Prosodic Characteristics of Feedback Expressions in Distracted and Non-distracted Listeners

Zofia Malisz[*], *Marcin Włodarczak*[*], *Hendrik Buschmeier*[†], *Stefan Kopp*[†], *Petra Wagner*[*]

[*]Faculty of Linguistics and Literary Studies
[†]Sociable Agents Group, CITEC and Faculty of Technology
Bielefeld University, Bielefeld, Germany
{zofia.malisz,petra.wagner,mwlodarczak}@uni-bielefeld.de
{hbuschme,skopp}@techfak.uni-bielefeld.de

Abstract

In a previous study [1] we investigated properties of communicative feedback produced by attentive and non-attentive listeners in dialogue. Distracted listeners were found to produce less feedback communicating understanding. Here, we assess the role of prosody in differentiating between feedback functions. We find significant differences across all studied prosodic dimensions as well as influences of lexical form and phonetic structure on feedback function categorisation. We also show that differences in prosodic features between attentiveness states exist, e.g., in overall intensity.

Index Terms: communicative feedback; prosody; dialogue; distraction; engagement; attention

1. Introduction

In spoken dialogue the behaviour of the interlocutor who is currently listening is characterised by short feedback signals (e.g., "uh-huh", "m", "yeah", "okay"). These signals minimally communicate presence, perception, understanding, acceptance as well as higher feedback functions such as agreement and attitudinal reactions to the speaker [2]. Feedback signals play an important role in grounding and coordination of the interaction as they allow listeners to inform speakers of their state of perception, understanding, etc. without interrupting the ongoing turn. At the same time speakers can estimate online how successful their utterance has been in communicating the intended message via received feedback. Therefore, feedback is used to adapt communicative behaviour to the listener's needs.

As a result, communication becomes difficult when feedback is inappropriately timed or expressed. In [3], listeners were induced to produce less context-specific feedback, which had a substantial influence on the speakers' behaviour and the quality of their storytelling. Similarly, in [4], speakers told more vivid stories when they expected an attentive listener and in fact interacted with one. Speakers also spent more time telling their stories when their expectations of listeners' attention states matched reality. Both studies showed that distractedness in listeners had an influence on speakers and their behaviour.

Conversational situations exist, where listeners are being distracted by simultaneous tasks (browsing the Internet, reading documents, etc.) or disengaged for other reasons. Speakers are then mostly able to notice that their dialogue partners are distracted and change their communicative behaviour accordingly. As [4, p. 582] note, speakers are "painfully aware when their conversational partners [. . .] are inattentive, and they can often tell when their partners are only pretending to pay attention." Consequently, it is reasonable to assume that distractedness manifests in the listeners' communicative behaviour in general and their feedback behaviour in particular. Dialogue partners should be able to perceive if listeners' behaviour deviates from the one expected of a fully engaged interlocutor.

Important engagement indicators are the timeliness and frequency of feedback signals in response to feedback elicitation cues produced by the speaker (e.g., [5]). In a multimodal context, listeners display mutual, joint and shared attention with gaze [6]. General presence, liveliness and readiness to cooperate is also signalled with posture shifts, appropriate head movements and manual gestures [7, 8].

The influence of prosody on the pragmatic function of feedback utterances has been a subject of study for some time. Syllabification, duration, loudness, pitch slope and pitch contour were identified as relevant for the discrimination of functional feedback categories in English [9]. A more detailed analysis in [10] found that English affirmative cue words are higher in pitch, intensity and pitch slope when used as a backchannel. Backchannels were also longer in duration, produced with shorter latencies and often preceded by a pitch rise in the interlocutor's speech. A study of Japanese backchannels, however, found that prosodic features marking interest and surprise vary depending on the backchannel's lexical realisation [11]. For German, a cluster analysis of the relation between the linguistic function and intonational form of the discourse particle "hm"revealed prototypical and functionally equivalent variants [12]. [13] used synthesised instances of the German backchannel "ja" with durational features and F_0 curves modelled after [14]. Subjects were asked to evaluate the backchannels along seven semantic dimensions (e.g.: *happy* vs. *sad*). The analysis identified prosodic features related to agreement, happiness, boredom, etc.

In previous work [1], we studied the distribution of functional feedback categories between distracted and attentive listeners in a dialogue corpus collected on the basis of the paradigm in [3]. We found that distracted listeners produced less feedback communicating understanding than attentive listeners. In the present paper, we analyse prosodic characteristics of the three most frequent feedback expressions in our corpus ("ja", "m", and "mhm") across their pragmatic functions. We also examine the differences in feedback produced by distracted vs. attentive listeners.

The first three authors contributed to the paper equally.

2. Data collection

To gather reliable data on feedback behaviour of distracted and attentive listeners, we carried out a lab-based face-to-face dialogue study. One of the dialogue partners (the 'storyteller') told two holiday stories to the other participant (the 'listener'), who was instructed to listen actively, make remarks and ask questions.

Listeners were engaged in a distraction task during either the first or the second story. Building upon the paradigm of [3], we instructed listeners to press a button on a hidden remote control every time the dialogue partner produced a word starting with the letter 's' (the second most common German word-initial letter usually corresponding to perceptually salient sibilants). In addition, they had to count the total number of 's-words'. Storytellers were informed that their partners would be listening for something in the dialogue, but they did not know during which of the two stories.

Participants were seated approximately three metres apart to minimise crosstalk. Close talking high-quality headset microphones were used. Furthermore, another microphone captured the whole scene and a fourth audio channel was used to record the 'clicks' synthesised by a computer when listeners pressed the button on the remote control. Interactions were recorded from three camera perspectives: medium shots showing the storyteller and the listener and a long shot showing the whole scene.

A total of fifty students (34 female and 16 male native speakers of German) were recruited at Bielefeld University to participate in the study, receiving either course credit or 4 euro as payment. They were assigned to one of 25 same-sex dyads. In all but four participant pairs, dialogue partners were unacquainted.

3. Annotation

Many annotation schemes distinguish only between two broad feedback function categories such as 'generic' vs. 'specific' [3].

For our more detailed analysis of listeners' behaviour, an annotation scheme distinguishing between subtler pragmatic variants of feedback signals was needed. The annotation scheme, discussed in detail in [1], is based largely on the framework of [2, 15] ascribing up to four basic functions to feedback signals: *contact, perception, understanding* and *attitudinal reactions*. In communicating one of these functions, listeners express their willingness and ability to continue the interaction, perceive or understand the message or to respond to it. In the present work, we focus on the three affirmative functions, named P1, P2 and P3, where P1 can be seen as what is usually called a backchannel or a 'continuer'. Category P2 signals successful interpretation of the message, and category P3 indicates acceptance, belief and agreement. These levels can be treated as a hierarchy with increasing value of judgement and 'cognitive involvement' or 'depth' of grounding. See Table 1 for an overview.

Feedback utterances in 14 sessions (i.e., 28 dialogues) were segmented and transcribed according to German orthographic conventions (where existent). A total of 1003 feedback functions were annotated, each independently by three annotators taking communicative context into account. Majority labels between annotators were calculated automatically and problematic cases (110; roughly 10%) were discussed and resolved.

4. Feature extraction and analysis

The 28 annotated dialogues have a total length of 180 minutes and each dialogue has a mean length of 6:25 minutes (Min = 2:16; Max = 14:29; SD = 2:31). On average 36 feedback signals were produced per dialogue (Min = 7; Max = 93; SD = 23.1).

Table 1: A subset of the feedback functions inventory. A detailed description can be found in [1].

C	Definition of category
P1	The partner signals perception of the signal. *'I hear you and please continue.'*
P2	The partner signals perception and understanding of the message content. *'I understand what you mean.'*
P3	The partner signals perception, understanding and acceptance of the message or agreement with the message. *'I accept / agree / believe what you say.'*

Duration in milliseconds was calculated for each feedback signal. Pitch and intensity values were extracted using Praat[1]. In order to avoid tracking errors, pitch was extracted in two steps with the floor and ceiling values for the second run set at the 15th percentile times 0.83 and the 65th percentile times 1.92 of the values in the initial run [16]. All measurements were converted to z-scores to normalise the differences between dialogues.

We calculated mean, standard deviation, and slope of pitch and intensity in each feedback signal. Next, we split each feedback signal into three parts of equal length and calculated the mean and standard deviation for each of these parts. Similarly, slopes (from linear regression) were calculated over the first and second half. The procedure yields the following features for each feedback signal: i) dialogue act label, ii) orthographic transcription, iii) duration, iv) mean.[pitch, intensity], v) sd.[pitch, intensity], vi) slope.[pitch, intensity], vii) segment.[1,2,3].mean.[pitch, intensity], viii) segment.[1,2,3].sd.[pitch, intensity], ix) segment.[1,2].slope.[pitch, intensity].

Two separate analyses using Generalised Linear Mixed Models (GLMM) were conducted for (a) feedback function differences and (b) distractedness-related differences. A dataset was used with expressions "ja", "mhm" and "m" combined. The prosodic feature vector exhibited high collinearity even after centring and scaling of the variables. Since high correlations between variables influence the validity of regression estimates for individual predictors, we performed a Principal Component Analysis to deal with multicollinearity. The procedure reduced the feature vector from 23 to 9 dimensions, which were chosen according to the cumulative level of variance explained by the components, here set at 0.94. The Varimax-rotated components were entered into the GLMMs.

A GLMM (with cumulative logit link function) was fitted with Feedback Function (P1, P2 and P3) as a dependent ordinal multinomial variable[2]. The variable Feedback Expression was entered as a fixed factor and in an interaction term with all prosody-based components to account for variability that is due to the phonetic structure of the different expressions. Other fixed factors included Task Order and Experimental Condition. The only random effect entered was Session, equivalent to speaker differences in laboratory designs.

A second GLMM (with logit link function) was fitted with Experimental Condition (distracted vs. non-distracted) as a binomial dependent variable[3]. All other terms were specified as in the model above with the exception of Feedback Function included here as a fixed factor.

[1] `http://www.fon.hum.uva.nl/praat/`

[2] Using the GENLINMIXED command in IBM SPSS Statistics 20.0.

[3] Using the lme4 R package, version 0.999375-42.

Table 2: The first nine components (accounting for 92% of the variance in the dataset; ordered by standardised loadings and proportional variances) alongside the prosodic features with high loadings on each component.

RC_i	Load	Var	Prosodic feature	Load
RC_1	3.60	0.16	segmented.slope.pitch.2	0.92
			segmented.sd.pitch.3	0.84
			segmented.mean.pitch.3	0.75
			slope.pitch	0.75
			sd.pitch	0.71
RC_2	3.24	0.14	mean.intensity	0.98
			segmented.mean.intensity.1	0.86
			segmented.mean.intensity.2	0.85
			segmented.mean.intensity.3	0.77
RC_7	2.90	0.13	segmented.mean.pitch	0.92
			mean.pitch	0.89
			segmented.mean.pitch.1	0.83
RC_3	2.44	0.10	sd.intensity	0.94
			segmented.sd.intensity.3	0.81
RC_4	2.28	0.10	segmented.slope.pitch.1	−0.90
			segmented.sd.pitch.1	0.87
RC_5	2.14	0.09	slope.intensity	0.92
RC_8	1.92	0.08	segmented.slope.intensity.1	−0.87
			segmented.sd.intensity.2	0.70
RC_9	1.35	0.06	segmented.sd.pitch.2	0.84
RC_6	1.23	0.04	duration	0.92

The statistically significant components resulting from each model were interpreted in terms of prosodic features with high loadings on a component (tabulated in Table 2). Thresholds for choosing a feature as relevant were set at the point of clear discontinuity within each component. Notably, all components are interpretable in terms of disjoint and coherent feature sets.

5. Results

5.1. Differences in prosody between feedback functions

Table 3 presents the main effects and interactions found to significantly differentiate between feedback functions in the GLMM described in Section 4.

Following the proportional odds assumption of multinomial ordinal logistic regression, we can interpret the proportional odds of choosing lower categories against higher categories given any partitioning category, i.e., P1 versus P2 and P3 combined, and P1 and P2 combined versus P3. Consequently, as RC_1 increases by one unit, the odds of choosing lower categories increases by 1.325. By contrast, a one-unit change in components RC_2, RC_9, RC_6 decreases the odds of choosing lower categories by 0.793, 0.685 and 0.770 respectively. Moving from "mhm" to "ja" decreases the odds of choosing lower categories. For RC_4 the odds of choosing lower categories depend on the lexical form and decrease by 0.563 for "ja" and increase by 2.002 for "mhm."

Interpreting components in term of prosodic features (Table 2), RC_1 expresses pitch variability especially in the last part of the expression (recall that slope values were calculated for the expression cut in two segments; the other values, SD and mean were calculated for the expression cut in three segments). RC_2 expresses mean intensity. The next component which had a significant main effect on prosodic categorisation of Feedback Function was RC_9, reflecting the variability of pitch in the middle of the expression. RC_6 is essentially duration. RC_4, i.e., pitch variability and the magnitude of the variability (slope) in the first part of the expression interact with the phonetic form of the expression itself. In other words, the effect of RC_4 depends on the particular expression e.g. "ja" and "mhm" or "m". At the same time there is a main effect of the feedback expression form on the feedback function classification, where "ja" highly significantly distinguishes between the categories.

Table 3: Fixed coefficients of the multinomial GLMM for Feedback Function (reference category: P1). LO: log odds; PO: proportional odds; SE: standard error; significance codes: 0.05: $*$, 0.01: $**$.

Model term	LO	PO	SE	t	p
$RC_1 *$	0.281	1.325	0.122	2.306	0.022
$RC_2 *$	−0.232	0.793	0.101	−2.305	0.022
$RC_9 *$	−0.378	0.685	0.117	−3.224	0.001
$RC_6 *$	−0.262	0.770	0.109	−2.404	0.017
"ja" $*$	−1.799	0.165	0.333	−5.398	0.000
"m"	−0.130	0.878	0.284	−0.458	0.647
$RC_4 \times$"ja" $*$	−0.574	0.563	0.223	−2.573	0.010
$RC_4 \times$"m"	−0.050	0.951	0.146	−0.342	0.733
$RC_4 \times$"mhm" $**$	0.694	2.002	0.264	2.633	0.009

5.2. Differences in prosody between conditions

Table 4 presents model estimates of predictors found to differentiate between feedback signals produced by listeners in the distracted and non-distracted condition. The value of the C index measures the concordance between the predicted probability in the model and the observed response. From a value of $C = 0.8$ a model exhibits real predictive power, given the subtle phenomena we are dealing with here, our value of $C = 0.77$ is strong.

The estimate for Feedback Function as a predictor of attentiveness confirms our previous results [1], where a decrease in the frequency of signalling understanding (function P2) was found in distracted listeners. The present model predicts that a unit increase in the P2 function increases the odds of the listener being attentive by 1.716.

As far as prosodic correlates of attentiveness are concerned, the results show that RC_2 (defined by the overall mean intensity measures) is highly significant; attentive speakers tend to speak more loudly. Energy is also less variable in the non-distracted case (RC_3 estimate). RC_1 defined by pitch variability measures is positively related to attentiveness (one-unit increase in RC_1 increases the odds of an attentive state by 1.823). It can be expected that feedback delivered with less intonational variability is indicative of lower engagement.

Significant interactions between the phonetic form of the expression and some of the pitch and intensity measures on the one hand allow for the fine-tuning of potential recognition of attention states in particular expressions and on the other hand confirm the differences depending on phonetic structure.

6. Discussion

The results presented in the previous section indicate that prosody plays a role in distinguishing between different functions of communicative feedback. However, the significant interactions with the lexical form in our results suggests that it is import-

Table 4: Fixed coefficients of the binomial GLMM for Condition. LO: log odds; PO: proportional odds; SE: standard error; significance codes: 0.05: $*$, 0.01: $**$, 0.001: $***$. Predictive strength measures: C = 0.77, $D_{xy} = 0.55$.

Model term	LO	EO	SE	z	p
"m"	0.451	1.570	0.365	1.237	0.216
"mhm"	−0.341	0.711	0.396	−0.862	0.389
RC_1*	0.600	1.823	0.284	2.113	0.035
RC_3*	−0.565	0.568	0.244	−2.319	0.020
RC_4**	−0.853	0.426	0.319	−2.672	0.007
RC_2***	0.472	1.603	0.114	4.144	0.000
RC_6*	−0.260	0.772	0.119	−2.181	0.029
$P2*$	0.540	1.716	0.255	2.120	0.034
P3	0.161	1.175	0.381	0.422	0.673
order$*$	1.109	3.031	0.442	2.507	0.012
$RC_1\times$"m"$**$	−0.965	0.381	0.332	−2.902	0.004
$RC_1\times$"mhm"	−0.517	0.596	0.347	−1.491	0.136
$RC_3\times$"m"	0.437	1.548	0.303	1.441	0.149
$RC_3\times$"mhm"$*$	0.774	2.169	0.316	2.448	0.014
$RC_4\times$"m"$*$	0.838	2.312	0.354	2.365	0.018
$RC_4\times$"mhm"$*$	0.861	2.365	0.403	2.137	0.033

ant to take the phonetic structure of particular expressions into account: prosodic features may strongly depend on segmental structure, e.g.: nasality vs. orality in "m" vs. "ja" and syllabic structure in "mhm" vs. monosyllabic expressions such as "ja". Consequently, in addition to general prosodic features distinguishing between feedback functions there are strategies specific to both the feedback function and the particular expression.

Concerning the prosodic characteristics of distracted listeners that could be detected instrumentally and, possibly, also by the interlocutors, some features that might help detect the listener attention state were found. Additionally, the frequency of signalling understanding [1] remains a consistent non-prosodic cue to distractedness in the listener.

7. Conclusions and future work

This work reported on the prosody of German feedback expressions "ja", "mhm" and "m" in a dialogue corpus where listeners' attention was manipulated by an ancillary task. In this study we have taken first steps towards prototypical prosodic profiles of German feedback functions and diverse predictors of attention. However, the complex interaction between prosody, feedback expressions and pragmatic functions needs to be disentangled possibly using more data and automatic classification techniques.

Findings like these could inform automatic methods for detecting listener's attentive states and the communicative intentions they convey via feedback. This information could be used by artificial conversational agents such as spoken dialogue systems to adapt their communicative behaviour to the needs and expectations of the user.

Acknowledgements – This research is supported by the Deutsche Forschungsgemeinschaft (DFG) at the Center of Excellence in 'Cognitive Interaction Technology' (CITEC) as well as at the Collaborative Research Center 673 'Alignment in Communication'. We thank anonymous reviewers for their helpful comments.

8. References

[1] H. Buschmeier, Z. Malisz, M. Włodarczak, S. Kopp, and P. Wagner, "'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness," in *Proceedings of INTERSPEECH 2011*, Florence, Italy, 2011, pp. 2057–2060.

[2] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.

[3] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of Personality and Social Psychology*, vol. 79, pp. 941–952, 2000.

[4] A. K. Kuhlen and S. E. Brennan, "Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling," *Discourse Processes*, vol. 47, pp. 567–587, 2010.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[6] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents*, Kos, Greece, 2005, pp. 229–240.

[7] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, pp. 140–164, 2005.

[8] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue*, London, UK, 2009, pp. 225–234.

[9] N. Ward, "Pragmatic functions of prosodic features in non-lexical utterances," in *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004, pp. 325–328.

[10] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007, pp. 1065–1068.

[11] T. Kawahara, Z.-Q. Chang, and K. Takanashi, "Analysis of prosodic features of Japanese reactive tokens in poster conversations," in *Speech Prosody 2010*, Chicago, IL, 2010, pp. 1–4.

[12] J. E. Schmidt, "Bausteine der Intonation?" in *Neue Wege der Intonationsforschung*, ser. Germanistische Linguistik, J. E. Schmidt, Ed. Hildesheim, Germany: Georg Olms Verlag, 2001, vol. 157-158, pp. 9–32.

[13] T. Stocksmeier, S. Kopp, and D. Gibbon, "Synthesis of prosodic attitudinal variants in German backchannel "ja"," in *Proceedings of Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1290–1293.

[14] K. Ehlich, *Interjektionen*. Tübingen, Germany: Max Niemeyer Verlag, 1986.

[15] S. Kopp, J. Allwood, K. Grammar, E. Ahlsén, and T. Stocksmeier, "Modeling embodied feedback with virtual humans," in *Modeling Communication with Robots and Virtual Humans*, I. Wachsmuth and G. Knoblich, Eds. Berlin: Springer-Verlag, 2008, pp. 18–37.

[16] C. De Looze and S. Rauzy, "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration," in *Proceedings of INTERSPEECH 2009*, Brighton, UK, 2009, pp. 2919–2922.