# Deciphering the Silent Participant

## On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions

Catharine Oertel
KTH
Royal Institute of Technology
Linstedtsvägen 44
Stockholm, Sweden
catha@kth.se

Kenneth A. Funes Mora
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne (EPFL)
Switzerland
kfunes@idiap.ch

Joakim Gustafson
KTH
Royal Institute of Technology
Linstedtsvägen 44
Stockholm, Sweden
jocke@speech.kth.se

Jean-Marc Odobez
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne (EPFL)
Switzerland
odobez@idiap.ch

## ABSTRACT

Estimating a silent participant's degree of engagement and his role within a group discussion can be challenging, as there are no speech related cues available at the given time. Having this information available, however, can provide important insights into the dynamics of the group as a whole. In this paper, we study the classification of listeners into several categories (attentive listener, side participant and bystander). We devised a thin-sliced perception test where subjects were asked to assess listener roles and engagement levels in 15-second video-clips taken from a corpus of group interviews. Results show that humans are usually able to assess silent participant roles. Using the annotation to identify from a set of multimodal low-level features, such as past speaking activity, backchannels (both visual and verbal), as well as gaze patterns, we could identify the features which are able to distinguish between different listener categories. Moreover, the results show that many of the audio-visual effects observed on listeners in dyadic interactions, also hold for multi-party interactions. A preliminary classifier achieves an accuracy of 64%.

## Categories and Subject Descriptors

H5.3 [**Information Interfaces and Presentation**]: Group and Organisation Interfaces—*Theory and models*

## Keywords

listener categories; non-verbal cues; eye-gaze


*ICMI'15,* November 09-13, 2015, Seattle, WA, USA

ACM 978-1-4503-3912-4/15/11 ...$15.00.
http://dx.doi.org/10.1145/2818346.2820759.

## 1. INTRODUCTION

### 1.1 Motivation

In recent years, there has been a growing interest in studying the interaction between humans and robots [22, 27, 16]. In particular, in order for a human and machine to seamlessly communicate, a good understanding and modelling of the turn-taking phenomenon that underpin human communication [17] is essential. In this direction, there have been several attempts at building statistical or computational models of how turn-taking is coordinated (e.g., [21, 20]). On the output side, several virtual avatar systems have made use of a combination of cues from speech, posture shifts and head movements in order to generate convincing interactive behaviours including active listener ones [19, 7].

Yet, turn-taking modelling in itself is not sufficient when it comes to designing collaborative robots which detect and react appropriately to different dynamics in interactions. For example, participants might loose interest in the conversation. They might disengage and avert their attention and then suddenly, due to a shift in topic, re-engage again. All of these dynamics are encoded in multi-modal social signals that need to be well understood and decoded to make the appropriate interaction move. There are several studies which try to model the engagement of the individual participant andor the general involvement of the group e.g. [25, 12]. To our knowledge there is however, no study which focuses on the listener in a multi-party setting and tries to model the dynamics in listener states.

One can think of many applications where it would be useful to characterize the members of a group discussion that are not currently speaking. One is embodied or virtual tutors in collaborative settings with groups of students e.g. as in [18, 31]. Knowing more about the cognitive state of the participant could help to for example better target teacher interventions. A complicating factor for these kinds of applications is that a system in such contexts would need to handle multiple users and open domains [8]. The use of

more robust low level cues could be therefore advantageous in such scenarios.

In the current paper, we therefore focus on the analysis of low-level audiovisual cues to characterize and classify silent participants in multiparty interactions.

## 1.2 Background and related work

In the following section, we are going to give a brief overview on literature concerning participation categories as well as research on audio-visual cues in conjunction with interaction modelling.

### *1.2.1 Participation Categories*

Clark [9] building on Goffman [14] described participation categories in the following way. First of all he distinguishes between participants and non-participants. The former include all people taking part in the conversation, such as the speaker, the current addressee, but also people which are not currently being addressed but still belong to the circle of ratified participants; they are referred to as side–participants. Clark calls everyone else an "overhearer", "Overhearers" can be divided into two sub–categories namely "bystanders" and "eavesdroppers". A"Bystander" is a person who is openly present, but not part of the conversation whereas an "Eavesdropper" is overhearing the conversation without the other participants' awareness.

In the current paper, we are building up on these definitions but are adapting them for the specific task of listener classification. Therefore, we are exchanging the term addressee for "attentive listener"', and are otherwise only using the terms "side-participants" and "bystanders" in order to distinguish between different listener types. For the exact definition of listener categories given to the annotators please refer to section 2.4

### *1.2.2 Gaze Patterns*

Concerning visual cues or more specifically gaze patterns we know from dyadic interactions that the speaker and listener are asymmetrical in that the listener looks at the speaker for long periods of time, while the speaker looks at the listener in short, but rather frequent periods [3]. Vertegaal [30] also found this pattern in multiparty interaction where participants gazed at an interlocutor 1.6 times more often while listening than while speaking. This means that it typically is the speaker that controls when mutual gaze occurs. Bavelas et al [4] found that listeners' verbal and nonverbal feedback was most common during periods of mutual gaze (gaze window). They also found that listener feedbacks often led to gaze aversion from the speaker, thus ending the periods of mutual gaze.

Gaze patterns are also reported to be related to participants' engagement level. Oertel and Salvi [26] for instance found that modelling whether participants were gazing at other participants or downwards described well the engagement level of participants. Moreover, Bednarik et al [5] found that participants with low engagement levels had long gaze durations at the same interlocutors (few gaze target shifts), while the engaged participants had shorter on-target gaze durations, but at a larger number of interlocutors (many gaze target shifts).

### *1.2.3 Verbal and Non-verbal backchannels*

Up to this point in time, only few studies have been carried out exploring the relationship between visual and verbal backchannels. Truong et al. [29] for example find that in face-to-face conversation eye-gaze appears to be a cue for backchannels. They show that mutual gaze occurs significantly more often during visual backchannels. In a preliminary study Bertrand et. al.[6] noticed that when the speaker is gazing at the interlocutor, the latter produces a succession of gestural BCs. In their data they however did not find gestural backchannels followed by vocal backchannels during speaker's gaze towards the interlocutor.

None of these studies to our knowledge, however, investigates the occurrences of verbal and visual backchannel in relation to various listener types.

### *1.2.4 Dyadic versus multiparty conversations*

Multiparty interaction differ from dyadic interaction in several regards [28]. One distinguishing feature is that in dyadic interactions there are only two different roles that and interlocutor can take on: speaker and listener. In multiparty interaction, humans may take on many different roles, such as speaker, addressee, side-participant and bystander or overhearers, as described above [9].

## 1.3 Contribution and paper organization

In the current paper we investigate human assessments of the listener categories attentive listeners, side-participants and bystanders and relate these to third-party assessment of listener's degree of engagement. First, we investigate whether gaze patterns found for dyadic interactions also hold for multi-party interactions. Second, we investigate to what degree it is possible to distinguish between different listener roles using only the low-level audio-visual cues *speech/no-speech*, *gaze* and *audiovisual backchannels*. Finally, we propose to use the different cues to train a classifier aiming at classifying the different listener categories and report on these results. The rest of the paper is organized as follows. In section 2 we present our corpus and thin slice perceptual test experiments. In section 3, we present the feature extracted from the corpus. Section 4 present our results while section 5 concludes the work.

# 2. OBSERVER ANNOTATIONS

In this section, we describe and motivate the experimental protocol used to collect people impressions on listeners. We start by describing the corpus of interactions we used, and then introduce the different elements of the web-based annotation experiments.

## 2.1 Corpus

In order to be able to investigate group dynamics we recorded a corpus of four-party interactions with a large number of sensors [24]. The main features are briefly summarized below.

**Set-up and sensors.** The set-up is shown in Fig. 1. Four people are interacting around a table, and their behaviours are recorded using several per-participant synchronized sensors: close-talking microphones, Windows Kinect 1 sensors positioned at around 0.8 meters of participants, and high resolution GoPro cameras.

Figure 1: The interaction setup.

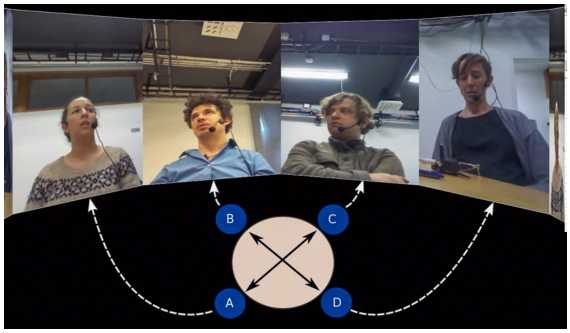


Figure 2: Annotation input. Observers were provided with a 15 second video clip, as shown above.

**Scenario.** The multiparty conversation data is in the domain of group-interviews. More precisely, each interview session involved four participants: an interviewer also called moderator (post-doc) and three interviewees (PhDs). All participants were made aware of the interview goals: PhD students were told that the moderator's purpose in the interview would be to find out who would be the most qualified for a prestigious scholarship. They were told that the interviewer could either choose all of them, two, one or none.

The interviews consisted of different phases, with different interaction dynamics: self-introduction, PhD work description along with potential impact on society, and brainstorming on proposing a joint project. It has to be noted that the group never splits into subgroups. There are only few stretches of speech in which participants overlap, but these stretches are mainly due to simultaneous speaker speech and listener feedback. There are some stretches of one-to-one interactions, however, these cannot be called "dyadic" as such. A typical example of such a situation would be the participants are formally introducing themselves to the moderator, where, in actual fact, the participants do not only introduce themselves to the moderator but to the group as a whole. There was also quite some variation in the moderators behaviour. Some moderators behaved as "Attentive Listeners" while others purely as "Bystanders".

**Dataset.** The corpus consists of five interactions of groups of four. Each interaction lasted for about an hour, which results in approximately 5 hours of recordings of multimodal and multiparty data.

## 2.2 Thin slices

To obtain annotations about listeners, we adopt a thin slice approach [1]. Sample video clips containing at least one non-speaking person are extracted from the corpus and shown to the annotators who then have to fill a questionnaire. The slice duration was selected as a compromise between having segments long enough to base judgement on sufficient evidence, and short enough to avoid mixed behaviours. Following previous works on engagement [5] or interest-level inference [13], we opted for a duration of 15 seconds.

## 2.3 Visual clip content

The video content displayed to annotators is shown in Fig. 2. It is composed of the video data from the Kinect sensors of each participant, transformed and joined together in order to provide viewers with a feeling of the 3D setup, and in particular get a better grasp of the gaze attention targets. A top-view table layout at the bottom of the display further allows to identify the geometric relation between participants as well as the listener to annotate.

As an important aspect of the experiment, no audio was added to the video content. This served two main purposes. First, this removed the influence of the semantic speech content on the understanding of the listeners' types, and helped viewers to concentrate on the non verbal behaviour content. Note however that the video resolution was high enough for people to identify who was speaking. Secondly, since annotators would watch several windows from the same interactions, this was avoiding people to identify the role played by participant, and in particular, the moderator role.

## 2.4 Questionnaire, protocol and annotators

Our primary goal was to obtain annotations about the listener type in each clip. As motivated earlier the proposed categories and their definition as given to annotators were the following:

- Attentive Listener (ALi): An attentive listener is a person, who is most likely to start speaking after the current speaker.
- Side Participant (SPa): A side participant is a person who is part of the group of potential future speakers, but is probably not the next speaker.
- Bystander (Bys): A bystander is a person who the group of potential future speakers is aware of, but who they do not expect to speak in the near future. The bystander acts as an observer rather than a potential future speaker.

In addition, to assess whether the perception of the listener type matched other dimensions, the following ratings were also asked:

- Engagement (Eng): rate the listener's degree of engagement in the conversation using the seven point scale, ordered from not engaged (1) to highly engaged (7);
- Expected time to speak (ETS): rate at what point in time you expect the listener to speak next using the seven point scale: now (1); soon in the next 30 seconds (3); within the next minute (5) never (7).

Finally, besides the description of the overall task, interaction setup, and questions and ratings, annotators were also instructed to preferably watch the clip at least twice, e.g. once focusing on the specified participant and the second time focusing on the group, and only afterwards make their ratings. In addition, 5 samples selected to present a diversity of listener behaviours were asked to be annotated

(but not used in analysis) to have users getting accustomed to the content and display. In total there was a pool of 25 annotators. They were not compensated for the time spent. We allotted round 40 minutes worth of annotations per annotator. All of the annotators were naive to the research topic and mainly sampled from a pool of post-graduate students who shared a similar cultural background to that of the participants.

## 2.5 Annotation window selection

As not all the 5 hours of data can be annotated, windows were selected to maximize diversity of situation and listeners as follows. In each window, the role of each participant was defined as the moderator ($Mo$), and for each interviewee, as speaker ($Sp$) if they spoke (which excluded backchannels), or as listener. The latter was further separated as $La$ and $Lb$ listeners according to whether the duration between the last time that they spoke and the start of the window (denoted **LT**) was lower or greater than 20 seconds. Window were then assigned a category according to the role combinations of the 3 interviewees: ($Sp$,$La$,$La$), ($Sp$,$La$,$Lb$), ($Sp$,$Lb$,$Lb$), ($Sp$,$Sp$,$La$), ($Sp$,$Sp$,$Lb$).

Using these definitions, 600 windows were sampled according to the following rules: i) windows where speakers spoke for less than 3 seconds were discarded; ii) windows were sampled uniformly across the 5 interviews and the 5 category types; iii) given a window, the listener to be annotated was sampled uniformly amongst the pool of listeners, which included the moderator.

# 3. DATA PROCESSING

In the following subsections we first describe the audio and visual cues which were extracted automatically or semi-automatically by processing the audio-visual data. In a second step, we introduce the different features that were computed from these cues to characterize the listeners in each temporal window of the thin slice experiments.

## 3.1 Automatic and semi-automatic cue extraction

### 3.1.1 Speaking turns and audio backchannels

**Voice Activity Detection** Voice Activity Detection (VAD) of the interlocutors was carried using a speech recognizer. The obtained voicing segments were manually checked by a phonetician for their accuracy. In the current paper we distinguish between two types voicing segments: "backchannels" and "normal speech". A voicing segment had to fulfil two conditions to be annotated as a backchannel: 1) it had to be a very short utterance as defined by [15] and 2) it had to be surrounded by "other speech". All other segments were classified as "normal speech".

**Turn Detection** Turns are inferred from the speaking activity. Only one participant can hold the floor at a given time time. A turn is then maintained until another participant talks for more than 1 second. In that case, the turn is assigned to that participant from the moment he/she starts speaking. The time in between is not assigned to any participant. In the case of an overlap, i.e., the turn is handed to the second participant only if the first participant stops talking before the second finishes. In that case, the turn change is defined at the instant the first participant ends talking.

### 3.1.2 Gaze annotations and labels.

As manual annotation of gaze is tedious, we followed a similar approach to [10] that was shown to provide up to 90% gaze coding accuracy in dyadic interactions. We extended this method to the group case, and used as gaze estimation method the code of an improved version of [11].

First, we calibrated the setup to obtain a single world coordinate system in which it is possible to refer all measured quantities. This was achieved in a similar way than [10], by exploited fitting planes to wall measures and the prior on sensor pose and locations.

**Multi-party gaze coding.** Looking at other participants from participant $i$ was measured by leveraging the availability of 3D information as follows. We defined a gaze reference vector $\hat{\mathbf{v}}_i^k$ for looking at each other person $k$. Provided that the gaze estimation output is $\mathbf{v}_i^{\mathbf{h}}$, we define the gaze angle from the participant $k$ as:

$$\psi_i^k = \arccos\left(\mathbf{v}_i^{\mathbf{h}} \cdot \hat{\mathbf{v}}_i^k\right) \tag{1}$$

from which we can derive the gaze target $T_i$ for the participant $i$ as:

$$T_i = \begin{cases} c & \text{if } \psi_i^c < \tau \\ -1 & \text{otherwise} \end{cases}, \tag{2}$$

where $c = \arg\min_k\{\psi_i^k\}$, i.e., the closest participant in terms of gaze direction, and "$-1$" is the *background* class, which indicates that the participant is not looking at any of the targets.

**Final gaze label set.** So far we have discriminated between looking at any of the participant or elsewhere (background). However, further splitting the background class into sub-classes such as: "up", "down", or "mid-targets" could be informative, as, for instance, [2] showed that different gaze-away directions could be more characteristic of different functions (turn taking, management, cognitive load, intimacy, etc). In the current case, we further added looking down as a label which may indicate disengagement. This was obtained by monitoring the pitch angle in the world coordinate system. So, in summary, the gaze of people was labelled according to the following set $\mathcal{G}$ of labels: each of the three other participants, Up, Down, Others.

### 3.1.3 Visual backchannel (nodding)

To annotate the visual backchannels, we used an approach similar to [23] developed for nodding recognition from video data. More precisely, we used the head orientation (pan, tilt, roll) from the head pose tracker to derive head pose dynamics. The resulting time-series were used as input to a set of Gabor filters and then classified as nod or not nod using a SVM classifier using radial-basis functions. While the method has difficulties to identify nods of speakers, its reliability to infer nods from listeners is usually high, with an accuracy of more than 80% measured on natural interactions.

## 3.2 Listener characterization

To characterize the listeners and interaction situations, we extracted for each listeners and for each clip a set of audio-visual (AV) features related to gaze and backchannels. More precisely, we computed:

- Visual backchannels (VisBack): the number of nods detected in the window duration;

- Audio backchannels (AudioBack): the number of verbal backchannel uttered by the listener in the window duration;
- Gaze at speaker(s) (GazeAtSpeak): percentage of window frames in which the listener is looking at the person currently speaking;
- Gaze received from speaker(s) (GazeFromSpeak): percentage of window frames in which the actual speaker look at the listener;
- Mutual gaze with speaker (MutGazeSpeaker): percentage of window frames in which the listener and the current speaker look at each other;
- Gaze down (GazeDown): defined as the percentage of window frame where the listener look down/in front of him. Similarly to Andrist et al [2] who showed that different gaze-away directions could be more characteristic of different functions, here we hypothesized that looking down (rather than elsewhere or at any participant) could be a visible indicator of bystanders or side participants disengagement.

# 4. RESULTS AND DISCUSSIONS

In this Section we present our results. We first comment on the results of the perception annotations. In Section 4.2, we analyse which of the cues distinguish the different listener types, and comment on the relationship with findings in the literature. In Section 4.3, we analyse verbal and visual backchannel during periods of mutual gaze. We relate those findings to findings reported for dyadic findings in the literature. Finally, in Section 4.4, we present some classification results.

## 4.1 Perception Test Analysis

Video frames were annotated by different amount of people, ranging from 5 to no annotators. The majority of video samples received 4 annotations. We only considered the samples which received at least 3 annotations, and assigned them to different categories according to the annotation configuration: single class (ALi, SPa, Bys) if at least 3 annotators agreed on this class. In-between categories if there was a tie between the 2 categories, and no vote for the 3rd one. This corresponds only to 2-2 ties in 4 annotator samples. The "no majority" class was assigned to the other cases.

Results are shown in Table 1. As can be seen, in 67.9% of the cases, there was a majority for a class. In addition, in the case of ties, these corresponded mainly to the plausible ones: ALi-SPa and SPa-Bys for 15% of the cases, as opposed to confusing an attentive listener with a bystander (ALi-Bys). Samples with at least one vote in each category were observed in 15% of the cases.

These results show that 1) people are usually able to agree on different listener categories; 2) do so based on visual information alone; and 3) these categories form some kind of continuum. Indeed, as will generally be shown below, the measured cues for these in-between categories often correspond to values falling in between the values taken by their 'pure' listener counterparts.

We further analysed the variation of a) Eng and b) ETS for the different listener categories. Results are depicted in Figure 3. ANOVA tests revealed that the listener categories had an effect on both ENG $F(4,314)=114.8$ ($p<0.001$) as well as ETS $F(4,314) = 309.2$ ($p<0.001$). A Bonferroni post-hoc test corroborated that the ALi was rated significantly higher in terms of ENG than the SPa ($p<0.001$), which was himself rated significantly higher than the Bys ($p<0.001$). Similarly, the ALi was estimated to speak significantly sooner than the SPa($p<0.001$) and the SPa significantly sooner than the Bys($p<0.001$).

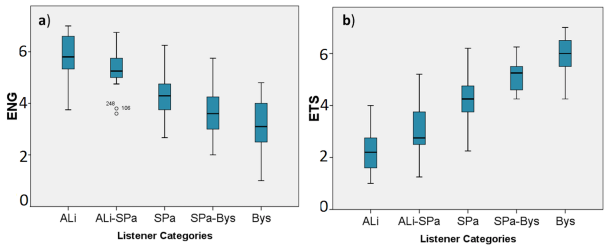


Figure 3: Eng and ETS results for different listener categories.

These results show that, as expected, observers were consistent in evaluating the degree of engagement and the time-to-speak. Still, the large variance measured also shows that there was quite some variability amongst them in general.

## 4.2 AV Analysis of Prototypical Listeners

**Methodology.** In this section, we evaluate how much different audio, visual, or audio-visual cues are characteristic of listeners. To do so, we compute the mean and standard deviations of the cue for each of the annotation configuration (five first columns of Table 1). Note that this comprises the in-between listener categories configuration, to better illustrate that cue values measured in these cases are indeed a mix of the values measured from the prototypical samples. Note however that to test the significance of the potential differences of the cues between the different prototypical listeners, we applied significance tests only on the purer samples (i.e having at least 3 agreeing annotators).

**Gaze pattern analysis.** We first evaluate how much gaze is different across the listener categories. Figure 4 depicts gaze distributions for the different features. Surprisingly, an ANOVA test revealed that there was no significant effect of the listener categories on the amount of "Gaze At Speaker", indicating that observers did not distinguish listeners based on their lack of attention to the speaker. However, there was a significant effect of the listener category on the amount of gaze from the active speaker $F(4,314)=34.5$, $p<0.001$. A Bonferroni post-hoc test revealed that the ALi receives significantly more gaze from the speaker than the SPa ($p<0.001$), and the SPa more than the Bys ($p<0.001$). Moreover, there was significant effect of the listener categories on the Mutual Gaze with the Speaker $F(4,314)=20.2$ $p<0.001$. A Bonferroni post-hoc test revealed that the ALi had significantly more mutual gaze with the speaker than the SPa($p<0.001$) and that SPa has share significantly more mutual gaze with the speaker than the Bys ($p<0.01$).

Overall, this is in line with Vertegaal [30] who found that in multiparty interaction participants gazed at an interlocutor 1.6 times more often while listening than while speaking, meaning that it typically is the speaker that controls when mutual gaze occurs. However, it also extends Vertegaal's findings in that it further describes the gaze distribution between the speaker and different listener categories. While our findings confirm that it is indeed the speaker who controls when mutual gaze occurs, they also show that the

Table 1: Distribution of samples across the different annotation configuration (denoted AConfig). NotEnAnnot stands for not enough annotation (i.e. the sample was annotated by less than 3 people). The percentages are given with respect to the number of samples with enough annotation.

| AConfig | ALi | SPa | Bys | ALi-SPa | SPa-Bys | ALi-Bys | No Maj. | NotEnAnnot |
|---|---|---|---|---|---|---|---|---|
| Amount | 62 (16.1 %) | 82 (21.2%) | 118 (30.6%) | 21 (5.4%) | 37 (9.6%) | 7 (1.8%) | 59 (15.3%) | 214 |

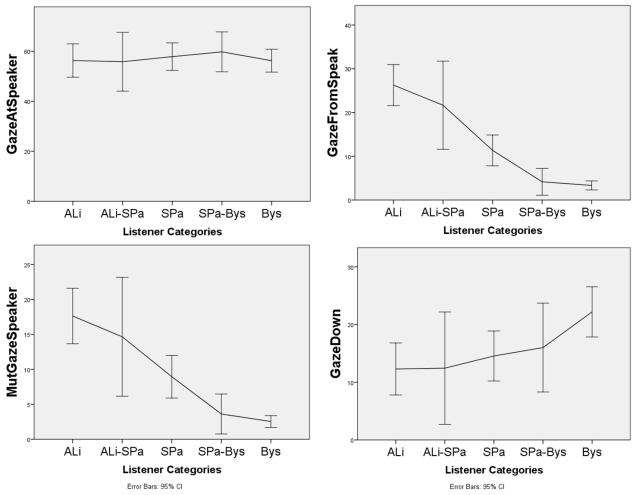


Figure 4: Gaze in Prototypical Listener Categories. See Section 3.2 for the definition of these measures.

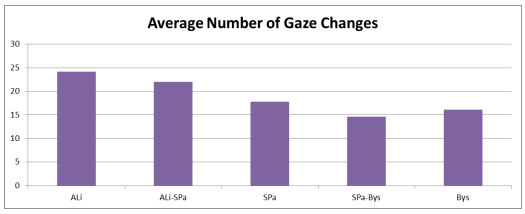


Figure 5: Number of Gaze Changes

speaker does not establish mutual gaze with all possible interlocutors equally but establishes mutual gaze most often with the ALi, who is the listener who is the most engaged of all listeners, and also indeed the one who most probably has spoken last (as will be shown later, cf Fig. 7).

Finally, there was a significant effect of the listener categories on GazeDown $F(4,314)=2.9$. A Bonferroni post-hoc test revealed that the ALi looks significantly less down than the Bys($p<0.05$). This is in line but also extends the findings of [26]. Oertel and Salvi found that in comparison to any other feature, "presence" (the feature describing whether participants were gazing at other participants or averting their gaze upwards or downwards) was the feature which distinguished best between low and high group involvement. We make a similar finding in that, also in this study, the more engaged participants are the ones who look more towards other participants than averting their gaze downwards. However, Oertel and Salvi did not distinguish between speakers and different listener categories and when participants are estimated to speak again. Interestingly, this study thus adds that silent participants who are more often averting their gaze downwards are judged to speak further in the future than silent participants who gaze towards another participant.

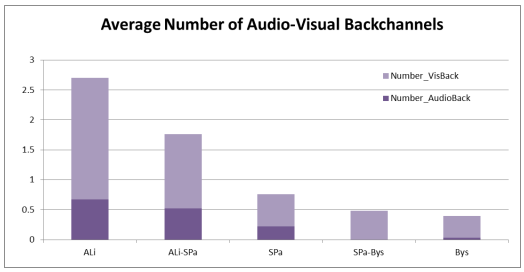


Figure 6: Number of Audio and Visual Backchannels

**Gaze changes.** Figure 5 depicts the number of gaze changes per listener category. As can be seen the Attentive Listener has the highest number of gaze changes, followed by the Side Participant and then finally by the Bistander. Although there is an apparent trend the $\chi$ 2 test does not show a significant effect.

Nevertheless, this trend, while not significant, is in line with findings of Bednarik et al [5] who found that participants with low engagement level had long gaze durations at the same interlocutors (few gaze target shifts), while the engaged participants had shorter on-target gaze durations, but at a higher number of interlocutors (many gaze target shifts). The fact that the number of gaze changes across listener categories, and indirectly listener engagement level, was not significant in our study might be due to two reasons. First of all, given the limited group size (three main interlocutors engaged in the conversation in general), showing too obviously disengagement might have been perceived as being rude by the others and might have lead the participants to try and minimize such disengagement cue. The second reason might be due to the fact that participants knew that they were being monitored and evaluated by the moderator and thus should try to appear as best as they could.

**Backchannels.** Figure 6 depicts the number of audio and visual backchannels per listener category. It can be observed that both the number of non-verbal backchannels A $\chi$ 2 (36, N=319)=107.338 ($p<0.001$) as well as verbal backchannels decreases significantly $\chi$ 2 (20, N=319)=74.070 ($p<0.001$) from Attentive Listener to Bystander. It is also noticeable that in all listener categories the number of visual backchannels is higher than the number of verbal ones. This fact is most pronounced for the Attentive Listener.

## 4.3 Backchannels during Mutual Gaze

Given the differences observed in amount of backchannels, we further analysed during which periods the production of backchannel in the corpus was more important. Following Bavelas et al [4] who found that listeners' production of verbal and non-verbal feedback was around height times higher than change during mutual gaze in dyadic situations, we extended this work to our multi-party case and counted the

Table 2: Listener backchannel behaviours according to the gaze activity with respect to the ''active" speaker I) Mutual gaze between the listener and the speaker. II) The listener gazes at the speaker. III) The speaker gazes at the listener. IV) Other. Table a) shows the estimated number of backchannels per minute; b) reports the total count of backchannels in all 5 interviews; while c) reports the total cumulated time of the different gaze situations states.

a)

| Modality | I | II | III | IV | Chance |
|---|---|---|---|---|---|
| Audio | 2.728 | 0.378 | 2.627 | 0.508 | 0.59 |
| Visual | 5.233 | 1.601 | 4.970 | 3.327 | 2.44 |
| AudioVisual | 7.960 | 1.980 | 7.597 | 3.835 | 3.03 |

b)

| Modality | I | II | III | IV |
|---|---|---|---|---|
| Audio | 147 | 90 | 83 | 69 |
| Visual | 282 | 381 | 157 | 452 |
| AudioVisual | 429 | 471 | 240 | 521 |

c)

| | I | II | III | IV |
|---|---|---|---|---|
| - | 3233.5 | 14274.8 | 1895.4 | 8151.7 |

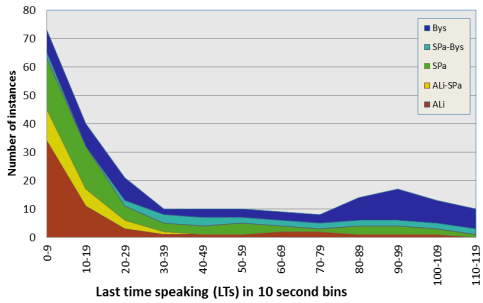


Figure 7: LTS Distribution

number of backchannels happening during specific periods of gaze patterns between the speaker and listener, as reported in Table 2. Using these numbers, we found out that in our data, the same effect was observable, though less pronounced: comparing columns (I) and the Chance one in Table 2a, we can note that it is 4.6 times higher than chance to have an audio verbal backchannel during mutual gaze, and 2.14 times higher for the visual (i.e. nods) one.

However, from Table 2a, an interesting finding that actually differs from Bavelas is that, in fact it is the speaker's gaze, that controls the amount of backchannels: backchannels are most common when the speaker looks at the listener - regardless if the backchanneling listener is gazing back at the speaker or not (around 7 AV backchannel per minute on average), whereas the least amount of backchannels (around 2 per minute) are obtained when the backchanneling listener gazes at the speaker, who looks at somebody else.

## 4.4 Predicting Listener Categories

In this section, we present our early experiment on classifying the different listeners.

**Prediction using LT**, defined as the duration between the last time when the listener last spoke and the start of the analysis window. Indeed, one hypothesis of the study was that the more recently a listener had spoken, the more attentive he would be. Surprisingly, given that observers did not have this information, this hypothesis was verified in practice. Figure 7 illustrates the distribution of LTs across the different listener categories. It can be observed that the greatest accumulation of ALis can be found in the period between 0 to 30 seconds, whereas the reverse is true for Bys. The majority of listeners rated as Bystanders spoke last between 70 to 120 seconds before the start of the analysis window. The distribution of SPa is more evenly spread with a slightly higher percentage in the period of 0 to 30 seconds. We used LT alone to perform classification using a decision tree. We obtained a classification accuracy of 52.8%.

**SVM prediction.** To test whether the found effects of the different cues depending on the prototypical listener categories would also be sufficient for classification, we used Support Vector Machine (SVM) classifiers based on the RBF kernel. We applied a grid-search with 10-fold cross validation to identify the hyper-parameters. We used one-vs-one classifiers. As features, in addition to LT, we used all those discussed in the above sections which were shown to have a significant effect (GazeFromSpeak, MutGazeSpeaker, MutGazeListeners, GazeDown, AudioBack, VisBack). We obtained an accuracy of 64.4%. It has to be noted, however, that a complicating factor was the fact that listener categories were not evenly balanced in terms of sample size. The sample size of Bys instances, for example, was higher than the sample size of the SPas. Moreover, the category of SPa lay in-between the other two categories which made classification of the category of Spa more challenging. When evenly balancing the classes in terms of sample size, by reducing the sample set to 62 samples per class, the accuracy is reduced to 54.1%, which corresponds to an 21% improvement over majority class classification (33.3% acc.), which was nevertheless disappointing.

# 5. CONCLUSION AND DISCUSSION

This paper investigated the characterization of listeners categories in multi-party situations. To this end, we relied on a corpus that we recorded using a range of sensors (close-talking microphones, Kinect), allowing for the automatic retrieval of accurate voice activity, eye gaze and visual backchannels annotations. This avoided the need for costly and time consuming manual annotations. We further proposed a thin-slice approach to collect observer annotations about three listener categories. Results showed that observers usually agreed on their labelling, making it possible to identify these categories from discriminative low-level audio visual cues. The advantage of low-level cues in comparison to high-level is that they are more robust and domain independent.

Our work builds on research carried out for dyadic scenarios, and extends it the multi-party case, using substantially more amounts of data than reported in previous findings about listeners. For instance, we extend Vertegaal's findings by further describing the gaze distribution between the speaker and different listener categories, and showed that silent participants who are more often averting their gaze downwards are judged to speak further in the future than silent participants who gaze towards another participant. Or we were able to extend to the multiparty case Bavelas findings that listener backchannel occur more during mutual gaze between a speaker and a listener. Indeed, in this later case, as one specific finding of our work, we showed that what might matter even more to trigger listener's backchannels is the speaker gaze, regardless of the listener's own gaze. Finally, a preliminary classifier achieves an accuracy of 64.4% in distinguishing between prototypical

listener categories.

The work can be improved in several ways. For instance, while we believe that the audio-visual models of listeners we have extracted are quite representative of listeners in general discussions, it would be interesting to verify whether they hold (and which feature are affected) in other settings (e.g. standing people) or other scenarios. On the classification side, performance could be increased by increasing the sample size, by adding prosodic analysis of the verbal backchannels, by extending the feature sets to include further multi-modal cues (and better characterizing the dynamics as compared to the aggregation statistics currently used), or by studying the impact of changing the analysis window size.

**Acknowledgments**. The authors would like to acknowledge the support from the Swedish Research Council Project VR(2013-4935) and the Swiss National Science Foundation (Project G3E, 200020_153085) `www.snf.ch`.

# 6. REFERENCES

[1] N. Ambady, F. J. Bernieri, and J. A. Richeson. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in experimental social psychology*, 32:201–271, 2000.

[2] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu. Conversational gaze aversion for humanlike robots. *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*, pages 25–32, 2014.

[3] M. Argyle and M. Cook. Gaze and mutual gaze. 1976.

[4] J. B. Bavelas and J. Gerwing. The listener as addressee in face-to-face dialogue. *International Journal of Listening*, 25(3):178–198, 2011.

[5] R. Bednarik, S. Eivazi, and M. Hradis. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, page 10. ACM, 2012.

[6] R. Bertrand, G. Ferré, P. Blache, R. Espesser, and S. Rauzy. Backchannels revisited from a multimodal perspective. In *Auditory-visual Speech Processing*, pages 1–5, 2007.

[7] E. Bevacqua, M. Mancini, and C. Pelachaud. A listening agent exhibiting variable behaviour. In *Intelligent Virtual Agents*, pages 262–269. Springer, 2008.

[8] D. Bohus and E. Horvitz. Dialog in the open world: platform and applications. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 31–38. ACM, 2009.

[9] H. H. Clark. *Using language*, volume 1996. Cambridge university press Cambridge, 1996.

[10] K. A. Funes Mora, L. S. Nguyen, D. Gatica-Perez, and J.-M. Odobez. A Semi-Automated System for Accurate Gaze Coding in Natural Dyadic Interactions. In *ICMI*, Sydney, Dec. 2013.

[11] K. A. Funes Mora and J.-M. Odobez. Gaze Estimation From Multimodal Kinect Data. In *Computer Vision and Pattern Recognition Workshops*, pages 25–30, June 2012.

[12] D. Gatica-Perez, I. McCowan, and S. Bengio. Detecting Group Interest-Level in Meetings. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages 489–492. Ieee, 2005.

[13] D. Gatica-Perez, I. A. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[14] E. Goffman. *Interaction Ritual: Essays in Face to Face Behavior*. AldineTransaction, 1967.

[15] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski. Very short utterances and timing in turn-taking. In *INTERSPEECH*, pages 2837–2840, 2011.

[16] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro. A two-month field trial in an elementary school for long-term human–robot interaction. *Robotics, IEEE Transactions on*, 23(5):962–971, 2007.

[17] A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63, 1967.

[18] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati. Comparing models of disengagement in individual and group interactions. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 99–105. ACM, 2015.

[19] R. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In *Intelligent Virtual Agents*, pages 25–36. Springer, 2005.

[20] R. Meena, G. Skantze, and J. Gustafson. Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech & Language*, 28(4):903–922, 2014.

[21] L.-P. Morency, I. de Kok, and J. Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84, 2010.

[22] B. Mutlu and J. Forlizzi. Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pages 287–294. IEEE, 2008.

[23] L. Nguyen, J.-M. Odobez, and D. Gatica-Perez. Using Self-Context for Multimodal Detection of Head Nods in Face-to-Face Interactions. In *ACM Int Conf. on Multimodal Interaction (ICMI), Santa Monica*, Oct. 2012.

[24] C. Oertel, K. Funes, S. Sheikhi, J.-M. Odobez, and J. Gustafson. Who will get the grant ? In *Int. Conf. on Multimodal Interaction Workshop on Understanding and modeling multiparty, multimodal interactions*, 2014.

[25] C. Oertel and G. Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*, pages 99–106, New York, New York, USA, 2013. ACM Press.

[26] C. Oertel and G. Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *International Conference on Multimodal Interaction*, 2013.

[27] A. M. Sabelli, T. Kanda, and N. Hagita. A conversational robot in an elderly care center: an ethnographic study. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, pages 37–44. IEEE, 2011.

[28] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 766–773. ACM, 2002.

[29] K. P. Truong, R. Poppe, I. de Kok, and D. Heylen. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proc. INTERSPEECH, Florence*. International Speech Communication Association, 2011.

[30] R. Vertegaal. The gaze groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 294–301. ACM, 1999.

[31] R. Zhao, A. Papangelis, and J. Cassell. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Intelligent Virtual Agents*, pages 514–527. Springer, 2014.