal signals were observed after the leader's instruction ended (about 1 s later, on average), with little effect of the degree of miscommunication on their onset time (cf. Figure 4). Contrary to our expectations, nonverbal signals (including "Making eye contact") were not produced immediately after the disruption, during the leader's instruction, but also after the end of the instruction, in a similar or even later time frame than verbal signals. Thus, "Horizontal nods" not only almost always occured with verbal signals, as described in the previous section, but also synchronously with

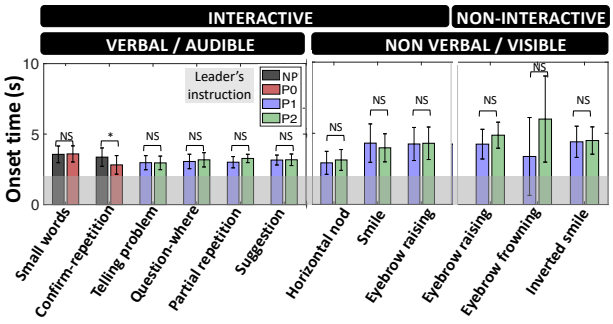

Figure 4: *Average delay of appearance of verbal and nonverbal feedback signals, from the onset of the leader's instruction.*

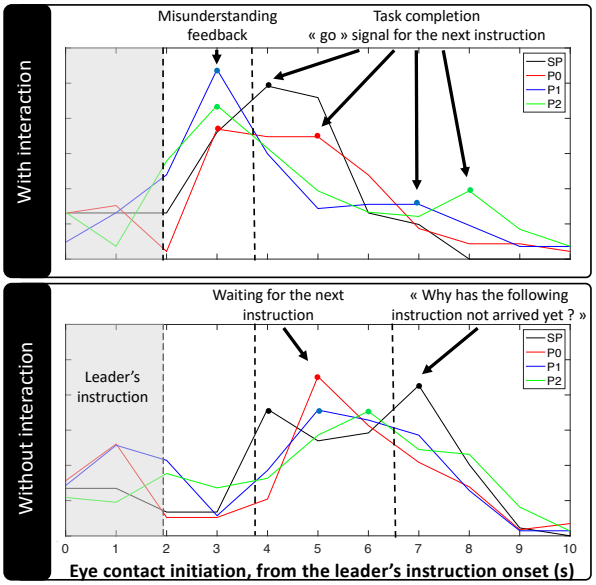

Figure 5: *Distribution of the eye contact initiation times, relative to the onset of the leader's instruction, for disrupted and undisrupted instructions received in either the interactive or non-interactive condition.*

them. This supports further that they may have the same function and meaning as the verbal repair initiators they accompany.

Despite a high percentage of co-occurence with the verbal signals, the "Smiles", however, were produced about 1s later, supporting the idea that they may have a different communicative function. Finally, in the interacive situation, the eye contact initiation times showed a bimodal distribution after a disrupted utterance (cf. Figure 5): a first peak was observed about 1 s after the end of the leader's instruction, i. e. , within a similar time frame as "Horizontal nods" and verbal responses (cf. Figure 4), and thus probably expressing listener uncertainty or misunderstanding; a second late peak was observed between 2 s and 8 s after the end of the instruction, with a delay that increased with the degree of miscommunication (NP $\rightarrow$ P0 $\rightarrow$ P1 $\rightarrow$ P2). In this case, "Making eye contact" seems to correspond rather to the completion of the task (immediately, or after the repair of the utterance), and could be considered as a turn-taking signal to invite the leader to deliver the next instruction [22]. In the non-interactive situation, on the contrary, a single late peak was observed in the distribution of eye contact initiation times, following a disrupted instruction. Coupled with the fact that the cumulated eye contact duration was also significantly greater in the non-interactive situation than in the interactive situation ($+29.4 \pm 14.6\%$, $p = 0.04$), this may indicate that the main function of eye contact in the non-interactive condition is to retrieve visual information in order to better understand speech, especially when disruptions may occur and perturb this understanding. "Making eye contact" in the non-interactive situation could therefore indicate that the listener is ready to receive the next instruction [22]. The second additional peak observed in an undisrupted situation (NP), approximately 3 s after the first eye contact initiation, could reflect a form of impatience or surprise that the leader has not given yet the next instruction.

3.4. Amplitude of facial movements

Only two feedback signals showed a significant increase in the activation of their corresponding FAU from a moderate (P1) to a severe (P2) level of miscommunication: "Eyebrow frowning" ($+0.54 \pm 0.15$, $p = .0006$) and "Inverted smile" ($+0.55 \pm 0.23$, $p = .029$).

4. Conclusions

We showed that listeners produce a set of verbal and nonverbal feedback signals in task-oriented conversations, all of which occur after the end of the received instruction, rather than immediately after the disruption, and only after moderate to severe levels of disruptions. They therefore appear to be related to task success or failure, rather than to "sensory disruption". Nevertheless, some differences were observed in the frequency, timing or magnitude of these signals between moderate and severe levels of miscommunication : for example, "Partial repetitions" and "Smiles" were more frequent, "Making eye contact" occured later, and "Eyebrow frowning" and "Inverted smiles" were more pronounced after severe disruptions. Moreover, based on their different contextual use, i. e. , either in interactive situations only, in non-interactive situations only, or both, and according to their timing and co-occurence, these different signals seem to convey different functions and meanings: some of them, like "Eyebrow raising", "Eyebrow frowning", or "Inverted smile" do not seem to be deliberately addressed to the interlocutor. Rather, they seem to be personal reactions of uncertainty, incomprehension, or inability to complete the task. Others, such as verbal signals, "Horizontal nods", "Smiles", or "Making eye contact" that immediately follow disrupted utterances, can be considered as repair initiators, i. e. , communicative signals deliberately addressed to the interlocutor to ask for repetition or clarification. Finally, "Making eye contact" seems to have multiple functions: when it occurs a few seconds after the end of the leader's instruction, it rather seems to indicate that the task has been correctly accomplished, or that the listener is ready and waiting for the next instruction (turn-taking signal). These different feedback signals and usages were observed here for task-oriented conversations. Further exploration would be needed to extend and/or complement these observations to other types of dialogues, such as those observed in the context of health care [23].

5. Acknowledgements

This project was supported by a local grant of the Pole Grenoble Cognition. We thank Frederic Elisei and Xavier Laval for their help with the experimental setup.

6. References

[1] A. Axelsson and G. Skantze, "Multimodal user feedback during adaptive robot-human presentations," *Frontiers in Computer Science*, vol. 3, p. 135, 2022.

[2] T. S. Curl, "Practices in other-initiated repair resolution: The phonetic differentiation of'repetitions'," *Discourse Processes*, vol. 39, no. 1, pp. 1–43, 2005.

[3] G. B. Bolden, "Speaking 'out of turn': Epistemics in action in other-initiated repair," *Discourse Studies*, vol. 20, no. 1, pp. 142–162, 2018.

[4] M. Dingemanse and N. J. Enfield, "Other-initiated repair across languages: towards a typology of conversational structures," *Open Linguistics*, vol. 1, no. 1, 2015.

[5] M. R. J. Purver, "The theory and use of clarification requests in dialogue," Ph.D. dissertation, University of London, 2004.

[6] N. Ward, "Non-lexical conversational sounds in american english," *Pragmatics & Cognition*, vol. 14, no. 1, pp. 129–182, 2006.

[7] R. F. Young and J. Lee, "Identifying units in interaction: Reactive tokens in korean and english conversations," *Journal of Sociolinguistics*, vol. 8, no. 3, pp. 380–407, 2004.

[8] M. Dingemanse, S. G. Roberts, J. Baranova, J. Blythe, P. Drew, S. Floyd, R. S. Gisladottir, K. H. Kendrick, S. C. Levinson, E. Manrique *et al.*, "Universal principles in the repair of communication problems," *PloS one*, vol. 10, no. 9, p. e0136100, 2015.

[9] R. Fusaroli, K. Tylén, K. Garly, J. Steensig, M. H. Christiansen, and M. Dingemanse, "Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions," in *the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)*. Cognitive Science Society, 2017, pp. 2055–2060.

[10] K. H. Kendrick, "Other-initiated repair in english," *Open Linguistics*, vol. 1, no. 1, 2015.

[11] M. Dingemanse, K. H. Kendrick, and N. J. Enfield, "A coding scheme for other-initiated repair across languages," *Open Linguistics*, vol. 2, no. 1, 2016.

[12] D. Heylen, "Challenges ahead: Head movements and other social acts in conversations," *Virtual Social Agents*, pp. 45–52, 2005.

[13] F. Oloff, ""sorry?"/"como?"/"was?"–open class and embodied repair initiators in international workplace interactions," *Journal of Pragmatics*, vol. 126, pp. 29–51, 2018.

[14] M. Colman and P. Healey, "The distribution of repair in dialogue," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33, no. 33, 2011.

[15] J. B. Bavelas, N. Chovil, L. Coates, and L. Roe, "Gestures specialized for dialogue," *Personality and social psychology bulletin*, vol. 21, no. 4, pp. 394–405, 1995.

[16] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2007.

[17] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multimedia/multi-modal resources with elan." in *LREC*, 2004, pp. 2065–2068.

[18] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[19] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions," in *Proceedings of the 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013*. Shanghai, China: IEEE, April 2013.

[21] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, March 2021.

[22] E. Krahmer, M. Swerts, M. Theune, and M. Weegels, "The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates," *Speech communication*, vol. 36, no. 1-2, pp. 133–145, 2002.

[23] F. Tarpin-Bernard, J. Fruitet, J.-P. Vigne, P. Constant, H. Chainay, O. Koenig, F. Ringeval, B. Bouchot, G. Bailly, F. Portet, S. Alisamir, Y. Zhou, J. Serre, V. Delerue, H. Fournier, K. Berenger, I. Zsoldos, O. Perrotin, F. Elisei, M. Lenglet, C. Puaux, L. Pacheco, M. Fouillen, and D. Ghenassia, "THERADIA: Digital Therapies Augmented by Artificial Intelligence," in *Proceedings of the International Conference on Applied Human Factors and Ergonomics: Advances in Neuroergonomics and Cognitive Engineering)*. New York, USA: Springer, August 2021, pp. 478–485.