3rd party observer gaze during backchannels

Jens Edlund[1], *Mattias Heldner*[2], *Anna Hjalmarsson*[1]

[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2]Linguistics, Stockholm University, Stockholm, Sweden

edlund@speech.kht.se, mattias.heldner@ling.su.se, annah@speech.kth.se

Abstract

This paper describes a study of how the gazes of 3rd party observers of dialogue move when a speaker is taking the turn and producing a back-channel, respectively. The data is collected and basic processing is complete, but the results section for the paper is not yet in place. It will be in time for the workshop, however, and will be presented there, should this paper outline be accepted..

Index Terms: speech synthesis, unit selection, joint costs

1. Introduction

The most common view of face-to-face communication, strongly influenced by [1], is that interlocutors *take turns* speaking. The principles that guide this turn-taking is a well-studied topic in spoken communication research and application development alike since decades. And although there are different opinions as to how precise and mandatory turn-taking is, there is no doubt that the most common observation in spoken dialogue is *one speaker at a time*, interspersed by transitions from the vocalizations of one speaker to those of another. In this paper, we focus on a subset of these transitions, namely when the incoming speaker gives a backchannel. Backchannels, as coined by [2], are brief feedback utterances generally described as being somehow produced in the background and often not taken to constitute a speaking turn or to claim the floor. While carrying little propositional information and being unobtrusive in character, it has been shown that these short interjections play a significant role in the collaborative processes of dialogue [e.g. 3]. Analyses of the segments preceding backchannels further show that there is a versatile set of multimodal behaviours that affects the probability of a backchannel [e.g. 4]. The motivation of the current work is to better understand backchannel behaviour in dialogue. More specifically, we aim to learn more about the timing and the conspicuousness of these events by analyzing the gaze patterns in 3rd party observers.

2. Background and related work

The flow of interaction in face-to-face communication is a multifaceted process that involves a complex set of behaviours and different modalities [1]. Many researchers approach this by first identifying appropriate places to take the turn. One way to do this is to pick those places where speaker changes in fact occurred [e.g. 5]. This method results in an objective and repeatable selection, particularly if automatic speech activity detection is used to decide when participants are speaking and when they are silent. An inherent problem with the method, however, is that it only captures actual speaker changes; never possible but unrealized speaker changes, or potential transition relevance places (TRPs) in the terminology of [1]. Another common method is to have one or more judges subjectively identifying places where a speaker change could occur [e.g. 6; 7]. The method has advantages. It potentially captures not only places where real speaker changes occurred, but also places where speaker changes might have occurred without harm to the flow of the interaction, but did not. The method might also leave out those places where inappropriate speaker changes actually occurred. An objection – possibly the strongest objection – to the method is its lack of ecological validity. It is debateable if people do the same thing when asked to for example press a button while listening to a dialogue as they would do when they contribute their voices as participants in conversation.

In the present study, we explore a novel method of identifying places where a speaker could have entered the conversation. The method,is based on [8, 9], who use gaze patterns and gaze shifts of non-participating listeners to study turn-boundary projection. The method relies on the intuition that 3rd party observers of a conversation tend to direct their gaze at the current speaker in the conversation [e.g. 10]. One end goal of this effort is to be able to judge, for each frame or segment of a dialogue, how appropriate it is for another speaker to start speaking.

2.1. 3rd party gaze

Gaze patterns of speakers and their addressees is a relatively well-explored research area [10, 11]. For example, it has been shown that listeners gaze almost twice as much on speakers in dyadic dialogue than vice versa [12] and the interactive gaze patterns between listeners and speakers play a significant role in controlling the flow of interaction [10].

In the present study, we use the gaze behaviour of 3rd party observers – overhearers – of a dialogue. The motivation of this method is to obtain a fine-grained measure of listeners' ongoing focus of attention which is directly time-aligned with events in the dialogue. The term 3rd party observers is used to refer to listeners that are not directly addressed by the speaker. Consequently, when a listener becomes an active party of the ongoing conversation, that person is per definition no longer a 3rd party observer. Based on the hypothesis that dialogue is a collaborative process and that the degree of participation affects comprehension, it has been shown that the processes of understanding differ between addressees and overhearers [3]. The 3rd party observers in the present study, however, are not co-present, but attending to a pre-recorded video of a dialogue, making their role as overhearers static. While the behaviour of 3rd party observers and their role in the dialogue may not be representative of a co-present active listener, we have previously

shown that 3rd party observervers of videos of pre-recorded dialogues largely look at the same thing, the speaker [13].

2.2. Backchannel feedback

A large number of vocalizations in everyday conversation are traditionally not regarded as part of the information exchange, but have important communicative and interactive functions. Examples include confirmations such as "yeah" and "ok" as well as traditionally non-lexical items, such as "uh-huh", "um", and "hmm". Vocalizations like these have been grouped in different constellations and called different names, for example backchannels (i.e. back-channel activity, [2]), continuers [14], feedback and grunts, and attempts at formalizing their function and meaning have been made [e.g. 15]. We follow [16], who argue that the term backchannel feedback is relatively neutral, and henceforth use the term backchannel.

In the present study, we investigate backchannels by analysing to what extent 3rd party observers gaze at speakers who produce backchannels and when this gaze shift is done relative to the offset of the previous speaker's turn. It has previously been shown that 3rd party observers occasionally appear to anticipate speaker changes, shifting their gaze to the other speaker before the new turn is initiated, sometimes even before the end of the original speaker's turn [9]. This finding supports the claim that listeners to some extent can anticipate the ends of speaker turns. In the current work, we focus on speaker changes when the incoming speaker gives backchannels. By analysing the gaze patterns of 3rd party observers, we will be able to make in-depth analyses of the nature of these events. That is, whether backchannels are events to which 3rd party observers pay little attention, or whether these events can be anticipated in advance and is attained to by listeners to similar extents as other types of speaker changes.

3. The Spontal corpus

3.1. Corpus description

The Spontal corpus contains in excess of 60 hours of dialogue: 120 nominal half-hour sessions (the duration of each dialogue is minimally 30 minutes). The subjects are all native speakers of Swedish. The subjects were balanced (1) as to whether the interlocutors are of same or opposing gender and (2) as to whether they know each other or not. The recordings contain high-quality audio and video. Spontal subjects were allowed to talk about anything they wanted at any point in the session, including meta-comments on the recording environment. Four segments of five minutes each were randomly chosen from the development set of the most recent Spontal recordings (SpontalIDs 09-20; 09-28; 09-30; 09-36), but in such a manner that they were taken from different balance groups: Spontal dialogues are balanced for same/different gender and for whether or not the participants knew each other before the recording. The segments included one known and one unknown same gender (male) pair, as well as one known and one unknown opposing gender pair. Each segment consisted of the first five minutes of the dialogue – that is the first five minutes of the official recording following the moment when the recording assistant told the participants that the recording had started. The segments were manipulated such that the front facing videos of both participants were displayed simultaneously next to each other, as seen in Figure 1.

Figure 1. Still-image from one of the front facing videos of both participants.

3.2. Speech/non-speech decisions

The analyses presented here were based on an operationally defined model of interaction. This interaction model is computationally simple yet powerful and uses boundaries in the conversation flow, defined by the relative timing of speech from the participants in the conversation, as the only source of information. In particular, we annotate every instant in a dialogue with an explicit interaction state label; states describe the joint vocal activity of both speakers, building on a tradition of computational models of interaction [17].

As a basis for the interaction model, we first performed automatic speech activity detection (SAD) (for a detailed description of this procedure see 18). The SAD produced a segmentation of each speaker state sequence into TALKSPURTS and PAUSES. TALKSPURTS were defined as a minimum of two contiguous speech frames (i.e. 200 ms, as enforced by the decoding topology) by one party that were preceded and followed by a minimum of two contiguous silence frames from the speaker. Similarly, PAUSES were defined as a minimum of two contiguous silence frames from that speaker. Based on these segments, we extract speaker changes (SC): those places where one solitary speaker speaks, followed by solitary speech from another.

3.3. 3rd party Gaze annotation

Eight subjects participated in the third-party observer gaze data collection. Each subject was placed in front of a monitor on which the side-by-side videos of Spontal dialogues could be shown in a sound-proofed studio. Sound was replayed through stereo loudspeakers. Throughout each session, a Tobii T120 gaze tracker was used to determine where the subjects were looking.

In order to motivate the subjects to pay close attention to the interactions, they were told that their task was to analyze the personalities of participants in each dialogue. They were given a questionnaire with questions about the topic of the conversation and of the "big five" personality traits of each participant. After each of the three five-minute dialogue segments, they filled in a questionnaire. Although the participants were aware that their gaze was being tracked, they had no knowledge of the purpose of this tracking, nor were they instructed at any point to pay special attention to the person speaking.

Gaze data is processed in a simple but robust manner. We used the fixation point data delivered by the system, rather than the raw data. For each frame, we count the number of subjects

whose fixation point rests on the left half and the right half of the monitor, respectively, and normalize this to a number between -1 and 1, where -1 means that every subject whose gaze was captured looked at the left half of the monitor, and 1 means that they all looked at the right half. More details on the collection of third-observer gaze data is presented in [13].

The timing of their shifting their gaze from a previous speaker to a next speaker has been shown to vary, and occasionally their gaze will shift only to shift back again when no speaker change occurs. By averaging the gaze target (speaker A, speaker B, elsewhere) from a number of 3rd party observers and normalizing the results, we get a number from -1 (everybody looks at speaker A) to 1 (everybody looks at speaker B). The number reflects who the 3rd party observers think is going to be the speaker in the near future, and plotted over time, provides insight about actual speaker changes, with which it is highly correlated, but also of moments in time where some or many observers expected a speaker change.

4. Method

4.1. Backchannel annotation

As a basis for further analysis, the Spontal dialogues used in the gaze data collection were manually annotated for verbal backchannels. The annotation was done on the talkspurt level, where a segment was considered to be a backchannel if that segment's (only) function was to provide feedback to the other interlocutor's speech, without providing any new propositional information. Using this guideline as principle for the annotation, two annotators labelled the three dialogues independently with high annotator agreement. In total there were 5 disagreements between the annotators, but all were solved in agreement after discussion.

In addition to the manual annotation of backchannels, the talkspurts were subdivided into very short utterances (VSUs) and their complement (NONVSUs) based on their duration. Talkspurts between 2 and 10 frames in duration (i.e. 200 ms to 1000 ms) were labelled VSUs and those longer than 10 frames (i.e. ≥ 1100 ms) were labelled NONVSUs [19].

4.2. Selection and alignment

For this investigation, we chose to look at the onsets of talkspurts - the transitions between silence and vocalization in one speaker's channel. We characterize these transitions based on whether the new talkspurt is a BACKCHANNEL or a NONBACKCHANNEL and whether the transition begins in OVERLAP, after a GAP, or (perceptually) with NOGAPNOOVERLAP. We also include the onset of CONTINUING talkspurts where the same speaker was the last to speak before a preceding silence - a pause. The resulting 8 combinations and there respective frequencies are shown in Table 1.

Table 1. Frequencies of different types of transitions from one speaker from another.

Talkspurt type	Transition type	Frequency
Backchannel	Overlap	
Backchannel	NoGapNoOverlap	
Backchannel	Gap	
Backchannel	Continuing	
NoBackchannel	Overlap	
NoBackchannel	NoGapNoOverlap	
NoBackchannel	Gap	
NoBackchannel	Continuing	

We then calculate how the gaze distribution - the number of 3rd-party observers watching the incoming speaker vs. the number of speakers watching the other speaker for each 100 ms frame up to ten frames before and after the talkspurt begins. We sum all of these distributions so that we get the average gaze distribution at T for T = -1s to T = 1s in relation to talkspurt beginnings. By splitting this data on the categories defined above, we hope to see not only to what extent 3rd-party observers look at incoming speakers under different conditions, but also how quickly and robustly they are attracted to the new speaker.

4.3. Grouping of categories

The backchannels were subsequently automatically categorized as overlapping or non-overlapping. The overlap categorization was based on whether the VAD (described in section 3.2) had detected speech in both channels in at least two adjacent frames. The minimum criterion of two frames overlap is used since [20] shows that about 130 milliseconds of simultaneous speech is needed for speech to be perceived as overlapping.

5. Results (pending)

Report for: BC/non-BC, non-bc after pause; overlap, gap, no-gap-no-overlap; perceptual gap/overlap/no gap no overlap.

5.1. Descriptive statistics of categories

Pending.

5.2. Gaze targets overall

Pending.

5.3. Timing of gaze shift

Pending.

6. Discussion

Pending.

7. Acknowledgements

The work was supported by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E Prosody in conversation, the EU project Get Home Safe, and the Swedish Research Council (VR) project and 2011-6152.

8. References

[1] Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

[2] Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567-578). Chicago.

[3] Schober, M., & Clark, H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

[4] Gravano, A., & Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of Interspeech 2009* (pp. 1019-1022). Brighton, U.K.

[5] Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

[6] Heldner, M., Edlund, J., & Carlson, R. (2006). Interruption impossible. In Bruce, G., & Horne, M. (Eds.), *Nordic Prosody, Proceedings of the IXth Conference, Lund 2004* (pp. 97-105). Frankfurt am Main, Germany.

[7] de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language, 82*(3), 515-535.

[8] Tice, M., & Henetz, T. (2011). The eye gaze of 3rd party observers reflects turn-end boundary projection. In *Procs. of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2011/Los Angelogue)* (pp. 204-205). Los Angeles, CA, US.

[9] Tice, M., & Henetz, T. (2011). Turn-boundary projection: Looking ahead. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Boston, Massachusetts, USA.

[10] Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

[11] Bavelas, J., Coates, L., & Johnson, T. (2002). Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication, 52*(3), 566-580.

[12] Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze.* Cambridge University Press.

[13] Edlund, J., Alexandersson, S., Beskow, J., Gustavsson, L., Heldner, M., Hjalmarsson, A., Kallionen, P., & Marklund, E. (2012). 3rd party observer gaze as a continuous measure of dialogue flow. In *Proc. of LREC 2012*. Istanbul, Turkey.

[14] Schegloff, E. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Tannen, D. (Ed.), *Analyzing Discourse: Text and Talk* (pp. 71-93). Washington, D.C., USA: Georgetown University Press.

[15] Ward, N. (2004). Pragmatic functions of prosodic features in non-lexical utterances. In *Proceedings of Speech Prosody* (pp. 325-328).

[16] Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 32*(8), 1177-1207.

[17] Norwine, A. C., & Murphy, O. J. (1938). Characteristic time intervals in telephone conversation. *The Bell System Technical Journal, 17*, 281-291.

[18] Heldner, M., Edlund, J., Hjalmarsson, A., & Laskowski, K. (2011). Very short utterances and timing in turn-taking. In *Proceedings of Interspeech 2011* (pp. 2837-2840). Florence, Italy.

[19] Edlund, J., Heldner, M., & Pelcé, A. (2009). Prosodic features of very short utterances in dialogue. In Vainio, M., Aulanko, R., & Aaltonen, O. (Eds.), *Nordic Prosody - Proceedings of the Xth Conference* (pp. 57 - 68). Frankfurt am Main: Peter Lang.

[20] Heldner, M. (2011). Detection thresholds for gaps, overlaps and no-gap-no-overlaps. *Journal of the Acoustical Society of America, 130*(1), 508-513.