In: *Proceedings of the 16th International Conference on Intelligent Virtual Agents, (IVA 2016)*, pp. 96–109, Los Angeles, CA, USA. doi:10.1007/978-3-319-47665-0_9
The final publication is available at http://link.springer.com/

This is what's important – using speech and gesture to create focus in multimodal utterance

Farina Freigang and Stefan Kopp

Social Cognitive Systems Group, Faculty of Technology
Center of Excellence "Cognitive Interaction Technology" (CITEC)
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld, Germany
farina.freigang@uni-bielefeld.de** and skopp@techfak.uni-bielefeld.de

Abstract. In natural communication, humans enrich their utterances with pragmatic information indicating, e.g., what is important to them or what they are not certain about. We investigate whether and how virtual humans (VH) can employ this kind of meta-communication. In an empirical study we have identified three modifying functions that humans produce and perceive in multimodal utterance, one being to create or attenuate focus. In this paper we test whether such modifying functions are also observed in speech and/or gesture of a VH, and whether this changes the perception of a VH overall. Results suggest that, although the VH's behaviour is judged rather neutral overall, focusing is distinctively recognised, leads to better recall, and affects perceived competence. These effects are strongest if focus is created jointly by speech and gesture.

1 Introduction

In natural communication, humans do not only transport propositional meaning. They add many signals to *modify* this message in order to help the addressee arrive at the correct interpretation of the speaker's intended meaning. However, albeit its prominence and importance, this meta-communication has not received much attention so far. A special role plays nonverbal communication [1], which speakers use to subtly indicate, e.g., what is important to them, or what their stance or epistemic state is about a fact. The synchronization of speech and gesture plays a key role in forming this multimodal utterance [2]. We are interested in how speech and gesture work together in such modifications in multimodal utterance.

Gestures contribute to the meaning of an utterance not only by *adding* information (semantics) but also by modifying the gestural or verbal content on a pragmatic level. In this case, the gesture may carry a *modifying function* (MF), which we investigated in previous work [3]. We created a corpus of natural communicative gestures and body movements and conducted a video rating study. Participants evaluated video snippets of multimodal utterances in two conditions: speech-and-gesture and gesture-only (with muted speech and cropped

** Corresponding author

head). The utterances were evaluated in terms of 14 adjectives assumed, first, to be intuitively understandable and, second, to correspond to the range of possible combined meanings that can be related back to specific MF, developed within the research of gesture studies building on work by [4–7]. Results show that index-finger-pointings are perceived to emphasise and affirm an uttered content. Brushing gestures change the utterance in a discounting or downtoning way. In further work [8], we conducted an exploratory factor analysis of the ratings of 14 adjectives, in order to analyse the underlying structures. Three main factors were found in the gesture-only condition. One distinct factor with high positive adjective loadings (.981 to .719) relates to **positive focusing** (or highlighting), a second with relatively high negative adjective loading (-.934) corresponds to **negative epistemic** functions (marking uncertainty). Another adjective linked to remaining factors was identified as **negative focusing**. The adjectives supported by the factors suggest which MF a gesture may be associated with.

With this work, we tackle two main issues. First, we want to gain insights into the role of speech and gesture in conveying pragmatic information (MF). Is the meanining understood and, in particular, what do MF in gesture contribute to it? We can measure this by assessing the recognition of MF and the recall of what a virtual human (VH) says in different modalities. Based on previous work [3] we expect that gestures with a highlighting function are particularly well recognised, and we hypothesise that highlighted messages are better remembered than downtoning or uncertain messages. Further, we hypothesise to get the strongest effect when MFs are conveyed in both speech and gesture, followed by speech, gesture, and a non-modified utterance. A second research question is how a VH is perceived more generally when its multimodal expressiveness is augmented with pragmatic aspects. After all, our aim is to create more communicative and accessible VH. In order to investigate these research questions, we synthesizing particular gestural behaviour and add it to specific verbal material. Note that modifications, and meta-communication more generally, are hardly conventionalized, standardized, nor clearly marked. In contrast, this information is often highly ambigue, vague, and subjectively interpreted. Hence, studying these questions raises considerable methodological challenges. In the next section, we explain how we arrived at the present experiment design (using a number of pretests). After describing the study procedure in Section 3, we present and discuss results in Section 4.

2 Experiment Design

On the basis of the insights that we gained from the factor analysis [8], we designed the current experimental setup. The first question at hand was **which gestures** should be tested? We decided on the following procedure. As described above, some adjectives were particularly meaningful for the factors of our MF. Thus, we filtered out the adjectives that had the highest ratings (1.4 to 2.8 on a 7-point Likert scale), retaining the three MF under discussion. The following categories emerged: the adjectives *affirmative*, *emphasising*, *focusing*, *opinion-*

ative, *classifying* and *relevant* represent the positive focusing (or highlighting) function (`Foc+`), the adjective *discounting/downtoning* represents the negative focusing function (`Foc-`), and the adjective *uncertain* stands for the negative epistemic function (`Epi-`). As a result of matching the adjectives back to the corresponding video snippets, ten videos could be selected which represent the three MF, each depicting a distinct gesture: pointings for `Foc+`, brushings for `Foc-` and palm up open hand gestures for `Epi-` (for examples, cf. Figure 1). A particular strength of this work is that the gestures with associated MF, which we wanted to test in a VH, are selected on the basis of empirical findings.

The second issue for the experimental setup was the **stimulus context**. Since this is a first test of gestures with MF, we did not plan a human-VH interaction, but just the presentation of videos of a VH. We designed a short story[1] in which our VH called Billie narrates about his life as a virtual character. The story was designed to have three parts, each with one topic. The parts were designed to be long enough to have an effect on the observer, as well as brief enough as not to become tedious. In the first story Billie talks about VH and his research institute, in the second one he talks about himself and for which reasons VH are used, and in the third story, Billie talks about the technical details of the software architecture underlying his behaviour. Each text contains 100 words (+1/−2), is structured into eight sentences, and was written in a neutral tone. In total the short story consists of 24 sentences and 299 words. As described below, the text was later enhanced by expressions for each category of MF.

Concerning the topic of MF, so far, we solely considered natural data of human interactions dripping with naturalness and modifications of all kind. The aspect of naturalness will be discussed further down (natural VH). The following paragraphs, first, will deal with the third issue of the experimental setup, namely, the **application of MF** in utterances. Our MF in gesture are taken from our corpus and are implemented in an VH, thus they are scripted gestures which can be easily controlled. In order to test what is in the gesture and what is in the speech of a VH, and since we cannot assume that a gesture on its own can convey the intended meaning, we needed modifications in speech to make a condition more obvious for a naïve observer. Possible linguistic options of modification include words choice, intonation, sentence structure and speech acts, among others. The control over prosody would be a desired modification, since prosodic and gestural highlighting may highly correlate in natural human interactions, and we would like to investigate this aspect in future work. To remain in control, we use particular words as markers for our MF (for the final version of "keywords" cf. Section 3). Different view points exists on which lexemes highlight, understate or make an utterance uncertain, and only few of have studied words connection to gesture. One example is [9], who discusses the relationship between modal particles and gestures in German. Opposite to his approach, we do not analyse the co-occurrence of speech and gesture in humans, but investigate modal particles and sentential adjectives that best match our MF in gesture

[1] The short story can be accessed at https://pub.uni-bielefeld.de/publication/2903503

in a VH. Since we do not want to rely solely on the existing (and in parts quite theoretical) literature, we tested the words in two pretest iterations.

In the first iteration of testing, the designed short story was enhanced with keywords, which we considered to have a MF similar to our MF in gestures. We collected keywords for three conditions (`Foc+`, `Foc-`, `Epi-`) and no keywords were added for a neutral condition. In this pretest only every other sentences included a keyword and only those sentences were tested, plus the same sentences in the neutral condition. The three parts of the short story were chunked sentence-wise and recorded with the synthesised voice of our VH Billie. The final 48 audio files of all four conditions had a duration of 6 to 19 seconds and were randomly ordered in a SoSci Survey [10] questionnaire (with which also the second pretest and the final experiment were conducted). The test was presented to three participants, partly aware of the research question, who classified the utterances of the audio playbacks to be one of the following: "Billie's utterance is ..." *emphasising*, *understating*, *uncertain*, or *neutral*. The options were visible during the playback. In order for a word to be accepted, two of three persons had to match the utterance to the correct category, i.e., any of the three MF or neutral. As a result, we kept half of the words (18) and replaced the others (18; 12 were neutral).

In a second pretest iteration, the three parts of the short story were each played in one piece and in all four conditions, accumulating to a rating test of 12 audio playbacks. Each playback was between 49 and 69 seconds long and presented in a randomised order. Again, participants had to rate the utterances of the audio playbacks according to the four options which were visible during the time of playback. At the end of this pretest, however, detailed questions about the sentences were asked on three pages. On one page, all sentences in one condition were presented in written form and the participants were asked which of the utterances were *not* either *emphasising - highlighting - focussing* or *discounting - understating - defocussing* or *uncertain*, depending on which page they were on. Eight subjects unfamiliar with the research questions participated in this pretest. The results of the audio tests (cf. Table 1) show that in the neutral condition, 67% were correctly identified and the rest were rated `Foc+`. In the `Foc+` condition, 71% of the ratings were correct and the rest were rated as neutral. In the `Foc-` condition, only 42% were rated correctly, the same amount rated neutral and the rest `Foc+`/`Epi-`. In the `Epi-` condition, 50% matched the desired category, and the rest was quite random: 21% were rated neutral, 17% `Foc+` and 12% `Foc-`. In conclusion, all conditions included neutral ratings (in total, 40% of the 96 ratings), in the neutral condition (and slightly `Foc-` and `Epi-`) some focusing elements were observed, and the `Epi-` condition was the one with the highest variance. These results indicate that even in spoken/written language there is a lot of ambiguity regarding words with modification. But when looking at the big picture, the `Foc+` and neutral conditions are rather clear. The results of the second part, the written-sentences test, indicated which keywords needed to be changed in order to give a sentence a certain modification. In order for a keyword to be changed, at least two of the eight participants had to state

a misfit. Five, four and one keywords were changed in the `Foc+`, `Foc-`, and `Epi-` conditions, accordingly.

Table 1. Results of the second pretest: Counts of how much the three parts of the short story, which included modifying keywords (in all but the neutral condition) and were presented by a synthetic voice, match a condition. Eight participants rated three times per condition, once per part of the short story, and 18 times on the whole test. Numbers in bold indicate the highest ratings.

		N	Foc+	Foc–	Epi–	total		N	Foc+	Foc–	Epi–	total
condition match	#	**16**	**17**	10	12	55	%	**66.7**	**70.8**	41.7	50.0	57.3
no condition match	#	8	7	**14**	12	41	%	33.3	29.2	**58.3**	50.0	42.7
total	#	24	24	24	24	96	%	100	100	100	100	100

A fourth issue for the experimental setup is the creation of a **natural VH**. There are many options for designing the behaviour of a VH as natural as possible, include facial expressions, "idle" gaze and saccades, posture or "idle" body movements. One prerequisite was that the behaviour of the VH should be as natural as possible and as controlled as necessary. We recorded data with a fifteen-camera OptiTrack motion capture system to give Billie the appropriate naturalness. And to control for unnatural "idle" behaviour in between gesture performances, we recorded each part of the short story in one piece and in each of the four conditions. Since the corresponding audio file was played back while recording the motion capture behaviour, in the final stimulus videos, the VH made movements which fit the context of the story extremely well and, thus, may have increased the degree of presence of the VH. We disabled a few joints in order to control for too much movement of the VH skeleton. No further adjustments were made like facial expressions or gaze movements.

3 Stimuli and Procedure

Our research question is whether the identified MF in gesture are also perceived in a VH. For an accurate test of how a gesture is perceived and whether the MF is recognised correctly, we compared a gesture-only (`G`) and a speech-only (`S`) condition against a speech-and-gesture (`S+G`) condition and a neutral condition as baseline (`N`). The `G` condition contained gestures with MF and speech without keywords, in `S` there were modifying keywords in speech and only gestural idle behaviour (only a few not meaningful arm movements), in `S+G` there were keywords and gestures with MFs and, in the `N` condition, there was speech without keywords and gestural idle behaviour. The following keywords were used for `Foc+`: "concretely", "even", "exclusively", "in any case", "more precisely", "most important", "most notably", "particularly", "primarily", "totally important" and "totally obvious"; for `Foc-`: "any", "anyway", "just/plainly", "merely", "only",

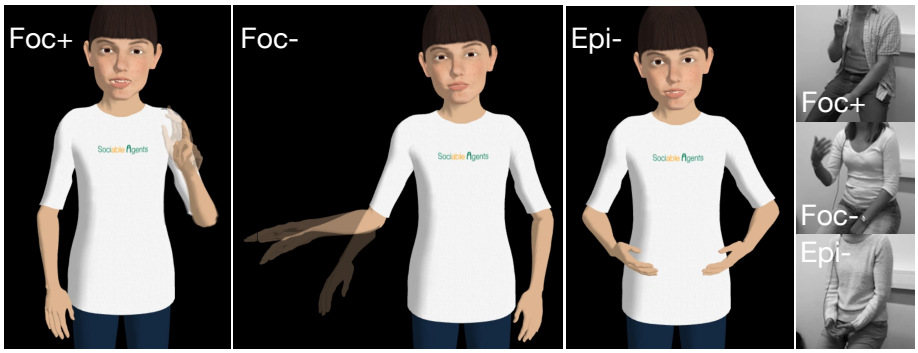

Fig. 1. Stills of the stimulus videos and stills of the snippets from our experimental setup with human gestures. From left to right and top to bottom: A gesture with a highlighting function (index-finger pointing), a gesture with a de-emphasising function (brushing away), and a gesture of uncertainty (palm up open hand).

"ordinarily", "solely", "(totally) spectacularly" and "trivially"; and for `Epi-`: "apparently", "maybe", "possibly", "potentially", "presumably", "probably", "seemingly" and "sort of".

Stimuli for all but the neutral conditions had to be created trice (`S+G`, `S`, `G`), accumulating to ten conditions in total: `N`, `Foc+SG`, `Foc+S`, `Foc+G`, `Foc-SG`, `Foc-S`, `Foc-G`, `Epi-SG`, `Epi-S`, `Epi-G`. The stimuli were videos of Billie telling three stories with and without keywords in speech and MF in gestures. Body movement and gestures were recorded using motion capture. From all recordings, only the most accurate twelve (three MF and `N` for three stories each) ones were kept for post-processing. Since the performance of a gesture critically depends on the shape of the hand (not recorded), one post-processing step was the definition of the hand shapes. Those were designed in the MURML Keyframe editor [11] and merged into the motion capture data. The sentences were aligned to the nonverbal-behaviour and eye blinking was added. In all conditions, the VH Billie is used, his behaviour is steered by AsapRealizer [12] and his speech is synthesised by the Text-To-Speech system CereProc[2] with the female voice Gudrun. In total, 30 stimulus videos of Billie were recorded, with a duration of 54 to 60 seconds.[3]

In order to investigate the research questions raised at the end of Section 1, a between-subject design was carried out. Each participant was shown the three stimulus videos of one condition in random order (cf. Figure 1). Four statements of various difficulty had to be evaluated into "correct", "wrong" or "I don't know" after each of the three stimuli videos. The statements were quite technical and detailed in order to estimate upper bounds. Five out of the 12 statements were wrong. Our hypothesis was that the participants remember facts about the narratives much better when they are highlighted by the VH as done in the `Foc+` conditions. Since Billie is supposed to convey uncertainty and de-emphasise

[2] www.cereproc.com

[3] Stimulus videos can be accessed at https://pub.uni-bielefeld.de/publication/2903503

content in the `Epi-` and the `Foc-` conditions, we expected less recall of the content in these conditions. After that, the participants were ask to rate Billie's behaviour according to a specific MF.[4] Each participant was asked to answer the MF question only once and had only one choice. Since also no slider was given, designing the query in this manner is similar to a forced-choice method. We expected good recognition of the `Foc+` function. Subsequently, 20 items had to be rated regarding the perceived competence, likeability and human-likeness of the VH as in [14, 13]. We hypothesised the association of strong competence with `Foc+` and less competence in `Epi`. Furthermore, questions about the observation of gestural and nonverbal behaviour of the VH were issued, about particular body movements [13], and how much these body movements and gestures helped in understanding the story. Due to space, not all analyses can be presented.

112 uninformed participants (52 female, 60 male, 0 other) with an average age of 24.4 (range [18,39], σ=3.8) took part in the experiment. They were not informed about the purpose of the study, we simply explained that we conducted the study to improve the VH. 104 of them took part locally, in a computer room of our research institute, and were predominantly from the University of Applied Science in Bielefeld. They were provided with headphones, the VH was presented in a video of the size 22 by 22 cm and the distance to the screen was approximately 40 to 50 cm. Those participants received a compensation of 2 Euros for an average test duration of 15 minutes. Eight participants took part online and could not be compensated. Taking part online was possible since we provided a link and a QR-code on the flyers that we distributed. The participants were distributed randomly across the seven conditions in the following way: N: 15, `Foc+SG`: 15, `Foc+G`: 15, `Foc-SG`: 18, `Foc-G`: 17, `Epi-SG`: 17, and `Epi-G`: 15. In a second elicitation, 45 uninformed participants (29 female, 16 male, 0 other) from Bielefeld University with an average age of 27.8 (range [20,79], σ=10.8) took part in three additional conditions with the same experimental setup, as recommended by the reviewers: `Foc+S`: 15, `Foc-S`: 15 and `Epi-S`: 15. This sums up to 157 participants and ten conditions in total.

4 Results

In this section, we will show results of whether the MF were matched to the correct conditions, how the content of the story was recalled and how the VH was perceived, each by analysing our MF and different modalities.

Effect of MF In the following, we evaluate whether the MF modelled in our VH Billie get across to humans. The participants were asked to categorise Billie's

[4] The exact wording of the question was: "In the following, please, determine Billie's utterances and communication. It is important *how* Billie uttered and communicated something. Make sure that the artificial pronunciation and speech melody does *not* have an effect on your assessment. Also, do *not* judge the relevance of things Billie talked about. Merely judge *how* Billie's utterances were: (A) emphasising and/or highlighting and/or focusing, (B) discounting and/or understating and/or defocusing, (C) uncertain and/or unknowing, (D) neutral."

utterances, which would ideally match our four broad conditions (MF and N). The results are presented in Table 2. With 37.5% (58 of 157 ratings) correct MF matchings vs. 62.5% (99) incorrect matchings, we already assume that the task was difficult and that other issues may be involved. A striking result is that the participants clearly recognised the Foc+ MF in the S+G condition from the stimulus videos with 73.3% of the counts and the neutral condition to the same extent. The Foc+S and the Foc+G conditions were still recognised by 53.3% and 46.7% of the counts, while almost the same amount has been distributed to the neutral condition. Similarly to the results of the second pre-study, the participants chose the neutral condition frequently: those ratings made up 48.9%. Unfortunately, Foc- and Epi- were poorly recognised. Possible reasons for N being chosen quite frequently are that the gestures and keywords are not as distinct as we had hoped, however, since the participants had no further option but to decide for any of the three MF and neutral, neutral may have been a fallback option. This may also indicate that the neutral condition is not as surpassing recognised as it seems.

Table 2. Contingency table of counts and the corresponding percentages of the ten conditions. Grey shaded numbers indicate the correct category of MF and numbers in bold indicate the highest result.

	N	Foc+SG	Foc+S	Foc+G	Foc-SG	Foc-S	Foc-G	Epi-SG	Epi-S	Epi-G	total
# N	**11**	2	6	7	**9**	**9**	**10**	**8**	5	**10**	77
# Foc+	4	**11**	**8**	7	4	4	3	5	1	4	51
# Foc-	0	2	1	0	5	2	4	0	3	1	18
# Epi-	0	0	0	1	0	0	0	4	**6**	0	11
total #	15	15	15	15	18	15	17	17	15	15	157
% N	**73.3**	13.3	40.0	46.7	**50.0**	**60.0**	**58.8**	**47.1**	33.3	**66.7**	48.9
% Foc+	26.7	**73.3**	**53.3**	46.7	22.2	26.7	17.6	29.4	6.7	26.7	32.9
% Foc-	0	13.3	6.7	0	27.8	13.3	23.5	0	20.0	6.7	11.1
% Epi-	0	0	0	6.7	0	0	0	23.5	**40.0**	0	7.0

In order to check how well our conditions were recognised, we merged the counts of Table 2 into "matched" (only grey shaded numbers) and "did not match condition" (sum of remaining three values) for each condition, giving us a 2-by-10 matrix. On this data, we applied Pearson's chi-square test using SPSS[5]. The assumptions for using categorical data were met: we ensured the independence of residuals in that each person contributed only to one cell of the contingency table and the values for each cell were sufficiently large. In avoidance of a Type I error, because of conducting 15 tests on this particular question, we calculate the Bonferroni correction for all tests that follow. With

[5] IBM Corp. Released 2015. IBM SPSS Statistics for Mac, Version 23.0. Armonk, NY: IBM Corp. This software and Microsoft Excel were used for all analyses in this work.

$\chi^2(9) = 35.11, p \leq .002$, the outcome was that there is a significant difference between the ten conditions and whether or not the participants rated that Billie's behaviour has a certain function, indicating that there is an association between the ratings and a particular condition. To check whether the ratings occurred due to chance, we calculated a 95%-confidence interval. With CI=[29.8;44.7], the overall rating results are clearly above a chance level of 25%.

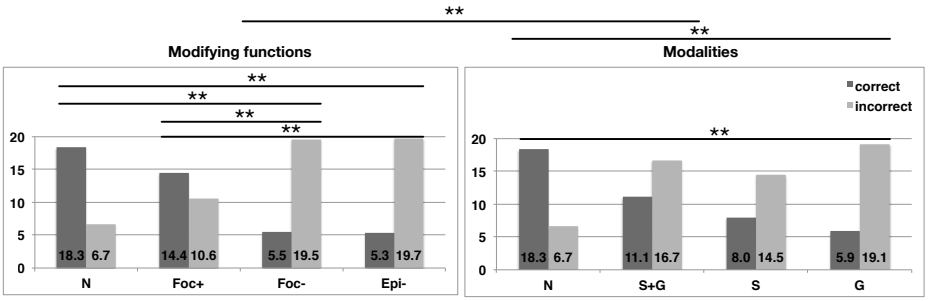

Fig. 2. Percentages of correctly and incorrectly rated MF, percentages sum up to 100 for MF and modalities each. Significant chi-square test results of correctly and incorrectly rated MF overall, for MF and modalities each, and between MF and modalities.

Using Pearson's chi-square test and the "match" vs. "no match" data preparation for more detailed analyses, we can find significant differences between single MF across modality conditions (cf. Figure 2): `N`+`Foc-`: $\chi^2(1) = 13.58, p \leq .002$; `N`+`Epi-`: $\chi^2(1) = 13.76, p \leq .002$; `Foc+`+`Foc-`: $\chi^2(1) = 12.75, p \leq .002$; and `Foc+`+`Epi-`: $\chi^2(1) = 12.86, p \leq .002$. Therefore, these conditions were perceived as rather different and, thus, `N`+`Foc+`, as well as `Foc-`+`Epi-` were perceived as rather alike. No significant difference for the merged categories `N`+`Foc+`+`Foc-`+`Epi-` could be found. The calculation of a 95%-confidence interval results only for the `N` and `Foc+` conditions in ratings above chance: CI=[48.0; 89.1] and CI=[43.3;71.0], respectively. This concludes that `N` and `Foc+` are well recognised, rated above chance level and rated differently than `Foc-` and `Epi-`, which are perceived less well in our VH Billie.

In a second step, we want to evaluate, if there are differences between the modality conditions. As Table 2 depicts, there is a trend between three conditions `S+G`>`S`>`G` (cf. Figure 2), indicating that MF are more clearly recognised if more modalities are involved, which is a clear statement in favour of multimodality. A significant effect between `S+G`+`G`: $\chi^2(1) = 3.07, p = .080$ was lost after correcting for type I error. However, since `N` is even better perceived than `S+G`, the difference of the categories `N`+`G` reaches significance: $\chi^2(1) = 12.38, p \leq .002$. Also, the 95%-confidence interval for modalities shows that the `S+G` condition has been rated above chance with CI=[48.8;80.8], `N` getting the same result as above. Finally, the merged categories for single modalities across MF conditions

reached significance: N+S+G+S+G: $\chi^2(3) = 26.66, p \leq .002$. For now, we can only report a trend between S+G>S>G, but it would be interesting to investigate this relationship further.

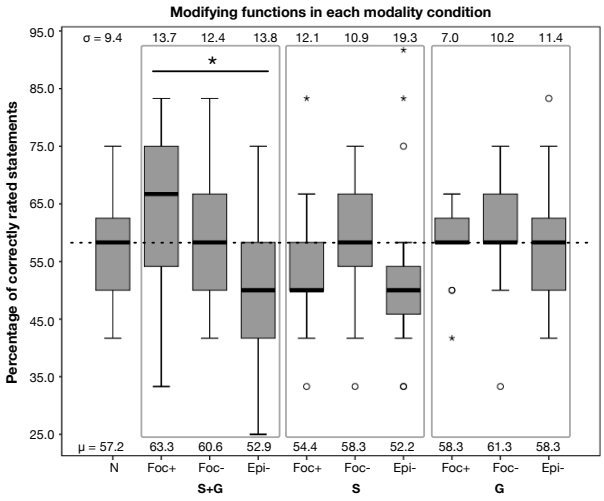

Fig. 3. Recall: Percentages of correctly rated statements in ten conditions. The baseline condition N is marked as a dotted line. The median values differ only in three numbers: 66.7, 58.3, and 50.0. In the multimodal S+G condition, a trend of decreasing recall between Foc+, Foc- and Epi- is visible. Recall in G seems slightly higher than in S.

Effect on recall We were interested in how much participants recalled from Billie's narration about his life as a VH. The results (cf. Figure 3) indicate that there is a trend of decreasing knowledge between the three S+G conditions: Foc+ (μ=63.3) > Foc- (μ=60.6) > Epi- (μ=52.9). The differences of correctly categorised statements between Foc+ and Epi- is 10.4%, this amounts to 114 correct answers in Foc+ vs. 108 correct answers in Epi-. Indeed, the main finding is that there is a significant difference between Foc+ and Epi- (p=.041, independent samples t-test, normally distributed). Unfortunately, the significant difference is too small as that it holds if we correct for type I error (we ran 18 tests).

A second trend is that the content in the respective G conditions is slightly better recalled than in the respective S conditions. However, this difference did not reach significance. Additionally, Foc-S was recalled better than Foc+S and Epi-S, although the keywords proved to be more complicated in Foc- in the pretests. Perhaps this indicates that obvious modifications in speech ('Foc+:"particularly" and Epi-: "I don't know") distract more from the content of the utterance than more subtle modifications (Foc-: "just") and gestures, being more subtle, distract less than speech. This assumption is supported by

the accumulated standard deviations: Foc+=15.5, Foc-=11.2, Epi-=14.8 and S+G=13.3, S=14.1, G=9.5. Compared to the baseline condition N=9.4, it seems that Foc- and in particular G can be easiest integrated, i.e., there is more certainty about how a statement is categorised.

Examining single conditions, we find that participants got the highest recall score in Foc+SG, probably due to a positive effect of the linguistic markers combined with gestures of focus and highlighting. In the condition Epi-G, there were more correct answers than in Epi-SG and Epi-S, maybe due to the fact that linguistic markers of uncertainty decreased recall. A possible explanation for many correctly identified statements in Foc- is that the gesture is quite prominent: it used the most amount of gesture space (see Figure 1) and may had a great visual effect on the participants, causing participants to pay extra attention and, thus, perhaps leading to better recall. To conclude, content recall is best in S+G, intermediate in G and more difficult to integrate in S. Furthermore, Foc+ triggers biggest recall and Epi- least.

VH perception In a third analysis, we evaluated how the VH was perceived by the participants. The question "How do you evaluate the agent?" was answered by 20 adjectives (e.g., "expert") on a 5-point Likert scale (5="very" to 1="not at all"). The VH was highly connoted "expert" (Mdn=4, SD=.91), "intelligent" (Mdn=4 SD=1.08) and "thorough" (Mdn=4 SD=.82) and least associated with "sensitive" (Mdn=2 SD=.93), "fun-loving" (Mdn=2 SD=.99), "lively" (Mdn=2 SD=.97) and "natural" (Mdn=2 SD=.89).

In order to make more general statements about the VH, we used a design carried out by [14, 13], in that we merged 17 items to three scales likeability, competence and human-likeliness. We calculated Cronbach's Alpha for the indices and the values for all three scales were above .7, which justifies the combination of these items into one mean value as a single index for this scale. Items for likeability (a=.833) are "pleasant", "sensitive", "friendly", "likeable", "affable", "approachable" and "sociable"; items for competence (a=.722) are "dedicated", "trustworthy", "thorough", "helpful", "intelligent", "organized" and "expert"; and items for human-likeness (a=.708) are "active", "humanlike", "fun-loving" and "lively".

Again, we carried out analyses on the differences between our MF (N, Foc+, Foc- and Epi-) and between the modality conditions (N, S+G, S and G). We calculated the non-parametric Mann-Whitney U test, as the distributions of all three scales deviated significantly from normal (Shapiro-Wilk test: $p \leq .000$). The assumptions for the test were met: the dependent variable is measured at the ordinal level (1 to 5) and the independent variable consists of two categorical, independent groups (10 conditions) with independent observations and all sample sizes were >30. There were small differences in the number of participants between the conditions, cf. end of Section 3. Comparing all categories of all scales and analyses, we get 36 comparisons in total and to avoid inflated error rates, we calculated the Bonferroni correction on all tests.

The results of the scales (cf. Figure 4) show similar results to those of the isolated items: the VH was perceived as rather competent (all Mdn=4), inter-

mediate likeable (all Mdn=3) and rather not human-like (all Mdn=2, but for `Foc+` Mdn=3) on the three scales and in the two analyses. Analysing differences between the MF, the VH was perceived as more likeable in `N`/`Foc+` compared to `Foc-`, since there is a highly significant difference between the conditions (each p=.004). The difference between `N`/`Foc+` and `Foc-`/`Epi-` also shows when looking at competence: There are highly significant differences between `N` and `Foc-` and between `N` and `Epi-` (each p=.004) and further between `Foc+` and `Foc-` (p=.004) and significant differences between `Foc+` and `Epi-` (p=.036). Thus, `N` in particular but also `Foc+` are perceived as more competent. On the scale of human-likeness, `Foc-` comes into focus, which is perceived as least humanlike, with differences to all other conditions and a significant difference to `Foc+` (p=.036).

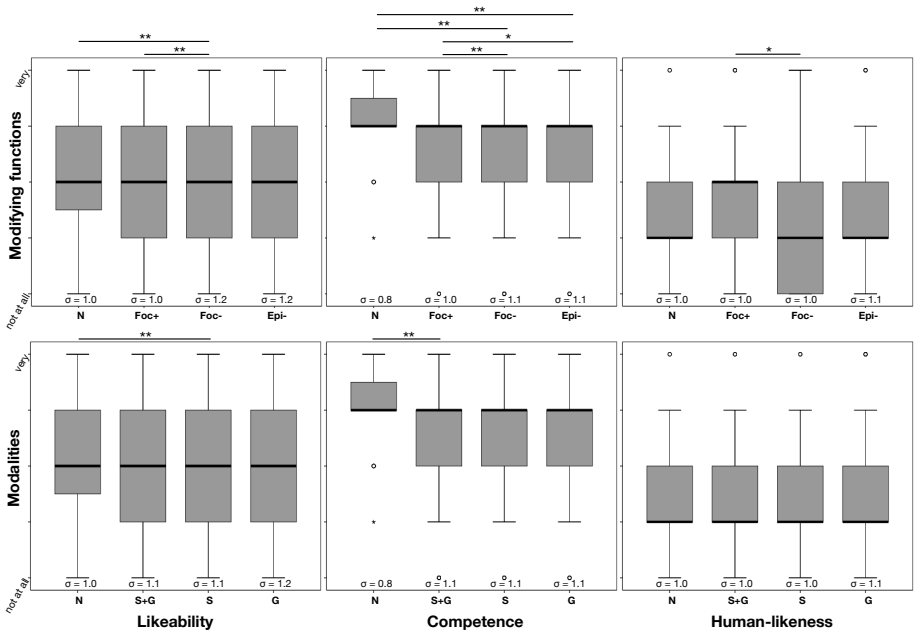

Fig. 4. Differences in the perception of the VH between the MF (`Foc+`, `Foc-` and `Epi-`) and between the modalities (`S+G`, `S` and `G`) on the three scales likeability, competence and human-likeness with significant results between conditions.

Analyses between the modalities are less diverse. `N` is perceived as more likeable with a highly significant difference to `S` (p=.004), thus, `S` is perceived as least likeable, followed by `G` and then `S+G`. As with MF, `N` is again perceived as most competent compared to the other modalities, with a highly significant difference to `S+G` (p=.004). Therefore, the condition `S+G` is perceived as least competent, followed by `G` and then `S`. For human-likeness, all conditions seem

to be perceived similar and there was only a significant difference between N and S before correcting for type I error. To sum up, the VH was perceived as competent (particularly N and Foc+) but rather not humanlike (especially Foc- and S+G) and rather not likeable in the S condition.

5 Conclusion

In this work, we investigated whether and how VH can use speech and gestures to add meta-communicative information to their utterances. Based on a study that identified three main modifying functions (Foc+/Foc-/Epi-), we tested whether such functions are also observed in VH, and whether this changes content recall and the perception of a VH. Our results suggest that, although the behaviour of the VH is generally judged rather neutral, Foc+ is distinctively recognised, may lead to better recall and affects the perceived competence of the VH. In contrast, Epi- triggers least recall and Foc- was perceived as least human-like. The high ratings of competence may be due to the partly detailed and technical descriptions given.

Effects were most pronounced in the S+G conditions, i.e., when speech and gesture acted together. A trend of S+G>S>G was found for the perception of MF and in parts in the recall analysis, suggesting that modification is multimodal and that this pragmatic level influences the processing of an utterance. Further, while keywords in speech may distract (recall in Foc+S), the integration of gesture into an overall meaning is more easily done (recall in Foc+G). In fact, many human gestures are non-representational and are assumed to be modulating or meta-communicative [2]. Thus, although the effect of gestures assuming such pragmatic MF is more subtle and partly weaker, the strength of them being perceived non-consciously should not be underestimated.

It is important to note that results are not fully unequivocal. Yet, we note that we have tackled a very difficult problem. Modifications and meta-communication on focus or epistemic state are hardly conventionalized and only rarely clearly marked. In contrast, this information is often "analogous" and strongly interpretative. We thus were faced with many methodological challenges, e.g., relating to the adjectives or keywords used to capture this phenomenon of pragmatic and content-modifying meta-communication. Yet, our corpus analyses imply that competent and cooperative speakers do use such markers to help their addressees arrive at the correct interpretation, and this behaviour should also be beneficial for VHs. Our results do seem to confirm this.

Possible reasons for the strong results of N were discussed in Section 4 (effect of MF). However, MF can be made significantly more salient using more distinct keywords in Foc-, more defined gestures in Epi- (and Foc-), adding intonation to the synthesised speech, and enabling more idle VH body movement. Regarding the procedure, the content questions should be queried only after all stimuli have been presented, since the use of content questions may have distracted from the nonverbal behaviour of the VH, similar to the selective attention test [15].

Acknowledgements

This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG). A special thanks goes to Iwan de Kok whose helpful comments made the paper clearer in various aspects.

References

[1] Wharton, T.: Pragmatics and non-verbal communication, Cambridge University Press (2009)
[2] McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago press (1992)
[3] Freigang, F. and Kopp, S., Analysing the Modifying Functions of Gesture in Multimodal Utterances Proc. of the 4th Conference on Gesture and Speech in Interaction (GESPIN), Nantes, France (2015)
[4] Kendon, A.: Gesture: Visible Action as Utterance, Cambridge Uni. Press (2004)
[5] Payrató, L., Teßendorf, S.: Pragmatic Gestures, In: C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill and S. Teßendorf (eds.), Body Language Communication: An Int. Handbook on Multimodality in Human Interaction. Handbooks of Linguistics and Communication Science 38(1):1531–1539 (2013)
[6] Lu, Y., Aubergé, V. and Rilliard, A.: Do You Hear My Attitude? Prosodic Perception of Social Affects in Mandarin, Int. Conf. on Speech Prosody Proc., 685–688 (2012)
[7] Kok, K., Bergmann, K., Cienki, A., and Kopp, S. Mapping out the multifunctionality of speakers' gestures. Gesture, 15(1), 37-59. (2016)
[8] Freigang, F. and Kopp, S., Modifying Functions of Gesture - Exploring the Dimensions of Function and Form (in prep.)
[9] Schoonjans, S.: Modalpartikeln als multimodale Konstruktionen. Eine korpusbasierte Kookkurrenzanalyse von Modalpartikeln und Gestik im Deutschen. KU Leuven: Dissertationsschrift (2014)
[10] Leiner, D. J.: SoSci Survey (Version 2.6.00-i) [Computer software]. Available at http://www.soscisurvey.com (2014)
[11] Kranstedt, A., Kopp, S., Wachsmuth., I.: MURML: A multimodal utterance representation markup language for conversational agents. Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents (2002)
[12] van Welbergen, H., Yaghoubzadeh, R., Kopp, S.: AsapRealizer 2.0: The next steps in fluent behavior realization for ECAs. In: Intelligent Virtual Agents. LNCS, vol. 8637, pp. 449462. Springer, Berlin, Germany (2014)
[13] van Welbergen, H., Ding, Y., Sattler, K., Pelachaud, C., Kopp, S.: Real-Time Visual Prosody for Interactive Virtual Agents. Intelligent Virtual Agents. Springer International Publishing (2015)
[14] Bergmann, K., Kopp, S., Eyssel, F.: Individualized gesturing outperforms average gesturing: evaluating gesture production in virtual humans. In: Intelligent Virtual Agents. LNCS, vol. 6356, pp. 104117. Springer (2010)
[15] Simons, D. J., Chabris, C. F.: Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception 28.9: 1059-1074 (1999)