Analyzing Listener Gestures for ECAs

Iwan de Kok
i.a.dekok@student.utwente.nl
Teun Lucassen
t.lucassen@gmail.com

ABSTRACT
One of the most awkward parts of communicating with an embodied conversational agent (ECA) is still the part when the ECA listens to the user. We analyze multi party meetings from the AMI corpus to find out how gestures can contribute to the natural behavior of a listening agent. Head gestures are the main focus. Hand gestures do not seem to occur enough to use in this analyses. We perform a statistical analyses of annotated meetings. Most of the results turn out to be less representative for head to head conversations with an ECA, because the observations are based on multi party meetings.

Keywords
Listening behavior, head gestures, hand gestures, focus of attention, ECAs.

1. INTRODUCTION
In the development of embodied conversational agents (ECAs) one of the goals is to maximize the natural behavior. A lot of research has been done to enhance this behavior while an agent is speaking. Less focus is on the behavior when an agent is listening to the user. This results in a more awkward interaction which causes a negative effect on the realism.

Various researches show that responsiveness from listeners enhances the performance of the speaker. Multiple approaches have been taken in these researches.

Kraut et al [KRA82] have tested the quality of spoken summaries of a movie. One half of the speakers explained the movie to a non-responsive audience. The other half spoke to a fully responsive crowd. The speakers that received immediate feedback performed significantly better. The speakers and listeners communicated through a full-duplex audio channel. They were not in the same room. This means that the responses of the listeners were purely auditive.

Other responsiveness in terms of gestures may also be important in listening behavior. Most researches on gestures focus on the speaker. Gestures are used for various goals. They can be used to emphasize speech using rhythmic beat gestures. Other applications are iconic or symbolic gestures to illustrate the size or shape of objects. Turn management is also partly done by using hand gestures. An overview of all categories often used in the classification of hand gestures is given by Allwood et al. [ALL05].

Gestures of listeners is a relatively unexplored field in the relation between speakers and listeners. We study the hand and head gestures of listeners in a given corpus of multi-party meetings. The results of this study may be used in the development of the listening behavior of ECAs.

2. AMI MEETING CORPUS
The European-funded AMI project is a 15-member multi-disciplinary consortium dedicated to the research and development of technology that will help groups interact better [CAR06]. The AMI meeting corpus consists of hundred hours of annotated videos of meetings. These meetings are scenario-based. All meetings are attended by four persons with their own specific roles.

For our study we need annotations of the following aspects:

- Head gestures
- Hand gestures
- Focus of Attention
- Speech

Six meetings in the AMI meeting corpus are annotated on all these aspects. This gives us a total of two hours and fifteen minutes of video material.

We use the following meetings from the AMI meeting corpus.

Meeting	Duration
IS1008a, b, c, d	94 minutes 50 seconds
ES2008a	17 minutes 16 seconds
TS3005a	21 minutes 51 seconds
Total	2 hours 13 minutes 57 seconds

Table 1: Used Meetings from AMI corpus

We will refer to the four meetings of IS1008a, b, c and d as IS1008 and calculate corresponding statistics as being one single meeting. This is possible because all IS1008 meetings have the same participants.

3. HAND GESTURES
Before starting to calculate various statistics about gesturing we analyzed the video by watching all the material and the annotations. We conclude that hand gestures are virtually not used in listeners behavior.

Besides some isolated incidents the only time that listeners used hand gestures is when they try to get the floor. At this point the listener becomes the speaker when the attempt is successful.

Since this is no pure listener behavior we decided not to investigate the hand gestures any further. The annotations of the AMI meeting corpus are also not suitable for this kind of research at this time. Only pointing gestures are annotated, all other gestures such as rhythmic beat gesture are left out of the annotations. These should be included for a solid research.

4. HEAD GESTURES

Head gestures are annotated using the following five labels:

- Concord signal
- Discord signal
- Emphasis
- Other
- No Comment

Only concord and discord signals occur during listening behavior. These signals are mostly expressed in nodding and shaking. From now on we refer to these signals as head gestures.

4.1 Influence of Attention

When looking at the annotated video's we expect that attention is an important trigger for making concord or discord signals through nodding.

Based on this expectation we counted the nods when someone had the attention of someone else (A Gestures) and the nods when no one paid attention to the gesturer (NA Gestures). The results for the three meetings are listed below. The percentage of time that someone had the attention is also calculated.

Participant	Attention	A Gestures	NA Gestures
A	31.4%	196	201
B	33,7%	148	102
C	33,1%	176	88
D	27,6%	104	41
Total:		**624**	**432**

Table 2: IS1008 meetings

Participant	Attention	A Gestures	NA Gestures
A	44,3%	31	22
B	33,2%	24	17
C	31,4%	15	8
D	23,8%	17	15
Total:		**87**	**62**

Table 3: ES2008a

Participant	Attention	A Gestures	NA Gestures
A	12,8%	6	10
B	26,6%	14	16
C	27,2%	26	10
D	19,5%	12	17
Total:		**58**	**53**

Table 4: TS3005a

From this data we observe that A Gestures occur more then NA Gestures like we expected. This observation is reinforced by the fact that most participants only have attention of someone else one third of the time. Despite the fact that the attention time is shorter than the time without attention most gestures occur with attention.

Table 5 lists the observed gestures per minute with and without someone's attention. Note that the average is not simply the average of the listed numbers. The average is corrected with the length of all meetings. From this data we see that head gestures occur 2,5 times as much with attention than without.

Meeting	A Gestures/min	NA Gestures/min
IS1008	6,0	2,6
ES2008a	4,0	1,6
TS3005a	3,7	0,9
Average	**5,2**	**2,1**

Table 5: Gestures/min

4.2 The Timing Aspect

Now that we know that head gestures occur more often when having someone's attention we try to find out when these gestures occur. To do so we find out for each head gesture when this occurs in the time span of someone's attention. We first generate an overview of the length of the attention time spans. This is shown in the histogram in Figure 1.

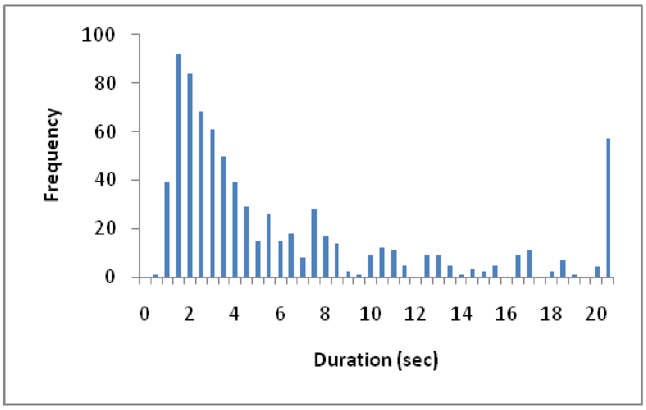

Figure 1: Durations of Attention Time Spans

The histogram shows that although most time spans last for less than 4 seconds there is a wide spread of durations. This gave us the idea to measure the occurrences of the head gestures relative to the duration of the time span. For instance if a nod is done 2 seconds after the start of an attention time span and this time span lasts for 10 seconds, the observation would be 0,2. This leads to the histogram in Figure 2.

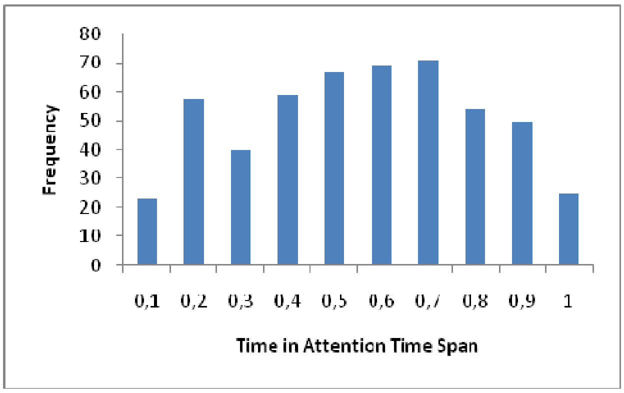

Figure 2: Head Gestures in Relative Time

From this figure we see that the head gestures are quite evenly spread over the time spans. Exceptions are the beginning and end of the spans. This is due to the fact that someone will not do a head gesture immediately when he is being looked at. On the other side a participant will not look away immediately when he received a concording or discording signal.

When we look at absolute time since the start from a attention time span we get the histogram in Figure 3.

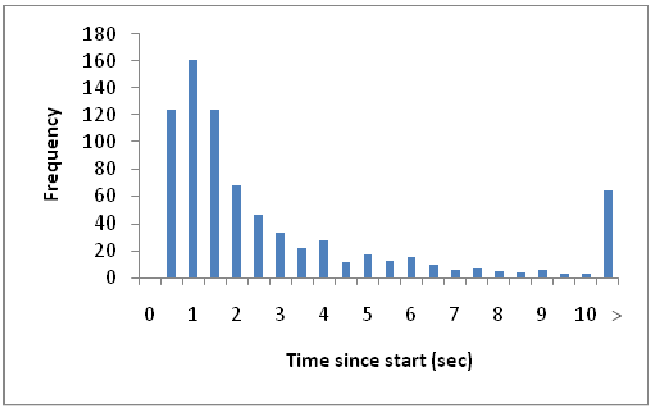

Figure 3: Head Gestures from Start in Absolute Time

We see that most of the gestures occur within 2 seconds from the start. This raises the suspicion that people tend to nod as a reaction to someone paying attention to them.

However, when we use the same methodology, but now for remaining time until the end of the attention we see a similar histogram. This is shown in Figure 4.

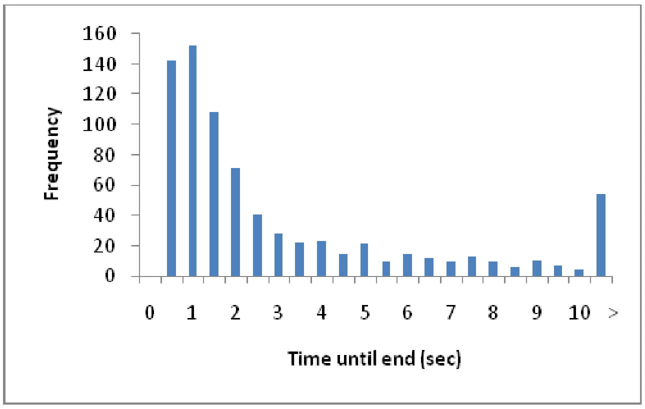

Figure 4: Head Gestures until End in Absolute Time

The fact that most head gestures occur within 2 seconds from the start of the attention time span and within 2 seconds from the end of this span complies with the fact that most spans are shorter than 4 seconds. It however does not say anything about the timing within the time spans.

4.3 Influence of Mutual Attention

Most of the head gestures that occur are done while someone is looking at the gesturer. This is the case in 56,8% of all head gestures.

When we look at the head gestures with someone's attention we see that in most of the cases the gesturer also pays attention to the attendee himself. We call this MA Gestures. When one of the two attending participants is the speaker at that time, we count this as a MAS Gesture.

The speaker is defined as the participant which is according to the speech annotation speaking some words. This implies that when someone takes a short pause, for example for a short breath, this person is no longer the speaker for this short period. It would be better to use some 'floor' annotation for this, but unfortunately this annotation is not available in this corpus.

Table 6 shows the counts for MA and MAS Gestures, compared to A and NA Gestures. These counts show that 69,2% of all gestures with someone's attention are done with mutual attention. This is 40,4% overall. 61,7% of these head gestures with mutual attention have one of the two participants as the speaker.

Meeting	**A Gestures**	**MA Gestures**	**MAS Gestures**	**NA Gestures**
IS1008	624	433	274	432
ES2008a	87	63	41	62
TS3005a	58	36	13	53
Total	**769**	**532**	**328**	**547**

Table 6: Mutual Attention

4.4 Mimicry

Something that was not directly observed in the annotated video's but that is suggested by literature as trigger for head gestures is mimicry. Maatman et al. suggest several reasons why mimicry is an important factor in natural behavior of ECAs [MAA05].

Before being able to count the cases in which mimicry applies we need to define mimicry in terms of occurrences in time. A head gesture of a listener is counted as a mimic if the speaker has done a similar head gesture in the last 2 seconds.

Table 7 shows the counts of the mimicked concord signals. Discords signals are not mimicked in our data.

Meeting	**Speaker concords**	**Mimics**
IS1008	271	187
ES2008a	51	53
TS3005a	37	19
Total	**359**	**259**

Table 7: Mimics

From this table we can extract that 72,1% of all speaker concords are mimicked. This figure is not exactly right. As one can see in Table 7 for ES2008a, more mimics occur than speaker concords. This is possible because there are four participants, making three listeners when one person is speaking. It would be more accurate to say that 359 speaker concords have caused 259 mimics when three listeners are attending.

5. DISCUSSION

During this research it became clear that it is not easy to create a model for natural listening behavior for ECAs. Nevertheless we did some observations, which we list here.

Listeners tend to nod more often when they have someone's attention. In most cases the listener also pays attention to that person. Unfortunately we were not able to conclude anything about the timing aspect. Mimicry is also an important trigger for head gestures. We see that in some cases more mimicked concords occurred than speaker concords, because the meetings were multi party conversations.

The fact that all meetings in the AMI corpus have four participants causes some problems in this research. We assume that conversations with ECAs are head to head, thus having two parties. A lot of observations in this research are caused by the fact that there are more than two parties and are therefore less representative for a dual party conversation. For instance mimicry gives false results because when more people are available to mimic someone, he will be mimicked more often.

Another problem with the AMI corpus is that all meetings are scenario-based. All participants are given a role description without any explicit personal goals. Since only a common goal (the design of a remote control) exists and the made decisions have no consequences whatsoever people tend to agree a lot. Some people did not seem to care about the project at all and were not motivated to contribute to the conversation in a significant way and even tried to avoid disagreement. We see that of all given head signals only 3,6% are discord signals. Of course this is not a realistic situation. The risk of creating a model not reflecting real life is fairly large when one uses this corpus.

Hand gestures are only annotated on pointing and emphasis gestures. The fact that this does not occur often in listening behavior has kept us from further research in this direction. In fact, pointing and emphasis gestures are only a small subset of all hand gestures. Occurrences like someone writing something down or beat gestures could be very interesting in the research of natural behavior in meetings. It is disappointing that the annotations do not contain this information.

Another issue with the annotations is the timing. When we look at the focus of attention, it is important in our research that the beginning and ending of the attention time spans are accurate. Also the timing of head gestures is important when looking at mimicry. We noticed that the timing of these annotations were not always strictly matching to the corresponding video. This causes unreliable results.

During our research we encountered the problem that there is no annotation of ‘floor’. As a workaround we used the speech annotation as floor annotation, but is not accurate in situations like small pauses in spoken utterances or when people are back channeling. We tried to avoid false speaker annotation of back channels by only classifying someone as a speaker when he speaks more than two words. Of course this is not foul proof. The annotation of ‘floor’ could give valuable information about dominance in meetings and the methods participants use to obtain the ‘floor’.

It was a bit disappointing that from such a huge corpus as the AMI corpus only 2 hours and 15 minutes of material was suitable for our research. This makes the results less substantial. The reason that a lot of material was not of any use was the fact that only a few meetings are annotated on all required aspects.

When trying to implement some of our findings in an embodied conversation agent some problems will be encountered. When an agent has to give concord or discord signals during the utterance of a user, real-time understanding and evaluation of the user’s speech is required. This is not available yet. Therefore this type of signaling can not be fully implemented without making guesses about the agreement between the user and the ECA.

Something that can be implemented without too many problems is mimicking. This would require real-time analysis of the user’s head movements which is already available and applied to cases like The Virtual Dancer [REI06]. Of course the agent would still not have any information about the level of agreement.

REFERENCES

[ALL05] Allwood, J., Nivre, J. and Ahlsén, E. On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics*, Vol. 9 No. 1 (1993).

[CAR06] Carletta, J. Announcing the AMI Meeting Corpus. *The ELRA Newsletter,*11-1 (2006), p. 3-5.

[KRA82] Kraut, R.E., Lewis, S.H. and Swezey, L.W. Listener Responsiveness and the Coordination of Conversation. *Journal of Personality and Social Psychology,* Vol. 43 (1982), p. 718-731.

[MAA05] Maatman, R.M., Gratch, J., Marsella, S. Natural Behavior of a Listening Agent. *Proceedings of the International Conference on Interactive Virtual Agents* (2005).

[REI06] Reidsma, D., Nijholt, A., Poppe, R., Rienks, R., Hondorp, H. Virtual Rap Dancer: Invitation to Dance. *CHI ’06 Extended Abstracts on Human Factors in Computing Systems* (2006), p.263-266.