

# MAGiC: A Multimodal Framework for Analysing Gaze in Dyadic Communication

Ülkü Arslan Aydın  
Cognitive Science Program  
Middle East Technical University  
Ankara, Turkey

Sinan Kalkan  
Computer Science Department  
Middle East Technical University  
Ankara, Turkey

Cengiz Acartürk<sup>†</sup>  
Cognitive Science Program  
Middle East Technical University  
Ankara, Turkey

The analysis of dynamic scenes has been a challenging domain in eye tracking research. This study presents a framework, named MAGiC, for analyzing gaze contact and gaze aversion in face-to-face communication. MAGiC provides an environment that is able to detect and track the conversation partner's face automatically, overlay gaze data on top of the face video, and incorporate speech by means of speech-act annotation. Specifically, MAGiC integrates eye tracking data for gaze, audio data for speech segmentation, and video data for face tracking. MAGiC is an open source framework and its usage is demonstrated via publicly available video content and wiki pages. We explored the capabilities of MAGiC through a pilot study and showed that it facilitates the analysis of dynamic gaze data by reducing the annotation effort and the time spent for manual analysis of video data.

Keywords: Gaze analysis, speech analysis, automatic face detection, automatic speech segmentation

## Introduction

In face-to-face social communication, interlocutors exchange both verbal and non-verbal signals. Non-verbal signals are conveyed in various modalities, such as facial expressions, gestures, intonation and eye contact. Previous research has shown that non-verbal messages prevail synchronous verbal messages in case of a conflict between the two. In particular, interlocutors usually interpret non-verbal messages rather than verbal messages as a reflection of true feelings and intentions (Archer & Akert, 1977; Mehrabian & Wiener, 1967). Therefore, an investigation

of the structural underpinnings of social interaction requires the study of both non-verbal modalities and verbal modalities of communication. In the present study, we focus on gaze as a non-verbal modality in face-to-face communication. In particular, we focus on eye contact and gaze aversion.

Eye contact is a crucial signal for social communication. It plays a major role in initiating a conversation, in regulating turn taking (e.g., Duncan, 1972; Sacks, Schegloff, & Jefferson, 1974), in signaling topic change (e.g., Cassell et al., 1999; Grosz & Sidner, 1986; Quek et al., 2000, 2002) and in adjusting the conversational roles of interlocutors (e.g., Bales, et al., 1951; Goodwin, 1981; Schegloff, 1968). Moreover, interlocutor's putative mental states, such as *interest*, are usually inferred from gaze (Baron-Cohen, Wheelwright, & Jolliffe, 1997). In particular, eye contact is a fundamental, initial step for capturing the attention of the communication partner and establishing joint attention (Fasola & Mataric, 2012; Kleinke, 1986).

<sup>†</sup>Corresponding Author: [acarturk@metu.edu.tr](mailto:acarturk@metu.edu.tr)  
Received May 13, 2018; Published November 12, 2018.  
Citation: Arslan Aydın, Ü., Acartürk, C. & Kalkan, S. (2018).  
MAGiC: A Multimodal Framework for Analysing Gaze in Dyadic  
Communication. *Journal of Eye Movement Research*, 11(6):2  
Digital Object Identifier: 10.16910/jemr.11.6.2  
ISSN: 1995-8692. This article is licensed under a [Creative Commons  
Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). 

Gaze aversion is another coordinated interaction pattern that regulates conversation. Gaze aversion is the act of intentionally looking away from the interlocutor. The previous research has explored the effects of gaze aversion on avoidance and approach. These studies have shown that an averted gaze of an interlocutor initiates a tendency to avoid, whereas a direct gaze initiates a tendency to approach (Hietanen, et al., 2008). Similarly, the participants give higher ratings for likeability and attractiveness when picture stimuli involve a face with a direct gaze contact, compared to the stimuli that involve a face with averted gaze (Mason, Tatkov, & Macrae, 2005; Pfeiffer, et al., 2011).

The conversational functions of gaze aversion are also closely related to speech (Abele, 1986; Argyle, & Cook, 1976; Kendon, 1967). In particular, gaze provides repeating, complementing, regulating and substitution of a verbal message. Speech requires complementary functions, such as temporal coordination of embodied cognitive processes including planning, memory retrieval for lexical and semantic information, and phonemic construction (Elman, 1995; Ford & Holmes, 1978; Kirsner, Dunn, & Hird, 2005; Krivokapić, 2007; Power, 1985).

A closer look at speech as a communication modality reveals that speech carries various useful signals about the content or quality of speech itself, such as intonation, volume, pitch variations, speed and actions done through speech (viz. speech acts). In the present study, we focus on *speech acts* due to its salient role as the speech modality in conversation. According to the speech act theory (Austin, 1962; Searle, 1969), language is a tool to perform acts, as well as to describe things and inform interlocutors about them.

The speech act theory is concerned with the function of language in communication. It states that a speech act consists of various components that have distinct roles. For analyzing language in communication, discourse should be segmented into units that have communicative functions. The relevant communicative functions should be identified and labelled accordingly. The speech acts are usually identified by analyzing the content of speech. However, temporal properties of speech convey information to the interlocutor, too. For instance, the analysis of a pause may be conceived as a signal for a shift in topic (Krivokapić, 2007). Similarly, a pause may be an indicator of speaker's fluency (Grosjean & Lane, 1976) and even for and indicator of a speech disorder (Hird, Brown, & Kirsner, 2006). The framework that we present in this

study (viz. MAGiC) enables researchers to perform analyses by employing both content of speech and its temporal properties. In the following section, we present a major challenge that MAGiC proposes a solution, namely gaze data analysis in dynamical scenes.

## Gaze Data Analysis in Dynamical Scenes

Eye tracker manufactures have been providing researchers with the tools for identifying basic eye movement measures, such as gaze position and duration, as well as a set of derived measures, such as Area of Interest (AOI) statistics. The study of gaze in social interaction, however, requires more advanced tools that would enable the researcher to automatically analyze gaze data on dynamical scene recordings. The analysis of gaze data in dynamical scenes has been a well-acknowledged problem in eye tracking research (e.g., Holmqvist et al., 2011) largely due to the technical challenges in recognizing and tracking objects in a dynamic scene. This is because eye trackers generate a raw data stream, which contains a list of points-of-regard (POR) during the course of tracking the participant's eyes. In a stationary scene, it is relatively straightforward to specify sub-regions (i.e., Areas of Interest, AOIs) of the stimuli on the display. This specification is then used for extracting AOI-based eye movement statistics. In case of a dynamical scene (cf. mobile eye-trackers), the lack of predefined areas leads to challenges in automatic analysis of gaze data. A number of solutions have been proposed to improve dynamic gaze data and to make it more robust against human errors in manual data annotation, such as using infrared markers, employing inter-rater analysis and combining the state-of-the-art object recognition techniques for image processing. However, each method has its own limitations (Brône, Oben, & Goedemé, 2011; De Beugher, S., Brône, G., & Goedemé, 2014; Stuart, et al., 2017). For instance, infrared markers may lead to visual distraction. In addition, in case of multiple object detection, markers are not economically or ergonomically feasible since they should be attached to each individual object to be tracked as reported by the previous research (e.g., Munn, Stefano, & Pelz, 2008; Stuart, Galna, Lord, Rochester, & Godfrey, 2014). To the best of our knowledge, there is no commonly accepted method for achieving eye movement analysis in dynamic scenes as reported by the previous research. In this study, we propose

a solution to this problem in a specific domain, i.e., dynamic analysis of face, as presented in the following section.

We focus on a relatively well-developed subdomain of object recognition: Face recognition. The recognition of faces has been subject to intense research in computer vision due to its potential and importance in daily life applications, e.g. in security. Accordingly, MAGiC employs face recognition techniques to automatically detect gaze contact and gaze aversion in dynamic scenarios, where eye movement data are recorded. It aims at the analysis of dynamic scenes by reducing the effort on time-consuming and error-prone manual annotation of gaze data.

MAGiC also provides an environment that facilitates the analysis of audio recordings. Manual segmentation of audio recordings into speech components and pause components is not efficient and reliable, since it may exclude potentially meaningful information from the analyses (Goldman-Eisler, 1968; Hieke, Kowal, & O’Connell, 1983). In the following section we report a technical overview of the framework by presenting its components for face tracking and speech segmentation.

## A Technical Overview of the MAGiC Framework

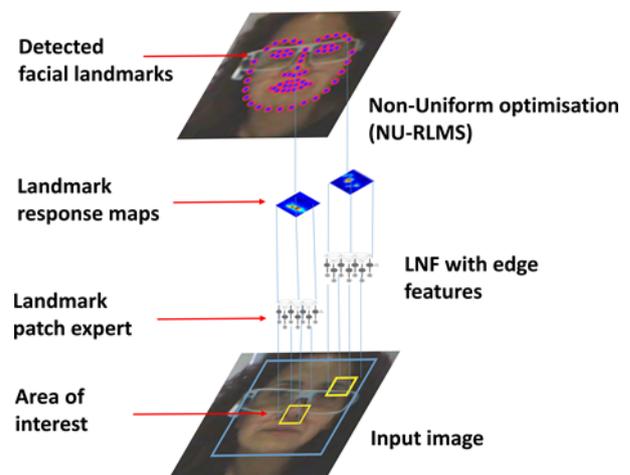
In the two subsections below, we present how *face tracking* and *speech segmentation* are conducted by MAGiC through its open source components.

### Face Tracking

Face tracking has been a challenging topic in computer vision. In face tracking, a face in a video-frame is detected first, and then it is tracked throughout the stream. In the present study, we employ an established face tracking toolkit called *OpenFace*, an open source tool for analyzing facial behavior (Baltrušaitis, Robinson, & Morency, 2016). *OpenFace* combines out-of-the-box solutions with the state-of-the-art research to perform tasks including facial-landmark detection, head-pose estimation and action unit (AU) recognition. The MAGiC’s face tracking method is based on Baltrušaitis et al. (2016), Baltrušaitis, Mahmoud, & Robinson, 2015, and Baltrušaitis, Robinson, & Morency (2013).

*OpenFace* utilizes a pre-trained face detector (trained in *dlib*), which is an open source machine-learning library

written in C++ (King, 2009). The Max-margin object-detection algorithm (MMOD) of the face detector uses Histogram of Oriented Gradients (HOG) feature extraction. The face detector is trained on sub-windows in an image. Since the number of windows may be large even in moderately sized images, relatively small amount of data is enough for training (King, 2009; 2015). After detecting a face for detecting the facial landmarks, *OpenFace* utilizes an instance of Constrained Local Model (CLM), namely Constrained Local Neural Field (CLNF), to perform feature detection problems even in complex scenes. The response maps are extracted by using pre-trained patch experts. Patch responses are optimized with a fitting method, viz. Non-Uniform Regularized Landmark Mean-Shift (NU-RLMS, see Figure 1).



**Figure 1.** A demonstration of *OpenFace* methodology, adapted from Baltrušaitis et al. (2013) It is intentionally limited to two landmarks patch expert for the sake of clarity (all photos used upon the permission of the participant).

The CLM (Constrained Local Model) is composed of three main steps. First, a Point Distribution Model (PDM) extracts the mean geometry of a shape from a set of training shapes. A statistical shape model is built from a given set of samples. Each shape in the training set is characterized by a set of landmark points. The number of landmarks and the anatomical locations represented by specific landmark points should be consistent from one shape to the next. For instance, for a face shape, specific landmark points may always correspond to eyelids. In order to minimize the sum of squared distances to the mean of a set, each training shape is aligned into a common coordinate frame by rotating, translating and scaling them. The Principal Component Analysis (PCA) is used for picking out the correlations between groups of landmarks among the

trained shapes. At the end of the PDM step, patches are created around each facial landmark. The patches are trained with a given set of face-shapes.

**Patch Experts**, also known as *local detectors*, are used for calculating response maps that represent the probability of a certain landmark that is being aligned at image location  $x_i$  (Eq. (1)), from Baltrušaitis et al. (2013). A total of 68 patch experts are employed to localize 68 facial landmark positions, as presented in Figure 2.

$$\pi(x_i) = C_i(x_i; I), \quad (1)$$

where  $I$  is an intensity image, and  $C_i$  is a logistic regressor intercept with a value between 0 to 1 (0 representing no alignment and 1 representing perfect alignment). Due to its computational advantages and implementational simplicity, Support Vector Regressors (SVR) are usually employed as patch experts. On the other hand, the CLNF (Constrained Local Neural Field) model uses the LNF approach, which considers spatial features that lead to fewer peaks, smoother responses and reduced noises.

**Regularised Landmark Mean Shift (RLMS)** is the next step of the CLM (Constrained Local Model). RLMS is a common method to solve the fitting problem. It updates the CLM parameters to get closer to a solution. An iterative fitting method is used to update the initial parameters of the CLM, until achieving a convergence to an optimal solution. The general concept of iterative fitting is defined in Eq. (2), adapted from Baltrušaitis et al. (2013):

$$p^* = \arg \min_p [R(p) + \sum_1^n D_i(x_i; I)], \quad (2)$$

where  $R$  is a regularization term and  $D_i$  represents the misalignment measure for image  $I$  at image location  $x_i$ . Regularizing model parameters is necessary to prevent overfitting (overfitting causes a model perform poor on data not used during training). RLMS does not discriminate between confidence levels of response maps. Due to noisy response maps, a novel non-uniform RLMS weighting mean-shifts is employed for efficiency.

At the end RLMS, the *OpenFace* toolkit detects a total of 68 facial landmarks (Figure 2). The detection of the face boundaries based on facial landmarks enables more precise calculations than using a rectangle that covers the face region.



**Figure 2.** A total of 68 landmark positions on a face.

We extended the *OpenFace* source code by making a set of improvements, which allowed the user to perform manual AOI annotation, generate visualizations that employ proposed input parameters, build a custom face detector and then use the detector to track the face, and generate separate output files depending on the input parameters. In the following section, we present the speech segmentation module.

## Speech Segmentation

Speech is a continuous audio stream with dynamically changing and usually indistinguishable parts. Speech analysis has been recognized as a challenging domain of research, since it is difficult to automatically identify clear boundaries between speech-related units. Speech analysis involves two interrelated family of methodologies, namely *speech segmentation* and *diarization*. Speech segmentation is the separation of the audio recordings into units of homogeneous parts, such as speech, silence, and laugh. Diarization is used for extracting various characteristics of signals, such as speaker identity, gender, channel type and background environment (e.g., noise, music, silence). The MAGiC framework addresses both methodologies, since both segmentation and identification are indispensable components of face-to-face conversation.

In MAGiC, we employed the *CMUSphinx* Speech Recognition System (Lamere et al., 2003) by extending it for the analysis of recorded speech. *CMUSphinx* is an open source, platform-independent and speaker-independent speech recognition system. It is integrated with *LIUM*, an open source toolkit for speaker segmentation and diariza-

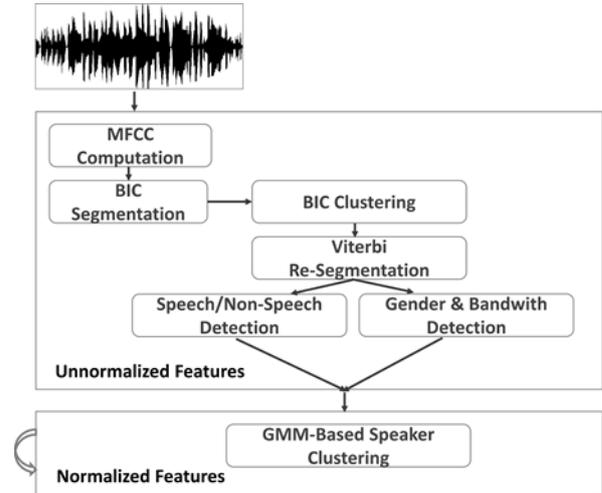
tion. The speech analysis process starts with feature extraction. *CMUSphinx* functions extract features, such as Mel-frequency Cepstral Coefficients (MFCC), which collectively represent power spectrum of a sound segment. It then performs speech segmentation based on Bayesian Information Criterion (Barras, Zhu, Meignier, & Gauvain, 2006; Chen & Gopalakrishnan, 1998).

The MAGiC framework performs two passes over the sound signal for speech segmentation. In the first pass, a distance-based segmentation process detects the *change points* by means of a likelihood measure, namely Generalized Likelihood Ratio (GLR). In the second pass, the system mixes together successive segments from the same speaker. After the segmentation, Bayesian Information Criterion (BIC) hierarchical clustering is performed with an initial set, which consists of one cluster per each segment. At each iteration, the  $\Delta BIC_{ij}$  values for two successive clusters  $i$  and  $j$  are defined, as described by Meignier and Merlin (2010), as follows:

$$\Delta BIC_{ij} = \frac{n_i+n_j}{2} \log|\Sigma| - \frac{n_i}{2} \log|\Sigma_i| - \frac{n_j}{2} \log|\Sigma_j| - \lambda P, \quad (3)$$

where  $|\Sigma_i|$ ,  $|\Sigma_j|$  and  $|\Sigma|$  are the determinants of the Gaussians associated to clusters  $i$ ,  $j$  and  $(i + j)$ ;  $n_i$  and  $n_j$  refer to the total lengths of cluster  $i$  and cluster  $j$ ;  $\lambda$  is the smoothing parameter that is chosen to get a good estimator, and  $P$  is the penalty factor. The  $\Delta BIC$  values for each successive cluster are calculated and they are merged when the value is less than 0.

As the next step of the speech analysis, Viterbi decoding is applied for re-segmentation. A Gaussian Mixture Model (GMM) with eight components is employed to represent the clusters. The parameters of the mixture are estimated by Expectation Maximization (EM). To minimize the number of undesired segments, such as long segments or the segments that overlap with the word boundaries, the segments were slightly moved to their low energy states, and the long segments were cut iteratively to create segments that are shorter than 20 seconds. Until this stage in the workflow, non-normalized features that preserve background information are employed during segmentation and clustering. This method facilitates differentiating speakers and assigning one single speaker to each cluster. On the other hand, it may also lead to allocation of the same speaker in multiple clusters. To resolve this issue, GMM-based speaker clustering is performed with normalized features to assign the same speaker to the same cluster. The GMM iterates until it reaches a pre-defined threshold value. Figure 3 shows the workflow of speaker diarization.



**Figure 3.** Typical workflow for speaker diarization and segmentation, adapted from LIUM Speaker Diarization Wiki Page (<http://www-lium.univ-lemans.fr/diarization/doku.php/overview>)

We extended the *CMUSphinx* source code and made the following additions. *CMUSphinx* does not generate segments for the whole audio. For instance, it does not generate segments for the parts when the speaker could not be identified. However, those non-segmented parts might contain useful information. Thus, we carried out additional development to automatically generate audio segments of non-segmented parts. To do this, the time interval of each successive segment was calculated. If there existed a time difference between the end of the previous segment and the beginning of the next one, we created a new audio-segment that covered that time-range. We also added a new functionality for segmenting audio with specified intervals.

## Demonstration of the MAGiC Framework: A Pilot Study

This section reports a pilot study that demonstrates the functionalities and benefits of the MAGiC framework. The setting is a mock job interview setting, where a pair of participants wear eye glasses and conduct a job interview. The gaze data and the video data are then analyzed by MAGiC.

## Participants, Materials and Design

Three pairs of male participants (university students as volunteers) took part in the pilot study (mean age 28, SD = 4.60). The task was a mock job interview. One of the participants was assigned the role of an interviewer and the other an interviewee. The roles were assigned randomly. All the participants were native Turkish speakers and they had normal or corrected-to-normal vision. No time limit was introduced to the participants.

At the beginning of the session, the participants were informed about the task. Both participants wore monocular Tobii eye tracking glasses with a sampling rate of 30 Hz with a 56°x40° recording visual angle capacity for the visual scene. The glasses recorded the video of the scene camera and the sound, in addition to gaze data. The IR (infrared)-marker calibration process was repeated until reaching 80% accuracy. After the calibration, the participants were seated on the opposite sides of a table, approximately 100 cm away from each other. A beep sound was introduced to indicate the beginning of a session, for synchronization in data analysis.

Eight common job interview questions, adopted from Villani, Repetto, Cipresso, & Riva(2012), were presented to an interviewer on a paper. The interviewer was instructed to ask the given questions, and also to evaluate the interviewee per each question by using paper and pencil.

## Data Analysis

We conducted data analysis using the speech analysis module, the AOI analysis module and the summary module in MAGiC. As a test environment, a PC was used with an Intel Core i5 2410M CPU at 2.30 GHz with 8 GB RAM running Windows 7 Enterprise (64 bit).

**Speech Analysis.** First, a MAGiC function (“Extract and Format Audio”) was employed to extract the audio and then to format the extracted audio for subsequent analysis. This function was run separately for each participant in the pair. Therefore, in total, six sound (.wav) files were produced. Each run took one to two seconds for the extraction. Second, the formatted audio files were segmented one by one. Audio-segments and a text file were created. The text file contained the id number and the duration of each segment. The number of segments varied depending on the length and the content of the audio (Table 1). Each run took one to two seconds for the analysis.

Table 1. Audio length and the number of segments for each participant’s recording.

	Interviewer/ Interviewee	
	Audio Length (m:ss.ms)	Number of Segments
<b>Pair-1</b>	3:46.066/ 3:57.00	170/176
<b>Pair-2</b>	5:25.066/ 5:40.00	120/200
<b>Pair-3</b>	5:28.000/ 5:09.00	246/208

Third, time-interval estimation, synchronization and re-segmentation were performed for each pair by using an interface that we call the “Time Interval Estimation” panel.

When the experiment session is conducted with multiple recording devices, one of the major issues is synchronization of the recordings. Currently, eye tracker manufacturers do not provide synchronization solutions. In most cases, the device clocks are set manually. MAGiC provides a semi-automatic method for synchronizing multiple recordings from a participant pair. In this method, the user is expected to specify the initial segment of the session in both recordings. Since, user identifies the beginning of sessions by listening to automatically created segments instead of a whole speech, this results in more accurate time estimation. Then, MAGiC calculates the time offset to provide synchronization by taking the time difference of the specified initial segments. After performing the re-segmentation process (by utilizing synchronization information and by merging segments from both recordings), we end up with equal-length session intervals for participants within each pair. The closer the microphone is to a participant, the cleaner and better the gathered audio recording is. Thus, segmentation of multiple recordings from the same session may result in different number of segments. A re-segmentation process merges segments from different recordings in order to reduce data loss. Table 2 presents the experiment duration in milliseconds and the number of segments produced after re-segmentation in our pilot study. Each run took one to two seconds.

Table 2. Audio length and number of segments for each participant’s recording. The number of segments increased after re-segmentation. (see Table 1)

	Exp. Duration (m:ss.ms)	Number of Segments
<b>Pair-1</b>	3:02.40	261
<b>Pair-2</b>	5:05.40	282
<b>Pair-3</b>	4:42.60	406

Finally, speech annotation was performed. A list of pre-defined speech acts was prepared as the first step of the analysis: Speech, Speech Pause, Thinking (e.g., “uh”, “er”, “um”, “eee”, for instance), Ask-Question, Greeting (e.g., “welcome”, “thanks for your attendance”), Confirmation (e.g., “good”, “ok”, “huh-huh”), Questionnaire Filling (Interviewer filling in questionnaire), Pre-Speech (i.e., warming up the voice), Reading and Articulation of Questions, Laugh, Signaling end of the speech (e.g., “that is all”).

The next step was the manual annotation process. For the end user, this process involved selecting the speech act(s) and annotating the segments. At each annotation, a new line was appended and displayed, which contained the relevant segment's time-interval, its associated participant (if any) and user-selected speech-act(s). AOI analysis was performed for the three pairs of participants separately. Each run took ten to twenty minutes, depending on the session-interval.

**AOI Analysis.** All six video recordings of the pilot study were processed with *OpenFace's default-mode face detector*. The tracking processes produced two-dimensional landmarks on the interlocutor's face image. The process took 4 to 10 minutes per video. Then, the gaps with at most two frames-duration were filled in by linear interpolation of raw gaze data. The raw gaze data file included the frame number, gaze point classification (either *Unclassified* or *Fixation*), and x-y coordinates. The processed data comprised 2% of total raw gaze data (Table 3). The gap filling process took less than a second per pair.

Table 3. The number and ratio of the filled gaps for each participant's raw gaze data.

Interviewer/ Interviewee		
	Number of filled gaps	Ratio of filled gaps (%)
<b>Pair-1</b>	146 / 236	2.15 / 3.32
<b>Pair-2</b>	171 / 236	1.75 / 2.31
<b>Pair-3</b>	157 / 335	1.60 / 3.61

After the gap filling process, we performed AOI detection by setting the parameters for eye tracker accuracy and image resolution. In the present study, the size of the captured images for face tracking was  $720 \times 480$  pixels, while the eye tracker image-frame resolution was  $640 \times 480$ . The eye tracking glasses had a reported degree of accuracy of half a degree of visual angle. The built-in scene camera recording angles of the eye tracking glasses were 56 de-

grees horizontal and 40 degrees vertical. The seating distance between the participants was approximately 100 cm. Accordingly, the eye tracker accuracy was 4.84 pixels horizontal and 5.34 pixels vertical. The AOI detection took a couple of seconds. Table 4 presents the number and the ratio of image-frames that AOI detection failed due to undetected face. The results indicate that higher undetected-face rates were observed at the interviewer's recordings. Nevertheless, face detection was performed with more than 90% success on average.

Table 4: The number and ratio of image-frames that face could not be detected.

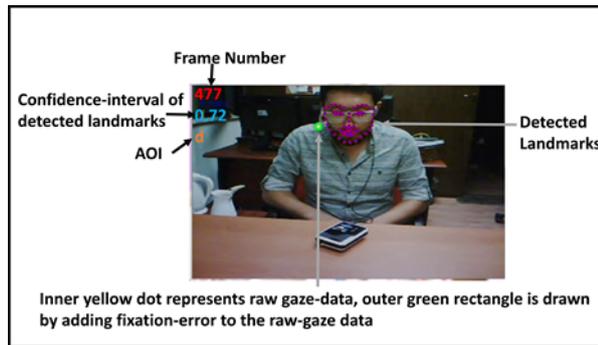
Interviewer/ Interviewee		
	Number of undetected	Ratio of undetected (%)
<b>Pair-1</b>	570 / 173	10.4 / 3.16
<b>Pair-2</b>	2113 / 488	23.1 / 5.33
<b>Pair-3</b>	1251 / 117	14.8 / 1.38

The absence of gaze data is another issue that leads to failure in AOI detection. Table 5 shows the ratio of undetected AOIs due to the absence of gaze data.

Table 5. The number and ratio of image-frames that raw gaze data were absent.

Interviewer/ Interviewee		
	Number of undetected	Ratio of undetected (%)
<b>Pair-1</b>	3237 / 392	59.1 / 7.16
<b>Pair-2</b>	4762 / 1050	52.0 / 11.50
<b>Pair-3</b>	4010 / 1732	47.3 / 20.40

The failure in AOI detection on the interviewer's side was approximately 50%. This is due to the experimental setting, where the interviewer looked at the questions to read them. This is a situation that experiment designers face frequently in dynamic experiment settings. The MAGiC framework's interface allows the user to detect the source of the problem and to annotated it by a label through a panel interface that we name “Visualize Tracking”. The panel interface displays the recording by overlaying the detected facial landmarks, raw gaze data and gaze annotation (looking at the interlocutor's face, i.e., in, or looking away the interlocutor's face i.e., out) on top of the video recording for each frame, as shown in Figure 4.



**Figure 4.** A snapshot from the visualize-tracking panel.

The analysis of the scenes by the “Visual Tracking” panel revealed that the missing raw gaze data were due to interviewer’s reading and articulation of the questions, and evaluating the interviewee’s response by using paper and pencil. In those cases, the interviewer looked outside of the glasses frame to read the questions on the notebook. In our pilot study, the manual annotation took 15 to 20 minutes per pair, on average.

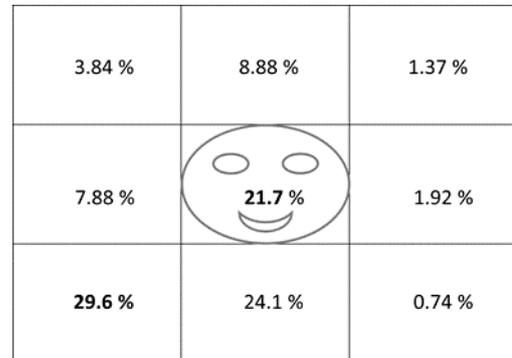
The final step in the AOI-analysis was composed of two further functions provided by the MAGiC framework: The re-analysis step merged automatically-detected AOIs with manually extracted AOI-labels. After then, the detection ratio was compared with the previous outcomes.

Table 6 shows face-detection and gaze-detection accuracies for the interviewer’s recordings. The results reveal an improvement of more than 30% after the final step, compared to the previous analysis steps (cf. Table 4 and Table 5).

Table 6. The number and ratio of the image-frames that face and gaze could not be detected.

Face		
Id	Number of undetected face	Ratio of undetected face (%)
1	4	0.07
2	38	0.41
3	5	0.06
Gaze		
Id	Number of undetected gaze	Ratio of undetected gaze (%)
1	1508	27.55
2	1292	14.10
3	1143	13.48

The analyses also revealed the distribution of interlocutor’s gaze locations. The findings showed a tendency of more frequent gaze aversion on the right side, especially to the right-bottom (see Figure 5).



**Figure 5.** At the 21.7% of the dwell time, participants looked at interlocutor’s face. The bottom left corner with the 29.6% was the most sighted region.

The rightward shifts are usually associated with verbal thinking, whereas leftward shifts are usually associated with visual imagery (Kocel, Galin, Ornstein, & Merrin, 1972). On the other hand, more recent studies report that the proposed directional patterns do not consistently occur when a question elicited verbal or visuospatial thinking. Instead, the individuals are more likely to avert their gaze while a listening to a question from the partner (see Ehrlichman & Micic (2012) for a review).

A further investigation of mutual gaze behavior of the conversation pairs and speech acts was conducted by a two-way ANOVA. The speech-acts had eleven levels (Speech, Speech Pause, Thinking, Ask-Question, Greeting, Confirmation, Questionnaire Filling, Pre-Speech, Reading Questions, Laugh and Signaling End of the Speech) and the mutual gaze behavior had four levels (Face Contact, Aversion, Mutual Face Contact, Mutual Aversion).

The analysis with normalized gaze distribution frequency revealed a main effect of gaze behavior,  $F(3,72) = 58.3, p < .05$ . The Tukey post hoc test was performed to establish the significance of differences in frequency scores with different gaze behavior and speech-acts. It revealed that the frequency of Gaze Aversion ( $M=0.5, SD=0.12$ ) was significantly larger than the frequency of Face Contact ( $M=0.1, SD=0.19, p < .05$ ), the frequency of Mutual Face Contact

( $M=0.02$ ,  $SD=0.06$ ,  $p<.05$ ), as well as the frequency of Mutual Aversion ( $M=0.38$ ,  $SD=0.15$ ,  $p<.05$ ). Moreover, the frequency of Mutual Aversion was significantly larger than the frequency of Face Contact ( $p<.05$ ) and the frequency of Mutual Face-Contact ( $p<.05$ ), while there was no significant difference between the frequency of Face Contact and the frequency of Mutual Face Contact ( $p=0.31$ ).

Finally, the interaction between speech-acts and gaze behavior was investigated. The results indicated that when the participants were *thinking*, there was a significant frequency difference between the frequency of Mutual Aversion ( $M=0.58$ ,  $SD=0.07$ ) and the frequency of Face Contact ( $M=0.03$ ,  $SD=0.05$ ,  $p<.05$ ), as well as significant difference between the frequency of Mutual Aversion and the frequency of Mutual Face Contact ( $M=0.01$ ,  $SD=0.02$ ,  $p=.02$ ).

## An Evaluation of the Contributions of the MAGiC Framework

In this section, we report how the MAGiC framework facilitated gaze analysis in the reported pilot study. MAGiC reduced the amount of the time spent for preparing manually annotated gaze and audio data for each image-frame of a scene video. To manually identify gaze contact and gaze aversion, and its location, a researcher would annotate 36,000 image-frames for a 10-minute session recorded by a 60 Hz eye tracker. Assuming that it takes 1 second to manually annotate a frame, the annotation would last 10 hours. MAGiC took approximately 5 to 10 minutes when it was run on a typical personal computer in today's technology (Intel Core i5 2.3 GHz CPU and 8 GB of RAM.) The time spent for the Area of Interest (AOI) and audio annotation was also reduced. The automated annotation improved the quality of annotated data. It is difficult for human annotators to detect speech instances at this level of temporal granularity. Since full annotation is the holy grail of gaze data in dynamic analysis scenes, MAGiC also offers an interface to make manual AOI annotation to the user. This component of MAGiC is one of the pillars for improvement for future versions.

MAGiC provides the functionality for visualizing face tracking data and AOI annotation frame-by-frame. It overlays the detected facial landmarks, the raw gaze data, and the status of gaze interaction in a single video recording. It also displays the ratio of non-annotated gaze data (thus, the success level of face detection) as a percentage of total data to the user. The absence of raw gaze data or undetected faces are major reasons for the failure of an automatic AOI annotation. The user can introduce training to create a custom face detector for better face detection performance. The MAGiC software is licensed under the GNU General Public License (GPL). Therefore, the source code of the application is openly distributed and programmers are encouraged to study and contribute to its development. In addition to MAGiC, we also provide the modified component toolkits (OpenFace for face tracking, dlib for training of a custom face detector, and CMUSphinx for speech segmentation) on MAGiC's github repository: MAGiC\_v1.0: <https://github.com/ulkursln/MAGiC/releases>

## Usability Analysis of MAGiC

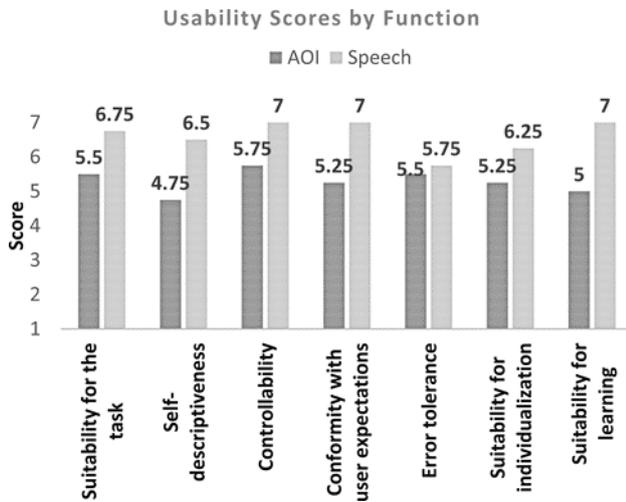
This section reports a usability analysis of the MAGiC framework. For the analysis, the AOI Analysis interface and the Speech Analysis interface were randomly assigned to a total of eight participants. The participants performed data analysis by using publicly available sources (see Supplementary material<sup>1</sup>). The usability analysis was conducted in three steps, as described below:

- (1) Perform the analysis manually,
- (2) Perform the analysis by using MAGiC,
- (3) Assess the usability of MAGiC using 7-point scale ISO 9241/10 questionnaire.

The Usability test scores are presented in Figure 6.

---

<sup>1</sup> See the MAGiC App Channel under Youtube, <https://www.youtube.com/channel/UC2gvq0OluwpdjVKGSgGg-vaQ>, and MAGiC App Wiki Page under Github



**Figure 6.** All of the usability metrics were scored higher than an average.

We recorded the time spent to perform data analysis, and then we compared it to the average duration when the participants performed the same analysis manually. In the AOI analysis, the mean duration to annotate a single frame decreased from 29.1 seconds ( $SD=22.7$ ) for manual annotation to an average of 0.09 seconds ( $SD=0.02$ ) in MAGiC. In the speech analysis, the mean duration for a single annotation decreased from 44.5 seconds ( $SD=8.8$ ) for manual annotation to an average of 7.1 seconds ( $SD=1.4$ ) in MAGiC.

## Discussion and Conclusion

In the present study, we introduced the MAGiC framework. It provides researchers an environment for the analysis of gaze behavior of a pair in conversation. Human-Human conversation settings are usually dynamic scenes, in which the conversation partners exhibit a set of specific gaze behavior, such as gaze contact and gaze aversion. MAGiC detects and tracks interlocutor's face automatically in a video recording. Then it overlays gaze location data to detect gaze contact and gaze aversion behavior. It also incorporates speech data into the analysis by means of providing an interface for annotation of speech-acts.

MAGiC facilitates the analysis of dynamic eye tracking data by reducing the annotation effort and the time spent for frame-by-frame manual analysis of video data. Its capability for automated multimodal (i.e., gaze and speech-act) analysis makes MAGiC advantageous over error-prone human annotation. The MAGiC interface allows

researchers to visualize face tracking process, gaze-behavior status and annotation efficiency on the same display. It also allows the user to train the face tracking components by providing labelled images manually.

The environment has been developed as an open source software tool, which is available for public use and development. MAGiC has been developed by integrating a set of open source software tools, in particular *OpenFace* for analyzing facial behavior, *dlib* for machine learning of face tracking and *CMUSphinx* for the analysis of recorded speech and extending their capabilities further for the purpose of detecting eye movement behavior, and for annotating speech data simultaneously with gaze data. MAGiC's user interface is composed of a rich set of panels, which provide the user an environment to conduct a guided, step-by-step analysis.

MAGiC is able to process data from a single eye tracker or data in a dual eye tracking setting. We demonstrated MAGiC's capabilities in a pilot study, which was conducted in a dual eye-tracking setting. We described MAGiC's data analysis capabilities by describing the analysis step on the recorded data in the pilot study. We intentionally employed a low-frequency eye-tracker, with a relatively low video quality, and a low-illuminated environment, since these are typical real-environment challenges that influence face tracking capabilities. Our analysis revealed that MAGiC is able to exhibit acceptable success ratios in automatic analyses, with an average Area of Interest (AOI) labelling (i.e., gaze contact and gaze aversion detection) efficiency of approximately 80%. Likely improvements in eye tracking recording frequency, eye tracking data quality, and image resolution of video recordings have the potential to increase the accuracy of MAGiC's outputs to better levels. We also note that MAGiC's speech analysis component, namely *CMUSphinx* provides several high-quality acoustic models, although there is no pre-build acoustic model for Turkish. Despite this challenge, MAGiC returned successful results for the speech analysis, too. The speech-act annotation also helped us overcoming speech segmentation issues by providing sub-segments for speech.

All the data analyses were completed in approximately two hours for the three pairs of participants. Our usability analyses revealed that the time and effort spent for manual, frame-by-frame video analysis and speech segmentation takes much longer to complete, in addition to being prone to human annotator errors.

Recently, MAGiC is in its first version. Our future work will include making improvements in the existing capabilities of MAGiC, as well as developing new capabilities. For instance, face-detection ratio may be increased by employing recently published OpenFace 2.0. Also, in its current version, MAGiC sets an AOI-label on the interlocutor's face image. We plan to expand this labelling method so that it processes other objects, such as the objects on a table. This will expand the domain of use of MAGiC into a broader range of dynamic visual environments not limited to face-to-face communication. However, this development would require training a detector for the relevant objects, which is a challenging issue for generalization of the object recognition capabilities. Moreover, the face-tracking function of MAGiC already makes it possible to extract facial expressions, based on the Facial Action Coding System (FACS). As a further improvement, MAGiC may automatically summarize facial expressions during the course of a conversation.

Finally, for speech analysis, MAGiC provides functions for semi-automatically synchronizing recordings. Further development of MAGiC may address improving its synchronization capabilities, its capability to transcribe speech into text and its capability for training speech-act annotation with pre-defined speech acts and automating subsequent annotations.

## Ethics and Conflict of Interest

The author(s) declare(s) that the contents of the article are in agreement with the ethics described in <http://biblio.unibe.ch/portale/elibrary/BOP/jemr/ethics.html> and that there is no conflict of interest regarding the publication of this paper.

## References

- Abele, A. (1986). Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior*, 10(2), 83–101. <https://doi.org/10.1007/bf01000006>
- Archer, D., & Akert, R. M. (1977). Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of Personality and Social Psychology*, 35, 443–449. <https://doi.org/10.1037//0022-3514.35.6.443>
- Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press. <https://doi.org/10.2307/3032267>
- Austin, J. L. (1962). *How to do things with words*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198245537.001.0001>
- Bales, R., Strodtbeck, F., Mills, T., & Roseborough, M. (1951). Channels of communication in small groups. *American Sociological Review*, 16(4), 461–468. <https://doi.org/10.2307/2088276>
- Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic Action Unit detection, in Facial Expression Recognition and Analysis Challenge, In *Proceeding of the 11th IEEE International Conference Automatic Face and Gesture Recognition* (Vol. 6, pp. 1-6). IEEE. <https://doi.org/10.1109/fg.2015.7284869>
- Baltrušaitis, T., Robinson, P., & Morency, L. P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 354-361). IEEE. <https://doi.org/10.1109/iccvw.2013.54>
- Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision* (pp. 1-10). IEEE. <https://doi.org/10.1109/wacv.2016.7477553>
- Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is there a 'language of the eyes'? Evidence from normal adults, and adults with autism or Asperger syndrome. *Visual Cognition*, 4(3), 311–331. <https://doi.org/10.1080/713756761>
- Barras, C., Zhu, X., Meignier, S., & Gauvain, J. L. (2006). Multi-stage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5), 1505–1512. <https://doi.org/10.1109/tasl.2006.878261>
- Brône, G., Oben, B., & Goedemé, T. (2011, September). Towards a more effective method for analyzing mobile eye-tracking data: integrating gaze data with object recognition algorithms. In *Proceedings of the 1st international workshop on pervasive eye tracking & mobile eye-based interaction* (pp. 53-56). ACM. <https://doi.org/10.1145/2029956.2029971>

- Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., & Yan, H. (1999). Embodiment in conversational interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp 520–527). ACM. <https://doi.org/10.1145/302979.303150>
- Chen, S., & Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA.
- De Beugher, S., Brône, G., & Goedemé, T. (2014, January). Automatic analysis of in-the-wild mobile eye-tracking experiments using object, face and person detection. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on* (Vol. 1, pp. 625-633). IEEE. <https://doi.org/10.5220/0004741606250633>
- Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283–292. <https://doi.org/10.1037/h0033031>
- Ehrlichman, H., & Micic, D. (2012). Why Do People Move Their Eyes When They Think?. *Current Directions in Psychological Science*, 21(2), pp.96-100. <https://doi.org/10.1177/0963721412436810>
- Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 195–226). MIT Press.
- Fasola, J., & Mataric, M. J. (2012). Using socially assistive human-robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE*, 100(8), 2512–2526. <https://doi.org/10.1109/jproc.2012.2200539>
- Ford, M., & Holmes, V. M. (1978). Planning units and syntax in sentence production. *Cognition*, 6, 35–53. [https://doi.org/10.1016/0010-0277\(78\)90008-2](https://doi.org/10.1016/0010-0277(78)90008-2)
- Goldman-Eisler, F. (1968). *Psycho-linguistics: Experiments in spontaneous speech*. New York, NY: Academic Press.
- Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York, NY: Academic Press.
- Grosjean, F., & Lane, H. (1976). How the listener integrates the components of speaking rate. *Journal of Experimental Psychology*, 2(4), 538-543. <https://doi.org/10.1037//0096-1523.2.4.538>
- Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175–204.
- Hieke, A. E., Kowal, S., & O’Connell, D. C. (1983). The trouble with “articulatory” pauses. *Language and Speech*, 26, 203–215.
- Hietanen, J. K., Leppänen, J. M., Peltola, M. J., Linna-Aho, K., & Ruuhiala, H. J. (2008). Seeing direct and averted gaze activates the approach-avoidance motivational brain systems. *Neuropsychologia*, 46(9), 2423–2430. <https://doi.org/10.1016/j.neuropsychologia.2008.02.029>
- Hird, K., Brown, R., & Kirsner, K. (2006). Stability of lexical deficits in primary progressive aphasia: Evidence from natural language. *Brain and Language*, 99, 137-138. <https://doi.org/10.1016/j.bandl.2006.06.083>
- Holmqvist, K., Nyström, N., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (Eds.) (2011). *Eye tracking: a comprehensive guide to methods and measures*. Oxford University Press.
- Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26(1), 22–63. [https://doi.org/10.1016/0001-6918\(67\)90005-4](https://doi.org/10.1016/0001-6918(67)90005-4)
- King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10, 1755-1758. Retrieved from <http://jmlr.csail.mit.edu/papers/v10/king09a.html>
- King, D. E. (2015). Max-Margin Object Detection. *arXiv preprint arXiv:1502.00046*. Retrieved from <http://arxiv.org/abs/1502.00046>
- Kirsner, K., Dunn, J., & Hird, K. (2005). Language productions: A complex dynamic system with a chronometric footprint. *Paper presented at the 2005 International Conference on Computational Science*, Atlanta, GA.
- Klinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological Bulletin*, 100(1), 78–100. <https://doi.org/10.1037//0033-2909.100.1.78>
- Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, 35, 162–179. <https://doi.org/10.1016/j.wocn.2006.04.001>
- Kocel, K., Galin, D., Ornstein, R., & Merrin, E. (1972). Lateral eye movement and cognitive mode. *Psychonomic Science*, 27(4), pp.223-224. <https://doi.org/10.3758/bf03328944>

- Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., ... Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In Proceedings of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing. Hong Kong.
- Mason, M. F., Tatlow, E. P., & Macrae, C. N. (2005). The look of love: Gaze shifts and person perception. *Psychological Science*, 16, 236–239. <https://doi.org/10.1037/e633912013-450>
- Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, 6, 109-114. <https://doi.org/10.1017/cbo9780511575945.011>
- Meignier, S., & Merlin, T. (2010). LIUM SpkDiarization: an open source toolkit for diarization. In *Proceedings of the CMU SPUD Workshop* (pp. 1-6). Retrieved from [http://www-gth.die.upm.es/research/documentation/referencias/Meignier\\_Lium.pdf](http://www-gth.die.upm.es/research/documentation/referencias/Meignier_Lium.pdf)
- Munn, S. M., Stefano, L., & Pelz, J. B. (2008). Fixation identification in dynamic scenes. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization - APGV '08* (pp. 33-42). ACM. <https://doi.org/10.1145/1394281.1394287>
- Pfeiffer, U. J., Timmermans, B., Bente, G., Vogeley, K., & Schilbach, L. (2011). A non-verbal Turing test: differentiating mind from machine in gaze-based social interaction. *PLoS One*. <https://doi.org/10.1371/journal.pone.0027591>
- Power, M. J. (1985). Sentence Production and Working Memory. *The Quarterly Journal of Experimental Psychology Section A*, 37(3), 367-385. doi:10.1080/14640748508400940
- Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X., Kirbas, C., ... Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions of Computer-Human Interaction*, 9(3), 171–193. <https://doi.org/10.1145/568513.568514>
- Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H., McCullough, K., ... Ansari, R. (2000). Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 247–254). IEEE. <https://doi.org/10.1109/cvpr.2000.854800>
- Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735. <https://doi.org/10.2307/412243>
- Schegloff, E. (1968). Sequencing in Conversational Openings. *American Anthropological* 70(6), 1075–1095. <https://doi.org/10.1525/aa.1968.70.6.02a00030>
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK.
- Stuart, S., Galna, B., Lord, S., Rochester, L., & Godfrey, A. (2014). Quantifying saccades while walking: Validity of a novel velocity-based algorithm for mobile eye tracking. In Proceedings of the IEEE 36th Annual International Conference Engineering of the Medicine and Biology Society, EMBC (pp. 5739-5742). IEEE. <https://doi.org/10.1109/embc.2014.6944931>
- Stuart, S., Hunt, D., Nell, J., Godfrey, A., Hausdorff, J. M., Rochester, L., & Alcock, L. (2017). Do you see what I see? Mobile eye-tracker contextual analysis and inter-rater reliability. *Medical & Biological Engineering & Computing*, 56(2), 289-296. doi:10.1007/s11517-017-1669-z
- Villani, D., Repetto, C., Cipresso, P., & Riva, G. (2012). May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview. *Interacting with Computers*, 24(4), 265-272. <https://doi.org/10.1016/j.intcom.2012.04.008>