3rd party observer gaze as a continuous measure of dialogue flow

Jens Edlund[1], Simon Alexandersson[1], Jonas Beskow[1], Lisa Gustavsson[2], Mattias Heldner[2], Anna Hjalmarsson[1], Petter Kallionen[2], and Ellen Marklund[2]

[1]KTH Speech, Music and Hearing, Stockholm, Sweden

[2]Stockholm University, Linguistics, Stockholm, Sweden

E-mail: edlund@speech.kth.se, simonal@kth.se, beskow@speech.kth.se, lisag@ling.su.se, mattias.heldner@ling.su.se, annah@speech.kth.se, lisag@ling.se.se, petter.kallionen@lucs.lu.se, ellen@ling.su.se

Abstract

We present an attempt at using 3rd party observer gaze to get a measure of how appropriate each segment in a dialogue is for a speaker change. The method is a step away from the current dependency of speaker turns or talkspurts towards a more general view of speaker changes. We show that 3rd party observers do indeed largely look at the same thing (the speaker), and how this can be captured and utilized to provide insights into human communication. In addition, the results also suggest that there might be differences in the distribution of 3rd party observer gaze depending on how information-rich an utterance is.

Keywords: gaze tracking, dialogue flow, turn-taking

1. Introduction

The study of the flow of the interaction in dialogues – of *who speaks next* and *how speaker changes are timed* – has received much attention both from an application perspective and for basic research into spoken communication ever since Sacks et al. (1974) and before. A widespread starting point for studies of interaction flow is to designate a finite number of delimited speech segments from recorded dialogues that are to be investigated. Different methods have been used to make this selection.

A common selection method is to simply pick those places where speaker changes in fact occurred (e.g. Duncan, 1972). The 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. Given that we strive to avoid subjectivity and interpretation at the level of data selection, this is a strength. However, subjectivity at the level of data selection is often introduced even with this method, either by using manual speech/non-speech labels or by removing a subset of the speech segments from the data by manually labelling them as backchannels. If the removal of backchannels is essential to the study, simply removing speech segments under a certain duration threshold creates a similar effect without introducing an element of human judgement in the data selection step (e.g. Heldner et al., 2011). Naturally, there is a flip side to any automated method: repeatability and objectivity come at the expense of control. The resulting data is noisy and in some meaning uncontrolled – exactly what, semantically, syntactically or pragmatically, is being selected is difficult to know. Automatic methods generally do not distinguish between speaker changes that are in some meaning "good" or "bad", nor do they allow for judgments often found when speaker changes are concerned – such as "appropriate" or "competitive". They merely pick segments where one speaker speaks first, followed by a next speaker. It's worth noting that proponents of the automated methods would argue that the lack of interpretation and assessment at the data selection stage is a strength rather than a weakness, and that the inclusion of all kinds of dialogue segments is in fact what permits us to explore the appropriateness of a speaker change experimentally. A graver problem with the method is that it only captures *actual* speaker changes, never possible but unrealized speaker changes or potential *transition relevance points* (TRPs) in the terminology of Sacks et al. (1974).

Another common method is to have one or more judges select places where a speaker change *could* occur. This can be done either by a trained expert, as is the case in the CA tradition (e.g. Local et al., 1986), or by one or more lay judges. One method of doing this is to ask subjects to press a button whenever it would be appropriate to speak (e.g. Heldner et al., 2006; De Ruiter et al., 2006). The method has advantages. It evades a possible objection to the CA tradition: that it seems odd that we would need trained experts to point out segments that any actual speaker can clearly and easily find when engaging in conversation. It also potentially captures not only places where real speaker changes occurred, but places where speaker changes might have occurred without harm to the flow of the interaction, but did not. Finally, the method might 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 press a button while listening to a dialogue as they would do when engaging in a dialogue and taking turns. It is also a rather costly method, where each subject can label a dialogue at real time at best, although in reality it is likely slower than that by some factor. More sophisticated varieties are possible. Hjalmarsson (2011) incorporates reaction times of judges to create a continuum describing how appropriate a moment in time is as a TRP, rather than a binary decision. Hjalmarsson's method, however, again relies on a pre-selection of a limited number of potential places: those places where the current speaker becomes silent.

Apart from problems stated above, these methods share an assumption that there is a discrete unit – the *turn* in most accounts – at the beginning and end of which the

only possible places for speaker changes are tethered. The transition relevance points – notably called *points*, although in most cases, they seem to be interpreted as discrete intervals – are a consequence of this segmentation of the speech signal into discrete turns. The task of defining turns in an objective and unambiguous manner, however, has science beat. Instead, speech emerges incrementally and dynamically, and speech seemingly end at all possible places in utterances, albeit with different probabilities. In the same vein, speaker changes occur everywhere, but are more likely to occur at certain places. For this reason, limiting our investigations of interaction flow to the positions where *turns* or the less theory-laden *talkspurts* (Norwine & Murphy, 1938; Brady, 1968) end or to the positions where speaker changes actually occurred introduces unacceptable limitations. Instead, we strive to investigate the dialogues in their entirety, but for that, we need a method to deem how suitable or appropriate each frame or segment of a dialogue is for a speaker change.

In this paper, we investigate a method of acquiring a continuous measure of the how likely human 3rd party observers think that a speaker change is, based on their largely subconscious gaze behaviour. The method was proposed by Tice & Henetz (2011), who test it with 32 observers of two split-screen dyadic dialogues taken from a Hollywood film, using manual labels of gaze direction from two annotators. The work presented here differs in several ways: (1) the data used is conversations from the Spontal corpus (Edlund et al., 2010a), which consists of dyadic, unscripted, task-free half-hour conversations; (2) the gaze is tracked using a Tobii gaze tracker (http://www.tobii.com/); and (3) the results are reported differently: where Tice and Henez report results for five question-answer pairs, we report a continuum over the whole data.

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. 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.

The working hypothesis is (a) that 3rd party observer gaze falls on the speaking person in a dialogue most of the time; (b) that 3rd party observer gaze moves from the speaking person to the presumed next speaker in anticipation of speaker changes; and (c) that because of this, potential but unrealized speaker changes can be traced in 3rd party observer gaze. We investigate this by (a) correlating different 3rd party observers' gaze target and (b) inspecting the speech signals from both speakers and the gaze shifts that occurred in dialogues. A secondary and weaker working hypothesis is that since (d) backchannels are produced with the goal of being unobtrusive, the extent to which utterances trigger gaze shifts in 3rd party observers should be negatively correlated to the extent to which the utterances are deemed to be backchannels.

2. Method

Figure 1: Graphical layout of the dialogue videos

2.1 Data

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.

Three segments of five minutes each were randomly chosen from the development set of the most recent Spontal recordings, 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. One of the dialogues contain an unknown male-male pair, one an unknown female-male, and one a known male-male pair. Each segment consists 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.

2.2 Processing

Spontal videos are head on, one for each participant in a dialogue. Side-by-side synchronized versions of these videos were created, with high-quality microphone audio from each speaker added to the left and right channel in such a manner that the stereo effect matched the video.

2.3 Subjects

We used eight subjects. All were connected to the linguistics department at Stockholm University in some way, but none had any knowledge of the experiment or its goals.

2.3 Experimental setup

The experimental setup was deliberately kept as simple as possible. 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 Tobi 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.

2.4 Processing

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.

2.5 Visualization

For this initial analysis, we experimented with several different visualizations. We found that resampling the gaze tracker data to 10 Hz (the system provides 60 Hz) presents a reasonable compromise between high granularity and low noise. We overlaid the 10 Hz on top of the speech waveforms from each channel.

3. Results

One of the eight subjects was captured less than 10 % of the time. Although this results in a very small amount of data that does not have a noticeable effect on the combined data, we have opted to remove it from the material as it is unreliable. Due to this, we have a maximum of seven gaze fixation points for each frame. The remaining seven participants were captured reliable most of the time: a fixation point was captured in 90, 88, 86, 84, 83, 76 and 54 % of the frames, respectively.

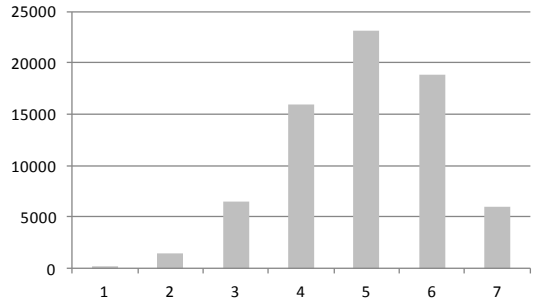

Figure 2: Histogram over number of successfully captured fixation point from subjects (X axis) for a frame. The histogram shows that for most frames, five out of the seven subjects were captured.

Only a very small number of frames (1576 out of 72000, or 9 %) got less than fixation points from less than three subjects. For a large proportion (47983, or 67 %), five or more fixation points were acquired. The distribution of number of successfully captured fixation points per frame is shown in Figure 2.

We found no particular bias to look at either side of the monitor from the subjects. Figure 3 shows, for each subject, the distribution of gaze on the left and right side of the monitor.

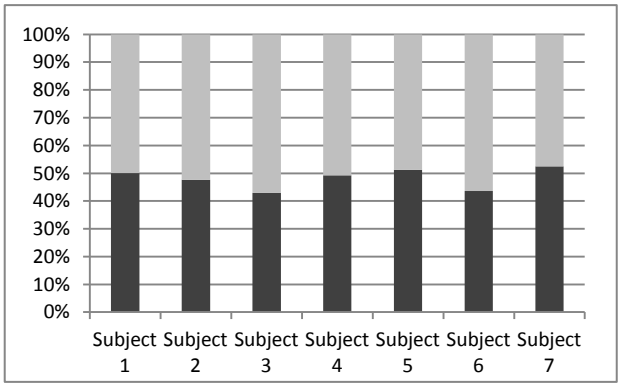

Figure 3: Proportion of gaze to the right (dark bars) and left (light bars) for each subject.

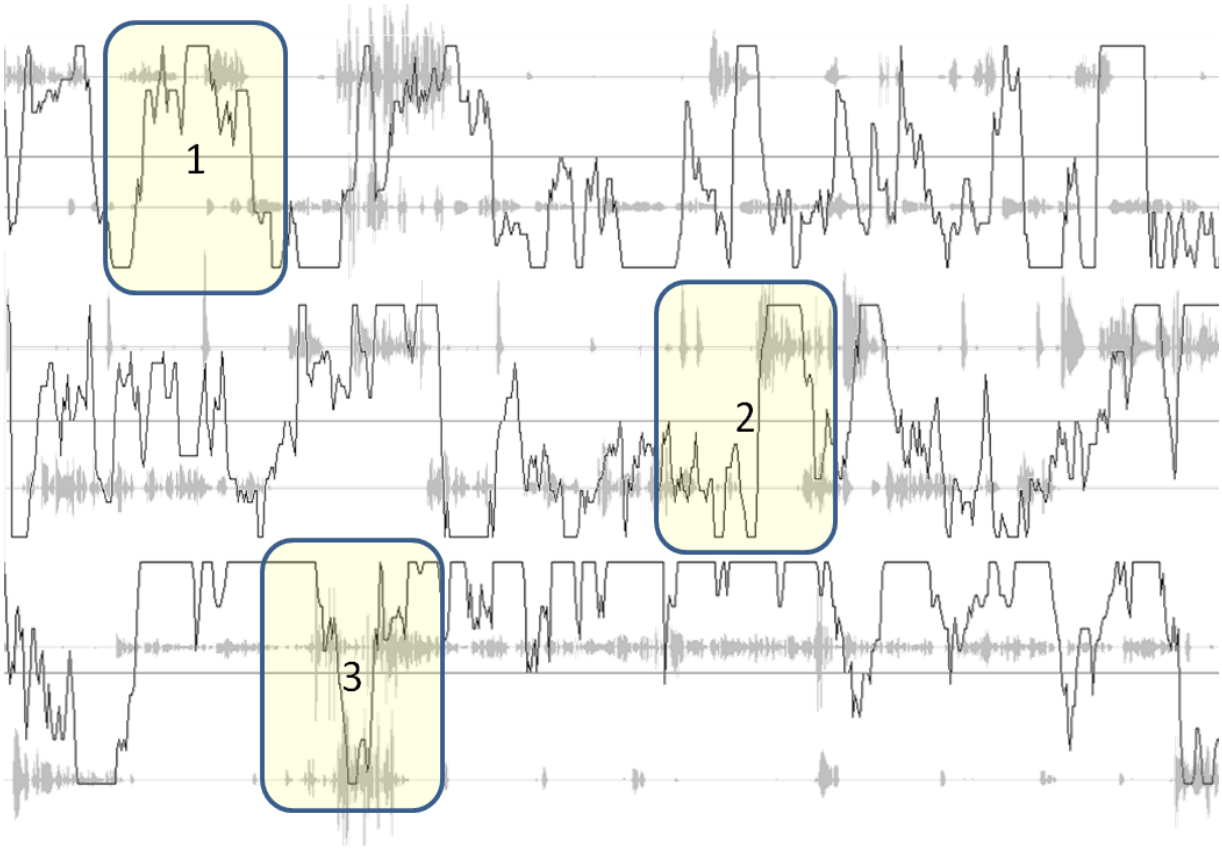

Figure 4: The first minute of each of the three dialogues investigated. For each of the three rows, the upper half shows the left speakers waveform and the lower half shows that of the right. The overlaid graph is the gaze fixation of the subjects, with high values representing a large proportion of gaze on the upper (left) speaker and a low value a large proportion of gaze on the lower (right) speaker.

All subjects looked at the same side of the monitor a large proportion of the time. 18 % of the time, all captured fixation points were on the left side of the monitor, and 18 % of the time, all captured fixation points were on the right side, making for a total of 36 % of the frames where all successfully captured subjects gazed on the same side of the monitor in unison. Based on the distribution of successfully captured fixation points per frame presented in Figure 2, the random chance of all fixation points being on the same side is less than 10 %.

Finally, Figure 4 shows the first minute of each of the three dialogues, with the waveform of the dialogue participant to the left in the video in the upper half of each dialogue and that of the participant to the right in the lower half. The overlaid graphs show the proportion of gaze fixation points to the left and to the right. The numbered rectangles are points of interest.

4. Discussion and future work

The preliminary study presented here show the potential of automatically tracked 3rd party observer gaze as an annotation method for spoken dialogue. Gaze tracking (a) is reliable enough to acquire annotation with high temporal resolution, with fixation points from more than half of the subjects being reliably captured for 2/3 of the frames; (b) captures onlooker behaviours that are similar between subjects without the need to give instructions to "look at the speaker", as evidenced by the fact that all subjects gazed at the same side of the monitor in 36 % of the frames. Further, (c) 3rd party observer gaze does follow the speaker, as seen in Figure 4.

The three highlighted areas in Figure 4 points to future research. In the area marked 1, there is a plateau with everyone looking at the top (left) speaker in spite of that speaker being silent. Upon inspection, the segment turns out to be a place where the left speaker pauses in search for a word. The right speaker even attempts to barge in, but only after the left speaker brings his point to

conclusion does the average gaze fixation shift to the right speaker. This suggests the possibility to use gaze to distinguish between intended pauses - silences within an utterance - and gaps between utterances.

The area marked 2 shows a different event. As the lower (right) speaker finishes, there is a silence during which there is uncertainty who will speak next. As the upper (left) speaker begins speaking, all subjects very quickly look over that way. This can be contrasted with the many occasions where a backchannel (most of the very brief talkspurts seen in the figure) are given almost no attention by the subjects.

Finally, the area labelled 3 shows an segment where the top (left) speaker speaks, and in the middle of this, both speakers laugh simultaneously. In the middle of this laughter, the lower (right) speaker gives some information. The segment where the subjects' gaze go to the lower (right) speaker coincides perfectly with this utterance, whereas their gaze remain on the original speaker throughout the mutual laughter.

We also note a number of things throughout the segments in Figure 4 (as well as the remaining four minutes of each of the dialogues). The event that is highlighted in (1) is very uncommon - whenever all observers gaze at the same side, the person on that side is nearly always speaking. Analysing the places where everyone looks at a person who is not speaking is a priority for future work. Furthermore, most short vocalizations are accompanied or preceded by a shift in gaze towards their speaker, but very few have all observers shift gaze. Another high priority for future work is to correlate the number of observers shifting their gaze towards the speaker of a short vocalization with the contents of that vocalization, for example a backchannel or a response to a question. Finally, we see clear differences in the precision and timing with which observers move from one speaker to another. In some cases the shift is simultaneous for all speakers and precisely timed with the speaker change, in others the gaze shift is more hesitant and differentiated between observers. Investigating the cause of these differences is a third area we will address early on.

Our results so far are encouraging enough to develop the technology further to make it simpler to use and more robust. Our current impression is that the method, even if we develop the analyses to be more or less automatic, is not particularly cheap, but may still be worthwhile as it is a means at getting to information that is otherwise not accessible to us. The impressionistic interpretation of the 3rd party observer gaze patterns we have so far, and one that is illustrated by the three highlighted areas, is that 3rd party observer gaze coincides not just with who is speaking at the moment, but with who is providing the most information. Semantically rich talkspurts seem to draw more gaze attention than more pragmatically and interactionally motivated talkspurts such as laughter and feedback.

5. Acknowledgements

The research presented here was funded in part by the Swedish Research Council project Samtalets Rytm (The rhythm of conversation; 2009-1766).

6. References

Brady, P. T. (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal, 47*, 73-91.

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.

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

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010a). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 2992 - 2995). Valetta, Malta.

Edlund, J., Heldner, M., Al Moubayed, S., Gravano, A., & Hirschberg, J. (2010b). Very short utterances in conversation. In *Proc. of Fonetik 2010* (pp. 11-16). Lund, Sweden.

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.

Heldner, M., Edlund, J., Hjalmarsson, A., & Laskowski, K. (2011). Very short utterances and timing in turn-taking. In *Proc. of Interspeech 2011*. Florence, Italy.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication, 53*(1), 23-35.

Local, J., Kelly, J., & Wells, W. (1986). Towards a Phonology of Conversation: Turn-Taking in Tyneside English. *journal of Linguistics, 22*(2), 411-437.

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

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

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.