A MULTIMODAL ACCOUNT OF LISTENER FEEDBACK IN FACE-TO-FACE INTERACTIONS

Martina Rossi[1], Marin Schröer[2], Bogdan Ludusan[2], Margaret Zellers[1]

[1]Kiel University, Germany, [2]Bielefeld University, Germany
mrossi@isfas.uni-kiel.de, marin.schroeer@uni-bielefeld.de, bogdan.ludusan@uni-bielefeld.de, mzellers@isfas.uni-kiel.de

ABSTRACT

In face-to-face interactions, the conversational feedback produced by the listener to signal attention and participation to the current speaker is multimodal: in the vocal channel, it consists of verbal expressions (e.g., "yes" or "exactly") and vocalizations without lexical content, such as non-lexical backchannels (e.g., "mhm") and laughter; in the visual channel, listener feedback includes movements of the head, such as nods or tilts. In the current research, we investigate the frequency and the distribution (i.e., the location and the transition type with respect to the other interlocutor's turn) of lexical and non-lexical items, laughter and head movements, as well as the phonetic variation of vocal feedback, in face-to-face dialogues in German.

We find that the feedback type influences the distribution and the variation of intensity values and voice quality, and, for multimodal items, it also influences the temporal alignment of the head movement with the vocal component.

Keywords: multimodal feedback, backchannels, head movements, laughter, German conversation

1. INTRODUCTION

In conversation, listeners tend to provide feedback to their conversational partner, i.e., they signal their active attention or react to what is being said, without interrupting the speaker's stream of talk [1, 2, 3, 4]. In face-to-face interactions, listener feedback phenomena, also called "backchannels" [1] or "listener responses" [5], are observed in both the vocal and the visual modalities, i.e., through verbal and non-verbal vocalizations, and through gestures.

Verbal feedback is constituted by lexical items or phrases with a very short constituent size [6]. In German, typical lexical feedback expressions are, e.g., "ja" "genau" "okay" "eben" "achso" [7]. Moreover, non-lexical vocalizations involving nasals, such as "mmm" or "mhm", are often also candidates for backchannels, in German as well as cross-linguistically [8]. Laughter is another non-verbal vocalization employed for backchanneling purposes, which often occurs in spontaneous interactions [9]. It has several functions in conversation, including linguistic ones (e.g., for discourse organization [10]). Previous work has shown that laughter is used as a feedback signal across languages, with the proportion of laughter out of the total backchannels varying between a few percentage points and almost a quarter of all backchannels [11, 12, 13].

Finally, head movements, such as head nods, are another feedback type occurring in face-to-face conversations [1, 14]. Being communicated via the visual channel, head movements are even less disruptive than vocal feedback and can potentially occur more often and in more locations then vocal ones without being judged inappropriate [15]. They can occur on their own, but are very often observed in co-occurrence with both verbal and non-verbal feedback expressions, or in their vicinity, often preceding vocalizations [16, 5].

Previous experimental research on listener responses in conversation has focused mostly on the identification and description of feedback inviting cues (e.g., [17, 18]), while less attention has been put on the phonetic features of feedback items and how they vary depending on their type and their placement in the ongoing talk [13]. In particular, while there is a growing body of research targeting multimodal backchannels (e.g., [4, 15]), non-verbal vocalizations tend not to be included in the investigations, so that not much is known about how and to what extent they differ from each other and from other types of feedback.

2. METHODS

2.1. Dataset

We investigated face-to-face dialogues between German native speakers using audio and video files

taken from the German subset of the multimodal DUEL corpus [19]. We analyzed 12 dyads involving 24 different speakers for a total of 2 hours and 15 minutes of dialogue. The recording setup for the DUEL corpus consisted of lapel microphones in front of each subject and two video cameras capturing the gesture space and the faces of the subjects. Participants were involved in different interaction scenarios designed to have them start talking without having to select a subject, while also allowing a free discussion, and to ensure the presence of laughter in different contexts. The scenarios analyzed for this study were "Film script" and "Dream apartment". The corpus includes the full transcription of the dialogues, their subdivision in conversational turns and utterances, as well as laughter occurrences.

2.2. Annotation

Listener feedback was identified in the dyads using the corpus' orthographic transcription, the audio and the video files. We consider as listener feedback all those lexical and non-lexical backchannels, laughter and head movements produced/performed by one of the interlocutors as an optional response to what the current speaker is saying (e.g., answers to questions are not considered as listener feedback) [17], and which are not part of a full-fledged turn (e.g., turn-opening, turn-closing items or discourse markers are not considered as listener feedback).

In Praat [20] we isolated lexical and non-lexical backchannels, and carried out a further annotation of laughter, in order to consider only those instances which corresponded to our feedback identification guidelines. The category of verbal feedback includes all the items that functioned as feedback, e.g., "ja", "genau", "okay", "das stimmt", while non-lexical feedback includes items such as "mhm". No form distinction will be used for the purposes of the current study. For the annotation of gestural feedback, we used the software ELAN [21] and the guidelines provided by the M3D annotation scheme [22] for head movements. First, all head movements were annotated using the video without the audio; then, in Praat, additionally considering the corresponding interlocutor's turns, a further annotation was carried out to isolate the movements which occurred as a stand-alone listener response, or in connection with a verbal, non-verbal or laughter type feedback.

Moreover, to investigate feedback distribution, we observed their location with respect to the current speaker's turn. If the feedback occurred completely in overlap with the other interlocutor's speech, it is considered "turn-internal", while if it occurred right after the interlocutor's turn end or it finished outside of it, followed by a silence, it is considered "turn-external" [3]. Finally, our study included a further analysis of turn-external feedback by categorizing the type of transition with which it occurred. Specifically, we labelled the transition type of feedback that occurred during an overlap with the other speaker's turn as "overlap," and feedback that occurred after a silent gap as "gap." [23, 24]. If the silent gap or portion of overlapped speech was less than 120 ms, the transition was considered "no-gap-no-overlap," as previous research has established 120 ms as the detection threshold for silences and overlaps [23].

2.3. Phonetic features

To investigate the phonetic variation of both laughter and verbal feedback we focus on the features of loudness and voice quality, using the values of maximum intensity (intmax) and the mean of the cepstral peak prominence (CPP). In contrast with other features, such as F0 and duration, intensity and voice quality have yet to be well investigated in feedback expressions. For lexical and non-lexical backchannels, [25] and [26] report that backchannels tend to be louder than the other affirmative words in American English, and that non-lexical items tend to be significantly less loud then short lexical expressions in Slovak. Using the parameter of jitter, [27] studied voice quality in backchannels in Ruruuli/Lunyala, finding higher values, correlated with lower periodicity, in longer lexical items, while non-lexical items displayed lower jitter. Intensity and voice quality have been studied also for laughter (e.g., [28]), with a higher intensity and a less modal production being found for laughter, compared to speech. For the current study, intmax and CPP were both extracted from the entire interval constituting either the feedback item (lexical, non-lexical, laughter) using a Praat script. Intensity values were normalized with the individual speaker's mean.

3. RESULTS

A total of 771 feedback items were extracted from the 12 dialogues. In terms of frequency, lexical items and head movements (without speech) are the most frequent feedback types in our data (with, respectively, 294 and 240 occurrences), followed by laughter (n=158), and lastly by non-lexical backchannels (n=79). Statistical analyses were carried out using R [29].

3.1. Head movements

A total of 462 head movements were extracted, including both those cases which constituted a feedback item on their own and those which accompanied vocalizations. The most common type of head movement with a feedback function is the nod (n=383, 83%), followed by head tilts (n=37, 8%). We also extracted a few occurrences of protrusions (n=32, 7%), which mostly accompany laughter, and head turns (n=10, 2%). Out of the 531 vocal feedback items, the 42% of the lexical backchannels are accompanied by a head movement (n=136), as well as the 56% of the non-lexical ones (n=44) and 27% of the laughs (n=42). The onset of the head movement tends to occur mostly before the vocalization, specifically 84 ms before lexical backchannels, 200 ms before non-lexical items and 140 ms before laughs, while the movement offset is observed 340 ms before the offset for laughter, and 140 ms and 172 ms after the end of the vocalization for lexical and non-lexical feedback respectively. Linear mixed models, with the speaker as a random effect, and subsequent Tukey post-hoc pairwise comparisons indicate a significant difference between the timing of the head offsets for laughter and lexical feedback (β = -.4691, t ratio = -4.696 p < .0001) and laughter and non-lexical feedback (β = -.5110, t ratio = -4.219, p = .0001), while the difference between the head offset timing for lexical and non-lexical item is not significant (β = -.0419, t ratio = -.418, p = .9084).

item	gap	nogap	overlap
lex	89(2.17)	25(0.44)	68(−2.17)
non	18(0.62)	10(2.22)	12(−1.68)
laugh	23(−2.12)	9(−0.75)	60(2.28)
head	37(1.19)	10(−1.19)	69(1.68)

Table 1: Count of observations (and residuals) of the χ^2 test for the transition type relative to the feedback item (lexical, non-lexical, laughter, head movement).

3.2. Distribution

With respect to the other interlocutor's speech, feedback constituted only by a head movement tend to occur mostly turn-internally, while all the other types of feedback are mostly observed turn-externally, with the exception of non-lexical backchannels, which are observed in almost equal measure turn-internally and turn-externally (see Fig.1). A χ^2 test confirms the existence of a relationship between the feedback type and the location in which it is produced by the listener (χ^2 = 11.099, df = 3, p = .0112). There is also a relationship between the feedback type and the transition with which it occur considering the final boundary of the other interlocutor's turn, i.e., with a gap, an overlap or a "no-gap-no-overlap" (χ^2 = 33.869, df = 6, p < .0001). The Pearson residuals for the test indicate that lexical feedback occurs significantly more frequently with a gap, laughter more frequently with an overlap, and non-lexical backchannels with no-gap-no-overlap transition (see Tab.1).

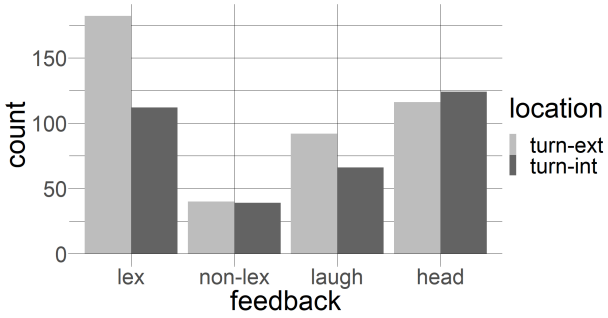

Figure 1: Distribution of feedback items (lexical, non-lexical, laughter, head movments) with respect to the interlocutor's current speaking turn (turn-external, turn-internal).

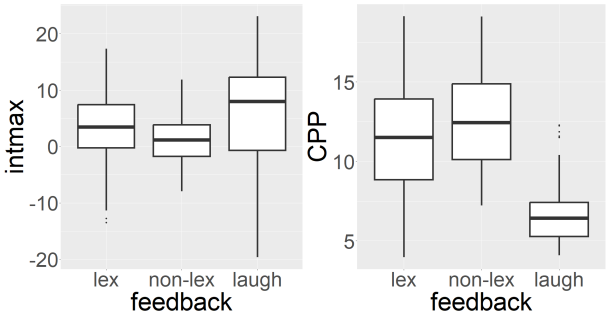

Figure 2: Maximum intensity (normalized with speaker's mean) and CPP values for vocal feedback items (non-lexical, lexical, laughter).

3.3. Phonetic variation

As for the phonetic features of the vocal feedback items, we find that laughter tends to have a higher intmax than the other types of backchannels, followed by lexical items and non-lexical items (see Fig.2). Conversely, mean CPP values are higher for non-lexical items, followed by lexical feedback and laughter (see Fig.2). Linear mixed models and Tukey post-hoc tests show that intensity values are

significantly different between lexical feedback and laughter (β = -2.75, t ratio = -4.062, p = .0002) as well as between non-lexical items and laughter (β = -4.59, t ratio = -5.132, p < .0001), while the difference is not significant between lexical and non-lexical items (β = 1.83, t ratio = 2.191, p = .0737). CPP values are significantly different between each pair in the post-hoc comparisons (lexical-non-lexical: β = -1.26, t ratio = -3.584, p = .0011; lexical-laughter: β = 4.94, t ratio = 17.305, p < .0001; non-lexical-laughter: β = 6.20, t ratio = 16.440, p < .0001).

Additionally, we tested if the presence of a head movement might also correlate with variation in the phonetic features of vocalizations. Except from some qualitative differences, feedback items with a co-occurring head movement do not have significantly distinct features from those where the movement is not present, with the exception of the intensity values for lexical items. In these, lexical items with a co-occurring head movement seem to be significantly louder than the unimodal ones, i.e., constituted only by the vocal element (β = -2.422, t ratio = -3.149, p = .0214).

4. DISCUSSION

We investigated the distribution and the acoustic realization of multimodal feedback in German face-to-face dialogues by observing lexical backchannels, head movements, laughter and non-lexical backchannels. Lexical items are the most frequent type of feedback in our dataset, followed by head movements. In line with previous studies (e.g., [4] for Dutch and [15] for British English), lexical backchannels tend to occur mostly turn-externally, and after a gap, while head movements, being the least disruptive type of feedback, occur more often turn-internally. Lexical items tend to be quieter than laughter, as previously observed for German by [28], but they are generally louder than non-lexical items. In terms of voice quality, they appear to be less periodic than non-lexical items, which could be motivated by their segmental content including voiceless segments or even short silences in complex lexical phrases.

Our results on head movements' temporal alignment with speech in multimodal feedback items show that the onset of the head movements precedes the vocalization by around 141 ms, which, as hypothesized by [16], is done by the listener to anticipate the start of a feedback response without interrupting the interlocutor. The alignment of the head movement with laughter is significantly different when compared to the other feedback types, since the offset of the movement occurs well before the laugh's end. This might be due to the fact that laughter is the type of feedback where the most protrusions are observed. A head protrusion would generally consist of a single movement of the head forward or backwards (and not in repeated back-to-back movements, as in e.g. head nods), so the head might tend to return to its resting position before the laughter is finished.

Moreover, we observe that multimodal lexical backchannels in our data appear to be louder than unimodal lexical ones. This might suggest a relation between the phonetic features of feedback items and head movements. For instance, [30] find that, for articulation rate and nodding, increased effort in speech production is accompanied by increased head movement. Further tests (e.g, logistic regression models) will be carried out on our data to find out whether this is the case for multimodal feedback as well.

Finally, we find that the two types of non-verbal vocalizations present some substantial differences from each other, both in terms of their distribution and their phonetic features. While non-lexical backchannels tend to occur at the same rate turn-internally and turn-externally, and are produced after a smooth transition at the end of the other interlocutor's turn, laughs mostly occur turn-externally, and often after a portion of overlapped speech at the end of the current speaker's turn.

5. CONCLUSIONS

With the current analysis we provided a description of the distributional features of different types of feedback, both unimodal and multimodal, and we tested how different factors, such as the feedback's form, location, its transition type and the presence of a co-occurring head movement, influence the acoustic realization of vocal and multimodal listener responses. Future research should include, e.g., the lexical and segmental content of the verbal backchannels to provide a more detailed description of their variation, and the phonetic features of the current speaker's turn corresponding to a listener response, in order to observe if determinate cues would elicit a specific type of feedback.

6. ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Council), project no. 444631148 for MR and MZ and project no. 461442180 for MS and BL.

7. REFERENCES

[1] V. H. Yngve, "On getting a word in edgewise," *CLS-70*, pp. 567–578, 1970.

[2] S. Duncan, "On the structure of speaker-auditor interaction during speaking turns," *Language in Society*, vol. 3, pp. 161–180, 1974.

[3] G. Kjellmer, "Where do we backchannel? On the use of mm, mhm, uh huh and such like," *International Journal of Corpus Linguistics*, vol. 14, pp. 81–112, 2009.

[4] K. Truong, R. Poppe, I. de Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogues," in *Proc. Interspeech 2011*, 2011, pp. 23–25.

[5] S. K. Maynard, "Conversation management in contrast: Listener response in Japanese and American English," *Journal of Pragmatics*, vol. 14, no. 3, pp. 397–412, 1990.

[6] F. E. Müller, "Affiliating and disaffiliating with continuers: prosodic aspects of recipiency," in *Prosody in Conversation: Interactional Studies*, M. Couper-Kuhlen, E. Selting, Ed. Cambridge University Press, 1996, pp. 131–176.

[7] B. Knudsen, A. Creemers, and M. A.Sn., "Forgotten little words: How backchannels and particles may facilitate speech planning in conversation?" *Frontiers in Psychology*, vol. 11, pp. 593–671, 2020.

[8] A. Liesenfeld and M. Dingemanse, "Bottom-up discovery of structure and variation in response tokens (backchannels) across diverse languages," in *Proc. Interspeech 2022*, 2022, pp. 1126–1130.

[9] J. Trouvain and K. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *Proc. of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 2012, pp. 36–39.

[10] B. Ludusan and B. Schuppler, "To laugh or not to laugh? the use of laughter to mark discourse structure," in *Proc. SIGDIAL*, 2022, pp. 76–82.

[11] S. Pammi and M. Schröder, "A corpus based analysis of backchannel vocalizations," in *Proc. Interdisciplinary Workshop on Laughter and other Interactional Vocalisations in Speech. Berlin, Germany*, 2009.

[12] P. Cutrone, "A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners," *Intercultural Pragmatics*, vol. 11, no. 1, pp. 83–120, 2014.

[13] V. Krepsz, V. Horváth, Á. Hámori, D. Gyarmathy, and C. I. Dér, "Backchannel responses in Hungarian conversations: a corpus-based study on the effect of the partner's age and gender," *Linguistica Silesiana*, pp. 113–140, 2022.

[14] E. McClave, "Linguistic functions of head movements in the context of speech," *Journal of Pragmatics*, vol. 32, no. 7, pp. 855–878, 2000.

[15] G. Ferré and S. Renaudier, "Unimodal and bimodal backchannels in Conversational English," in *Proc. SEMDIAL 2017*, 2017, pp. 27–37.

[16] A. T. Dittmann and L. G. Llewellyn, "Relationship between vocalizations and head nods as listener responses," *Journal of Personality and Social Psychology*, vol. 9, no. 1, pp. 79–84, 1968.

[17] N. Ward and W. Tsukahara, "Prosodic features which cue backchannel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[18] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[19] J. Hough, Y. Tian, L. de Ruiter, S. Betz, S. Kousidis, D. Schlangen, and J. Ginzburg, "DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter," *Proc. LREC' 16*, pp. 1784–1788, 2016.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2022. [Online]. Available: http://www.praat.org

[21] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proc. LREC 2006*, 2006.

[22] P. L. Rohrer, I. Vila-Giménez, J. Florit-Pons, G. Gurrado, N. Esteve Gibert, A. Ren, S. Shattuck-Hufnagel, and P. Prieto, "The multimodal multidimensional (M3D) labelling scheme for the annotation of audiovisual corpora," in *Proc. GESPIN 2020*, 2020.

[23] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *Journal of the Acoustical Society of America*, vol. 130(1), pp. 508–513, 2011.

[24] M. Rossi, K. Feindt, and M. Zellers, "Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish," in *Proc. Speech Prosody 2022*, 2022, pp. 185–189.

[25] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proc. ICPhS XVI*, 2007, pp. 1065–1068.

[26] S. Benus, "The prosody of backchannels in Slovak," in *Proc. Speech Prosody 2016*, 2016, pp. 75–79.

[27] M. Zellers, "An overview of forms, functions, and configurations of backchannels in Ruruuli/Lunyala," *Journal of Pragmatics*, vol. 175, pp. 38–52, 2021.

[28] B. Ludusan and P. Wagner, "ha-HA-hha? Intensity and voice quality characteristics of laughter," in *Proc. Speech Prosody 2022*, 2022, pp. 560–564.

[29] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[30] M. Tiede, C. Mooshammer, and L. Goldstein, "Noggin nodding: Head movement correlates with increased effort in accelerating speech production tasks," *Frontiers in Psychology*, vol. 10, 2019.