Analysis of gaze and speech patterns in three-party quiz game interaction

Samer Al Moubayed, Jens Edlund, Joakim Gustafson
KTH Speech, Music and Hearing, Sweden
sameram@kth.se, {edlund,jocke}@speech.kth.se

Abstract

In order to understand and model the dynamics between interaction phenomena such as gaze and speech in face-to-face multiparty interaction between humans, we need large quantities of reliable, objective data of such interactions. To date, this type of data is in short supply.

We present a data collection setup using automated, objective techniques in which we capture the gaze and speech patterns of triads deeply engaged in a high-stakes quiz game. The resulting corpus consists of five one-hour recordings, and is unique in that it makes use of three state-of-the-art gaze trackers (one per subject) in combination with a state-of-the-art conical microphone array designed to capture roundtable meetings. Several video channels are also included.

In this paper we present the obstacles we encountered and the possibilities afforded by a synchronised, reliable combination of large-scale multi-party speech and gaze data, and an overview of the first analyses of the data.

Index Terms: multimodal corpus, multiparty dialogue, gaze patterns, multiparty gaze.

10.21437/Interspeech.2013-302

1. Introduction

The speech group at KTH has a long-term goal of building computational models of patterns in multiparty face-to-face situated interaction among humans. The goal is to understand human-human conversational behaviours enough to build an artificial conversational partner that can behave like a human would, in similar conversational contexts [1]. In order to develop a conversational partner that can engage in multi-party interactions, the group has developed the robotic head Furhat that projects state-of-the-art facial animation onto a face mask [2]. Apart from speech recognition, the head is equipped with head, face and gaze trackers. This makes it possible for the system to use its gaze in a human-like manner, both to indicate mutual attention and for turn-taking in multi-party settings [3,4,5].

In order to model dynamics and speech patterns in spoken interactions, there is a need for conversational data. KTH has recently collected about 60 hours of audio, video, and motion capture data in two-party conversations within the project Spontal: Multimodal database of spontaneous speech in dialogue [6], and a database of multi-party human-human conversations in the D64 corpus [7]. Neither of these has made use of gaze-trackers, which means that analysis of gaze behaviours is very time consuming, and potentially limited and unreliable [8]. However, researchers have only recently started to record corpora on three-party interactions where at least one of the participants is equipped with a gaze tracker, e.g. [9].

Acquiring simultaneous speech and gaze data for all participants in multiparty interactions serves many potential functions. For example, involvement and engagement in conversation are important phenomena in spoken multi-party face-to-face interaction. Over the course of an interaction, the participants may be involved to a varying degree, and hence, conversational engagement is a very important measure in designing system's strategies and behaviours in human-machine spoken interaction. One of the salient cues of engagement is temporal gaze patterns [10]. But in order to elicit engagement in interaction, the task needs to be fun and interesting. One way of achieving this is to use games. For two-party dialogue corpora collections games have been used to collect, e.g. the Columbia Games corpus [11] and the MTD corpus [12] that contains multi-tasking dialogues that involved a poker game as the main task, and a picture game as an interrupting real-time task. The Speech group at KTH has initiated collections of a series of multi-party games for interactional corpus collections (The KTH games corpora). The first of these is the current data collection – the Three-party Estimation Quiz Game corpus – in which teams of three competed in a spoken quiz game. One of the main objectives of these recordings is to record massively multi-modal data-streams by employing the steadily progressing non-intrusive capture equipment. The current paper describes the quiz-game task, the recording setup, and the work done to process and analyse gaze and speech patterns with zero manual labour. Finally we will provide some preliminary results when using the data collection and analysis tools to collect five one-hour recordings of three-party quiz games.

2. The Three-party Estimation Quiz Game

The dialogue task was designed in the shape of a quiz-game competition between five teams of three. All team members worked at KTH, and only the winning team was rewarded for participating in the data collection. Apart from the honour of beating their colleagues, the winning team got to share 15 cinema tickets. The result was more than five hours of highly engaged three-party conversations, with a large number of speaker changes and very little silence. The teams were seated around a round table on which recording equipment was mounted (see Figure 2). The game master sat in another room where he could hear and see the game members via a web camera (Figure 3). He initiated all interactions with the team by pressing a push-to-talk-button, and spoke to the subjects using a small speaker placed on the round table.

The quiz game consisted of 10 3-part estimation questions of an encyclopaedic nature, designed so that it would be almost impossible to know the exact answer, but rather easy to speculate about it. The following is an actual example from the game: a) *what is the 8th largest country in terms of population?,* b) *how large is its population?,* and c) *what is its size in square kilometres?*. The teams were informed that they would be scored for each of the sub-question according to how their answers ranked in terms of distance to the correct answer, as compared to the other teams. For each question, the teams were given three minutes to discuss the three sub-questions and agree upon an educated guess before reporting the answer to the game master. After three minutes, the game master would interrupt them and request the answers. The game master then told the team the correct answers and left them to discuss this for about one minute before returning with the next question. This scenario results in an iteration of five

phases in the corpus: (1) the game master poses the questions, (2) the team discusses the questions, (3) one team member gives the answers, (4) the game master provides the correct answers; and (5) the game master allows the team to discuss their accomplishments for one minute.

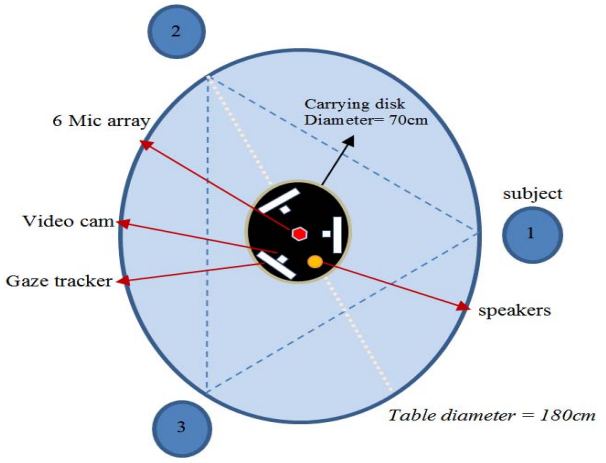

Figure 1: layout chart of the physical and equipment setup.

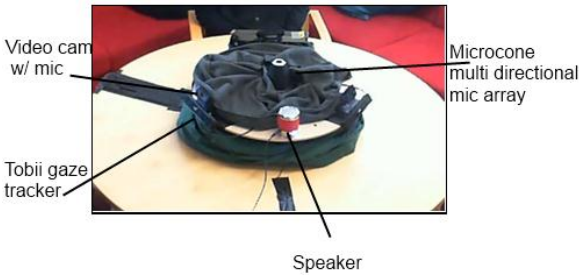

Figure 2: A chart of the recording equipment.

3. Three-Party Multimodal Data Collection Environment

3.1. Physical setup

The physical setup for capturing the three-party dialogue was designed so that the participants were at the same distance from each other, as well as at the same angles (sitting at the corners of an equilateral triangle, as seen in the diagram in Figure 1). This was achieved by placing them evenly around a round table where the recording equipment was placed at the same height and distance from the table border (see Figure 1). The table was rigged with stand-alone nonintrusive audio-video-gaze recording equipment that was placed so that each subject had a clear vision towards both other interlocutors. The table did not contain any other objects as can be seen in Figure 2 that shows a photograph of the recording setup.

In addition to the frame-level recording equipment (described in the next section), two overview video cameras were used to record the dialogues from two different points of view (Figure 3). One of these was also used by the Game Master in the adjacent room in order to get visual access to the game table. On the table there was a small-size mono-speaker that would communicate the questions posed by the Game Master to the game participants.

3.2. Equipment and Recordings

3.2.1. Gaze recording

In order to capture the gaze data simultaneously from all three interlocutors, three Tobii X2[1] stand-alone gaze trackers were placed on the table, and directed upwards towards the face of the interlocutors. The Tobii X2 tracker is a state-of-the-art gaze tracker that is widely used for research on gaze behaviour in spoken interaction [13,14,15,16]. It offers a 30 fps gaze data streaming, and is supported with a Software Development Kit (SDK) that allows for more controlled capture of the data in real-time. In this setup we decided that the calibration step usually done before recording would not be useful, since we do not have a pre-defined flat surface where we want to track the gaze, and opted for an approach of post-recording automatic clustering of gaze.

The choice of the physical setup having the subjects seated on static chairs was also made to allow for a more accurate gaze tracking since the tracker is optimized to capture gaze when the subject's head is placed in a small movement box, and hence not allowing for a large head movement by the subject (44cm*42 cm, at 65cm distance). One of the limitation of using the Tobii gaze tracker is that only a single tracker can be connected to a single computer (due to limitations in the SDK itself), and hence, each of the trackers was connected to a separate computer. Because of this, and to allow for an accurate synchronization of the three gaze trackers against each other and against other data streams, an audio channel was recorded by each of these computers (using the video camera's built-in stereo microphone). The audio was then used to synchronize the data across computers.

Figure 3: Photographs from the over-view video cameras.

3.2.2. Multichannel location and audio recording

One of the main motivations behind the recording of this corpus was to capture data using state-of-the-art equipment that would require minimal or no manual annotation of the data itself, and would allow for completely-automatic post analysis of gaze and speech patterns. Furthermore, we wanted the setup to be non-intrusive and fast-started – as natural as possible, which led to the decision to not use close microphones or headsets. However, recording multiparty dialogues with far-field microphones makes it hard to achieve reliable speaker and speech activity detection, and manual speaker diarization was not an option since it would require

[1] www.tobii.com

time-consuming manual annotation. The solution was to make use of a state-of-the-art high-quality microphone array called Microcone™[1]. Microcone™ is a conical microphone array that is made of 6 directed microphones to capture directed and de-noised audio, in addition to one omnidirectional stereo microphone to capture overall raw audio data, at a 48kHz sampling rate.

In addition to the audio data captured by the microphone array, Microcone™ employs a built-in audio processing unit that provides a measure of microphone activity (a binary measure) for each of the 6 microphones, and is triggered whenever the audio volume in one of the microphones is above a certain relative threshold. The array also provides an estimate of the loudness in each of the microphones that is used to calculate the exact angle from where the audio is coming in, for each of the microphones. These measures allow for an estimate of the location of the speakers (which is not used in this corpus since subjects were seated on non-moving chairs), and an automatic Voice Activity Detection VAD for each of its microphones.

To capture the data from the microphone array, an SDK that is supported by the Microcone™ device is used, in the same environment with the gaze and video recording systems, allowing for automatic synchronization of the audio, gaze and video for each of the computers used to capture the corpus. The device is shown in Figure 1 and Figure 2.

3.2.3. Video Recording:

A recording of the face of each of the interlocutors was also done by rigging high-resolution stable 20fps USB3 video cameras to each gaze tracker that captured the same field of view as the tracker. A video recording of the view of the tracker also allows for future analysis of facial movements (blinks, head-rotation, eyebrows, etc.). The rotation of each of the video cameras was adjusted before each recording session to accommodate for the height of the subjects. The video was captured in the same in-house recording environment used to capture the audio and gaze data, and hence that allowed for an automatic synchronization and alignment of the data on a frame-by-frame basis. Figure 4 shows photographs of the view from the 3 video cameras.

3.2.4. Synchronization

To capture the data from all the recording equipment, a Java based environment that utilizes the equipment's SDKs was developed.

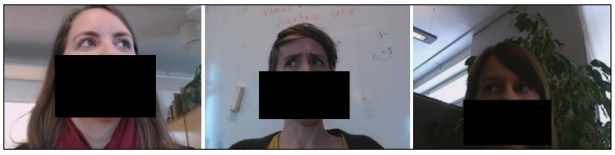

Figure 4: Example photographs from the video cameras.

[1] http://www.dev-audio.com/products/microcone/

Capturing the data using a unified single programming environment allows for automatic synchronization and alignment between the different data streams, especially when the different equipment have a varying and non-stable capture-frame-rate. However, since the Tobii gaze trackers cannot be attached to one single computer, a post-recording automatic offset synchronization between the gaze trackers was done using a simple timestamp solution and the sync-audio streams. The gaze data was also automatically synchronized with the microphone array data using timestamps provided along with the location and activity data. Figure 5 shows a flowchart of the data streams recorded by the environment.

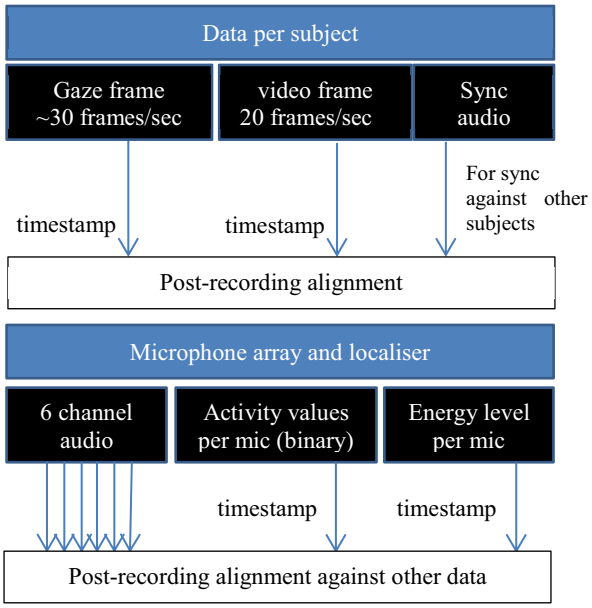

Figure 5: A Flow chart of the different streams of data and the corresponding synchronization.

4. Analysis

As mentioned earlier, 5 dialogues, consisting of 3 participants, and one Game Master were collected, resulting in a five-hour long corpus of gaze-rich audio-visual interactions. Table 1 shows the amount of data, in minutes, identified as speech (or silence) using the automatic microphone activation system. The numbers show that, in total, all participants contributed a similar amount of speech data to the conversation (105 minutes), while the whole 5.2 hours recording contained a total of 30 minutes of silence.

Table 1. The Quiz Game corpus.

Cone VAD	Minutes (position 1,2, 3)
silence	30
participant speech	105 (33, 37, 35)
master speech	117

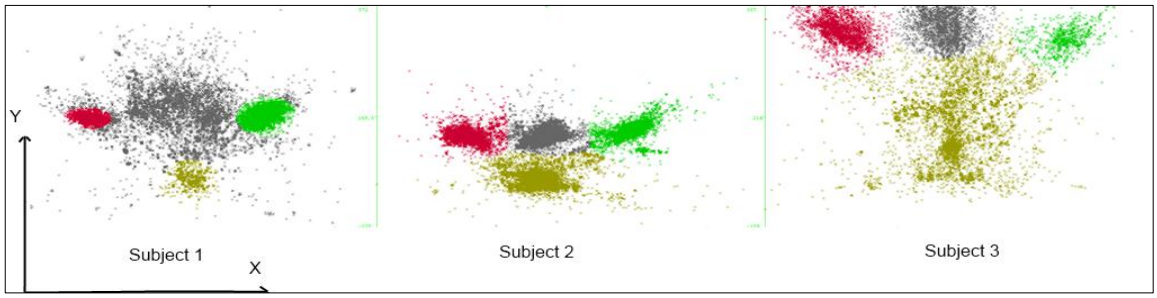

Figure 6: An example showing the gaze plots for one of the dialogues for the three participants, showing the resulting of the automatic clustering using a 4 2-dimensional Gaussian mixture model.

4.1. Gaze Data Preparation

One of the first steps to process raw gaze data during interaction is to cluster gaze vectors into recognizable gaze targets (e.g. on-head gaze and off-head gaze). This can either be done manually using subjective estimate on where the interlocutor subjects were sitting by visually looking at the gaze data, which can be highly biased and less reliable (especially when the subjects slightly change their positions). Another method that we briefly describe here is to automatically cluster the gaze data into a pre-defined number of clusters, which is assumed to divide the data into clusters depending on the density of gaze points on the XY axis, and then manually map these clusters into recognizable targets (such as Right Interlocutor, Left Interlocutor, Table, Ambient, etc.). To do this, a four, 2-dimensional Gaussian mixture model was trained using the Expectation Maximization method (EM) which was applied separately on each of the gaze data for each subject and recording session. The result is shown in Figure 6.

4.2. Gaze Analysis

Although this paper is intended as a description of the methodology, setup, and equipment to capture the first part of the KTH Games corpora, featuring gaze-rich audio-visual three-party spoken dialogue, in the following we present some initial examples of different statistics and analysis that are possible to investigate using our fully automated data collection and analysis tools. The purpose is to give examples of the advantages of doing simultaneous gaze recording of all participants during interaction. By combining the gaze clustered targets of each participant with the cone VAD values, we can directly investigate patterns across gaze and speech (e.g. when someone is speaking or listening, when all participants are silent, or even in overlaps). By doing this, we can for example quantify the amount of time when all participants avoided looking at each other. As can be seen in Figure 7, most of these instances occurred when the Game Master was talking, or during complete silence. By further analyzing the data when the participants were speaking, we found that the speakers on average looked at either of the two listeners 57% of the time, and they looked equally much at both interlocutors. When the participants were listening they looked at the speaker 47% of the time, at the co-listener 17% of the time, and at nobody the remaining 36% of the time.

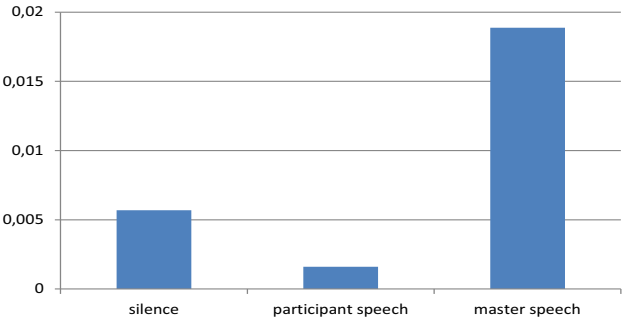

Figure 7: The proportion of frames where all three trackers detected off-head-gaze, depending on who was speaking.

5. Future Work

In this paper we presented an engaging quiz game task for three-party interaction collections, a recording setup for three-party speech and gaze collections, as well as a description of a corpus of non-intrusive gaze-rich audio-visual three-party spoken interaction. The setup made use of state-of-the-art capture equipment that allows for fully automatic recording and synchronization of the different data streams.

The possibilities allowed by this recording method give rise for future investigations of speech and gaze patterns in three-party interactions. Although we only presented initial data analysis as examples of how the corpus can be used for this purpose, we intend to study the dynamics of temporal speech-gaze patterns depending, for example, on overlaps, speaker changes, and third-party observer behaviour. We will also analyze how the gaze and speech patterns vary across the 5 different states of the current game-quiz corpus. We will also add facial analysis (such as head rotation, facial expressions, etc.) of the on-participant video recording, using automatic facial analysis software, such as SHORE©[1] or FaceAPI™[2].

6. Acknowledgements

The work was supported by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation* and the *Situated audio visual interaction with robots* project within the KTH Strategic Research Area *ICT - The Next Generation*.

[1] http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore.html
[2] http://www.seeingmachines.com/product/faceapi/

7. References

[1] Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

[2] Al Moubayed, S., Skantze, G., & Beskow, J. (2013). The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction. International Journal of Humanoid Robotics.

[3] Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B.(2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., & C. Müller, V. (Eds.), Cognitive Behavioural Systems. Lecture Notes in Computer Science. Springer

[4] Al Moubayed, S., Edlund, J., & Beskow, J. (2012). Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. ACM Transactions on Interactive Intelligent Systems, 1(2), 25.

[5] Al Moubayed, S., & Skantze, G. (2011). Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. In Proceedings of AVSP. Florence, Italy.

[6] Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). 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.

[7] Oertel, C., Cummins, F., Edlund, J., Wagner, P., & Campbell, N. (2012). D64: a corpus of richly recorded conversational interaction. *Journal of Multimodal User Interfaces*.

[8] Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012). Gaze Patterns in Turn-Taking. In *Proc. of Interspeech 2012*. Portland, Oregon, US.

[9] Jokinen, K. (2011). Turn taking, utterance density, and gaze patterns as cues to conversational activity. In *The ICMI workshop on Multimodal Corpora for Machine Learning*. Alicante, Spain.

[10] Oertel, C., Scherer, S., & Campbell, N. (2011). On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Interspeech 2011* (pp. 1541-1544). Florence, Italy.

[11] Benus, S., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics, 43*(12), 3001-3027.

[12] Yang, F., Heeman, P. A., & Kun, A. L. (2011). An investigation of interruptions and resumptions in multi-tasking dialogues. Computational inguistics, 27(1), 75-104.

[13] Al Moubayed, S., Beskow, J., & Granström, B. (2010). Auditory-Visual Prominence: From Intelligibilitty to Behavior.Journal on Multimodal User Interfaces, 3(4), 299-311.

[14] Bednarik, R., Eivazi, S., Hradis, M.: Gaze & Conversational Engagement in Multiparty Video Conversation: An annotation scheme and classification Gaze-In '12 Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, Article 10, 2012.

[15] Bailly, G., S. Raidt & F. Elisei (2010). "Gaze, conversational agents and face-to-face communication." Speech Communication - special issue on Speech and Face-to-Face Communication, 52(3): 598–612.

[16] Nakano, Y.I. & Ishii, R. (2010). Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations. in 2010 International Conference on Intelligent User Interfaces (IUI2010). Hong Kong.