Observations on Listener Responses from Multiple Perspectives

Iwan de Kok
Human Media Interaction
University of Twente
i.a.dekok@utwente.nl

Dirk Heylen
Human Media Interaction
University of Twente
heylen@utwente.nl

Abstract

In this paper we present three studies that investigate the individual differences in nonverbal listening behavior. Besides collecting a corpus of listener responses, we asked people to watch a video of a speaker and indicate where they would produce a listener response. Also we asked people to judge the appropriateness of listener responses that we generated using a virtual human. The combination of the multiple perspectives collected in these studies provides us with a rich data set in which different types of response opportunities are distinguishable. There are moments where there is *high agreement* between these multiple perspectives that a listener response is appropriate or inappropriate, moments where a listener response is *controversial* and moments neither a response was given nor a response was judged inappropriate (*neutral*). We will show that different contextual characteristics can be used to discriminate these response opportunities. Observations show relations to sentence structure, conversational structure and proximity of earlier responses.

1 Introduction

In a conversation humans highly coordinate their behavior to transfer information from one to another. In this interaction not only the behavior of the speaker guides the conversation, but the responses from the listener to the contributions of the speaker do so as well (Yngve, 1970; Kraut et al., 1982; Bavelas et al., 2000). These listener responses can take the shape of nonverbal behaviors such as head nods, head shakes and facial expressions, and verbal expressions, such as "hmmm" and "yeah". The function of these listener responses is to signal the state of mind of the listener towards the speaker, conveying whether the contributions of the speaker are attended to, understood, agreed upon and/or affective attitudes towards the contributions (Allwood et al., 1992; Clark, 1996).

Our interest in this behavior comes from the goal to build embodied conversational agents which can interact as if they are a human. A model of these listener responses is one of the components needed to achieve the same kind of coordinated interaction as humans have. A challenge in the achievement of this goal is the optional characteristic of listening behavior, which causes high variation in the type, timing and amount of listener responses between individuals. One missed opportunity for a listener responses will not immediately break the interaction, but the total absence of this behavior will. The question is which moments are essential to respond to as a listener and which ones can be passed up. And what are the characteristics of the moments where listener responses is inappropriate?

In this paper we will present three studies that capture the individual differences in nonverbal listening behavior by combining multiple (positive and negative) perspectives. In the first study a corpus is recorded with three listeners in parallel interaction with the same speaker, which gives us three positive perspectives on appropriate moments for listening behavior. In the second study we collect extra positive perspectives on appropriate listening behavior through the parasocial consensus sampling method. In the third and final experiment we collect multiple (negative) perspectives on *in*appropriate behavior by generating listening behavior and let participants judge the appropriateness of each individual listener response. By combining the data of these three studies some moments stand out by either *high agreement* between multiple perspectives (positive or negative), *controversial* perspectives on the appropriateness (positive and negative responses at the same mo-

ment) or *neutral* moments (neither positive nor negative responses). We end the paper with a discussion of these types of moments in our data and with recommendations based on our observations to improve the state-of-the-art of predictive models for listener responses.

2 Study 1: Parallel Recording

In the first study we recorded a corpus aimed at capturing the variation and similarities in listening behavior between people. In traditional corpora to study nonverbal listening behavior an interaction between two people is recorded. The listening behavior in reaction to the speaker is regarded as the ground truth. However another individual placed in the same interaction will most likely not act in the same way. He/She will provide listener responses at different times or use different type of listener responses.

By collecting multiple perspectives we are able to analyze the optionality of listener responses. Our hypothesis is that by combining multiple perspectives one can find moments where a response is given in all perspectives, moments where a response is given in some perspectives and moments where no response is given at all. In the first case, it is probably mandatory for a virtual agent to produce a response, in the second case it might be optional and in the third case it seems better to avoid giving a response. The following section explains the experiment resulting in the recording of the MultiLis corpus in which multiple listeners are recorded in interaction with the same speaker.

2.1 Procedure

The MultiLis corpus (de Kok and Heylen, 2011b) is a Dutch spoken multimodal corpus of 32 mediated face-to-face interactions totaling 131 minutes. Participants (29 male, 3 female, mean age 25) were assigned the role of either speaker or listener during an interaction. In each session four participants were invited to record four interactions. Each participant was once speaker and three times listener.

What is unique about this corpus is the fact that it contains parallel recordings of three individual listeners in interaction with the same speaker, while each of the listeners was tricked into believing to be the sole listener. The speakers saw only one of the listeners, believing that they had a one-on-one conversation. All listeners were placed in a cubicle and saw the speaker on the screen in front of them. The camera was placed behind an interrogation mirror, positioned directly behind the position on which the interlocutor was projected. This made it possible to create the illusion of eye contact.

To ensure that the illusion of a one-on-one conversation was not broken, interaction between participants was limited. Speakers and listeners were instructed not to ask for clarifications or to elicit explicit feedback from each other, so no turn-switching would take place. The speaker received a task of either watching a short video clip before the interaction and summarizing it to the listener, or learning a recipe in the 10 minutes before the interaction and reciting it to the listener. The listener needed to remember as many details of what the speaker told as possible, since questions about the content were asked afterwards.

2.2 Annotation

The recordings of each listener were annotated by one annotator on listening behavior. Each listener has her/his own (perspective on) listening behavior. To study the variety and similiarities in these perspectives one annotator grouped simultaneous listener responses in reaction to the same context. We call the timeframe they span from the first response to that context to the last response the *response opportunity*. Thus, response opportunity can be defined as the window of opportunity to provide a response to a specific context in an interaction.

2.3 Results

The MultiLis corpus contains 2796 listener responses. These listener responses are reactions to 1735 response opportunities. Of these responses opportunities 1142 have one response, 456 have two responses and 128 have responses from all three listeners.

Figure 1 represents a segment of 48 seconds from one of the interactions. It shows the distribution of response opportunities in this segment. On the horizontal axis time is represented. The response opportunities in these 48 seconds found in the MultiLis corpus are indicated with as magenta bars. The height of these bars represent the amount of recorded listeners that gave a response at this response opportunity.

The segment is taken from an interaction where agreement between listeners is quite high. In this

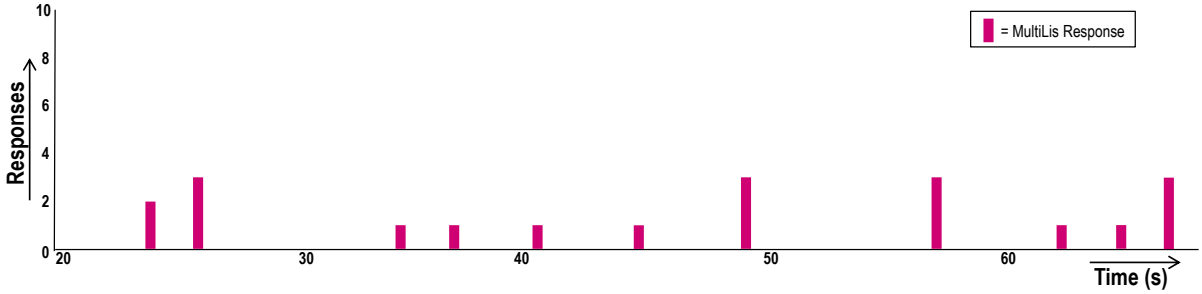

Figure 1: Sample of the distribution of responses in the MultiLis Corpus.

segment there are four response opportunities with three listener responses, one with two listener responses and six with one listener response. No listener has performed a listener response at all these response opportunities. This illustrates that with this corpus we have a more complete view of all the opportunities for a listener response.

In the remainder of the paper we will take a closer look at this segment. We will see how new perspectives correlate with the recorded behavior, how the response opportunities correlate with inappropriate moments and what the characteristics of these response opportunities are. Does the speaker explicitly elicit the listener responses at response opportunities with high agreement or are there other causes?

3 Study 2: Parasocial Consensus Sampling

In the previous study we recorded a corpus where three listeners listened and responded to the same speaker. What if we had more listeners? We would get an even more complete view of all the opportunities for a listener response. There may still be moments that all three listeners have passed up, while a listener response would still be appropriate. The discrimination between mandatory response opportunities, option response opportunities and inappropriate moments to provide a response would also be more clear.

With the Parasocial Consensus Sampling method (Huang et al., 2010b) this is actually possible. In this method multiple participants watch the video recording of the speaker and they indicate through a keyboard when they would give a listener response. We have used this method to collect 8 new (PCS-)perspectives for a subset of the MultiLis corpus.

3.1 Procedure

The collection of PCS perspectives is performed on 8 interactions from the MultiLis corpus. Ten months after the original MultiLis experiments we reinvited 6 of the original listeners in these 8 interactions to collect their PCS perspectives for the same interactions. While watching and listening to the 3 recordings of the same speakers they listened to earlier, they gave responses through the keyboard. Each time they would give a listener response they were instructed to press the spacebar of the keyboard.

Furthermore we invited 10 new participants to collect their PCS perspectives to these interactions. Each of these participants gave their PCS perspectives on 4 interactions. Thus, for each of the 8 interactions, we have 3 original listener perspectives and 7 or 8 PCS-perspectives. From these perspectives there are 5 perspectives from the new participants and 2 or 3 perspectives from the original listeners, depending on whether one of them was the speaker in that interaction or not.

3.2 Results

The 8 interactions used in this study contain 347 response opportunities of which 202 with one response, 98 with two responses and 47 with three responses as identified using the annotations of the three listeners in the corpus. Adding the new PCS perspectives increases the amount of response opportunities identified to 582 response opportunities. The distribution of the amount of responses to each response opportunity is shown in the histogram in Figure 3.

Most response opportunities have only a few responses, but there are still 15 response opportunities with 9 responses, 3 with 10 responses and 3 with 11 responses. We will take a closer look at these response opportunities in Section 5.

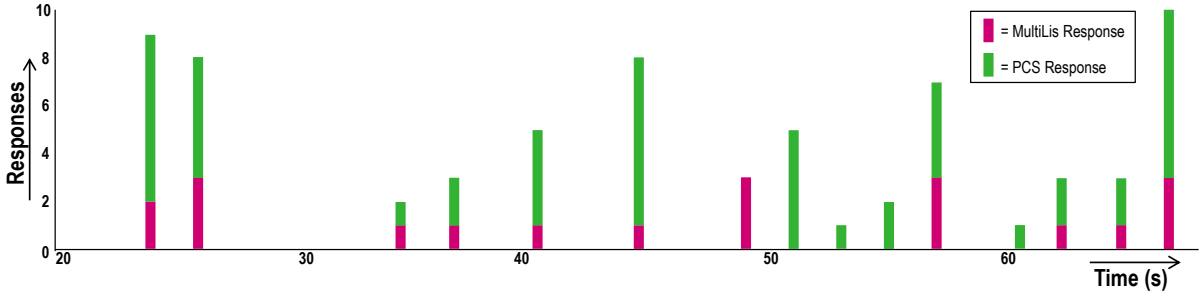

Figure 2: Sample of the distribution of responses in the MultiLis Corpus and PCS responses.

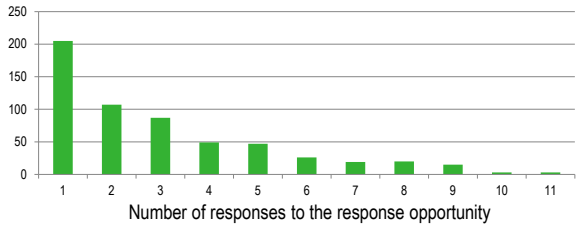

Figure 3: Histogram of the number of (MultiLis and PCS) responses to each response opportunity.

Figure 2 represents the same 48 seconds from the previous study. In green the responses from the collected PCS-perspectives are added to the responses from the MultiLis corpus. The participants provided a PCS-response to almost all the response opportunities found in the previous study with the exception of the response opportunity just before 50 seconds. Interestingly this response opportunity was responded to by each listener in the MultiLis corpus. Furthermore there are 4 new response opportunities of which one was responded to by 5 participants.

4 Study 3: Individual Perceptual Evaluation

With the previous two studies we have compiled a more complete picture of the response opportunities in the interactions than a traditional corpus does by collecting multiple (positive) perspectives. We have identified 582 moments where giving a listener response is appropriate according to at least one individual. Does this mean that every other moment is an inappropriate moment to give a listener response? And are listener responses given at these moments appropriate according to everyone?

To answer these questions we use the Individual Perceptual Evaluation method. In this method we generate virtual listening behavior in reaction to a recorded speaker and let participants judge for each generated listener response, whether this response was appropriate or not. We thus collect a negative perspective on listener responses, which tells us the inappropriate timing of listener responses. In the following we will explain the method and the used stimuli in more detail.

4.1 Stimuli

We presented subjects with clips of a speaker from the MultiLis corpus in interaction with a virtual listener, animated using the BML realizer Elckerlyc (van Welbergen et al., 2010). We used the same 8 interactions as in the previous study. The virtual listener performs only head nods (and everytime the same head nod). The timing of the head nods is based on the multiple perspectives from the previous studies.

182 head nods are generated at appropriate times and 90 head nods are generated at *not*-appropriate times according to these perspectives. The appropriately timed head nods (or *at-head-nods*) are performed at the times where at least 4 perspectives agreed that this is an appropriate time to provide a listener response. The 90 *not*-appropriately timed head nods (or *between-head-nods*) are placed in the biggest gaps between the *at-head-nods*. Within these biggest gaps they are placed in the biggest gap between the moments where at most 3 perspectives agreed to be an appropriate time to provide a listener response.

4.2 Procedure

We invited 8 participants to watch the interactions between the speaker and the virtual listener. They were asked to judge each head nod on appropriateness. When a head nod was inappropriate according to their judgment they pressed the spacebar on a keyboard (a *yuck response*). The participant had the option to replay the video.

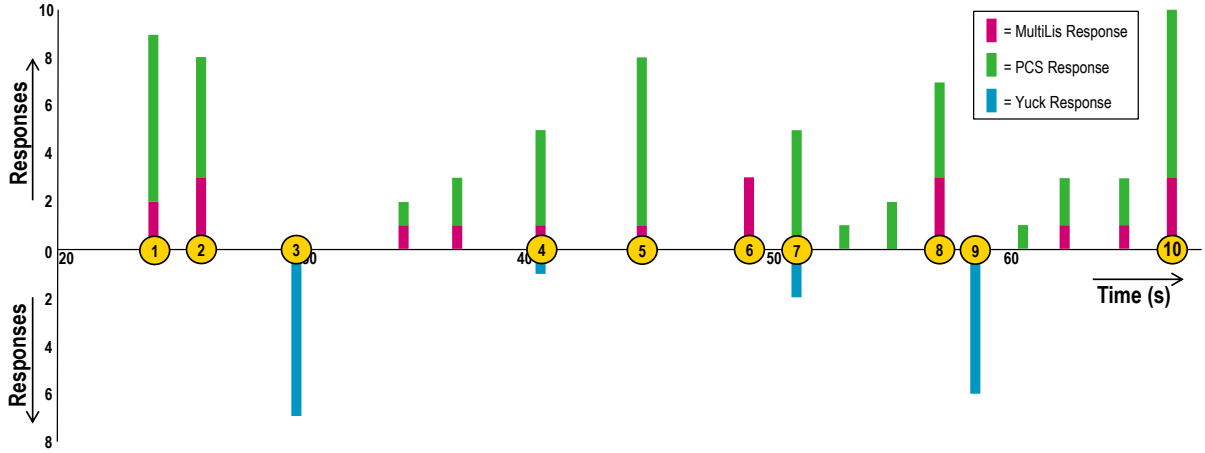

Figure 4: Sample of the distribution of responses in the MultiLis Corpus, PCS responses and the yuck responses. The numbers in the yellow circle correspond to the transcript in Table 1.

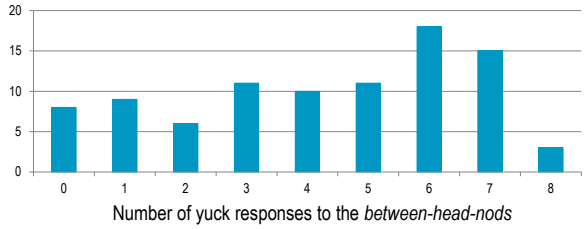

Figure 5: Histogram of the number of yuck responses to each *between-head-nod*.

4.3 Results

On average each participant judged 53 out of 272 head nods as inappropriate, for a total of 424 yuck responses. 42 yuck responses were in reaction to *at-head-nods* and 382 were in reaction to *between-head-nods*. The 42 yuck response in reaction to *at-head-nods* were in reaction to 29 individual *at-head-nods*. 4 of these *at-head-nods* were yucked 3 times, 5 were yucked 2 times and the other 22 were yucked once.

Figure 5 shows the histograms of the 379 yuck responses in reaction to the *between-head-nods*. For each of the 90 generated *between-head-nods* we counted the amount of yuck responses. Most of the *between-head-nods* (56 out of 90) get yucked by at least half of the participants. There were 3 *between-head-nods* which were yucked by each participant. There were 8 *between-head-nods* which were found appropriate by each participant, even though in the previous experiment none of the participants gave a response at that time.

Figure 4 represents the same 48 seconds from the previous studies. Now we added the yuck responses below the previous responses as negative responses. Note that only a head nod was generated and evaluated at response opportunities with at least four MultiLis or PCS responses. Moments 3 and 9 where the only generated *between-head-nods* in this segment. So, there were no head nods generated at *not*-appropriate times that nobody judged as inappropriate.

5 Discussion

In the previous studies we have collected positive responses (in the first two studies) and negative responses (in the last study). Combining these responses gives us three type of moments in our data. These types are *high agreement* (positive or negative), *controversial* and *neutral* moments. The *high agreement* moments have either positive *or* negative responses, the *controversial* moments have positive *and* negative responses and *neutral* moments have neither positive *nor* negative responses. In the following section we take a closer look at these type of moments. We do this by presenting several transcriptions of these moments and discussing the timing of the responses in relation to the context.

We first take a look at the response opportunities with *high agreement*; moments where most perspectives agree these are appropriate or inappropriate moments to provide a listener response. For this we take a look at the segment in Figure 4 and see what actually happens in the interaction. This segment is taken from an interaction where the speaker recites a recipe for risotto with mushrooms. In this segment the speaker is halfway through the ingredient list. The transcript is pre-

Table 1: Transcript of the segment displayed in Figure 4. The numbers in the rightmost column correspond to the response opportunities with the same number in Figure 4.

19.1 - 20.8	twee eetlepels	two tablespoons		
20.9 - 22.3	olie. Dus één liter	oil. So one liter		
22.5 - 23.3	twee en twee	two and two		
			24.1	**1**
24.3 - 25.2	olijfolie	olive oil		
25.3 - 25.8	natuurlijk	of course		
			25.9	**2**
27.9 - 28.7	euhm	euhm		
29.6 - 30.1	je hebt	you've got	29.9	**3**
30.3 - 32.9	verder voor de seasoning	furthermore for the seasoning		
33.4 - 34.5	één teentje knoflook	one clove of garlic		
35.6 - 36.4	één ui	one onion		
37.7 - 40.1	euh twee stengels bleekselderij	euh two sticks of celery		
			40.1	**4**
42.0 - 42.8	euh tijm	euh thyme		
42.9 - 43.9	één handjevol tijm	one handful of thyme		
			44.4	**5**
46.4 - 49.0	en natuurlijk euh heel veel paddestoelen	and of course a lot of mushrooms	49.0	**6**
49.1 - 50.1	500 gram	500 grams		
51.0 - 51.6	en	and	51.0	**7**
51.9 - 52.8	euhm	euhm		
53.2 - 54.4	natuurlijk de rijst	of course the rice		
55.2 - 57.0	400 gram rijst	400 grams rice		
			57.3	**8**
57.8 - 58.0	dus je hebt	so you've got		
58.4 - 61.9	euh 500 gram paddestoelen	euh 500 grams mushrooms	58.6	**9**
	400 gram rijst	400 grams rice		
62.4 - 65.3	en 100 gram parmezaanse kaas	and 100 grams parmesan cheese		
	dus in totaal	so in total		
65.4 - 65.7	mooi	nicely		
66.0 - 66.5	één kilo	one kilo	66.5	**10**

Table 2: Transcript of the most controversial response opportunity in the collected data, with 6 positive responses (3 MultiLis and 3 PCS) and 3 negative yuck responses.

29.0 - 31.1	het moment dat hij boven komt, euhm	the moment he arives at the top, euhm		
31.6 - 32.1	oh wacht	oh wait		
32.3 - 32.9	helemaal verkeerd	that's wrong		
			33.3	**11**

Table 3: Transcript of a neutral response opportunity where no positive and no negative responses are recorded.

30.5 - 34.1	euh, volgende list moet ie verzinnen	euh, he has to come up with a new trick		
	hij gaat vanaf	he goes from		
34.6 - 35.4	euh	euh		
35.5 - 36.1	een tegenoverliggend gebouw	an opposing building	36.1	**12**
36.1 - 40.8	via allemaal lijnen	across all those cables		
	die daar gespannen zijn	that are spanned there		

sented in Table 1. The numbers in the rightmost column correspond to the response opportunities with the same number in Figure 4. The *high agreement* moments in this segment are 1, 2, 5, 8 and 10 (positive), and 3 and 9 (negative).

The response opportunities 1 and 10 both are in reaction to a summarizing statement. Both statements summarize the previous ingredients with a mnemonic device to help them memorize the ingredients by summarizing the numbers mentioned (1) or by adding up the weights to a round figure (10). Beside the verbal cues, the speaker also makes iconic gestures to accompany the summarizing statements.

The other three *high agreement* response opportunities in this segment (2, 5 and 8) are all in reaction to a refining statement in which a previously mentioned ingredient is more precisely described: the oil is specified as being olive oil (2), the amount of the thyme is specified (5) and the precise weight of the rice (8). The other ingredients (like the garlic and onion) are also acknowledged with a listener response by some, but agreement between individuals is much lower in these cases (see the unnumbered response opportunities in Figure 4).

The moments with *high agreement* in negative yuck responses (3 and 9) are both mid sentence. They are not placed near or after the end of a grammatical clause, which is identified as a cue by Dittman and Llewellyn (Dittmann and Llewellyn, 1968), but instead are placed during or directly after the theme of the sentence. So, no new information has been mentioned by the speaker yet (rheme) and the listener response is premature. Furthermore, moments with *high agreement* in negative yuck responses are moments after long silences of at least 2 seconds, moments in between the article and the noun, and moments shortly (within 1.5 seconds) following another listener response.

An interesting case are the moments 6 and 7. The listeners in the corpus respond to "mushrooms", while the PCS responses are in reaction to the refining statement "500 gram". According to a previous study PCS responses are on average 220 ms slower (de Kok and Heylen, 2011a). Since the pause between the two statements is very short (a little over 100 ms), this delay would cause the PCS-er to place the PCS response during the "500 gram" statement. Instead they wait until the refining statement is finished. However, the faster responses from the listeners do not interfere with this statement and are made before the refining statement is started. Response opportunity 7 is a *controversial* moment since it is also yucked by two individuals. This is probably due to the timing, which is synchronous to the start of the word "and".

Besides response opportunities 4 and 7 there are other *controversial* response opportunities in the corpus. The most controversial moment has 6 positive responses (3 MultiLis and 3 PCS) and 3 negative yuck responses. The transcript of this moment is presented in Table 2. In this segment the speaker corrects himself. An acknowledgment from the listener through a listener response is valid according to six perspectives. The recorded listeners all responded to this moment, however two of them did not respond with a head nod, but with a polite smile (the speaker also smiles at this moment). However, the generated virtual agent in study 3 only performs a head nod. So it is likely that the response opportunity is not yucked because of the timing, but because of the type of listener response displayed.

Another reason for controversy in the corpus is that two response opportunities in quick succession (within 2 seconds) are individually regarded as good response opportunities (at least 4 positive response to each opportunity in the first two studies), but when generating a listener response at both moments in the third study, the second listener response gets yucked by some.

The last category of responses are the *neutral* responses. These are responses which are generated as *between-head-nods* in Study 3 at moments they received no positive responses in the first two studies. However, in the third study they were not seen as inappropriate responses and thus not yucked. In Table 3 one of these moments is transcribed. The head nod is placed mid sentence, not during a pause. The complete statement is not yet finished. However, it is placed directly after a vital piece of information within this statement ("an opposing building"), which is emphasized by the speaker and memorized after a short hesitation. A confirmation of this piece of information is appropriate according to Study 3 even though no other perspectives previously provided a response there. There are 7 *neutral* moments in our data (see Figure 5). In 5 of these moments the listener response

is placed mid sentence after a vital piece of information as in the previous example. In the other two cases the listener response is placed between sentences.

6 Conclusion

In this paper we have illustrated individual differences in nonverbal listening behavior. The combination of the multiple perspectives collected in these studies has provided us with a rich data set in which different types of response opportunities are distinguishable. There are moments where there is *high agreement* between these multiple perspectives that a listener response is appropriate or inappropriate, moments where a listener response is *controversial* and moments neither a response was given nor a response was judged inappropriate (*neutral*).

Analysis of the context of the different type of response opportunities has shown different contextual characteristics that should help discriminating these response opportunities. Observations have shown relations to sentence structure (listener responses before (part of) the rheme is completed are considered inappropriate), conversational structure (listener responses in reaction to summarizing or refining statement are more appropriate) and proximity of earlier responses (producing two similar listener responses in close succession is considered inappropriate).

So far these characteristics are not used in state-of-the-art predictive models for the timing of listener responses (Morency et al., 2010; de Kok et al., 2010; Huang et al., 2010a). We feel that, in order to push these predictive models beyond the state-of-the-art, these characteristics should be taken into account. An obstacle towards the use of these characteristics, is the absence of real-time recognition systems of these characteristics on output of speech recognition software, such as theme and rheme discrimination within sentence and classification of statements and their relation to earlier statements.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics*, 9(1):1–26.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Iwan de Kok and Dirk Heylen. 2011a. Appropriate and Inappropriate Timing of Listener Responses from Multiple Perspectives. In *Intelligent Virtual Agents*, pages 248–254. Springer.

Iwan de Kok and Dirk Heylen. 2011b. The MultiLis Corpus - Dealing with Individual Differences of Nonverbal Listening Behavior. In Anna Esposito, Antonietta Esposito, Raffaele Martone, Vincent C. Müller, and Gaetano Scarpetta, editors, *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, pages 374–387. Springer Verlag.

Iwan de Kok, Derya Ozkan, Dirk Heylen, and Louis-Philippe Morency. 2010. Learning and Evaluating Response Prediction Models using Parallel Listener Consensus. In *Proceeding ICMI-MLMI '10 International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM Press.

Allen T. Dittmann and Lynn G. Llewellyn. 1968. Relationship between vocalizations and head nods as listener responses. *Journal of personality and social psychology*, 9(1):79–84.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010a. Learning Backchannel Prediction Model from Parasocial Consensus Sampling : A Subjective Evaluation. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 159–172.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010b. Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In *Proceedings of Autonomous Agents and Multi-Agent Systems*, Toronto, Canada.

Robert E. Kraut, Steven H. Lewis, and Lawrence W. Swezey. 1982. Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 43(4):718–731.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84.

Herwin van Welbergen, Dennis Reidsma, Zsófia M. Ruttkay, and Job Zwiers. 2010. Elckerlyc - A BML Realizer for continuous, multimodal interaction with a Virtual Human. *Journal on Multimodal User Interfaces*, 3(4):271–284.

Victor H. Yngve. 1970. On getting a word in edgewise. In *sixth regional meeting of the Chicago Linguistic Society*, volume 6, pages 657–677.