A GAZE-CENTERED MULTIMODAL APPROACH TO FACE-TO-FACE INTERACTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ÜLKÜ ARSLAN AYDIN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

JANUARY 2020

A GAZE-CENTERED MULTIMODAL APPROACH TO FACE-TO-FACE INTERACTION

Submitted by ÜLKÜ ARSLAN AYDIN in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Prof. Dr. Cem Bozşahin
Head of Department, **Cognitive Science** _____

Assoc. Prof. Dr. Cengiz Acartürk
Supervisor, **Cognitive Science Dept., METU** _____

Assoc. Prof. Dr. Sinan Kalkan
Co-Supervisor, **Computer Engineering Dept., METU** _____

Examining Committee Members:

Prof. Dr. Cem Bozşahin
Cognitive Science Dept., METU _____

Assoc. Prof. Dr. Cengiz Acartürk
Cognitive Science Dept., METU _____

Assoc. Prof. Dr. Mehmet Serdar Güzel
Computer Engineering Dept., Ankara University _____

Assoc. Prof. Dr. Hatice Köse
Computer Engineering Dept., İTÜ _____

Assist. Prof. Dr. Umut Özge
Cognitive Science Dept., METU _____

Date: **13.01.2020**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ülkü Arslan Aydın

Signature : _____

ABSTRACT

A GAZE-CENTERED MULTIMODAL APPROACH TO FACE-TO-FACE INTERACTION

Arslan Aydın, Ülkü

Ph.D., Department of Cognitive Sciences

Supervisor: Assoc. Prof. Dr. Cengiz Acartürk

Co-Supervisor: Assoc. Prof. Dr. Sinan Kalkan

January 2020, 167 pages

Face-to-face conversation implies that interaction should be characterized as an inherently multimodal phenomenon involving both verbal and nonverbal signals. Gaze is a nonverbal cue that plays a key role in achieving social goals during the course of conversation. The purpose of this study is twofold: (i) to examine gaze behavior (i.e., aversion and gaze on face) and relations between gaze and speech in face to face interaction, (ii) to construct computational models to predict gaze behavior using high-level speech features. We employed a job interview setting, where pairs (a professional interviewer and an interviewee) conducted mock job interviews. Twenty-eight pairs of native speakers took part in the experiment. Two eye-tracking glasses recorded the scene video, the audio and the eye gaze position of the participants. To achieve the first purpose, we developed an open-source framework, named MAGiC (A Multimodal Framework for Analyzing Gaze in Communication), for the analyses of multimodal data including video recording data for face tracking, gaze data from the eye trackers, and the audio data for speech segmentation. We annotated speech with two methods: (i) ISO 24617-2 Standard for Dialogue Act Annotation and, (ii) using tags employed by the previous studies that examined gaze behavior in a social context. We then trained simplified versions of two CNN architectures (VGGNet and ResNet) by using both speech annotation methods.

Keywords: Mobile Eye Tracking, face-to-face interaction, gaze analysis, ISO 24617-2 standard, CNN for time series

ÖZ

YÜZ YÜZE İLETİŞİME BAKIŞ MERKEZLİ
ÇOK MODLU YAKLAŞIM

Arslan Aydın, Ülkü
Doktora, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Doç. Dr. Cengiz Acartürk
Ortak Tez Yöneticisi: Doç. Dr. Sinan Kalkan

Ocak 2020, 167 sayfa

Yüz yüze iletişim, doğası gereği, etkileşimin, hem sözsel hem de sözsel olmayan sinyallerini içeren çok modlu bir yaklaşımla karakterize edilmesini gerektirir. Bakış, iletişim sürecinde, sosyal hedeflere ulaşmada kilit rol oynayan sözsüz bir ipucudur. Bu çalışmanın amacı iki yönlüdür: (i) bakış davranışını (göz kaçırma ve yüze bakma) yüz yüze iletişimdeki bakış ve konuşma arasındaki ilişkilerle incelemek, (ii) bakış davranışlarını tahmin etmek için, üst seviye konuşma özellikleri kullanan hesaplamalı modeler oluşturmak. Çiftlerin (mülakatı yapan bir profesyonel ve iş başvurusu yapan aday) sahte iş görüşmeleri yaptığı iş görüşmeleri ayarladık. Deneyde anadil konuşanlarından oluşan 28 çift yer aldı. İki göz izleme gözlüğü, çevredeki görüntü, ses ve katılımcıların baktıkları pozisyonları kaydetti. İlk amaca yönelik olarak, yüz izlemede kullanılan görüntü, göz izleme cihazlarından bakış ve konuşma segmentasyonunda kullanılan sesi içeren, çok modlu verilerin analizleri için MAGiC (İletişimde Bakışları Analiz Etmek için Çok Modlu Çerçeve) adlı açık kaynaklı bir çerçeve geliştirdik. Konuşmayı iki yöntemle etiketledik: (i) Diyalog Eylemi Etiketleme için ISO 24617-2 standardı ve (ii) sosyal bağlamda bakış davranışlarını inceleyen önceki çalışmalarda kullanılan etiketleri kullanma. Daha sonra her iki etiketleme yöntemini kullanarak iki CNN mimarisinin, VGGNet ve ResNet, basitleştirilmiş versiyonlarını eğittik.

Anahtar Sözcükler: Mobil Göz İzleme, yüz yüze iletişim, bakış analizi, ISO 24617-2 standardı, zaman serileri için CNN

To all women who struggle to survive and
be existent in this male-dominant world

ACKNOWLEDGMENTS

I am more than grateful to my supervisor Assoc. Prof. Dr. Cengiz Acartürk and co-supervisor Assoc. Prof. Dr. Sinan Kalkan for their patience, support and guidance through such a long and challenging journey. I gained a lifetime experience and tried to learn from their valuable feedbacks. Even in the days when things were not so bright for me, their advice helped me to focus on my studies.

Besides my supervisor and co-supervisor, I would like to thank my thesis monitoring committee members Prof. Dr. Aydan Erkmen and Assoc. Prof. Dr. Hatice Köse for their feedbacks and insightful comments. They always made me feel more relieved and confident with their kind and inspiring words.

I would also like to express my heartfelt gratitude to Prof. Dr. Deniz Zeyrek Bozşahin for sharing her knowledge and study, which was of importance in shaping the scope of this dissertation. She is one of the best scientists to be taken as a role model for both academic and professional development. I am also grateful to Prof. Dr. Cem Bozşahin. He is the one who made me love philosophy and science. It was always a great pleasure to attend his classes.

Beyond all, I should express my deep gratitude to my husband Murat, who was always with me in my hard days and was one of my biggest supporters in finishing this thesis, to my dearest son Kerem Uraz, with the birth of whom I questioned life and its meaning again, to my mother, whose perseverance and personality I admire, to my father, who has always done his best to make me relieved and happy and to my little brother, Emre, who is one of the biggest reasons behind completing my thesis by his scientific and operational guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AOI	Area of Interest
BIC	Bayesian Information Criterion
CNN	Convolutional Neural Network
CLM	Constrained Local Model
DAMSL	Dialogue Act Markup using Several Layers
DiAML	Dialogue Act Markup Language
DAD	Direction and Attention Detector
ECA	Embodied Conversational Agent
EDD	Eye Direction Detector
EDU	Elementary Discourse Unit
FACS	Facial Action Coding System
FCL	Fully Connected Layer
GMM	Gaussian Mixture Model
GPF	General Purpose Function
GUI	Graphical User Interface
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IR	Infrared
ML	Machine Learning
MAGiC	A Multimodal Framework for Analyzing Gaze in Communication
NLP	Natural Language Processing
OpenCV	Open Source Computer Vision Library
PDTB	Penn Discourse Treebank
POR	Points-of-Regard
RLMS	Regularised Landmark Mean Shift
ReLU	Rectified Linear Unit
RST	Rhetorical Structure Theory
RST-DT	Rhetorical Structure Theory Discourse Treebank
SoC	Separation of Concerns

CHAPTER 1

INTRODUCTION

"When Gregor Samsa woke up one morning from unsettling dreams,
he found himself changed in his bed into a monstrous vermin".
(Franz Kafka, The Metamorphosis)

The skills of conversation using language along with the accompanying non-verbal signals set us apart from other species. Hence, conversation is considered to be one of the important indicators of humanness and human interaction. An influential figure in this sense was Alan Turing, who proposed keyboard conversation between machine and a human as a method for evaluating the ability of a computer to mimic a human (Turing, 1950). Nowadays, Embodied Conversational Agents (ECAs) are becoming more common. As Cassell (2000) stated, perhaps we are in the age of thinking about "face-to-face Turing test". Face-to-face conversation implies that interaction should be characterized as an inherently multimodal phenomenon, instead of speech in isolation (e.g., Kendon, 2004, Levinson & Holler, 2014; Mondada, 2016) This is because we, as human beings, have an ability to send and receive information by means of nonverbal cues such as facial expressions, gestures, gaze, and posture, during a social conversation. In particular domains, they even correspond to 50% - 70% of the entire messages that the speaker conveyed (Gerwing & Allison, 2009; Holler & Beattie, 2003). Listeners comprehend the speakers' messages by integrating multiple nonverbal and verbal channels (Kelly, Healey, Özyürek, & Holler, 2015; Willems, Özyürek, & Hagoort, 2007).

Gaze is an important nonverbal cue that plays a key role in achieving natural social interaction. Although it varies depending on different personalities and cultural backgrounds, we usually make eye contact with the interlocutor which, for instance, facilitates joint and shared attention. Even though we have such a tendency, face-to-face conversation is not just an interactive communication where partners constantly sustain eye contact, instead, it involves a sort of transition between gazing towards and away from the communication partner(s). In his landmark study, Kendon (1967) identified the differences in gaze behavior between the speaker and the listener. Speakers shifted their gaze away from the listeners more frequently, while listeners tend to keep an eye on the speakers (Bavelas, Coates, & Johnson, 2002; Ho, Foulsham,

& Kingstone, 2015; Kendon, 1967). Kendon (1967) ascribed three fundamental functions to the gaze behavior: (i) regularity function, (ii) monitoring function, and (iii) expressive function. First, gaze behavior has a regularity role in coordinating turns between speakers. Just before starting the speech, speakers direct their gaze away to indicate that they want to be the next speaker, i.e., taking the turn. Similarly, speakers avert their gazes from their partners to inform them that he/she continues speaking. On the other hand, the speaker looks at the recipient to show his/her intent to yield the turn. Secondly, speakers look at the others and try to interpret the recipient's gestures, intentions, attentional states and so on. Kendon (1967) suggested that speakers do not focus on recipients while speaking since they probably think about what they will say rather than interpreting the others' states. Lastly, mutual gaze expresses particularly the level of emotion and arousal. For instance, when the emotional level between two interacts is high, the mutual gaze will be less, as an indicator of embarrassment. Conversely, the desire to cooperate leads to an increase in the mutual gaze.

Research on gaze have attracted considerable attention since the 1960s (Klienke, 1986). Especially in recent decades, the development of eye-tracking technologies has enabled more accurate measurements and various experimental designs in this field (Gredeback, Johnson, & von Hofsen, 2010). However, most of the studies were performed in a laboratory by adopting static eye-tracking methods (Pfeiffer, Vogeley, & Schilbach, 2013), in which participants often monitor the stimulus presented to them on the computer screen. Although such experimental designs are advantageous in allowing one to provide a controlled procedure, the findings lack generalizability. Eye movements in the field might be different from those in studies conducted with static stimuli in a highly controlled laboratory environment (Risko, Richardson, & Kingstone, 2016). This difference can be explained by the two-way function of gaze in social communication. While gaze sends messages about, for instance, floor management or the desire to work together, we also gather information on emotions, intention or attentional states of others by gazing on them. Since we are somehow aware of this dual function of gaze, it causes an individual to be influenced by the presence of another person in the environment, and in terms of eye movements, individuals tend to behave differently compared to an environment where they are alone (Gobel, Kim, & Richardson, 2015; Risko et al., 2016). Studies have reported that people follow other's gaze more frequently and for a longer duration when they are not visible by their interlocutors. (Foulsham, Walker, & Kingstone, 2011; Gallup, Chong, & Couzin, 2012).

Describing gaze behavior in natural communication is an appropriate starting point to examine the underlying modalities and their relevance in face-to-face conversation. Advances in mobile eye-tracking technology have opened the door to researchers who study social interaction in real-life situations. Eye-tracking glasses are capable of recording participants' eye movements while they interact with the environment, without requiring participants to sit in front of a computer. This technology allows obtaining a rich data-set in dyadic interaction where both participants wore eye-tracking glasses. Since these glasses are available with relatively new technologies and data analysis is more challenging than the static experimental designs, there are not

many publications present in the literature yet. Rogers, Speelman, Guidetti, and Longmuir (2018) summarized two studies published in this area, one utilizing dual eye-tracking paradigm with Applied Science Laboratory (ASL) (Broz, Lehmann, Nehaniv, & Dautenhahn, 2012) and the other utilizing Dikablis eye-tracking glasses (Ho et al., 2015). Broz et al. (2012) recorded 15-minute conversations between 37 pairs and reported that during 46% of the entire conversation, participants involved in mutual face gaze. The drawback of this study was the significant loss in gaze data. If the participant was not looking through the center of the glasses, eye movements could not be recorded properly. In order to minimize data loss, Ho et al. (2015) performed manual coding by replaying the synchronized recordings exported from the glasses of both participants. They studied the timings of gaze behaviors in turn-taking mechanisms. Rogers et al. (2018) extend gaze on face analysis one step further by dividing the face area into five regions; eyes, nose, mouth, forehead and other parts, and the off-face area into four regions; off-left, off-right, up and down. They examined gaze patterns in a face to face conversation during which they utilized Tobii Pro Glasses2 along with Mangold INTERACT as the behavioral coding software. They discussed the necessity of more research to make a more accurate estimation of gaze patterns in dyadic conversation.

Meanwhile, studies of Natural Language Processing (NLP) involving text mining, automated question answering and machine translation have gained momentum as a reflection of the developments in Machine Learning (ML) technology (Meyer and Popescu-Belis 2012; Popescu-Belis 2016; Sharp, Jansen, Surdeanu, Clark, 2015). Hence, researchers' attention to discourse analysis has increased in parallel. There is a distinction between the usual meaning of a word or a sentence and the meaning it implies in specific circumstances. We need to distinguish between the direct and implied meanings of the texts. Sometimes we ask a question with the implicit intention of request. For instance, when one goes to a restaurant, one of the likely questions that would be asked to the waiter is "Can I see the menu?". In fact, nobody expects to hear "yes" or "no" as the answer indicating the ability to see a menu, instead this is a kind way of requesting the menu. This is the implicit intention of the speaker. This dichotomy, meaning and pragmatics on the one hand, the use on the other, is controversial among linguists regarding the specifying of relational arguments of a speech and in broader terms, of a discourse. Discourse relations might ground in lexical items, be driven by semantics, or consist of both intentional and semantic relations. In the last few decades, a variety of discourse annotation schemas were proposed involving RST (Rhetorical Structure Theory), RST Treebank, SDRT, DISCOR, ANNODIS and PDTB (Penn Discourse Treebank) In addition, an ISO standard for dialogue act annotation, namely ISO standard 24617-2 was developed (ISO DIS 24617-2; 2010). Behind these efforts, researchers have the goal of automating discourse analysis in accordance with technological developments, as well as the intention to characterize the high-level features of discourse.

Assuming that we are in the age of "face-to-face Turing test", we are expected to perform analysis of dyadic conversation with a multimodal approach and, hence create automation based on the performed analysis. Studies examining the relationship

between gaze and language processing showed the significance of that relation. Prasov and Chai (2008) demonstrated the importance of gaze for reference analysis in multiple interaction environments. In another study, Qu and Chai (2009) showed that the coupling of speech and gaze significantly makes word acquisition performance better. From this point of view, are proposed discourse annotation schemes, in particular ISO DIS 24617-2, proper and sufficient to be considered in coupling with another crucial modality in face-to-face communication, namely gaze, or can similar coupling performances between gaze be achieved with a simpler and hence cost-effective annotation method?

1.1. Significance and Scope of the Thesis

The studies in face-to-face communication are not new to the literature, yet the increasing interest in ECAs draw more researchers' attention to the field. In the present study, we investigate the relation between speech, particularly high-level language processing, instead of low-level features like acoustics properties, and gaze behavior, specifically face contact and gaze aversion, in a dyadic conversation. The main motivation behind the present study is to explore such relations in a more nuanced and comprehensive manner through employing state of the art technologies and by taking into account the limitations of the previous studies in the field. The main constraints encountered in the previous studies that we study are as follows:

i. Most of the eye tracking studies in the related research were conducted in a laboratory environment with highly controlled stimuli and subjects were generally asked to sit in front of the computer screen displaying the stimuli. However, this so-called static eye tracking method, is insufficient to reflect the underlying gaze behavior of face-to-face social interaction in real life.

ii. Eye trackers generate a raw data stream containing a list of points-of-regard (POR) while the subject is performing a task. Depending on the duration of a task and the sampling rate of an eye tracker, excess POR data can be produced. Fixation identification algorithms are employed to group the POR data within a specified neighborhood or velocity. Working on fixations rather than PORs not only decreases the amount of data to be analyzed but also eliminates the noise and saccadic movements. Most of the well-known fixation identification algorithms supposed that the scene viewed by the observer is stationary, however, wearable eye trackers capture dynamic scenes. There is still no commonly accepted method to extract fixation from POR data in dynamic scenes (Munn, Stefano, & Pelz, 2008; Stuart, Galna, Lord, Rochester, & Godfrey, 2014). Although there exist a few methods suggested by some commercial analysis frameworks, since these methods are generally not open source, we could not get detailed information about inner processing.

iii. It is more complex to perform Area of Interest (AOI) analysis in dynamic scenes extracted from the mobile eye-tracking devices compared to static ones. For this

reason, researchers often manually annotate the corresponding area where a subject is looking at. As mentioned above, the amount of data that researchers have to annotate may be largely depending on the duration of the study. Therefore, researchers might spend days, or even months, just for the annotation of gaze data. Also, human-related errors might occur when annotation is performed manually. In addition, because of the hardware or operational constraints, eye-tracking devices can estimate the gaze location with errors. Eye tracker manufactures provide the estimated error that is specific to device in degrees for the visual angle. It is not possible to annotate the area corresponding to eye gaze coordinates manually by taking into account this margin of error, unless the tool in which the researcher makes annotation, calculates the gaze location taking the specified margin of error into account and presents the updated location to the researcher.

iv. There exist studies that made operational assumptions for the gaze and speech relation in a conversation by proposing computational models that simulate the gaze behavior on humanoid robots through head movements alone, or by encoding the presence or absence of human speech rather than language processing. These operational assumptions can be considered as oversimplification compared to the real life settings, and to the extent allowed by technical capabilities, they should be replaced by advanced computational models.

The present study has a two-fold purpose: first, we examine the gaze and speech modalities and their relations in face to face social communication by considering the constraints mentioned above, and secondly, we construct a computational model to predict gaze behavior using high-level speech features. For these purposes we conducted human-to-human experiments in a mock job interview environment where both participants were wearing eye-tracking glasses, and then analyzed the frequency and duration of gaze behavior, speech instances and their relations. In order to overcome the methodological constraints mentioned above, we have developed an open-source framework, namely MAGiC (A Multimodal Framework for Analyzing Gaze in Communication) (Arslan Aydin, Kalkan, & Acarturk, 2018), for analyzing face contact and gaze aversion by incorporating speech. We annotated speech with two schemes, ISO 24617-2 standard for dialogue act annotation and a simple scheme consisting of tags that we identified by considering previous studies examined gaze behaviors in a social context. The reason we create an alternative speech tag set is not proposing a new scheme for discourse annotation. Our aim is to examine, in a sense, the ability of one of a current dialogue act annotation framework, which has major efforts behind, in the computational modeling of gaze behavior by comparing its performance with a simplified speech tag set.

1.2. Research Questions

The present study, in general, aims to investigate how people use face contact and gaze aversion mechanisms in face-to-face conversations to achieve conversational goals and convey their intentions in a social environment, and to find out whether gaze behavior can be predicted by employing speech modality. To this end, we will consider the following questions:

> **RQ1:** What are the underlying features of gaze behavior among humans and what is the relation between gaze and speech to achieve conversational goals in a specified face-to-face interaction environment, namely in a job interview?
>
> **RQ2:** How can we computationally model gaze behavior with the high-level features of speech and what is the appropriateness of employing discourse analysis scheme, namely ISO 24617-2 standard, in a computational model of gaze behavior?

1.3. Organization of the Dissertation

This dissertation is composed of seven chapters in total. The introduction chapter sets the significance, scope and aims of the thesis. Chapter 2 provides a theoretical background for the research questions. The literature review is presented under three main headings, studies in gaze, speech annotation and computational models of face to face interaction. Under the title of gaze studies, the role of gaze functions in social communication and the state of the art developments in eye tracking methodologies are represented. Then, under the title of speech, the frameworks proposed for dialogue-act and rhetorical relation (RR) annotations are reviewed. At the final section of Chapter 2, computational models of face-to-face interaction is summarized. In Chapter 3, we provided information about pilot study conducted between three pairs and we assessed the problems with the experimental design and analysis procedure in order to improve upon the design and analysis. In Chapter 4, we presented an open-source framework, namely MAGiC for analyzing gaze behavior in face-to-face communication by integrating eye-tracking, audio, and video data for investigating gaze behavior, speech analysis, and face tracking, respectively. The experiment which is conducted between 28 pairs with professional interviewers and the results of statistical analysis are presented in Chapter 5. In Chapter 6, the history of neural networks, the components of a basic Convolutional Neural Network (CNN) and two CNN models that we utilize in the present study, VGG and ResNet, which are well known for their high performance are summarized. We also reported the accuracy of developed models. Chapter 7 is the final chapter and a general discussion about the outcomes together with the aims and research questions of the dissertation is given. Contributions and limitations are also presented as well as possible future works for which the experience gained in the process of the present study has paved the way.

CHAPTER 2

LITERATURE REVIEW

The review of related literature includes separate sections for the studies in gaze, speech analysis and computational models. The role of gaze in the social context and dyadic interaction researches utilizing mobile eye tracking are presented in the first section. In the next section, developments in NLP schemes for the annotations of dialogue-act and RR are reviewed. We focus on the history and architecture of Convolutional Neural Network (CNN) in the last section.

2.1. Social Gaze

In our social lives, compared to nonhuman primates, the specialized morphology of the human eyes, which have a sharp contrast between the white sclera and darker pupil, indicates the special role of revealing gaze direction by the sender and, thus, enables those around the sender to acknowledge about the direction of his gaze. (Kobayashi & Kohshima, 1997). We have the ability to make a distinction between directed and averted gaze from a very young age. Farroni, Csibra, Simion and Johnson (2002) stated that even an infant can make such a distinction in the first days of his life. Following the gaze direction enhances cooperation. Moreover, in case of a discrepancy between the deceiver's verbal and gaze clues, children older than 3 years begin to prefer gaze cue in obtaining information from the interlocutor (Freire, Eskritt, & Kang, 2004; Tomasello, Hare, Lehmann, & Call, 2007).

The range of functions that the gaze fulfills in social interaction is extensive. Expressing emotions is one of the well-known function of gaze (Izard, 1991). An individual should perform eye movements in an appropriate way for the aim of conveying emotional states to an addressee successfully (Fukayama, Ohno, Mukawa, Sawaki, & Hagita, 2002). In addition, gaze takes part in regulation of conversation, transmitting the intention, coordination of turn taking, asserting uncertainty or dissatisfaction, regulation of intimacy, and, signaling the dominance and conversational roles (Argyle, Lefebvre, & Cook, 1974; Duncan, 1972; Ho et al., 2015; Kendon, 1967).

Moreover, shared and joint attention requires following the gaze of an interlocutor. In shared attention both individuals are aware of the other's direction of attention, whereas in joint attention, only one of the individuals observes the other's attention. Emery (2000) summarized the role of gaze to differentiate joint and shared attention, see Figure 1. In short, Baron-Cohen (1994) designed a system for modeling the theory of mind in human infants. His system consists of four components: Eye Direction Detector (EDD), Shared Attention Mechanism, Intentionality Detector and Theory of Mind Mechanism. Later on, Perrett and Emery (1994) proposed two additional components to the Baron-Cohen system: Direction and Attention Detector (DAD) and a Mutual Attention Mechanism. Activation of the EDD or DAD components is necessary to initiate joint attention, whereas in shared attention, Shared Attention Mechanism component could be activated when Mutual Attention Mechanism is activated as well as EDD or DAD components. Joint attention and hence the role of following other's gaze also studied in the literature of language learning and observational learning (Dunham, Dunham, & Curwin, 1993; Tomasello & Farrar 1986). Similarly, Otteson and Otteson (1980) revealed that students show a high level of understanding when a teacher makes eye contact with them.

As well as eye contact, gaze aversion functions a crucial role in social interaction as an important non-verbal cue. Gaze aversion is defined as the act of looking away from the interlocutor. There exist cognitive, psychological, sociological and neuropsychological studies conducted on gaze aversion. Hietanen, Leppänen, Peltola, Linna-aho and Ruuhiala (2008) claimed that averted gaze of another person initiates a tendency to avoid, whereas direct gaze would initiate a tendency to approach. In their study, participants viewed pictures of people either directing the gaze towards them or averting the gaze from them. The participants give higher ratings for likeability and attractiveness when the presented picture is combined with direct rather than averted gaze (Mason, Tatkow, & Macrae, 2005; Pfeiffer, Timmermans, Bente, Vogeley, & Schilbach, 2011). Furthermore, Adams and Kleck (2003, 2005) assumed that facial expressions of sadness and fear are associated with the avoidance-motivation, while happy and angry faces are associated with the approach-motivation. Participants recognize happy and angry faces faster when they are demonstrated with a direct gaze rather than averted gaze. On the contrary, sad and fearful faces are recognized faster when they are presented with averted gaze than they are recognized with a direct gaze. In the next sub-sections, the advantages of mobile eye tracking for researches in social gaze along with related studies on this subject are summarized.

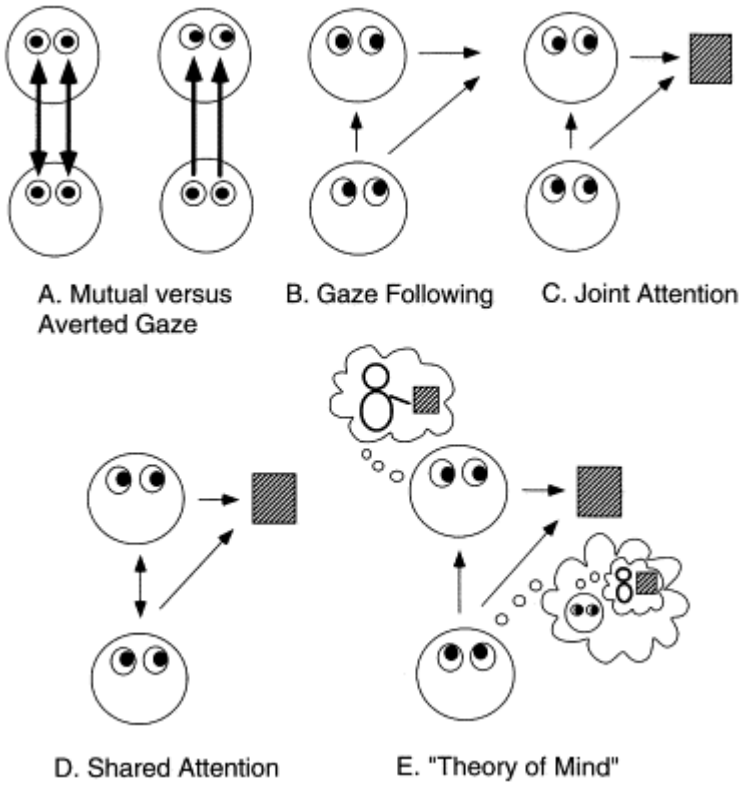


Figure 1: The role of following gaze direction. Gaze direction supplies some clues about the information on others or objects in an external world or enables to learn about the intentional states of other individuals. In Joint Attention (C), unlike the situation in Gaze Following (B), there exists something on which both people concentrate. Shared Attention (D), on the other hand, is a mixture of Gaze Following (B) and Joint Attention (C). In this case, not just one but both of them focus on each other. Lastly, individuals use their higher-level cognitive strategies during the attention process in Theory of Mind (E) (adapted from Emery, 2000).

2.1.1. Mobile Eye Tracking in Dyadic Interaction

Gaze behavior study is not a new topic in the literature. The first studies date back to the 1960s (Klienke, 1986). Research in the field thrived during the 1970s and 1980s with the developments in eye-tracking technology. Psychologists started to investigate the connection between cognitive processes and eye-tracking data with improved eye trackers that became less intrusive and provided better accuracy (Gredeback et al., 2010). The majority of this research have been conducted under highly controlled conditions in which participants were required to sit in front of a monitor and interact with screen-based stimuli. For social interaction, such an experimental design might result in somewhat divergent findings that are distant from real-life situations. Thus, it would be problematic to generalize the findings of this context constructed by an experimenter to real-life contexts (Pfeiffer et al., 2013). This is because static stimuli cannot provide a proper situation to observe the dual function of the human gaze. In a natural social interaction, an individual directs attention on a particular object or situation to receive information, i.e., encoding function of gaze, while communicating to others and revealing information about himself, i.e., a signaling function of gaze

(Risko et al., 2016). Moreover, studies have shown that eyes also transfer information as well as collecting it, for instance, if an individual is informed that his or her eye movements are being observed by others, he or she behaves differently than the case they are not being observed (Myllyneva & Hietanen, 2016).

The advent of mobile eye-tracking offers new opportunities for studies in the real, dynamic world. Utilizing eye-tracking glasses in a real interactive face-to-face communication, allows researchers to examine how gaze information is conveyed between two individuals during a real-time social interaction. Rogers et al. (2018) designed an experiment that involved a face-to-face conversation between participants both wearing Tobii Pro glasses[1], and thus their gaze behavior could simultaneously be recorded. They examined two topics: (i) the personal differences in gaze patterns during dyadic interaction, (ii) the incidence percentages of mutual face gaze and mutual eye contact in a conversation. In line with the Kanan, Bseiso, Ray, Hsiao, and Cottrell (2015), they found some individual differences in the patterns of gaze scanning when looking at the face. The general trends observed were divided into three groups. The first group focused on the mouth, the second group on eyes and the third group spread their gaze on the mouth and eye region consecutively. Moreover, Rogers et al. (2018) found that the duration of mutual face gaze (i.e., when both participants were looking at each other's face at the same time) was shorter up to 1 second than the findings of previous studies (e.g., Binetti, Harrison, Coutrot, Johnston, & Mareschal, 2016), which was 3.3 seconds on average.

Through the usage of a pair of mobile eye trackers, Rogers et al. (2018) also measured the mutual eye contact duration in a dyadic conversation. On average, it lasted about 0.36 seconds and spanned up to 10% of the whole interaction. At the end of the session, the participants were asked to rate the frequency of mutual eye contact that they perceived during a conversation on a 6-point scale; *Never* represented the least and *Very Often* represented the highest frequency. They reported that there was a difference in the frequency of the mutual contact perceived and the measured values. People tend to estimate the frequency of mutual eye contact more than measured value. This failure in participants' estimation rates might be stemmed from limitations on cognitive resources that were allocated to the comprehension of conversation. In fact, participants' estimations were closer to the frequency of mutual face gaze. One might need further studies to understand whether people have the ability to differentiate between the perceived mutual face gaze and eye contact. In line with this information, we focus on the face gaze (viz. face contact) instead of eye contact in the present study.

2.1.2. The Role of Gaze in Conversation

As Kendrick and Holler (2017) stated, the practical nature of the human gaze is perhaps most apparent during a face-to-face conversation. The direction of the eye

[1] Tobii Pro Glasses 2: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

gaze undertakes an important function when initiating social interaction and while maintaining it (Gorman & Hall, 1964).

The landmark study by Kendon (1967) examined the role of gaze in a face-to-face conversation. He summarized the difference between a listener and a speaker regarding the gaze behavior in terms of frequency and duration. According to this approach, listeners tend to look at the interlocutor more frequently compared to speakers. Moreover, gaze contact of listeners took longer than the gaze contact of speakers. He also noted that speakers generally tend to look at their companion when they were close to finishing their speech. On the other hand, they averted gaze at the beginning of the speech. In agreement with the studies initiated by Kendon (1967); Vertegaal, Slagter, Van Der Veer, & Nijholt (2001) showed that by observing the eye movements of an individual in a face-to-face conversation, it is 88% likely to infer whether he or she is a speaker or a listener. Even though the context of a conversation drives the behavior of gaze, similar results are generally reported in the studies that examined the functions of gaze in turn-taking and regulation (Argyle et al., 1974; Goodwin, 1980; Kendon, 1967). However, as Rossano, Brown, and Levinson, (2009) pointed out, many of those studies have been carried out with the participants from western societies speaking English. Therefore, even though they did not express it clearly, those studies implicitly assumed that gaze behavior in a face-to-face conversation is independent of culture and language. Rossona et al. (2009) found some similarities as well as differences in the gaze behaviors of participants from different cultural backgrounds. For instance, the primary factor that drives the gaze behavior during conversations in Italian is the sequence of talk, instead of turn-taking. Similar to the gaze behavior in turn-taking, people tend to signal the start and the end of a sequence.

In another study, i.e., Bavelas et al. (2002), coordination role of gaze during a face-to-face conversation was examined. One of the participants told a story as the speaker and the other one listened to him. In line with the previous studies, the authors reported that listeners looked at the speakers more often than speakers did. The most remarkable finding of the study was that whenever the speaker asked for a response, first, a mutual gaze contact was established and then, during mutual face contact, listener gave a feedback with a verbal/vocal expression such as, “yes”, “mhm”, “okay” or non-verbal signals like head gestures, and, after that in a short period of time, speaker averted his gaze from the listener and continued the speech. Thus, it confirmed that gaze coordinated the speech and integrated into it during a conversation.

Srinivasan, Bethel, and Murphy (2014) summarized the literature to automate gaze behavior based on the structure of sentence and time intervals between certain structures. They argued that it might be possible to generate good enough autonomous head-gaze acts without semantic understanding. In order to generate autonomous head-gaze motion, they proposed analyzing structures for sentences and computing time intervals between certain structures. For instance, if it is the beginning of a new turn, the new speaker will say the very first word which can be easily detected in real-time. At the same time, the speaker would most probably avert his gaze from an addressee

to emphasize that he or she is about to speak. In another case, if it is the middle of a turn for the speaker, he would probably direct his gaze towards an addressee. Based on the previous studies, they proposed a behavioral framework in a particular sentence structure and with a corresponding social gaze act, four of these behaviors are as follows:

• **Start of Turn:** When the very first word of a turn is presented, speakers generally avert the gaze.

• **Middle of Turn:** It can be categorized under two types: (i) speakers tend to avert gaze more than a chance level, right after punctuation marks located between sentences. (ii) after punctuation marks, speakers fixate on an interlocutor with a probability of 70%, when more than half of the words, around 75% of them were presented.

• **End of Turn:** When the last sentence before the carriage return ends, speakers tend to fixate on an interlocutor

• **Robot manifesting interaction:** Robots fixate on an object, 800 ms to 1 s before their names are uttered.

In the proposed study, we will adopt an approach similar to the ones in Srinivasan et al. (2014) studies, but we will perform experiments with Turkish speaking participants. For this, one needs to discover behaviors and related sentence structures for Turkish dialogues. Srinivasan et al. (2014) research is based on corpus in English in which the theme of a dialogue or a simple sentence is generally given at the beginning and the rheme is generally presented towards the end of a dialog and sometimes as a simple sentence. However, the Turkish language is different than English as they belong to different language families. Therefore, at first, we will conduct human-human interaction experiments to discover such relations between sentence-structure and gaze behaviors in Turkish.

So far, the advantages that mobile eye-tracking devices provide in researches of social gaze, particularly investigation of gaze behaviors in face-to-face communication are summarized. In the present study, we examined speech-driven social gaze, therefore, we discussed speech in terms of dialogue and rhetorical relation annotations, in the following section.

2.2. The Annotation of Discourse Relation

Natural Language Processing (NLP) dates back to the 1950s. A landmark study in this domain was Alan Turing's pioneering works. Turing proposed an imitation game to test the ability of computers to exhibit an indistinguishable intelligent behavior of a human, in a real-time written conversation. In that study, conversation alone was assumed to be a sufficient tool for impersonating a human (Turing, 1950). Later, Chomsky proposed the existence of an innate language faculty which makes it easier

for children to learn how to speak, i.e., the theory of universal grammar (Lees & Chomsky 1957; Peter & Chomsky 1968). On the other hand, starting in the late 1980s, interest in NLP studies showed an increase with the introduction of ML (Machine Learning) algorithms. Until the late 1980s, there was relatively less research in the field of Machine Translation and NLP (Natural Language Processing). Some significant developments during this period were Augmented Transition Networks which is a sort of syntax processor that also provided a formalism to express domain-specific knowledge, Case Grammar which contributed especially to the translation of prepositions problems of Machine Translation and also to the semantic information with a little processing effort, and lastly developments in semantic research, e.g., Conceptual Dependency (Fillmore, 1968; Schank & Tesler, 1969; Woods, 1970).

The first instances of ML algorithms were based on hardly coded if-then rules similar to the hand-written rules that had been proposed up to the 1980s. However, later on, instead of hand-coding a large set of rules, researches focused on probabilistic models that automatically learn rules by analyzing real-world data. In parallel to the developments in approaches to ML technology, the sub-fields of NLP such as machine translation (Meyer & Popescu-Belis, 2012; Popescu-Belis, 2016), automated question answering (Sharp et al., 2015) and text mining along with improvements in their real-world applications like sentiment analysis, automatic text summarization, topic extraction, relationship extraction and so on, rise rapidly. Besides sentence-level analyses, in recent years we have also seen an increase in the attention paid to the discourse processing, especially to the field of discourse relation annotation. Collections of large-scale corpus annotated according to various schemes have fastened the progress in the field of discourse relation annotation[2]. In particular, PDTB (Prasad et al., 2008), Rhetorical Structure Theory Discourse Treebank (RST-DT; Carlson, Marcu, & Okurowski, 2001) and DialogBank (Bunt, Volha, Andrei, Alex, & Kars, 2018) include texts in English. There are also numerous resources developed for other languages (see, for example, Zeyrek, Demirşahin, & Bozşahin, 2018; Zeyrek et al., 2019, Oza, Prasad, Kolachina, Sharma, & Joshi, 2009).

There is still no agreement on a particular scheme of discourse relation annotation. In any annotation scheme, there are two subjects to be identified: the annotation unit and the labels. The annotation unit can be determined depending on the type of the word or sound, phrase, clause etc. Labels can vary in dimensions and the number of layers from scheme to scheme. The definitions of labels, annotation units and the features associated with these units should be as clear and operational as possible so that the labels assigned to the same piece of discourse do not change from annotator to annotator (Ide, 2017). In addition, the quality of these operational definitions will affect the success level of the model during the automatic annotation of discourse relations.

[2] Note that in the present study, terms of scheme, framework and taxonomy are used interchangeably.

The existing frameworks identify relational differences based on similar fundamental concepts. They differ in the way that they specify the relational arguments. For instance, discourse relations might either ground in lexical items, drive-by semantics, or consist of both intentional and semantic relations. Besides, discourse structure resulting from defining the relational arguments might be either tree or non-tree like (Demberg, Scholman, & Asr, 2019). Two of the most well-known frameworks are RST-DT and PDTB, on which many researchers have studied.

RST-DT is essentially an RST implementation like RST treebank. RST was proposed by Mann and Thompson (1988) while they were working on computer-based text generation. RST-DT follows the RST style annotation, however, it differs from other implementations in terms of the way the segments are specified and the whole set of labels. There are two main features of this framework. First, relational arguments are determined in such a way that no part of the text is left out and the resulting form of the discourse should be in the tree structure. Second, at least one of the relational arguments must be the key element, i.e., nucleus. If both arguments are equally important according to the type of the relation, as in the case of contrast relation, this relation is made up of two nuclei, otherwise one of the arguments becomes the nucleus and the other becomes the satellite.

Before the annotation process, the segments have to be extracted as it is the case in all other frameworks. In RST-DT, segmentation refers to the function of splitting text into a sequence of elementary discourse units (EDUs). EDUs are clause like units that serve as basic elements for discourse parsing in RST. Then, to make a label assignment, the nucleus is determined simultaneously with the label assignment. The determination of the nucleus is based on the intent of the sender. In order to understand this intention, it is often necessary to comprehend the context of the text. Discourse relations are established as recursive with a bottom-up approach, starting from EDUs. Consequently, discourse relations are in the structure of a hierarchical tree (Carlson et al., 2001).

The largest manually annotated discourse relation corpus is the PDTB 2.0 corpus (Prasad et al., 2007) created using the PDTB framework. Recently, PDTB 3.0 was introduced as a more operational and extended version of PDTB (PDTB 3.0, Prasad, Webber, & Lee, 2018). In contrast to RST-DT, PDTB makes no promise to the type of high-level structure that is built from the low-level annotation of relations. Furthermore, PDTB adopted a lexically-based approach for representing the discourse relations. Discourse annotated with the PDTB framework has either implicit or explicit relations. There is an explicit relation when there exist lexical items such as conjunctions inside a discourse, otherwise, the relation is implicit. If discourse connective does not explicitly exist, annotator is expected to enter the most appropriate connective as an implicit discourse connective. PDTB framework allows 3 specific labels as an implicit connective: *AltLex, EntRel* and *NoRel*. The PDTB annotator generally assumed to annotate each successive segment, while, not all successive segments need to be related. In such cases, the NoRel label is assigned as a connective. If the relation with the previous one is only entity-based, then EntRel is assigned

whereas if adding an explicit connective will result in redundancy because of the sentence structure, then AltLex is assigned.

These differences make it difficult for researchers to work on corpus annotated with different schemes. It also limits the number of available inputs provided for training the model during automatic labeling as the granularity levels and set of labels change scheme to scheme. It would be a hassle to find the corresponding labels from one scheme to another. Studies on the problem of mapping between discourse relations have gained interest in the last decade (Zitoune, & Taboada, 2015; Sanders et al., 2018). Bunt and Prasad (2016) proposed an ISO standard for the annotation of semantic relations in a discourse, namely ISO DR-Core, and they defined a mapping between ISO DR-Core and among most of the well-known taxonomies such as RST, RST Treebank, SDRT, DISCOR, ANNODIS and PDTB. In the present study, we employed ISO 24617-2 for dialogue-act annotation and ISO DR-Core for RR annotation.

2.2.1. Dialogue Act Annotation

The dialogue act is the act that the speaker is performing during a dialogue. In a simplified sense, it is a speech act used in a conversation. A dialogue act has a particular semantic content that specifies the objects, events and their relations. Furthermore, it maintains a communicative function intended to change the state of mind of an addressee by means of its semantic content. In practice, dialogue act annotation generally depends on the communicative function.

In the 1990s, a variety of domain-specific dialogue act annotation schemes such as TRAINS and Verbmobil were proposed (Allen & Core, 1997; Alexandersson et al., 1998). Although there were some common communicative functions in those schemes, there were also inconsistencies between. In order to overcome this difficulty, in the late 1990s, a domain-independent and multi-layered scheme, DAMSL (Dialogue Act Markup using Several Layers) were proposed (Allen & Core, 1997). Subsequently, many studies were carried out until the establishment of ISO standard for dialogue act annotation. Especially, two of them played a major role in the idea of building a standard framework. First, Bunt developed the DIT++ scheme (Bunt, 2006; Bunt, 2009) by combining the studies on the developed extensions of DAMSL and his previous work DIT (Bunt, 1994). DIT++ is multidimensional, and it is mutually consistent with the referenced schemes according to the communicative functions and dimensions (Bunt, 2006; Bunt, 2009). The second attempt was the LIRICS project, which identified data categories for manual annotation using some of the communicative functions proposed in the DIT++ scheme (LIRICS, 2006a, 2006b). As these studies were mature enough, efforts were made to establish an ISO standard for dialogue act annotation. Eventually, ISO standard 24617-2 “Semantic annotation framework (SemAF) – Part 2: Dialogue acts” was developed (ISO DIS 24617-2, 2010).

A turn represents the duration that the speaker is talking and it is an important organizational tool in spoken discourse. It is necessary to participate effectively in conversation without interrupting the person speaking. Turns can be rather long and complex, in this case, they cannot be taken as units to determine communicative functions. They need to be cut into smaller parts called functional segments. Functional segments supply information to determine both the semantic content, namely "dimensions", and communicative functions of a dialogue act.

In case an addressee does not understand an entire functional segment or just a chunk of it such as a single word or a sequence of words, he or she may want to verify the information when it is his or her turn by saying something related to the previous functional segment. ISO 24617-2 annotation scheme required to specify such relations as feedback dependence between the current dialogue act and the previous functional segment. In general, feedback dependences are involved with the perception, comprehension, and assessment of what was previously said. Therefore, it may be related to the previous dialogue act, as well as a previous functional segment. Moreover, most dialog acts are responsive in character and rely on one or more dialog acts previously performed in the dialog. This refers, for instance, to answers whose content depends fundamentally on the question. Similarly, returning to greeting, to self-introduction and to goodbye, accepting the suggestion, the offer and the request, agreement or disagreement to information and, confirming or disconfirming a yes-no-question are also responsive in nature and require the specification of functional dependence between the related dialogue acts. Furthermore, in Dialogue-act annotation, distinct roles are assigned to participants: (i) "sender" or "speaker" is the one whose communicative behavior will be interpreted by examining the purpose of his utterance rather than focusing on what he explicitly says, (ii) "addressee" or "recipient" is the participant whose mental state is tried to be influenced by a sender via communicative functions.

Dialogue act annotation can be done in three main steps: (i) the dialogue is the initial source and it is divided into two or more functional segments, (ii) one or more dialogue acts are associated with each functional segment, (iii) annotation components are assigned to dialogue acts, see Table 1 for the components.

Table 1: Annotation components. One and only one dimension, communicative function, sender and addressee should be attached to a dialogue act. On the other hand, there might be zero, one or more qualifiers, rhetorical relation, participant other than sender and addressee, and dependence relation. * Relation is between dialogue acts. ** Relation is between either dialogue acts or a dialogue act and a functional segment

Component	Number
Dimension	1..1
Communicative Function	1..1
Qualifier	0..N
Rhetorical Relation*	0..N
Participant	
sender	1..1
addressee	1..1
other	0..N
Dependence Relation	
feedback**	0..N
functional*	0..N

In successful communication, the listener understands what the speaker says, the way the speaker desires. In doing so, the listener takes into account the basic characteristics of the speaker's utterances, as well as the motivation behind the initiation and the history of the dialogue, and even his/her assumptions about the opinions and goals of the interlocutor. We cannot derive the communicative function of a dialogue act by considering only the surface form of utterances since the same utterance forms can have different meanings in different conversations between different people. The form-based dialogue act annotation is applied mostly by automatic annotation systems. Intention-based approaches, however, is more applicable for human annotators, as they are experienced in understanding the intention of others.

A general-purpose dialogue act annotation framework should provide communicative functions which require deep semantic knowledge that can be easily understood by humans and should support a form-based approach in order to enable automatic annotation. ISO standard 24617-2 introduced qualifiers and hierarchy of communicative functions to handle such requirements (Bunt, 2019). To further specialize the communicative function based on the speaker's presumed intention, this qualifier or a lower-level communicative function can be assigned. The set of communicative functions is illustrated in Figure 2 in a hierarchical tree structure, see Bunt (2012) for detailed information on each function.

Almost all dialogue act annotation frameworks neglect some minor nuances that the speaker intended to give. For instance, the communicative function of *Inform* would be assigned when the speaker is giving information. However, that annotation could not reflect whether the speaker is sure of the information she/he provided. The speaker may want to emphasize that he/she is not sure or very confident. Similarly, when the speaker accepts an offer, he may wish to emphasize that it makes him happy or he

conditionally accepts it. ISO standard 24617-2 recommended 3 qualifiers, see Table 2.

Table 2: Qualifier attributes, set of values and default values. * ISO standard 24617-2 does not provide a set of sentiment qualifiers, instead, the annotator is free to use whatever elements they deem appropriate with regard to the dialogue context.

Attribute	Values	Default value
Certainty	Uncertain, certain, quite certain	Certain
Conditionality	Conditional, unconditional	Unconditional
Sentiment*	Happiness, surprise, anger, sadness..	Empty

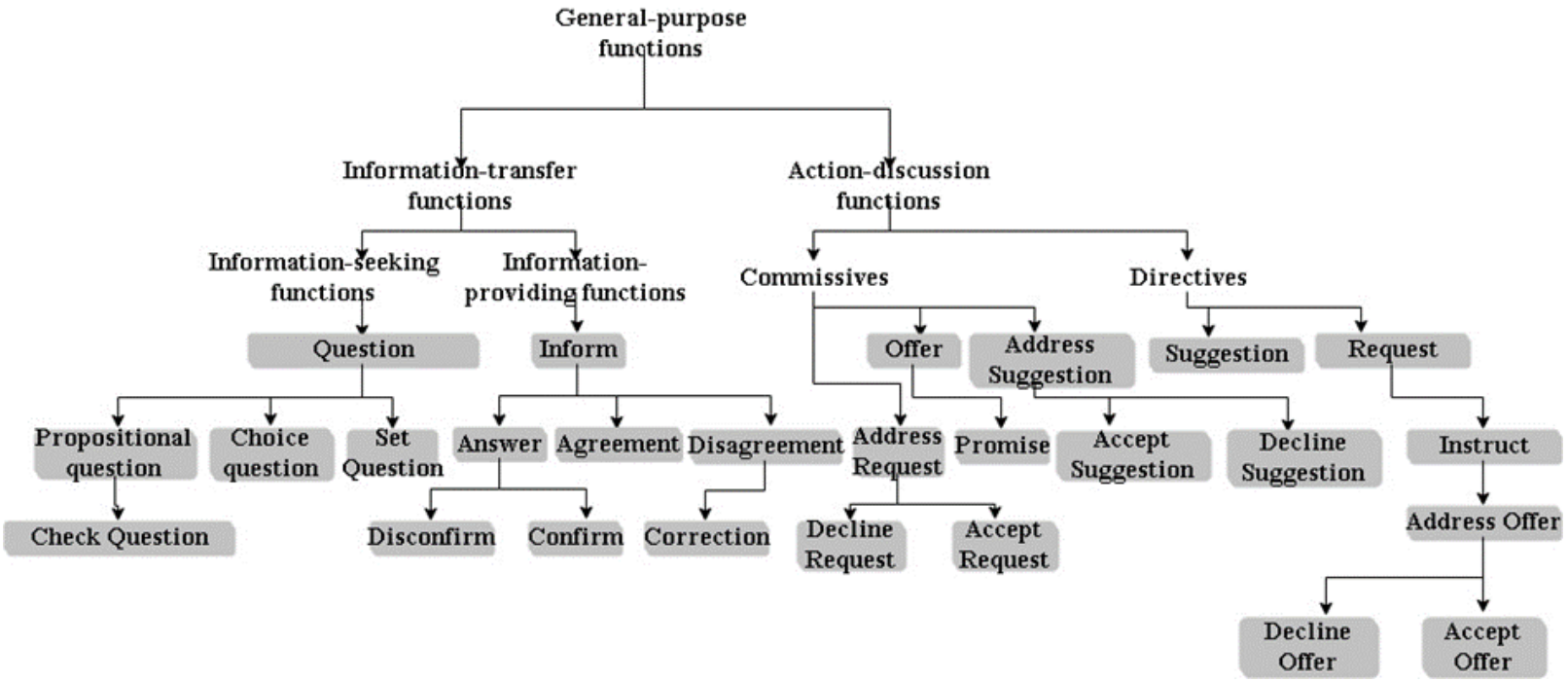


Figure 2: Hierarchy of general purpose functions. General purpose functions are presented with a gray background

Functional segments in a dialogue can comprise a single word or a sequence of words, or can be broken up into multiple of those. Then proper dialogue acts are assigned to each one. However, this does not require that the utterances of each dialogue act should be different from each other. Therefore, the same utterances could be related to more than a single communicative function. For instance, the speaker might repeat the utterances of a question as a response. That way, one conveys simultaneously that he or she has taken the turn and understands the question but needs some time to think about the answer. For such cases, ISO standard 24617-2 has adopted a multidimensional approach. In order to determine the "core dimensions" that can be used in a general purpose framework, Petukhova and Bunt (2012) examined related previous studies and set some criteria to determine nine basic dimensions. The functions listed in Figure 2 are general purpose functions and can be applied to any of these nine dimensions. The remaining functions, however, are dimension specific. These dimensions and a set of communicative functions that can be assigned under related dimension are presented in Table 3, see Bunt (2012) for detailed information and examples.

Dialogue Act Markup Language (DiAML), being a part of the ISO standard 24617-2, follows the ISO linguistic annotation framework and makes a distinction between representation and annotation. The term "annotation" indicates the linguistic information and is applied to a portion of dialogue regardless of the way it is represented. On the contrary, the word "representation" refers to the manner in which the information is presented. DiAML XML annotations can be created with the ANVIL annotation software and are ideal for computational processing. Nonetheless, for human inspection and alteration, other formats such as DiAML-TabSW and DiAML-MultiTab (Bunt, 2019) are more convenient. In the present study we use DiAML-MultiTab format for annotation.

Table 3: Dimensions and communicative functions defined in ISO 24617-2.

Dimension		Communicative Functions
Task:	Category of dialogue acts that helps to carry out the tasks or activities that inspire the dialogue	General Purpose Functions (GPFs)
Auto-Feedback:	Category of dialogue acts that take place, in which the sender addresses his processing of past dialogue.	AutoPositive, AutoNegative, GPFs
Allo-Feedback:	Category of dialogue acts that take place, in which the sender argues about the addressee's processing of past dialogue.	AlloPositive, AlloNegative, FeedbackElicitation, GPFs
Turn Management:	Category of dialogue acts that are intended to coordinate the role of the speaker	TurnAccept, TurnAssign, TurnGrab, TurnKeep, TurnRelease, TurnTake, GPFs
Time Management:	Category of dialogue acts that deal with the allocation of time during the speech	Stalling, Pausing, GPFs
Own Communication Management:	Category of dialogue acts where in the ongoing turn the speaker alters his own speech	SelfCorrection, SelfError, Retraction, GPFs
Partner Communication Management:	Category of dialogue acts where in the ongoing turn the speaker alters the speech of the previous speaker	Completion, CorrectMisspeaking, GPFs
Discourse Structuring:	Category of dialogue acts that organize the dialogue directly	InteractionStructuring, Opening, GPFs
Social Obligations Management:	Category of dialogue acts carried out to meet social responsibilities such as welcoming, thanking and apologizing	InitialGreeting, ReturnGreeting, InitialSelfIntroduction, ReturnSelfIntroduction, Apology, AcceptApology, Thanking, AcceptThanking, InitialGoodbye, ReturnGoodbye, GPF

ISO standard 24617-2 also supports the annotation of rhetorical relations (RR). Although this standard does not provide any specific set for RR, it suggests a specific standard, namely ISO 24617-8. In the present study, we adopted ISO 24617-8 or better known as ISO DR-Core for RR annotation as presented in the following section.

2.2.2. Rhetorical Relation Annotation

For understanding a discourse, it is not enough to understand individual sentences or clauses. The relationship between individual semantic units is called RR (also called "discourse relations" or "coherence relations") and it allows us to understand the discourse as a whole. Although semantic units associated with RRs such as cause, result, condition, dialogue act, usually correspond to a sentence; they may be even longer, like as paragraphs, or even shorter, like dialogue segments. Parallel to the increase in NLP studies in recent years, more studies to create resources annotated with RR are being carried out in order to meet the needs and demands in these areas. ISO standard 24617-2 has been developed in order to provide the theoretical and empirical background for semantic annotation of discourse relations by examining those studies in terms of their commonalities and differences (Prasad & Bunt, 2015; Bunt & Prasad, 2016).

Two of the most well-known frameworks in this field are; PDTB (Prasad et al., 2008, 2018) and RST Bank (Carlson et al., 2001) based on RST (Mann & Thompson, 1988), As we mentioned above in the dialogue act section, Prasad and Bunt (2015) summarized one of the most fundamental issues where frameworks differ from each other is the representation of discourse structure. For instance, RST based models aim to build a tree structure containing all discourse as a result of the annotation process. The tree structure adopted in these models varies: Nodes of a tree might have single or multiple-parents, there might be crossing edges (edge from vertex v to vertex u which is not an ancestor or a descendant of v) or graph might be acyclic (having no graph cycles). PDTB framework, however, does not force to build a tree-like structure at the end of an annotation. ISO DR-Core aims to provide interoperability with existing frameworks, it has adopted the principle of low-level annotation. Thus, if it is desired to be compatible with a framework that requires high-level annotation, such as a tree structure, annotated relations can be further processed to provide this structure. Another issue that differs between frameworks is the intention or information based definition of RRs. RST supports intention-based relations, while PDTB supports information-based ones. In many cases, the relation from one approach to another can be mapped. ISO DR-Core has adopted the information-based approach.

One or both two of the RR arguments might have an implicit belief beyond the semantic content. For instance, in the following example (1), the second sentence gives information about the act of offering itself, instead of information about the offer's content.

Do you want to drink coffee? Because you look sleepy. (1)

This distinction is referred to as the semantic-pragmatic distinction in the literature (Van Dijk, 1979; Miltsakaki, Robaldo, Lee, & Joshi, 2008). ISO DR-Core supports the semantic-pragmatic distinction, but not on the basis of relation, but in the sense of the role that the arguments of the relations take. Furthermore, the ISO standard imposes restrictions not on the syntactical form, but the semantic character of arguments. That is, an argument of discourse relation must imply some sort of abstract object. Therefore, non-clausal phrases, as well as clauses might be the arguments of a discourse relation. Lastly, regarding adjacency, some frameworks like RST require the corresponding arguments to be carried out with adjacent textual utterance, while others like PDTB only impose that limitation on implicit relations. In this respect, the ISO standard is noncommittal and does not impose any limitations on the context or adjacency of the arguments (Prasad & Bunt, 2015; Bunt & Prasad, 2016).

Almost all existing frameworks reflect the symmetrical and asymmetrical relations, that is to say, in the case of the relation REL and its arguments A and B, the discourse relation will be symmetric if (REL, A, B) substitutes (REL, B, A), and vice versa. For instance, the discourse relation of *Similarity* is symmetrical while the discourse relation of *Exemplification* is asymmetrical. The list of relations with their definitions and roles of the arguments if the relation is asymmetric is presented in the list below, where the first and the second arguments of discourse are represented by Arg1 and Arg2 respectively. For detailed information please see ISO (2016) and Bunt and Prasad (2016).

Cause: Arg1 is used for the interpretation of Arg2. It is an asymmetric relation with the roles of Reason and Result.

Condition: Arg1 is an unrealized condition that brings Arg2, if it is realized. It is an asymmetric relation with the roles of Antecedent and Consequent.

Negative Condition: Arg1 is an unrealized condition that brings Arg2, if it is not realized. It is an asymmetric relation with the roles of Negated-Antecedent and Consequent.

Purpose: Arg1 is used to let Arg2 occur. It is an asymmetric relation with the roles of Goal and Enablement.

Manner: Arg1 discusses how Arg2 happens. It is an asymmetric relation with the roles of Means and Achievement.

Concession: Arg2 cancels or refuses the anticipated causal relation between Arg1 and Arg2. It is an asymmetric relation with the roles of Expectation-raiser and Expectation-denier.

Exception: Arg1 refers to a number of circumstances where the status mentioned is present, whereas Arg2 refers to one or more cases in which it is not addressed. It is an asymmetric relation with the roles of Regular and Exclusion.

Substitution: Arg2 is the preferred or chosen one where alternatives are Arg1 and Arg2. It is an asymmetric relation with the roles of Disfavored-alternative and Favored-alternative.

Exemplification: A variety of circumstances are listed in Arg1 and Arg2 is a component of that set. It is an asymmetric relation with the roles of Set and Instance.

Elaboration: Arg1 and Arg2 represent the same situation but Arg2 provides more information. It is an asymmetric relation with the roles of Broad and Specific.

Asynchrony: Arg1 is before Arg2 in the time domain. It is an asymmetric relation with the roles of Before and After.

Expansion: Arg2 provides additional definitions of a certain entity/entities in Arg1. It is an asymmetric relation with the roles of Foreground and Entity-description.

Functional Dependence: In case, Arg1 is a responsive dialogue-act, the response to Arg1, i.e. Arg2 will functionally depend on Arg1. It is an asymmetric relation with the roles of Antecedent-act and Dependent-act.

Feedback Dependence: Arg2 is a dialogue act that produces information on the status or assessment of one of the dialog participants of Arg1's. It is an asymmetric relation with the roles of Feedback-scope and Feedback-act.

Contrast: This relation indicates the differences between Arg1 and Arg2, as a whole or in the context of a common entity they are referring to. It is a symmetric relation.

Similarity: This relation indicates the similarities between Arg1 and Arg2, as a whole or in the context of a common entity they are referring to. It is a symmetric relation.

Conjunction: Arg1 and Arg2 have the same relation with some other circumstances elicited in the discourse. This relation indicates that they either do the same thing or do it together with respect to these circumstances. It is a symmetric relation.

Disjunction: In case Arg1 and Arg2 are alternatives, this relation indicates that at least one of the arguments is carried out. It is a symmetric relation.

Restatement: Although Arg1 and Arg2 are the same states, they are defined from different perspectives. It is a symmetric relation.

Synchrony: This relation indicates that there is a certain degree of time overlap between Arg1 and Arg2. It is a symmetric relation.

Up to this point, we have summarized the studies in the literature on social gaze and discourse annotation. In the present study, we investigate the speech-driven gaze in

accordance with the multimodal nature of face-to-face interaction. The following section represents the computational models of verbal and nonverbal behaviors from the perspective of the multimodal approach.

2.3. Computational Models of Face-to-Face Interaction

Face-to-face communication involves some sort of harmony in which partners continuously adjust their behaviors according to verbal and non-verbal signals. Although interpersonal behaviors exhibited by interacts have long been studied in the literature, with the developments in the machine learning, signal processing, and pattern recognition, researchers get the opportunity to use these techniques for analyzing, recognizing and predicting individual's behavior during social interaction. This research direction has many practical applications. For instance, improvements in recognizing human behaviors would have impacts in many contexts including human interaction, medicine (Beck, Daughtridge, & Sloane, 2002), education (Skinner & Belmont, 1993), marketing and services (Gabbott & Hogg, 2000; Sundaram & Webster, 2000). Moreover, studies in human-computer interaction (Pantic, Pentland, Nijholt, & Huang, 2007), affective computing (Picard, 1999) and human-robot interaction (Fong, Nourbakhsh, & Dautenhahn, 2003) would also benefit providing a natural way to communicate with virtual agents and robots. Even, the related studies provide information for the diagnosis of autism spectrum disorders (Wall, Kosmicki, Deluca, Harstad, & Fusaro, 2012).

The multimodal nature of human communication makes it inherently challenging to identify underlying mechanisms of an individual's behaviors, clearly. Studies on understanding multi-modal behaviors differ in their approach to addressing the issue. According to an effective approach put forward by Ekman and Davidson (1994); and some later studies by other researchers (e.g., Jaimes & Sebe, 2007), it is possible to interpret human behaviors in the light of emotion experience. A similar line of approach is proposed to interpret human behaviors in the context of social signals (Vinciarelli, Pantic, & Bourlard, 2009). In this approach, automatic communication analysis uses social signal data to predict social emotions (e.g., happiness, anger), social activities (e.g., turn-taking and backchannel) and social relations (e.g., roles). In order to address these problems, various computer models have been proposed. The influence model which is proposed to model the interaction between individuals in a communication environment is one of them. This computer model is developed based on a term of influence in statistical physics and it aims to prevent the high parameter requirement of models such as Hidden Markov Models (HMMs) (Basu et al., 2001; Choudhury & Pentland, 2004). In another model, Otsuka, Sawada, and Yamato (2007) proposed to use Dynamic Bayesian Network (DBN) for modeling turn-taking mechanisms in communication. In this 3-layered method, the first layer is based on the external observation and the 2nd and 3rd layers are based on the estimation. First, speech and head movement data are taken, gaze patterns are predicted in the next layer, and in the final layer, the regime of the conversation is estimated. In the model they proposed to distinguish laughter from speech, they showed that using audio and visual modalities together presents better results than using speech. In this model, they

used AdaBoost for feature selection and neural network for classification. In addition, ANNA (Artificial Neural Network Assistant) (Fragopanagos & Taylor, 2005) and RNN (Recurrent Neural Network) (Karpouzis et al., 2007) are proposed to predict social emotion by using audio and visual data in a multimodal manner. In the present study, as described in the sixth chapter, we used state of the art network which is a particular type of Deep Neural Network known as CNN.

The multi-modal characteristics of human communication can be modeled by multi-modal machine learning that takes and processes information from various modalities. As Baltrusaitis, Ahuja, & Morency (2019) summarizes, multi-modal machine learning studies present a number of challenges for researchers. First of all, the heterogeneous data in multi-modal learning should be represented and summarized by highlighting the complementary context while avoiding redundancy. For example, the language is represented by symbols while the audio is indicated by signals and videos are composed of frames. Secondly, the way of mapping from one modality to another should be identified clearly. This is not only due to the heterogeneous nature of multi-modal data, but also it is the result of the open-ended and subjective interpretation of relations between modalities. In the third place, it is necessary to analyze and align the relations between modalities. For example, to align the steps of a recipe by watching a cooking video, we need to look at the interrelationships of different models and their interdependence, even if there is a long range between them. Next, information obtained from different modalities should be joined for prediction by considering their various predictive power and noise topology, as well as handling the possibly missing data. Lastly, it may be important how the information learned through one modality can be transferred to a computational model trained with another modality. This may be problematic especially when one of the modalities has a limited resource. We summarized the details of input features coming from speech and gaze modalities, their representations and the way we align them as a time series signal in the sixth chapter.

CHAPTER 3

ANALYSIS OF GAZE AVERSION AND SPEECH IN A FACE-TO-FACE INTERACTION: A PILOT STUDY

This study was conducted for improving the experimental design and data analysis. In this chapter, we report the pilot study and the experience gained through it. This chapter outlines, firstly, participants, apparatus and the experimental design employed during the study. Thereafter, the procedure followed during the analysis was introduced along with the results of the analysis. The analyses involved synchronization of multimodal data including video recording data for face tracking, gaze data from the eye trackers, and the audio data for speech segmentation. Lastly, we assessed the problems with the experimental design and analysis procedure in order to improve upon the design and analysis of the full-scale experiment.

3.1. Materials and Design

3.1.1. Participants

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. The participants were assigned the role of either an interviewer or an interviewee and the roles were distributed randomly. All the participants were right-handed, native Turkish speakers and had a normal or corrected-to-normal vision.

3.1.2. Apparatus

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. Each participant was positioned exactly one meter away from a wall. Then, we asked them to follow the IR (infrared) marker while wearing Tobii glasses. The IR marker calibration process was repeated until 80% accuracy is achieved.

3.1.3. Procedure

At the beginning of the experiment, participants were informed about the task. We asked an interviewee to think about a position that he is interested in, so as to motivate him for the interview. Eight common job interview questions, adopted from Villani, Repetto, Cipresso, & Riva (2012), were translated into Turkish and handed to an interviewer on a sheet of paper, (questions are listed in Appendix A). The interviewer was instructed to ask given questions, and also to evaluate the interviewee for each question right after the response, by using paper and pencil. The evaluation criterions are given in Appendix **B**. Ratings were on a scale of 1 to 7, where 7 was the highest score.

After calibration, the participants were seated on the opposite sides of a table, approximately 100 cm away from each other. The experimental protocol is adopted from the Andrist, Mutlu, and Gleicher (2013) study, and it is illustrated in Figure 3. Lastly, a beeping sound was generated to indicate the beginning of a session. The participants were left alone in the room throughout the experiment.

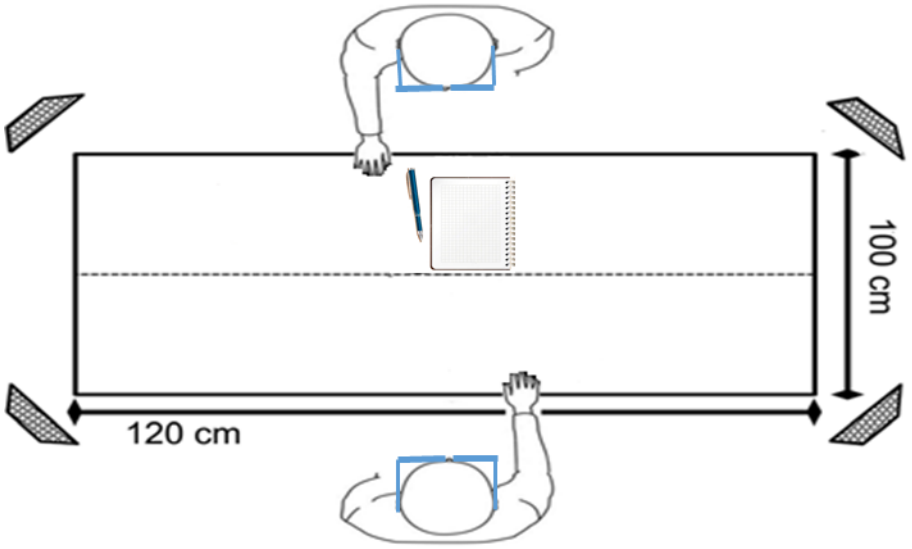


Figure 3: Schematics of the experimental setup

3.2. Data and Analysis

Data analysis consists of three main steps. In the first one, we extracted gaze aversions for each participant. We used OpenCV-3.0[3] (Open Source Computer Vision Library) libraries to detect and track faces in each video frame. As the next step, we analyzed audio data to

[3] OpenCV (Open Source Computer Vision Library) is an open-source computer vision and machine learning software library. The official web-site is: http://opencv.org/

recognize speakers and to segment the audio file into smaller chunks including sub-words and pauses by using CMU Sphinx4[4] libraries. We, then, manually annotated each speech segment using a predefined list of speech instances, hereinafter referred to as speech-acts or speech-tag set. In the final step, we synchronized gaze aversion data with speech annotations and performed statistical analysis on it.

3.2.1. Speech Analysis

Audio data were extracted from the video files. Since CMU Sphinx4 requires a 16 kHz, 16 bit, mono and little-endian audio format, we converted audio data into a supported format for CMU Sphinx4 input, command is given below.

```
ffmpeg -i input.wav  -acodec  pcm_s16le -ac 1 -ar 16000 output.wav        (2)
```

3.2.1.1. Speaker Recognition and Speech Segmentation

The CMU Sphinx4 libraries enabled us to obtain speech segments at millisecond precision. In order to store the starting time and duration of speech segments, we forked open-source Sphinx4 repository and then, implemented corresponding requirements.

As a result of pair recordings, we ended up with two audio files for each session, one was recorded by the interviewer's glass and the other from the interviewee's. Both recordings were processed in the same environment. We preferred to annotate the segments extracted from interviewers' audio recordings.

The LIUM tools embedded in Sphinx4, identify unique speakers in an audio file, viz. speaker recognition, and split the audio into distinct chunks, namely segments. We run both speaker recognition and speech segmentation functions on interviewees' recordings. Outputs were time intervals in which speakers are recognized in an audio stream, audio-segments and the text file containing the duration of each segment. For different pairs, the number of segments, which varied depending on the length and the content of the audio, is given along with the number of speech intervals in Table 4.

Table 4: The number of speech segments and recognized speech intervals.

Interviewer ID	Speech Segments	Recognized Speech Intervals
Interviewer-1	86	30
Interviewer-2	55	29
Interviewer-3	126	38

4 The Sphinx4 is a speech recognition system jointly designed by Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs, and Hewlett-Packard's Cambridge Research Lab. The Official web-site is: http://cmusphinx.sourceforge.net/

The duration of speech segments might be different from the recognized speech intervals. We merged both intervals in order to improve segmentation. The interval merge process is illustrated in Figure 4. We notice that Sphinx4 did not generate segments when the speaker could not be identified. However, those non-segmented parts may contain information that might be useful for researchers. Thus, we carried out additional development works to generate audio segments automatically from non-segmented parts in an audio file. In addition, the closer the microphone was to the participant, the cleaner and the better the gathered audio recording was. Therefore, we might miss data in case we annotated segments that were extracted solely from interviewer's recordings. In order to overcome this problem, we segment both interviewer's and interviewee's recordings from a session, and then, after synchronization (discussed in the next topic), we merged time intervals of segments originating from two distinct sources, for detailed information see the chapter 4).

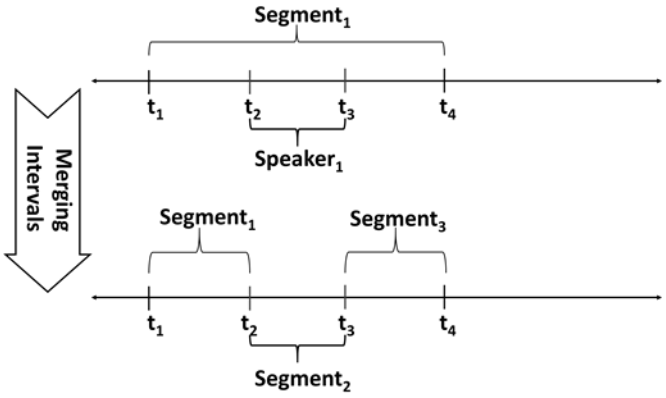


Figure 4: Merging intervals of segments and speakers

3.2.1.2. Synchronization

In an investigation of interactions, especially between participants in a pair, synchronization of the recordings is crucial. Since it is not practically possible to start to record at exactly the same moment on two devices, we had to synchronize their recordings. We signaled the start of the experiment by playing a distinguishable beeping sound not only to ease the determination of an initial segment but also to provide a reference point in time in the synchronization process.

After the segmentation of both interviewer's and interviewee's recordings in a pair was completed, we specified the beginning of the session for each participant by determining the audio-segments containing beeping sound. Then, the starting point of the next segment was assumed to be the initial time for the session. Time offset in a session, which is essential for synchronization of interviewer's and interviewee's recordings, was taken to be the time difference between the starting moments of two recordings in that session. The flow chart of the synchronization algorithm run for the first pair is given in Figure 5.

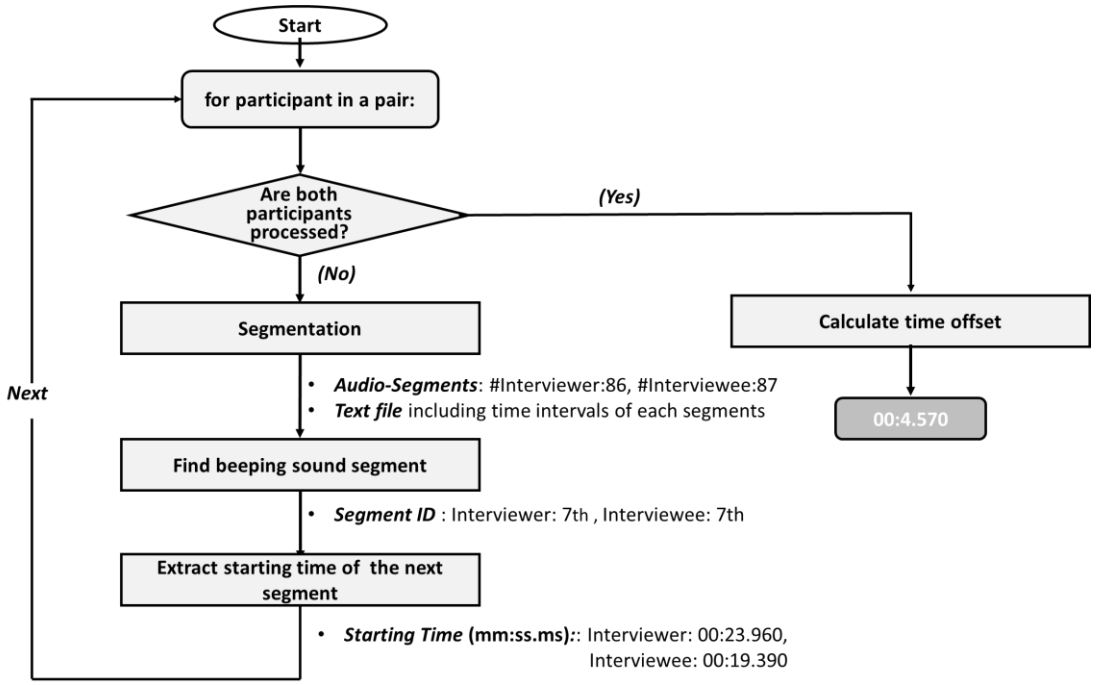


Figure 5: The flow chart of the synchronization process. Values of first pair are presented.

In addition, the final moment of a session was determined first by specifying the last segment of an interviewee that contained speech, and then by extracting the end of that particular segment. As a result, even if the initial times in interviewer's and interviewee's recordings differ for a session, the duration must be the same for both recordings in a pair, see Table 5.

Table 5: Time intervals of session recordings.

	Interviewer (mm:ss.ms)	Interviewee (mm:ss.ms)	Session duration (mm:ss.ms)
Pair-1	00:23.960 – 03:18.520	00:19.390 – 03:13.090	02:54.560
Pair-2	00:16.730 – 05:13.980	00:19.000 – 05:16.250	04:57.250
Pair-3	00:38.740 – 05:17.500	00:07.900 – 04:45.850	04:38.760

3.2.1.3. Annotation

The Speech-act theory is applicable to discourse analysis. It focuses on actions performed through speech and provides a framework to specify the conditions for understanding an utterance as a linguistically realized action. Searle classified this theory further. He states that the taxonomy of speech act is deficit as its original definition, and he proposed criteria for distinguishing one kind of illocutionary force from another. As we stated before, Searle divides illocutionary acts into the following types: *Directive*, *Commissive*, *Representative*, *Declarative* and *Expressive*.

As we investigate gaze aversion mechanism in accordance with speech modality, in addition to speech acts, we proposed additional speech-instances that they might have an effect on gaze aversion mechanisms. In the first place we proposed the following list of speech instances for speech annotation:

Speech: Includes the speech itself. It is a type of commissive or declarative speech-act.

Asking a Question: Speaker requests for information. It is a type of directive speech-act. This category was specific to interviewers.

Confirmation: Act of verifying or making something certain. It is a type of representative speech-act.

Pre-Speech: The non-speech instance which includes the silence before the speech and the sounds for warming up the voice.

Speech Pause: Includes the pauses during the course of speech.

Thinking: We named the conversation segment as thinking when it included filler sounds, such as uh, er, um, eee, the repetition of a question, and drawls – the nonphonemic lengthening of syllables.

Signaling End of Speech: The conversation segments that include phrases signifying the end of the speech, such as that's all

Questionnaire Filling: The interviewer evaluates the interviewee after each question by looking at the notebook and using a pen. This category was specific to interviewers.

After we reviewed data, we realized that the interviewer looked at the notebook while asking questions from it and evaluating the interviewee's response. Thus, in terms of generated gaze behavior, these actions generally caused the same behavior, namely gaze aversion. Consequently, we merged *Questionnaire Filling* and *Asking a Question* instances into a single instance called *Looking at the Notebook*. Furthermore, we eliminated speech-instances that show up less than 5%. As a result, we annotated segments with the following speech instances: *Pre-Speech, Speech, Speech Pause, Thinking, Signaling End of Speech, Looking at the Notebook.*

3.2.2. Gaze Analysis

We first exported videos from recordings by running the corresponding function of Tobii Studio and obtained six video-files for three pairs of participants. Tobii Studio supports AVI file format for movies, which contains both video and audio data, as well as information on audio-video synchrony. We converted AVI files into WMV prior to the remaining analysis.

3.2.2.1. Face Detection

We extracted frame images of each video stream. Tobii glasses recorded data with a sampling rate of 30 Hz., i.e., video stream had 30 Frames Per Second. Therefore, the duration of each frame was 33 milliseconds. Besides, the frame resolution was 640 x 480 pixels.

In order to detect faces in extracted frame-images, we developed a C# application that calls the OpenCV image processing library. At first, we employed the Viole-Jones method for face detection. However, most of the faces could not be detected because of the poor resolution, rapid head movements and/or the errors dependent on the technical constraints of eye-tracking glasses worn throughout the study. To overcome this problem, face detection and face tracking processes were combined so that when the face detection algorithm failed to detect a face, the application we had developed run Camshift face-tracking-method by passing the coordinates of the last detected face.

Camshift generally performs better for moving objects than the other face tracking methods such as meanshift. It achieves fairly good tracking results on a simple background as it considers the color histogram of the target. However, it is not robust against complex backgrounds containing noise and/or objects with the same color as the target. In such cases, the algorithm would fail to track the target (Stern & Efros, 2002; Wang & Yagi, 2008). Accordingly, we made a further improvement in the face detection application. Along with the Camshift algorithm, we used Kalman Filters which consider the direction and the velocity of the object and handles the loss of target on a complex background, as proposed by Kim and Kang (2015). The Algorithm is given as follows:

Algorithm 1: Face Detection

```
1:  function FaceDetection_VioleJones(frame-image)
2:  function Camshift (last-coordinates, KalmanFilter)
3:  function SaveFaceCoordinates (Lines[], fileName)
    Input: Sfi - Set including list of frame-images of each recording
    Output: Text files storing face coordinates
    Initialization: frame-index, Lines, face-coordinates, last-coordinates, fileID
4:  frame-index:=0
5:  fileID :=1
6:  for all s ϵ Sfi do
7:      for all frame-image ϵ s do
8:          face-coordinates:= FaceDetection_VioleJones(frame-image)
9:          if face-coordinates is empty then
10:             face-coordinates:= Camshift (last-coordinates, KalmanFilter)
11:         last-coordinates:=face-coordinates
12:         Lines[frame-index]:= face-coordinates
13:         frame-index:= frame-index +1
14:     SaveFaceCoordinates (Lines, fileID+» «+'.txt')
15:     fileID:= fileID +1
```

Figure 6: The algorithm of face detection

Face Detection algorithm detected faces in a rectangular shape and specify them with four values. First two of them represent coordinates of the top left corner and the last two indicate the width and the height of the rectangle. Nevertheless, as the data is reviewed we realized that locating the face in a rectangle caused an unreliable gaze-behavior estimation, especially when the raw gaze data was near the corners of the rectangle. The problem is illustrated in Figure 7. For this reason, later on, we adopted OpenFace framework which includes facial landmark detection and, hence, identifies the face boundary with a more realistic shape. (for detailed information see the chapter 4). At the end of the face detection phase, we had six text files storing 68 landmark positions which means face-boundary in each frame-image of the recording is identified.

Figure 7: Face detected either in a rectangular shape or with landmark points a)The previous method identified face boundary as a rectangle b)OpenFace detected 68 facial landmarks for positioning the face.

The yellow dot represents the gaze point on that particular frame-image. According to the previous method, since gaze point was inside the rectangle, it would be interpreted as if the interviewee was looking at the interviewer's face, i.e., there was no gaze aversion at that particular time, which was not true. OpenFace enabled us to identify the exact boundary of the face and, hence, a reliable decision about gaze behavior is possible.

3.2.2.2. Detection of Gaze Behavior

We exported raw data of eye movements obtained by the Tobii Glasses Eye Tracker. Tobii Glasses had just one camera positioned on the right-hand side, thereof, Tobii Studio generated an output file storing x and y positions of the right eye at a resolution of 33.33 milliseconds. Afterward, we developed a C# application to decide whether, at a particular time, a participant was looking at the interlocutor's face (viz. in) or looking away from it (viz. out). The inputs of the application were text files containing the face coordinates of each frame-image, which were generated in the previous face detection phase, and eye movements on these frame-images exported from Tobii Studio 3.3.1.

We realized that more than 50% of gaze aversions generated by interviewers lasted up to 33 msec, in other words, correspond at most one frame-image, for the numbers see Table 6. However, since previous studies reported longer fixation durations, we made further improvements in the detection of gaze aversion.

Table 6: Percentages of gaze aversions lasted 33 ms

	Interviewer	Interviewee
Pair1	22	29
Pair2	73.8	39.1
Pair3	50	37.3
Mean	43.8	33.1

Fixation identification algorithms may then be employed to determine whether raw data points accumulate into fixations during the course of gaze aversion. A challenge in the specification of fixations from raw data comes from the fact that wearable eye trackers capture dynamic scenes. Currently, there is no commonly accepted method for detecting eye movement events in dynamic scenes (Munn et al., 2008; Srinivasan et al., 2014). In the present study, we analyzed raw data after a cleansing process described in the following section.

In the detection of gaze aversion, we used cleansed raw gaze data as input. The cleansing process involved gap-filling via linear interpolation where at most two frames were filled. After detecting gaze aversions, we merged adjacent aversions between which there were at most two consecutive non-aversion frames. Finally, we eliminated short aversions that are less than 100 ms (Figure 8).

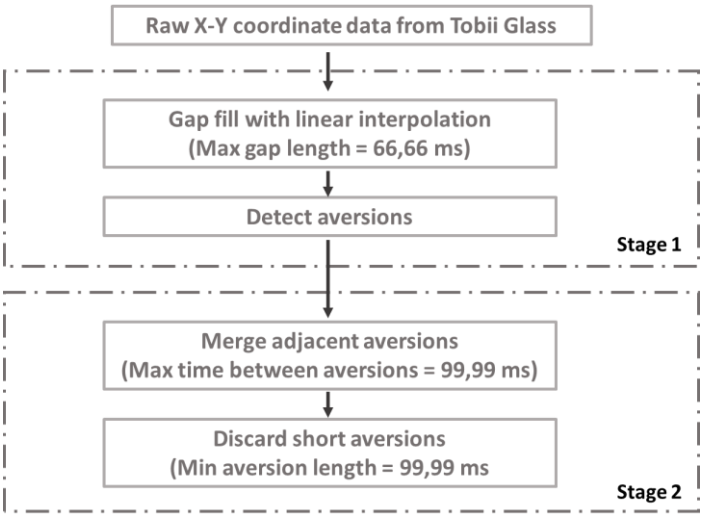


Figure 8: Process flow for detection of gaze aversion.

In addition, the application developed computed coordinates of gaze relative to the detected face of the interlocutor. As shown in Figure 9, frame-image was theoretically divided into 9 (3x3) areas of interests (AOIs). Each AOI might be different in size. If the gaze data was inside the detected face, 'E' is assigned as a label to the AOI, otherwise one of the eight characters, namely *a, b, c, d, e, f, g, h, i* was assigned. Characters in labels were determined according to the relative position of that particular area with respect to the face area. For instance, north-west of face-area was always labeled as *a*, and similarly south of face-area was labeled as *h*. The application produced text files containing frame-image IDs along with the corresponding AOI-labels, for each recording. Figure 9 shows detected facial landmarks and gaze data overlay on a sample image frame.

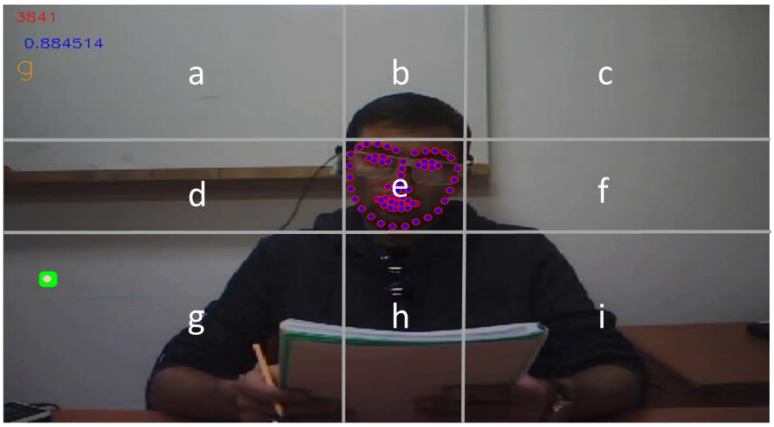


Figure 9: Gaze location relative to the face. The yellow dot represents the gaze data of an interlocutor, in this case, of an interviewee. The frame-image of an interviewer was divided into 9 (3x3) AOIs. The middlemost area was the detected face of an interviewer. An interviewee was looking at the south-west of an interviewer's face. Thus, this frame-image should be labeled as 'G'.

3.2.2.3. Gaze and Speech Multimodal Data

At the last phase of gaze analysis, we synchronized gaze behavior and annotated speech data obtained from the previous analysis. We iterated AOI labels, annotated-speech-segments and interlocutor's AOI labels, synchronously. Eventually, in each iteration, we ended up with sender information, AOI label, speech-instance and interlocutor's AOI label, for the particular frame-image. An iteration was assigned as the starting frame of the gaze aversion if its AOI label was different from *e* provided that the AOI label of the previous iteration had been *e*. Gaze aversion continued as long as the AOI label remained different from *e*. At the end, we kept the following information for each frame-image:

Gaze Behavior: It can be one of the following labels: *Aversion*, *Face Contact* or *Empty*. The *Empty* label was assigned, when raw gaze data of the participant could not be extracted and/or there was a problem in face detection. This value was handled separately for both interviewer and interviewee participants.

Gaze Behavior Onset: It is the duration of instant gaze behavior starting from its initial occurrence. This value was handled separately for both interviewer and interviewee participants.

Sender: It can be either an *interviewer* or an *interviewee*.

Speech Instance: It can be 1 of the following 6 items: *Pre-Speech*, *Speech*, *Speech Pause*, *Thinking*, *Signaling End of Speech* and *Looking at the Notebook*.

Speech Modality: It is a combined feature including both the sender (i.e. an interviewer or an interviewee) and the speech-instance.

Speech Modality Onset: It is the duration of instant speech-modality starting from its initial occurrence.

3.3. Results

We analyzed the mean number of gaze aversions per minute (i.e., gaze aversion frequency), the mean duration of gaze aversions and the timings of gaze aversion instances. All analyses were carried out in R programming language and environment (R Core Team; 2016). using `lme4` and `lmerTest` software packages. All data files and R scripts used during the analysis are publicly available.[5]

[5] Data files and R scripts are available under:
https://gist.github.com/ulkursln/9d14fe288471b9e83f845607d5c3045d

3.3.1. Gaze Aversion Frequency

The number of gaze aversions was closely related to the length of the corresponding session. Since no time limit was imposed in the experiment, we needed to calculate a normalized frequency per minute of gaze aversion. The analysis revealed that the interviewees performed more frequent gaze aversions (M = 27.95, SE = 8.53) when compared with the interviewers (M = 22.72, SE = 3.26).

3.3.2. Gaze Aversion Duration

The analysis revealed that gaze aversions of the interviewees took longer (M = 2207.9 ms, SE = 1291.2) than gaze aversions of the interviewers (M = 1860.0 ms, SE = 363.0). These numbers represent the analysis which covered all gaze aversion instances. However, as already mentioned above, the interviewers looked at the notebook while they filled in the questionnaire to evaluate the interviewee's response and while they articulated the questions. Therefore, we repeated the analysis by excluding those instances where the interviewer looked at the notebook, as they did not represent genuine cases of gaze aversions during the course of conversation. The renewed analysis resulted in a more salient difference in duration of gaze aversions between the interviewers (M = 1179.3 ms, SE =384.1) and the interviewees (M= 1802.3 ms, SE = 921). We also investigated the relation between gaze aversion and speech-instance type. A single gaze aversion might be related to multiple speech-instances. Figure 10 shows the average duration while a participant was averting his gaze from the interlocutor's face and performing the specific speech-instance.

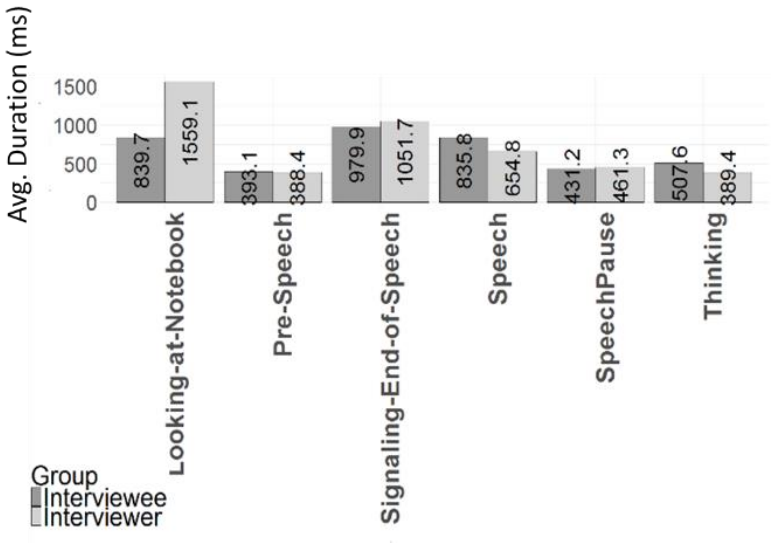


Figure 10: The average duration of gaze aversion for each type of speech-instances. Light gray bars represent interviewers and dark gray bars are for interviewees.

The durations of gaze aversions were analyzed via linear mixed effects regression, LMER, by using the lme4 package in R. We treated the participant pairs (viz., pair-id) as random

effects to control the influence of different duration values associated with the same pair. In a mixed-model, removing *Sender*, *Speech-Instance*, *Gaze-Behavior-Onset* and *Speech-Modality-Onset* significantly decreased the goodness of fit, as indicated by likelihood ratio tests – effect of *Sender* $\chi2(1) = 22.1$, $p < .000$; effect of *Speech-Instance* $\chi2(5) = 69.6$, $p < .000$; effect of *Gaze-Aversion-Onset* $\chi2(1) = 16.1$, $p = .000$ and effect of *Speech-Modality-Onset* $\chi2(1) = 20$, $p < .000$. A post hoc Tukey test performed on speech-instance category showed that *Looking-at-Notebook* (M =1067.8 ms, SE = 76.9) significantly (all *p*s<.000) increased aversion duration compared with *Speech* (M =742.6 ms, SE = 40.6), with *Pre-Speech* (M =398.2 ms, SE = 45.2), with *Speech-Pause* (M =432.2 ms, SE = 33.9) and with *Thinking* (M =477.7 ms, SE = 34.5). Moreover, the following pairs of instances found to be significantly different (all *p*s< .05): *Pre-Speech* and *Speech*, *Speech-Pause* and *Speech*, *Thinking* and *Speech*, *Signaling-End-of-Speech* and *Pre-Speech*, *Signaling-End-of-Speech* and *Speech-Pause*, and *Signaling-End-of-Speech* and *Thinking*.

A post hoc Tukey test performed on the *Sender* category showed that aversion duration significantly (*p*<.000) decreased when the speaker was the interviewee (M =629.6 ms, SE = 25.4) instead of the interviewer (M =918.8 ms, SE = 63.9). Finally, the lmer mixed-model showed that the duration of aversion was linearly related to *Gaze-Aversion-Onset* (*b =132.9 ms, SE = 32.9*), and *Speech-Modality-Onset* (*b =-102.5 ms, SE = 30.9*).

3.3.3. Occurrence of Gaze Aversion

We introduced mixed-effects-logistic-regression models, in order to investigate the effects that influence whether it is time to avert gaze by considering following aspects for every 30 milliseconds during the all three sessions: (The sample size was 23,115[6])

The first model was created to predict the interviewer's gaze-behavior-type (i.e., whether it was gaze aversion or not, in that particular time). As fixed-effects, we identified the interviewer's *Gaze-Behavior-Onset*, a correlated relation of *Sender*, *Speech-Instance* and *Speech-Modality-Onset* and lastly a correlated relation of interviewee's *Gaze-Behavior* and interviewee's *Gaze-Behavior-Onset*. As the random effect, we had *Pair-Id*, as mentioned in the previous section. In a mixed-model, removing the *Sender*, *Speech-Instance*, interviewer's *Gaze-Behavior-Onset*, *Speech-Modality-Onset*, interviewee's *Gaze-Behavior* and interviewee's *Gaze-Behavior-Onset* significantly decreased the goodness of fit, as indicated by likelihood ratio tests – effect of *Sender* $\chi2(1) = 2031.7$, $p < .000$; effect of *Speech-Instance* $\chi2(5) = 85.9$, $p < .000$; effect of *Gaze-Behavior-Onset* $\chi2(1) = 927.9$, $p < .000$; effect of *Speech-Modality-Onset* $\chi2(1) = 77$, $p < .000$; effect of interviewee's *Gaze-Behavior* $\chi2(1) = 35.4$, $p < .000$ and effect of interviewee's *Gaze-Behavior-Onset* $\chi2(1) = 6.25$, $p <= .01$.

[6] The link to access the data file:
https://drive.google.com/open?id=0B-DfZx3YFEzgRldUNm1fZ3ZPZDQ

A post hoc Tukey test was performed for making pairwise comparisons among the ratios of gaze aversion to face-contact (i.e., odd ratio) for several *Speech-Instances*. If the odd ratio of the first instance in the pair is larger than the second one, the confidence interval will be on the positive side, otherwise, it will be on the negative side. Moreover, results indicate that the following pairs did not significantly differ from each other (i.e., their confidence intervals include 0): *Speech – Pre-Speech*, *Speech Pause – Pre-Speech* and *Speech Pause – Speech*, and for all the other pairs the differences are significant (see Figure 11).

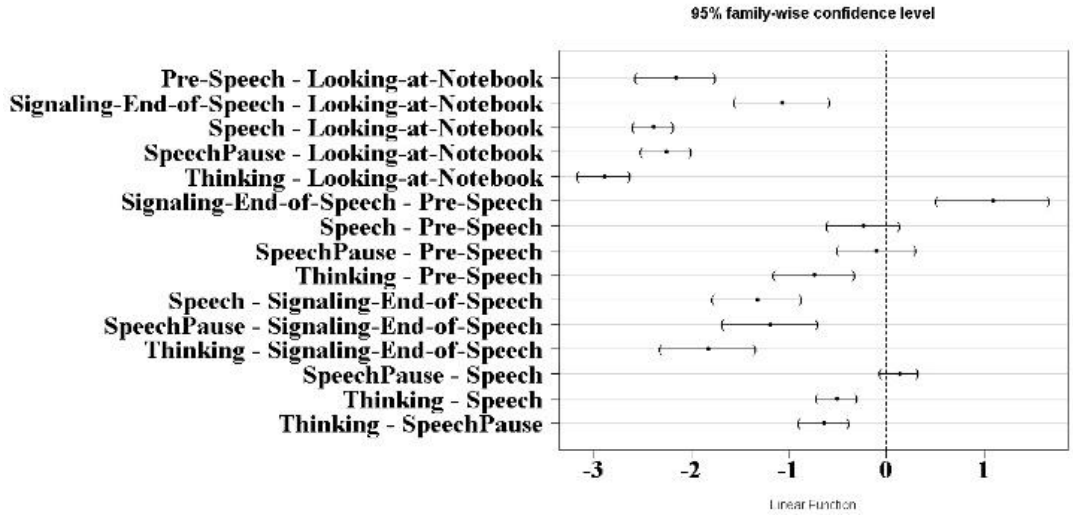


Figure 11: Interviewer's pairwise comparisons among the ratios of gaze aversion to face-contact for several *Speech-Instances*. The confidence intervals that do not include 0, point out a significant difference. For instance, an interviewer is more likely to avert his eyes while the *Speech-Instance* is *Signaling-End-of-Speech* rather than being *Pre-Speech***.**

We performed a similar analysis also for the interviewees. The second model was created to predict the interviewee's gaze-behavior-type (i.e., gaze aversion or not). We identified the interviewee's *Gaze-Behavior-Onset*, correlated relation of *Sender*, *Speech-Instance* and *Speech-Modality-Onset*, and lastly a correlated relation of interviewer's *Gaze-Behavior* and interviewer's *Gaze-Behavior-Onset*, as fixed-effects. As the random effect, we had *Pair-Id*. In a mixed-model, removing the *Sender*, *Speech-Instance*, interviewee's *Gaze-Behavior-Onset*, *Speech-Modality-Onset*, interviewer's *Gaze-Behavior* and interviewer's *Gaze-Behavior-Onset* significantly decreased the goodness of fit, as indicated by likelihood ratio tests – effect of *Sender* $\chi2(1) = 11.6$, $p < .000$; effect of *Speech-Instance* $\chi2(5) = 1020$, $p < .000$; effect of interviewer's *Gaze-Behavior-Onset* $\chi2(1) = 62.61$, $p < .000$; effect of *Speech-Modality-Onset* $\chi2(1) = 7.23$, $p < .000$; effect of interviewer's *Gaze-Behavior* $\chi2(1) = 27.01$, $p < .000$ and effect of interviewee's *Gaze-Behavior-Onset* $\chi2(1) = 25.22$, $p < =.000$.

A post hoc Tukey test was performed for making pairwise comparisons among the ratios of gaze aversion to face-contact (i.e., odd ratio) for several *Speech-Instances*. Results indicate that the following pairs did not significantly differ from each other: *Speech Pause*

– *Pre-Speech* and *Speech – Signaling End of Speech*, and for all the other pairs the differences are significant (see Figure 12).

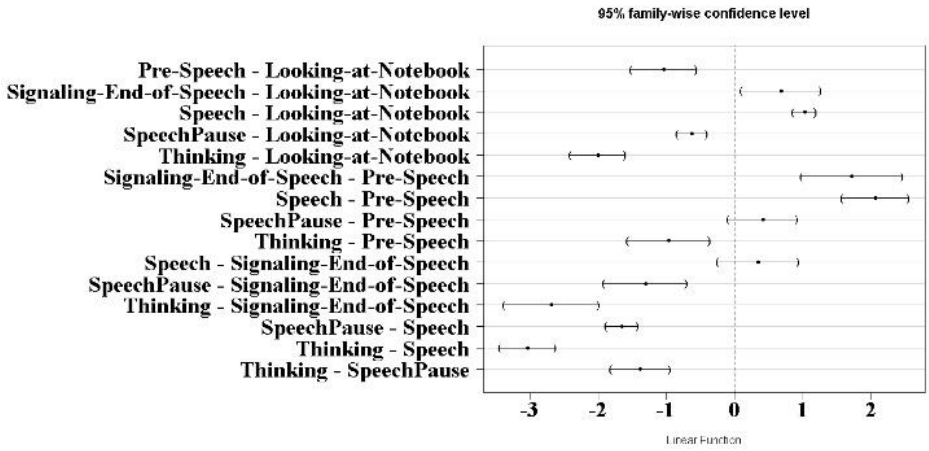


Figure 12: Interviewee's pairwise comparisons among the ratios of gaze aversion to face-contact for several *Speech-Instances*. The intervals that do not include 0 point out a significant difference. For instance, an interviewee is more likely to avert his eyes while the *Speech-Instance* is *Speech* rather than being *Pre-Speech*.

3.3.4. Relative Spatial Positions of Gaze Aversions

We calculated the relative spatial positions of gaze aversions with respect to an interlocutor's face. As illustrated in Figure 13, during gaze aversion, the interviewees frequently looked at the lower right-hand side of an interlocutor, whereas the interviewers looked straight down in the case of articulating questions or filling the questionnaire, as expected.

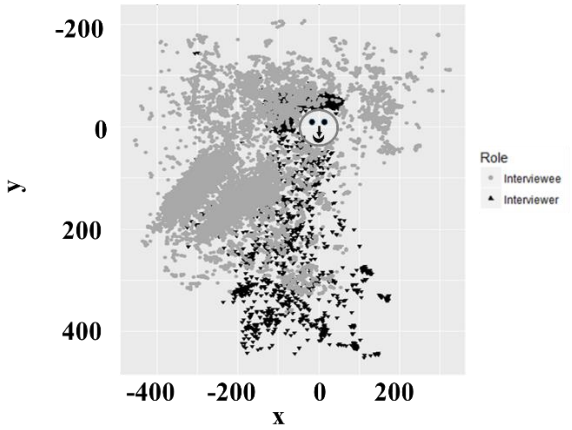


Figure 13: The distribution of gaze aversion's location relative to the location of interlocutor's face. Dots represent the relative positions of gaze aversions with respect to the interlocutor's face.

3.4. Discussion

The purpose of the pilot study was to improve the design and analysis methods, for the purpose of gaining an in-depth understanding of gaze behavior in a natural conversation of pairs. In this chapter, we investigated gaze aversion from a multimodal perspective, by employing face tracking and analyzing speech data as well as eye-tracking data in a mock job-interview task. Synchronous use of face tracking and gaze data overlay allowed us to detect gaze aversions of both communication partners.

The results of the study show that gaze aversion characteristics differ between interviewees and interviewers. In particular, the interviewees exhibited more frequent gaze aversions than the interviewers did. We also found that the interviewees and the interviewers employed different patterns of specific speech instances during the course of conversations.

In terms of improvement in design principle, we noted two important points. Firstly, we realized that the face detection algorithm performs suboptimal due to the noise and poor lighting conditions in the environment. Thus, we decided to perform the next study in a room with proper lighting and a clear background. Secondly, we realized that the interviewer looked at the notebook while asking a question and evaluating the interviewee's response. That also affected the performance of face detection. Face detection algorithms might miss the face when the head was tilted. Therefore, we abandoned the use of pen and pencil and decided to provide an alternative solution.

On the other hand, we observed the necessity of improvements in speech and gaze analysis. For speech analysis, we run speaker recognition and speech segmentation functions of Sphinx4, both. We, then, merged the speech intervals and segments generated as outputs of these two functions in order to improve segmentation. Nevertheless, Sphinx4 did not generate segments when the speaker could not be identified despite the fact that those non-segmented parts might contain information useful for researchers. Thus, we carried out additional development to generate audio segments automatically from non-segmented parts in audio recordings. In addition, the closer the microphone was to the participant, the cleaner and the better the gathered audio recording was. Therefore, in case we annotated segments that were extracted only from an interviewer recording, we might miss data. In order to overcome this problem, we segmented both interviewer's and interviewee's recordings for a session, and then, after synchronization, we merged intervals of segments coming from distinct sources. Lastly, after we reviewed the annotated speech data, we realized that it could be better to handle the proposed speech-instances with the perspective of functional roles of gaze in social communication.

The Gaze analysis phase was composed of face and gaze-aversion detection. We, first, employed the Viole-Jones method for face detection. Then, we made an improvement in case the face detection algorithm failed to detect the face. The application developed run Camshift face-tracking-method by passing the coordinates of the last detected face. Yet, Camshift is not robust against the complex backgrounds containing noise and/or objects

with the same color as the target. Therefore, we proposed further improvement in face detection and adopted the Kalman filter. Furthermore, Face Detection algorithm detected faces in a rectangular shape and this might cause unreliable estimation of gaze behavior, especially when the gaze data of a participant was near the corners of face-rectangle, which indeed should be the case of gaze aversion. Thus, we adapted OpenFace framework, which includes facial landmark detection and, hence, identifies the face boundary with a more realistic shape.

CHAPTER 4

MAGIC: A MULTIMODAL FRAMEWORK FOR ANALYSING GAZE IN COMMUNICATION[7]

This chapter presents an open-source framework, namely MAGiC, for analyzing gaze contact and gaze aversion in face-to-face communication. The analysis of dynamic scenes has been a challenging domain in eye tracking research. MAGiC provides an environment that is capable of detecting and tracking conversation partner's face automatically, overlaying gaze data on top of the face video, and incorporating speech through speech-act annotation. Specifically, MAGiC integrates eye-tracking, audio, and video data for gaze, speech segmentation, and face tracking, respectively. MAGiC has been developed as an open-source software tool, which is available for public use and development. Separation of Concerns design principle is adopted in order to address different concerns under separate modules. Moreover, MAGiC produces standard output files (such as wav or txt files) in each inner step. This helps researchers to understand inner processing and enables them to conduct further analysis. We demonstrate the capabilities of MAGiC through a pilot study and report the usability analysis.

4.1. 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 when there is any inconsistency between the messages simultaneously conveyed by non-verbal and verbal modalities, the former prevails the latter. 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). Accordingly, investigating the structural underpinnings of social interaction requires the study of non-verbal modalities as well as verbal modalities of communication. In the present chapter, we focus on gaze,

---

[7] This chapter, largely in its current form, is published as:
Arslan Aydin, Ü., Kalkan, S., & Acarturk, C. (2018). MAGiC: A multimodal framework for analysing gaze in dyadic communication. *Journal of Eye Movement Research, 11*(6). https://doi.org/10.16910/jemr.11.6.2

in particular the analysis of eye contact and gaze aversion, as the non-verbal modality in face-to-face communication.

Eye contact plays a crucial 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 managing the conversational roles of interlocutors (e.g., Bales, Strodtbeck, Mills, & Roseborough, 1951; Goodwin, 1981; Schegloff, 1968). Interlocutor's putative mental states, such as interest, are usually inferred from gaze (Baron-Cohen, Wheelwight, & Jolliffe, 1997). Eye contact is a fundamental, initial step for capturing the attention of the communication partner and establishing joint attention (Fasola & Mataric, 2012; Kleinke, 1986). Gaze aversion, complementary to eye contact, is another coordinated interaction pattern that regulates the conversation. Gaze aversion is defined as the act of looking away from the interlocutor, intentionally. In the literature, there are numerous studies concerning the effects of gaze aversion on avoidance and approach motivations. Hietanen et al. (2008) report that an averted gaze of an interlocutor initiates a tendency to avoid, whereas a direct gaze initiates a tendency to approach. In similar studies, the participants gave higher ratings of likeability and attractiveness when the picture stimuli included a face with a direct gaze contact, compared to the stimuli that included a face with averted gaze (Mason et al., 2005; Pfeiffer et al., 2011). These findings suggest that gaze aversion is expected to last shorter than eye contact in an efficient conversation. More generally, three conversational functions have been attributed to gaze aversion (Abele, 1986; Argyle & Cook, 1976; Kendon, 1967):

i. **Intimacy modulation:** The overall level of intimacy is influenced by periodic gaze aversions.

ii. **Floor management:** Gaze aversion occurs when the speaker takes a break by temporarily stopping the conversation during the course of speech.

iii. **Cognitive management:** The speaking partner conducts more gaze aversion than the listening partner to facilitate thinking and remembering. This eventually reduces the effort needed to pay attention to the listener.

As the above classification suggests, the conversational function of gaze aversion is closely related to speech. In other words, speech and gaze are closely connected modalities in social interaction. Similar to other non-verbal signals, gaze provides repeating, complementing, and substitution of a verbal message as well as regulating it. Speech requires temporal coordination of embodied cognitive processes: planning, phonemic construction, and memory retrieval for lexical and semantic information (Elman, 1995; Ford & Holmes, 1978; Kirsner, Dunn, & Hird, 2005; Krivokapić, 2007; Power, 1985). Speech involves various sorts of signals depending on its content or quality, such as intonation, volume, pitch variations, speed, and actions performed throughout it (viz. speech acts). We focus on speech acts due to 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 at least three components that have distinct functional roles:

i.*Locutionary* act refers to the act of saying something with its literal meaning

ii.*Illocutionary* act indicates the intent of the speaker

iii.*Perlocutionary* act is the effect of an utterance on the interlocutor.

For analyzing language in communication, discourse should be segmented into units that have communicative functions, and related communicative functions should be identified and labeled accordingly. For instance, the following labels are proposed by Searle (1969) to classify locutionary acts.

Directives: to make the listener perform a particular action (e.g., request, order, advice, etc.),

Commissives: speaker commits himself to take further action (e.g., promises, planning, etc.)

Assertives: speaker represents a state of affairs (e.g., concluding, suggesting, etc.)

Expressives: speaker express emotions and attitudes towards the situation denoted by the preposition (e.g., apologizing, congratulations, thanks, etc.)

Declaratives: speaker changes the world by uttering a locutionary act (firing, resigning, nominating, betting, etc.)

Speech-acts are identified by analyzing the content of a speech. However, not only the content but also the temporal properties of speech convey information to the interlocutor. For instance, the analysis of a pause may be taken into account for signaling a shift in topic (Krivokapic, 2007), or it may be used for estimating speech intent, evaluating speaker's fluency (Grosjean & Lane, 1976) and even detecting speech disorders (Hird, Brown, & Kirsner, 2006). MAGiC enables researchers to carry out analyses by employing both the content of speech and its temporal properties.

In the current state of technology, eye tracker manufactures provide 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) based statistics. The analysis of social behavior, however, requires more advanced tools that are able to overlay gaze data on top of 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) as there exist technical challenges in recognizing and tracking objects in a dynamical scene. This is because eye trackers generate a raw data stream, which contains

a list of points-of-regard (POR) during the course of task performance by the participant. In a stationary scene, it is relatively straightforward to specify subregions (i.e., AOIs) of the stimuli on the display, and then this information is used to extract AOI-based eye movement metrics. In the case of a dynamic scene, as in the case of mobile eye trackers, automatic detection of regions is a complex task. To the best of our knowledge, there is no commonly accepted method for achieving eye movement analysis in dynamic scenes (Munn et al., 2008; Stuart et al., 2014).

MAGiC focuses on face recognition, which is a relatively well-developed subdomain of object recognition. The recognition of faces has been subject to intense research in computer vision due to its potential importance to practical applications in daily life, such as its use in digital camera recordings for security purposes. MAGiC employs face recognition techniques to automatically detect gaze contact and gaze aversion in dynamic scenarios, where eye movement data are recorded. It aims to facilitate frame-by-frame analysis of dynamic scenes, thus reducing the effort for time-consuming and error-prone manual annotation 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 present major characteristics and the benefits of MAGiC in more detail.

4.2. An Overview of Characteristics

4.2.1. Reduced Annotation Effort and Time

MAGiC reduces the amount of time spent on preparing manually annotated gaze and audio data for each image frame of a scene video. Without MAGiC, in order to identify face contact, gaze aversion, and their location, a researcher would need to annotate 36,000 image-frames, for a 10-minute recording of a 60 Hz eye tracker. Assuming that one needs 1 second for annotating one frame, the duration would exceed 10 hours for a 10-minute recording. Fortunately, MAGiC significantly reduces the amount of time spent on annotation. The same process takes approximately 5 to 10 minutes per 10-minute recording, in a typical personal computer with Intel Core i5 2.3 GHz CPU and 8 GB of RAM. Likewise, the effort spent for the AOI annotation, the segmentation, and annotation of an audio recording have been significantly reduced by MAGiC. It automatically segments the audio file in a couple of seconds and also provides an interface to facilitate the annotation of audio segments.

4.2.2. Automated Multimodal Analysis

MAGiC provides functionalities for automatic analyses of both speech and gaze. In addition to saving time, automation enables researchers to obtain further information that

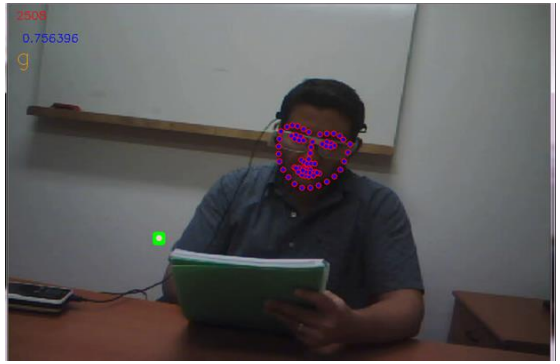
may not be extracted manually. For instance, OpenFace[8], an open-source facial behavior analysis toolkit utilized in MAGiC, detects the coordinates of 68 facial landmarks. MAGiC extracts the coordinates of some facial features such as eyes or mouth and then evaluates the relative coordinates of gaze location to extracted facial features. In addition, MAGiC employs CMUSphinx[9] framework for segmenting audio signals at millisecond precision. Speaker change, speech-pause and humming (e.g., sounds like "hmm", "mhm", "uh-huh") are some of the content or temporal based speech properties that might be taken into consideration in speech analysis. The automated annotation improves the quality of annotated data since it is virtually impossible for human annotators to detect speech instances at this level of temporal granularity. MAGiC also offers an interface to make manual AOI annotation.

4.2.3. Performance Improvement and Visualization

MAGiC has the functionality to visualize face tracking data and the AOI annotation frame-by-frame. It overlays the detected facial landmarks, the raw gaze data, and the status of gaze interaction (gaze-contact and gaze-aversion) in a single video recording. MAGiC displays the ratio of non-annotated gaze data (thus, the success level of face detection) as a percentage of total data. If the user is not satisfied with the face detection performance, one may employ MAGiC's training interface to improve face detection. The training interface aims to increase the average accuracy of face detection, see Figure 14.

[8] OpenFace: an open-source facial behavior analysis toolkit, https://www.cl.cam.ac.uk/~tb346/res/openface.html, retrieved on April 15, 2017.

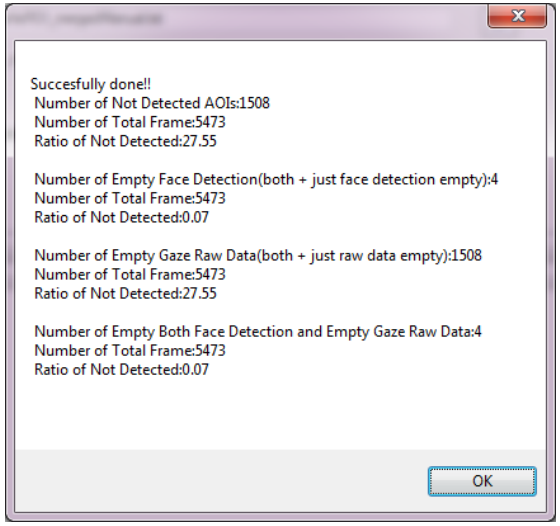
[9] CMU Sphinx, Open-Source Speech Recognition Toolkit, http://cmusphinx.sourceforge.net/, retrieved on April 15, 2017.


0.756396

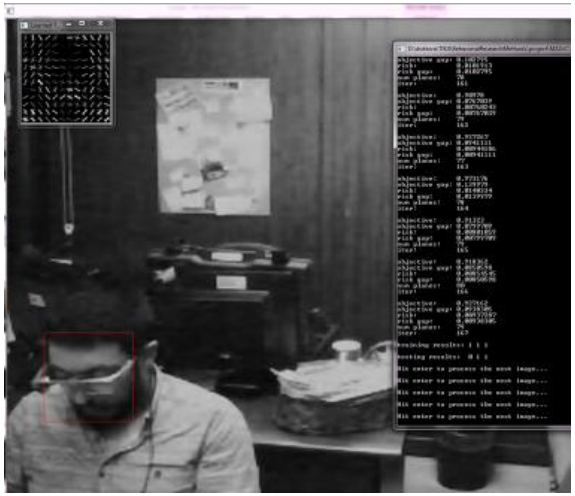
(a)


File Help
Save
Save As
Remove Selected Images
Exit
Next Label:
trainingImages\829.png
trainingImages\831.png
trainingImages\835.png
trainingImages\9.png

(b)


Succesfully done!!
Number of Not Detected AOIs:1508
Number of Total Frame:5473
Ratio of Not Detected:27.55
Number of Empty Face Detection(both + just face detection empty):4
Number of Total Frame:5473
Ratio of Not Detected:0.07
Number of Empty Gaze Raw Data(both + just raw data empty):1508
Number of Total Frame:5473
Ratio of Not Detected:27.55
Number of Empty Both Face Detection and Empty Gaze Raw Data:4
Number of Total Frame:5473
Ratio of Not Detected:0.07
OK

(c)



(d)

Figure 14: A set of screenshots taken from related MAGiC's components for visualization, training and performance monitoring. a) The set of facial landmarks are presented around the face of the interviewer with pink circles, b) The ratio of non-annotated gaze data along with their causes are displayed. The absence of raw-gaze data or undetected faces are the reasons behind the failure of AOI-annotation. c) During the training of a custom face detector, the user has to specify the boundaries of the face in the pre-defined set of training-images by drawing boxes around d) After training is complete, the performance of the custom-detector is displayed. Each detected face in the pre-defined set of test-images is displayed one-by-one.

4.2.4. Flexibility

The number of audio and video sources varies across experimental designs. The simplest set-up involves a single source, such as a single eye-tracker glasses in a conversation dyad. A more complex design may consist of two or more eye trackers, which requires synchronization of the data sources in multi-source recordings. MAGiC provides an interface to synchronize pair recordings semi-automatically. It is a semi-automatic process since once the MAGiC completed the automatic segmentation of audio files, a human annotator has to listen to audio-segments for specifying the first and the last segments of the recordings. Another aspect of flexibility in MAGiC is based on the adoption of the Separation of Concerns (SoC) design principle (Reade, 1989). It guarantees the modularity and the independence of the modules, in order to address different needs from users. Accordingly, each module in MAGiC can be used in isolation, without employing all the functionalities of the tool. For instance, it is possible to use MAGiC only for speech segmentation, face tracking, or synchronization of pair recordings.

4.2.5. Extensibility

The implementation of MAGiC has taken possible future improvements into consideration, in order to facilitate extensibility. MAGiC has been developed as an open-source software application, accessible as a public and non-commercial resource. Hence, open-source developers have the opportunity to contribute to the development of MAGiC which may eventually expand its capabilities to accommodate the requirements of researchers in various fields. MAGiC utilizes three open-source toolkits, namely OpenFace, dlib, and CMUSphinx; for face tracking, training of face detector and speech segmentation, respectively. The improvements in the component toolkits will lead to enhancing the performance of MAGiC. The loosely-coupled design of MAGiC to its component toolkits intends to reduce compatibility issues with newer versions of the component toolkits.

4.2.6. Ease of Use

MAGiC has been designed to serve researchers from different research domains. For this reason, its design has been based on minimizing the requirement for technical background. The required input from the user is video recording and raw gaze data. MAGiC generates standard output files (such as .wav and .txt files) to help researchers access data directly and to ease the data transfer to statistical software.

A technical overview of the component toolkits for face tracking and speech segmentation along with the further source code improvements of the toolkits for MAGiC are in Section 3. Section 4 reports the system requirements, the installation procedure, the phases of the software development, the design principles, and the guidelines for using the MAGiC interface. In Section 5, the capabilities of MAGiC are illustrated through a demonstration of data analysis in a pilot, experimental study. Section 6 concludes the article, and it discusses future work to improve MAGiC.

4.3. A Technical Overview of Components

In the next two subsections, we explain how face tracking, speech segmentation and annotation of segments are conducted by MAGiC by employing its open-source components.

4.3.1. Face Tracking

Face tracking has been a challenging topic in computer vision. In face tracking, firstly a face in a video-frame is detected and then it is tracked throughout the stream. In this chapter, we employ a face-tracking toolkit called OpenFace, which is an open-source tool for analyzing facial behavior. OpenFace combines out-of-the-box solutions with state-of-the-art research to perform some tasks including facial-landmark detection, head pose estimation, and action unit (AU) recognition. The face-tracking method used in our study (and presented in this section) is based on the studies conducted by Baltrušaitis, Robinson and Morency (2016) which are also connected to previous ones (Baltrušaitis, Robinson, & Morency, 2013; Baltrušaitis, Mahmoud, & Robinson, 2015).

OpenFace makes use of a pre-trained face detector, trained in dlib[10], which is an open-source machine-learning library written in C++. Max-margin object-detection algorithm (MMOD), using Histogram of Oriented Gradients (HOG) feature extraction, was employed to train a face detector with a relatively small amount of training data (King, 2009, 2015). After detecting a face, for detecting the facial landmarks, OpenFace utilizes a novel instance of Constrained Local Model (CLM), namely Constrained Local Neural Field (CLNF), to handle feature detection problems in complex scenes. The response maps are extracted using pre-trained patch experts, and patch responses are optimized with Non-Uniform Regularized Landmark Mean-Shift (NU-RLMS), which is a novel fitting method (seeFigure 15).

[10] Dlib C++ Library, http://dlib.net/, retrieved on April 15, 2017.

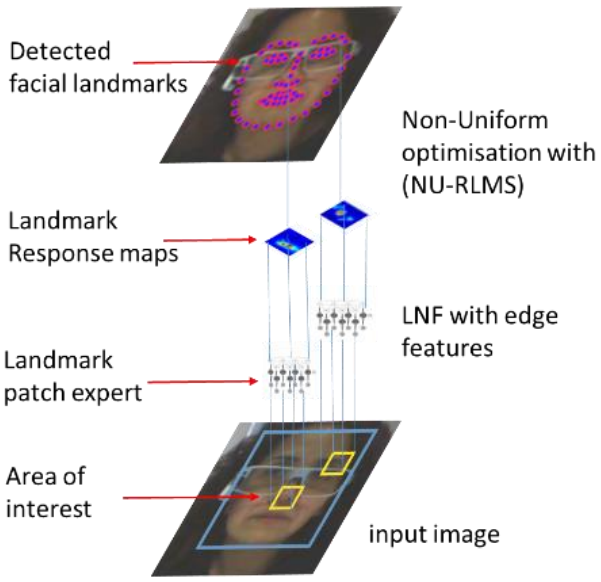


Figure 15: A demonstration of OpenFace methodology, adapted from Baltrušaitis et al. (2013). Here it is intentionally limited to two landmarks for illustrative purposes

The Constrained Local Model (CLM) is composed of three main steps, as described below.

(1) A Point Distribution Model 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 in successive shapes. For instance, for particular face shape, specific landmark points may always correspond to eyelids. Then, 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. Principal Component Analysis is used to pick out the correlations between groups of landmarks among the trained shapes. At the end of this step, patches are created around each facial landmark. The patches are trained with a given set of face-shapes.

(2) Patch Expert, also known as local detectors, are used for calculating response maps, which represent the probability of a certain landmark that is being aligned at point x_i (Equation 1), from Baltrušaitis et al. (2013).

$$\pi(x_i) = C_i(x_i; I), \qquad \text{(Equation 1)}$$

where I is an intensity image, and C_i is a logistic regressor intercept with a value between 0 to 1 (here 0 represents no alignment and 1 represents perfect alignment). Due to its computational advantages and implementation simplicity, Support Vector Regressors are usually employed as patch experts. On the other hand, the CLNF

model uses the Local Neural Field approach, which takes the spatial features that lead to fewer peaks, smoother response and reduced noise into account.

(3) Regularised Landmark Mean Shift (RLMS) is the third step of the CLM. It is a common method for solving data fitting. It updates the CLM parameters to get closer to a solution. An iterative fitting is performed to update the initial parameters of the CLM until convergence to an optimal solution is achieved. Adapted from Baltrušaitis et al. (2013), the general concept of iterative fitting is defined as

$$\arg\min_{\Delta p} [R(p0 + \Delta p) + \sum_1^n D_i(x_i; I)], \qquad \text{(Equation 2)}$$

where R is a regularization term to degrade complex deformations and D_i represents the misalignment measure for the image I at the image location x_i. RLMS does not discriminate between confidence levels of response maps. Since outcomes revealed that some response maps are noisier than the others, a novel non-uniform RLMS weighting mean-shifts is proposed to overcome this issue.

At the end of the third step, the OpenFace toolkit detects a total of 68 facial landmarks (Figure 16). Determining the face boundaries based on facial landmarks instead of a rectangle covering the face enables more precise calculations.

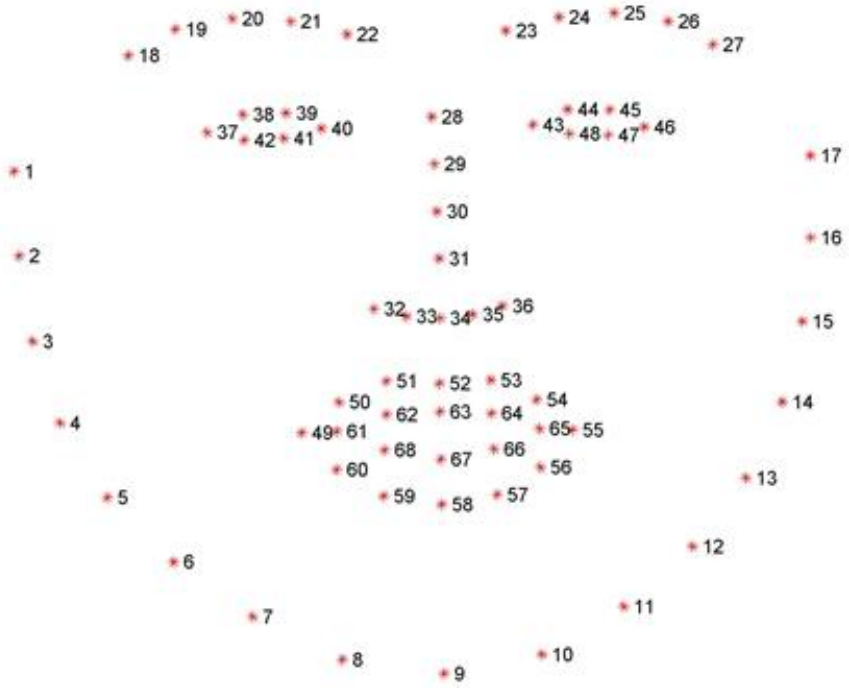


Figure 16: A total of 68 landmark positions on a face.

We extended the OpenFace source code by making a set of improvements, which allowed the user to conduct manual AOI annotation, to generate visualizations that employ new

input parameters, to build a custom face detector and then use it to track the face with that detector and to generate separate output files depending on the input parameters.

4.3.2. Speech Segmentation

Speech is a continuous audio stream with dynamically changing and not readily distinguishable parts. Speech analysis has been a challenging domain of research, due to the difficulty to identify clear boundaries between speech-related units, automatically. Speech analysis involves two interrelated families of methodologies, namely speech segmentation and diarization. Speech segmentation is the process of separation of the audio recordings into units of homogeneous parts, such as speech, silence, and laugh. Diarization, on the other hand, is used to extract various characteristics of signals such as speaker id, gender, channel type and the background environment (e.g., noise, music, silence). MAGiC addresses both methodologies since both segmentation and identification are indispensable components of face-to-face conversation.

We extended the CMUSphinx Speech Recognition System for analyzing recorded speech. CMUSphinx is an open-source, platform-independent and speaker-independent speech recognition system. CMUSphinx is integrated with LIUM[11], an open-source toolkit for speaker segmentation and diarization. The speech analysis process starts with feature extraction. As features, Mel-frequency Cepstral Coefficients, which collectively represent the power spectrum of a sound, are extracted by CMUSphinx functions. Afterwards, the speech segmentation, which is based on the Bayesian Information Criterion (BIC), is performed (Barras, Zhu, Meignier, & Gauvain, 2006; Chen & Gopalakrishnan, 1998).

Two passes are performed over the signal for the segmentation. In the first pass, a distance-based segmentation detects the change points by means of a likelihood measure, namely the Generalized Likelihood Ratio (GLR). In the second pass, the system mixes the successive segments of the same speaker (Meignier & Merlin, 2010) together. After the segmentation, BIC hierarchical clustering is performed with an initial set consisting of one cluster per each segment. At each iteration, ΔBIC_{ij} value for two successive clusters i and j is determined by Equation 3, from Meignier and Merlin (2010), as

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

where $|\Sigma_i|$, $\left|\Sigma_j\right|$ and $|\Sigma|$ are the determinants of Gaussians associated with the clusters i, j and $(i + j)$,. Here n_i and n_j refer to the total lengths of cluster i and cluster j. λ is the smoothing parameter to be chosen appropriately to get a good estimator and P is the penalty factor. ΔBIC values for each successive cluster are calculated and they are merged when the value is less than 0.

[11] LIUM Speaker Diarization Wiki,
http://www-lium.univ-lemans.fr/diarization/doku.php/welcome, retrieved on April 15, 2017.

As the next step of the speech analysis, the Viterbi decoding is applied for re-segmentation. A Gaussian Mixture Model (GMM) with eight components is employed to represent clusters (the parameters of the mixture are estimated using Expectation Maximization). However, the Viterbi decoding may not always determine the segment boundaries with high accuracy. This leads to some issues such as too long segments or segments that overlap within word boundaries. These issues are minimized by moving the segments slightly towards the low energy states and by cutting long segments iteratively so as to create segments shorter than 20 seconds. Up to this point, un-normalized features preserving background information are employed during segmentation and clustering processes. This facilitates to differentiate speakers and assign one and only one speaker to each cluster, however, one might need to place the same speaker in multiple clusters. GMM-based speaker clustering with normalized features is performed to assign the same speaker to the same cluster. GMM iterates until it reaches a pre-defined threshold value. Figure 17 shows the workflow of speaker diarization.

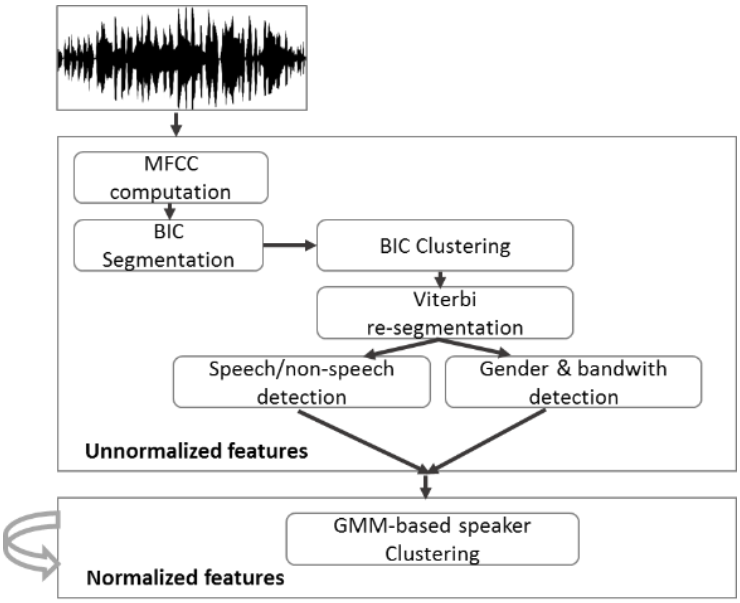


Figure 17: Classical process for speaker diarization and segmentation, adapted from LIUM Speaker Diarization Wiki Page[12]

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 for researchers. Thus, we carried out additional development to generate an audio segment from the non-segmented parts automatically. To achieve this, we compared the time interval of each successive segment.

[12]Wiki page of LIUM Speaker Diarization is available under:
http://www-lium.univ-lemans.fr/diarization/doku.php/overview

If there is a difference between the end of the previous segment and the beginning of the next one, we have created a new audio segment that covers this time range. We have also added a new functionality for segmenting audio with specified intervals.

4.4. Framework and Usage

4.4.1. The Development

MAGiC is developed as a desktop application in C# programming language and using it does not require any programming skills. The user can select the desired function and set the analysis parameters from a graphical user interface. MAGiC comes with detailed help pages describing the step-by-step procedure to run each function. Each page specific to a function briefly states the purpose of its use, characterizes the input parameters and if the parameter is a file, there exists a link to access a sample file, otherwise, the user is supplied with a sample value. Moreover, tooltips are created for almost all fields in the interface to enhance usability. Prior to running a function, data validation is performed in order to ensure the correctness and the consistency of data. If validation fails, user-friendly error messages are displayed close to non-validated fields. Similarly, the user is informed of the success status.

Separation of Concerns (SoC) design principle (Reade, 1989) is adopted to achieve high cohesion and low coupling between the features. As explained in the previous section, the SoC approach enables users to focus on specific processes without employing all functionalities of a software program. For instance, MAGiC can be solely used for speech segmentation or face tracking. In the following subsections MAGiC's software architecture, system requirements, installation procedure and availability are mentioned.

4.4.1.1. Software Architecture

Figure 18 provides an overview of the architecture used in MAGiC. The home screen is the main panel of the application. It loads the user interface and initializes the controllers. There are four main modules: *Speech Analysis*, *AOI Analysis*, *Summary*, and *Walkthroughs*. The Graphical User Interface (GUI) is stored under the *View* folder and back-end classes are collected under the *Controller* folder. There is a one-to-one relation between the GUI and the related controller class. The *AOI Analysis* includes OpenFace and dlib executable files with dependent libraries. Similarly, *Speech Analysis* has executable files of CMUSphinx in it. All the components of interfaces are members of the base user-interface class which is the *ParentUI*. The home screen implements the *Navigation Listener* interface and the *ParentUI* has a *Navigation Listener* as a field. Such a structure enables the user to navigate between walkthrough-pages and the related function, and vice versa.

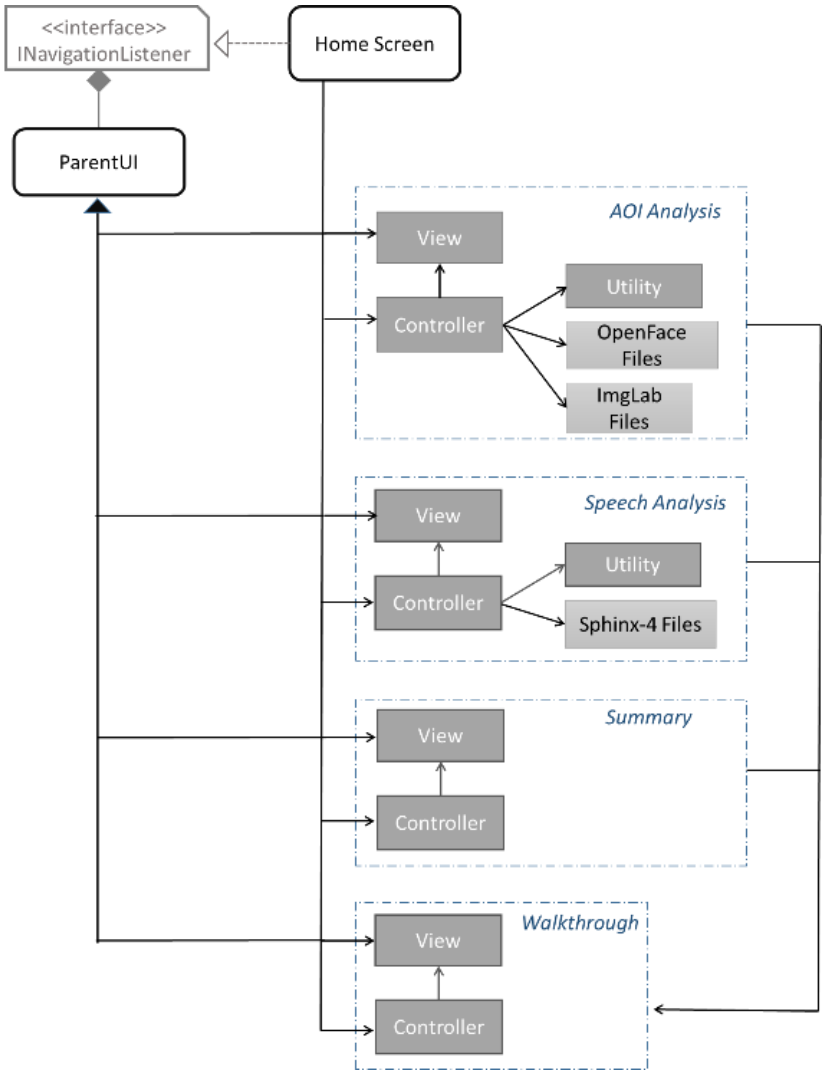


Figure 18: The Software architecture of MAGiC.

4.4.1.2. System Requirements

MAGiC is developed under .Net Framework 4.6 as a Windows Forms application and the operating system must be Windows 7 or later. There are different requirements for end-users and developers. System requirements for end-users to run MAGiC are as follows: (i) .Net Framework 4.6 (or a later version) which needs at least 512 MB Ram and 4.5 GB Disk Space, (ii) Visual C++ Redistributable for VS2015 (or later), requires 50 MB of available hard disk space, (iii) Java Runtime Environment-8 (or later) supposing that there is 128 MB memory, and (iv) MAGiC itself requires 250 MB of free disk space. For users who do not plan to make any improvement, these four items will suffice.

In order to contribute to the development of MAGiC, the developers should download the source code from the GitHub repository[13] and open it with VS2015 or with a more recent version. The .net Framework 4.6 and Visual C++ Redistributable for VS2015 are prerequisites. MAGiC also involves 3rd party toolkits such as dlib, OpenFace and CMUSphinx, as introduced in the previous sections.

4.4.1.3. Installation Procedure

An installation file is provided for end-users at the GitHub repository. The installation first checks whether the system meets the software prerequisites and it installs the missing ones in case it is needed. Once the initial requirements have been met, MAGiC is installed, a shortcut for Start menu item is created and MAGiC is launched automatically[14].

An open-source developer can re-deploy the employed 3rd party tools, then complete additional developments. CMUSphinx is available as a Gradle package. Among other things, the build.gradle file holds the project's description, version and its dependencies to external libraries. The Gradle FatJar plugin allows creating a JAR file with all dependencies. It is necessary to copy the generated JAR file (namely sphinx4-core-all-1.0) to the CMUSphinx under the speech_analysis folder in the parent directory. The feature extraction project in OpenFace and the imglab project in dlib are also employed in MAGiC. In order to reflect the OpenFace modifications to MAGiC, the user should copy the release version of FeatureExtraction.exe into OpenFaceFiles under the aoi_analysis folder in the parent directory. Similarly, changes in the dlib framework can be adapted to MAGiC by copying the release version of generated imglab.exe into imglabFiles under the aoi_analysis folder in the parent directory.

4.4.1.4. Availability

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 on it and contribute to its development. In addition to MAGiC, we also provide particular modified component toolkits (OpenFace for face tracking, dlib for the training of a custom face detector, and CMUSphinx for speech segmentation) on MAGiC's GitHub repository.

---

[13] MAGiC_v1.0, https://github.com/ulkursln/MAGiC/releases, retrieved on May 11, 2017.

[14] MAGiC App. Channnel, https://www.youtube.com/channel/UC2gvq0OluwpdjVKGSGg-vaQ, retrieved on May 11, 2017

4.4.2. Usage and Modules

4.4.2.1. The Main Control Panel

The main control panel shows up after launching MAGiC. It consists of two panels split by a collapsible splitter. On the left panel, links to the modules are listed and on the right one, all the functions of the module selected from the navigation pane are presented. The main workspace-area is under the right panel. The user can create a larger one on the right-hand side by collapsing the left panel (see Figure 19).

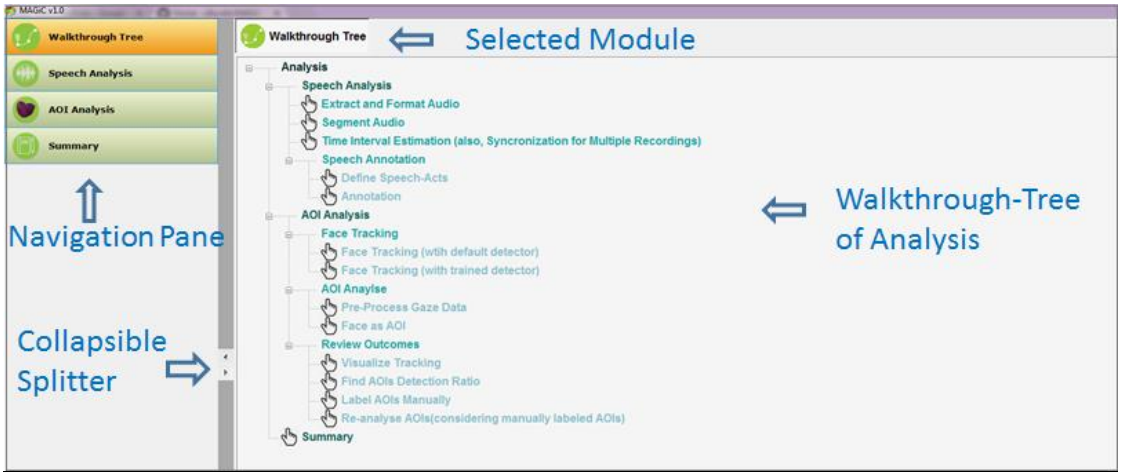


Figure 19: The main panel of MAGiC

4.4.2.2. The Walkthrough Module

The main page of the walkthrough module is in the form of a tree structure which represents the hierarchical nature of (cf. the right panel in Figure 19). The leaf nodes in blue color provide access to the related walkthrough page. Walkthrough pages provide a brief description of the corresponding function and step-by-step guidance on how to conduct analysis. It is possible to navigate from the walkthrough page to the related function and vice versa.

4.4.2.3. The Speech Analysis Module

For the sake of usability, the functionalities of each module are presented sequentially. Within each module, accordion panels are arranged in accordance with the procedure sequence (see Figure 7). For instance, the speech-analysis module consists of four consecutive processes: (i) formatting and extraction of audio, (ii) segmentation of audio, (iii) time interval estimation and (iv) speech annotation. In case of traveling the sequence in reverse order: for annotating speech data gathered from an experiment, time interval of an experiment must be specified (iv – iii); time-interval-estimation requires segmentation of audio (iii – ii); and at the top, for segmentation, the audio must be extracted from video and it must be formatted appropriately (ii – i). Each function produces the output files that

might be needed for the next step so that the user does not have to perform all the functions in a single session. Also, the user will have the opportunity to examine the outputs generated at each step and make further analysis, thus increasing the usage diversity.

(1) Formatting and Extraction of Audio: As the first step, the "Extract and Format Audio" panel is used to extract an audio file from an input video file in AVI format. It also allows formatting the audio for further CMUSphinx analysis. Sphinx4 requires '16 kHz', '16 bit', 'mono' and 'little-endian' wav files. The user has to select a video file and specify the output folder where the extracted and formatted audio file will be saved under. Both fields are mandatory and as is the case for all other functions, mandatory fields are validated when the operation button is pressed. Unless specified as "optional" in GUI, all the fields are mandatory and must be entered a value.

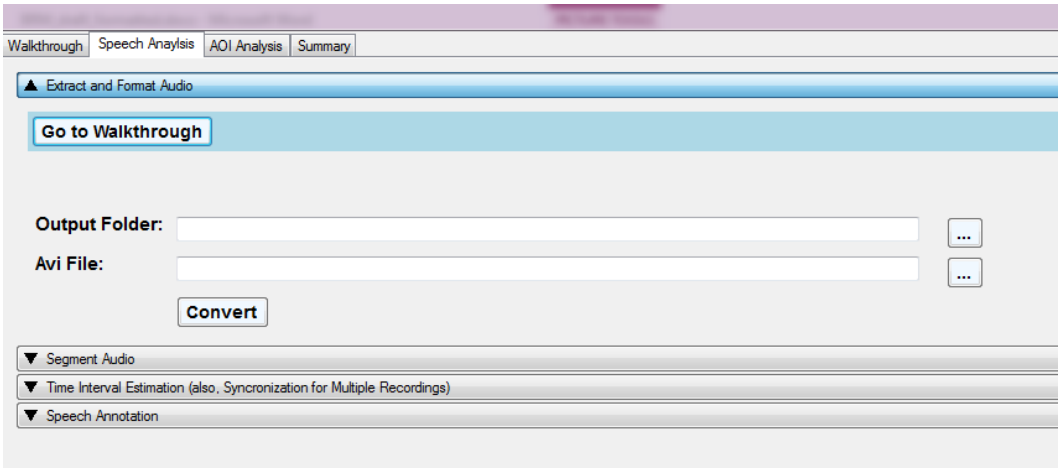


Figure 20: A sample interface showing the accordion panels.

(2) Segmentation of Audio: As the second step, the "Segment Audio" panel offers an interface to segment the audio file into smaller chunks including sub-words and pauses. The user has to select the formatted audio file. As the output, audio-segments and a text file are generated. The text file contains the duration of each segment in milliseconds. The audio-segments are named automatically starting from zero and incrementing by one, to preserve the sequence information. For instance, if the audio is divided into 300 segments, the name of the last segment will be 299.wav.

*(3) Time Interval Estimation***:** As the third step, the "Time Interval Estimation" panel can be used for performing two tasks: time interval estimation of an experiment session and synchronization of pair recordings. Even though the segmentation is completed automatically, a human annotator is still needed to listen to audio-segments, in order to specify the first and the last segments of an experiment session recording. It is a recommended practice to start an experiment session with a distinguishable bleep signal to facilitate the identification of the initial segment in

the analysis phase. This panel is also necessary to provide synchronization of participant pairs' session recordings. In particular, recordings from different sources have to be synchronized to conduct synchronous data analysis in a dyad conversation setting. The GUI provides an interface for performing the time-interval-estimation of either a single participant (or a participant pair session. In the case of a single-participant, the user is expected to enter the unique id of the participant, select the segments-interval file, which has already been created in the previous step, and enter the experiment interval by typing the name of the first and the last segment. In the case of a participant pair, the same procedure is repeated for both participants.

When the experiment session is conducted with multiple recording devices, one of the major issues is the synchronization of the recordings. A method to ensure synchronization is to employ audio alignment software, which synchronizes multiple audio tracks. An alternative method is to set and synchronize the internal clock of the recording devices over a network by Network Time Protocol. Although both solutions work with high accuracy, eye tracker manufacturers do not provide synchronization solutions yet. In most cases, the device clocks are set manually, causing errors in precision.

MAGiC overcomes the synchronization problem by following a data-driven approach instead of a device-driven approach. It provides a semi-automatic method for synchronizing multiple recordings from a participant pair. In this method, the user is expected to specify the beginning of the experiment session by listening to the segments, after automatic speech segmentation. The time offset which provides synchronization is calculated from the difference in time between the starting moments of two device recordings. In the case of a participant pair, a re-segmentation option is provided. After synchronization, the user obtains two recordings with the same length and processed in the same environment.

A re-segmentation function is provided as an optional choice. Re-segmentation aims at improving segmentation quality by merging information from different segmentation processes. This applies to situations in which each participant in a pair has his/her own microphone. The closer the microphone is to a participant, the cleaner and more accurate the audio recording becomes. The re-segmentation process involves selecting a base audio recording, generating a segments-interval file, and re-segmenting the base audio recording based on the merged segments-interval file.

(4) Speech Annotation: At the final stage of the analysis, the “Speech Annotation” panel involves two sub-panels, namely “Define Speech-Acts” and “Annotation”. From the former, the user may select a pre-defined set of speech-act items, devise a new speech-act list from scratch, or update a list for the specific analysis. Selecting the latter, the user annotates segmented audio files with the pre-defined speech-acts. To do this, the user has to upload the list of the speech-acts, the audio-segments and

the segments-interval file. The interface embeds a media player control to play audio files, which also displays the audio time in milliseconds. The annotation process involves listening to the audio file and selecting the appropriate speech-act(s) along with the speaker identity from the list. The interface automatically appends a new line comprising the speech-act(s), speaker identity and the time-interval that corresponds to the speech-act; and then it plays the next audio file.

4.4.2.4. The Area of Interest Analysis Module

The AOI Analysis Module consists of three main consecutive steps: (i) face tracking, (ii) AOI analysis, and (iii) Review Outcomes. Similar to the Speech-Analysis module, the outputs are generated after each step for further analysis.

*(1) Face Tracking***:** In the first step, the "Face Tracking" panel offers two sub-panel options. These are "Face Tracking with the default detector" (henceforth, the default mode) and "Face Tracking with the trained detector" (henceforth, the training mode). The default mode employs the default face detector propped by the dlib toolkit for tracking the face in the uploaded video. The interface offers an option to visualize detected faces during tracking. The user selects a video file as an input and specifies an output folder where the output files will be saved. The output file options are listed in Appendix C. The success of face detection in the default mode is subject to a set of technical challenges, such as the positioning of the light source (e.g., sun or lamps). For instance, if the light source is behind the participant, the face detection performance is adversely affected. MAGiC provides the user with the interface for monitoring the efficiency of face detection through the "Review outcomes" panel and the "visualize tracking" option, both of which serve for detecting the success ratio of face tracking (the Review Outcomes panel is explained in the next section). Whenever the user is not satisfied with the efficiency of face tracking in the default mode, the training mode provides the opportunity to design a custom face-detector. The training mode is composed of three stages: exporting image frames, training and face tracking. The image-frames are automatically extracted from a user-specified video, stored in a folder in sequential order, and the image names are displayed as a list on the panel. The user then selects a group of images from the list and these images are stored in an output folder. In the training stage, the user hits the train button and labels the face in each selected image by drawing boxes around (see Figure 14). In the background, a variant of Support Vector Machine with a usual C parameter is used for training. The C parameter represents the tolerance value to the outliers and its default value is 1. Developers may select higher or lower values to avoid overfitting or underfitting (i.e., losing the generalization property) by simply changing the parameter given to set_c function in the dlib environment. Likewise, developers may try different epsilon values for the specification of the risk gap. The default value for epsilon is 0.01. Smaller epsilon values yield a more accurate SVM optimization but it will take longer to train. The outcome of the training is a customized face-detector. At the last stage,

the user tracks faces on face tracking sub-panel and checks if the customized face-detector is able to detect the faces efficiently.

*(2) AOI Analysis***:** The first step in the AOI module was the face tracking step, as explained above. The second step is the AOI analysis. The "AOI Analyse" panel has two sub-panels: "Pre-process Gaze Data" and "Face as AOI". The former offers an option to complete the missing data in a raw-data file generated by an eye tracker. Missing data are mostly due to blinks or temporary problems in the eye tracker recording process. The fill-in function fills in the data gaps via linear interpolation. The user has the option to specify the maximum gap length to be filled. The user also has an option to make drift correction to handle systematic errors in the raw gaze data. The latter, namely "Face as AOI", can be used for defining the boundary of the face and then testing the conversation partner's gaze direction. The face tracking data and the gaze data are synchronized by overlaying a two-dimensional landmark file on raw gaze data. MAGiC automatically annotates each image frame to identify whether a participant is looking at the interlocutor's face (viz. *in*), or away from it (viz. *out*). The relative positions of gaze data with respect to the face location are also stored. As shown in Figure 21, the image frames are theoretically divided into 9 (3x3) AOIs. If the gaze of the participant is inside the interlocutor's face, an *e* character is assigned as an AOI-label to denote in. Moreover, if the participant is looking at the interlocutor's face, the face area is divided into three regions (mouth, nose and eyes) and the region at which the participant is looking is stored. On the other hand, if the gaze location is outside the face boundary, 1 of 8 character values, namely, *a, b, c, d, f, g, h, i* that corresponds to the gaze region is assigned as an AOI-label to denote *out*.

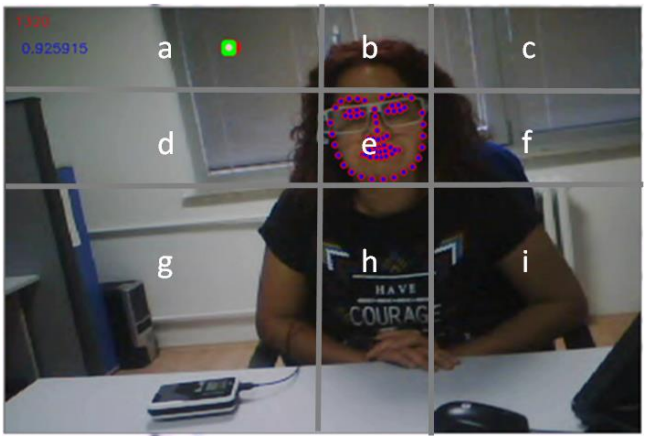


Figure 21: The AOIs specification. The green dot on the upper left AOI shows the gaze location of the conversation partner.

(3) Review Outcomes: In the AOI Analysis module, the third step allows the user to review the outcomes of the first "Face Tracking" and the second "AOI Analysis" steps, as briefly mentioned above. The "Review Outcomes" panel is composed of the following sub-panels: "Visualize Tracking", "Find AOIs Detection Ratio",

"Label AOIs Manually" and "Re-analyze AOIs (considering manually labeled AOIs)".

(3.1) Visualize Tracking Sub-panel: The interface can display the video file overlaid by facial landmarks, AOI-label and gaze point location. There is an option to specify and correct systematic drift errors of the eye tracker, another option to specify a confidence threshold for facial landmark accuracy, and a third one to visualize a specified recording session.

(3.2) Find AOIs Detection Ratio: This sub-panel indicates the number and the percentage of labeled image frames. For this, the user selects the AOIs file, the experiment-interval-file, enters the participant id and the frequency of an eye-tracker. In case the detection-ratio is below the expectations, the user may re-track the face in the training mode or label the AOIs manually, as described below.

(3.3) Label AOIs Manually: The interface for manual labeling provides the user with an environment for labelling the AOIs in a fast and efficient way, via a keypad (Figure 22). The keypad buttons correspond to the nine AOIs illustrated in Figure 21. Navigation between the frames is via arrow keys on the keyboard.

Figure 22: AOI labels associated with keypad numbers

(3.4) Re-analyze AOIs: This panel offers a function for merging automatically extracted and manually labeled AOIs.

4.4.2.5. The Summary Module

This module allows the user to combine the data obtained in speech and AOI analyzes into a single file. Each line of the generated text file corresponds to each frame of video-recording. It contains time interval in milliseconds, the speech act, speaker identity, AOI-label of the participant if there is only one or AOI-labels of each participant (for a pair) along with the coordinates of the minimum bounding rectangle of the detected faces and raw gaze data.

4.5. A Pilot Study

This section reports a pilot study to demonstrate the functionalities and benefits of MAGiC. The setting is a mock job interview setting, in which a pair of participants wear eyeglasses and conducts the interview. The gaze data and the video data are then analyzed by MAGiC.

4.5.1. Participants

Three pairs of participants (university students as volunteers) took part in the pilot study (mean age 28, SD = 4.60). The task was a mock job interview. The participants were assigned the role of either an interviewer or an interviewee and the roles were distributed randomly. All the participants were native Turkish speakers and they had a normal or corrected-to-normal vision. The participants were not restricted by any time constraints.

4.5.2. Materials and Design

At the beginning of the session, the participants were informed about the task. Both participants wore monocular Tobii eye-tracking glasses at 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-marker calibration process was carried out at 100 cm distance to the participant. After the calibration, the participants were seated on opposite sides of a table, 100 cm away from each other. A beeping sound was generated to indicate the beginning of a session.

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

4.5.3. 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 with an Intel Core i5 2410M CPU at 2.30 GHz with 8 GB RAM, running Windows 7 Enterprise (64 bit) was used.

4.5.3.1. Speech Analysis

Firstly, the "Extract and Format Audio" function 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 .wav files were created. Each run took 1 to 2 seconds. Secondly, 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 (see Table 7). Each run took 1 to 2 seconds.

Table 7: Audio length and number of segments for each participant's recording.

	Interviewer		Interviewee	
	Audio Length (m:ss.ms)	# Segments	Audio Length (m:ss.ms)	# Segments
Pair-1	3:46.066	170	3:57.000	176
Pair-2	5:25.066	120	5:40.000	200
Pair-3	5:28.000	246	5:09.000	208

Thirdly, time-interval estimation, synchronization and re-segmentation were performed for each pair through the "Time Interval Estimation" panel. At the beginning, we listened to audio-segments to specify the start and the end of the segments of the recording sessions. Whenever needed, the re-segmentation process guaranteed time synchronization of the segments by utilizing synchronization information and merging segments from both recordings. After the re-segmentation, we ended up with equal-duration of the session for participants within each pair. Table 8 lists the experiment duration in milliseconds and the number of segments produced after re-segmentation. Each run took 1 to 2 seconds.

Table 8: Experiment duration and corresponding segment-numbers

	Exp. Duration (ms)	# Segments
Pair-1	182,410	261
Pair-2	305,420	282
Pair-3	282,620	406

Finally, speech annotation was performed. A list of speech-acts was defined as the first step of the analysis. Below is a list of predefined speech-acts:

- 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 the end of the speech (e.g., "that is all")

The next step was the annotation process. This process involved selecting the Speech-act(s) and hitting the “annotate” or "annotate and play next" icons placed next to the speech-act list. At each annotation, a new line was appended and displayed, which contained the relevant segment's time-interval, the associated participant if any and selected speech-act(s). This step was repeated for all three pairs of participants. Each run took 10 to 20 minutes, depending on the session-interval.

4.5.3.2. AOI Analysis

All six videos were processed with a default-mode face detector. The tracking processes produced two-dimensional landmarks on the interlocutor’s face image. This process took 4 to 10 minutes per video, depending on its length. Then, the gaps with at most two frames duration in the gaze raw-data file were filled in by linear interpolation. The raw data file included the columns for the frame number, gaze point classification (either Unclassified or Fixation), and x and y coordinates. The processed data comprised 2% of the total gaze raw-data (see Table 9). The gap-filling process took less than a second per pair.

Table 9: The number and the ratio of the filled gaps for each participant’s raw gaze data.

	Interviewer/Interviewee	
	Number of filled gaps	Ratio of filled gaps (%)
Pair-1	146 / 236	2.15 / 3.32
Pair-2	171 / 236	1.75 / 2.31
Pair-3	157 / 335	1.6 / 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 during face tracking was 720 × 480 pixels, while the eye tracker image-frame resolution was 640 × 480. The eye-tracking glasses had a reported degree of accuracy of half a degree of visual angle. The recording angles of the built-in scene camera of the eye-tracking glasses were 56 degrees horizontal and 40 degrees vertical. The seating distance between the participants was approximately 100 cm. The corresponding eye tracker accuracy was 4.84 pixels in horizontal and 5.34 pixels in vertical. The AOI detection took a couple of seconds. After AOI detection, the “Find AOIs Detection Ratio” panel was used for reviewing the AOI detection accuracy. Table 10 lists 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 interviewers’ recordings. Nevertheless, face detection was performed with an average success rate over 90%.

Table 10: The number and the ratio of image-frames in which face could not be detected.

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

Another reason for AOI detection failure is the absence of gaze data. Table 11 shows the ratio of undetected AOIs due to the absence of gaze data.

Table 11: The number and the ratio of image-frames for which raw gaze data were absent

	Interviewer/Interviewee	
	Number of absent data	Ratio of absent data (%)
Pair-1	3237 / 392	59.1 / 7.16
Pair-2	4762 / 1050	52.0 / 11.5
Pair-3	4010 / 1732	47.3 / 20.4

The failure in AOI detection on the interviewer's side is about 50% or more owing to the experimental setting, which requires the interviewer to look at the questions to read them. This is a common situation that experiment designers face frequently in dynamic experiment settings. MAGiC's interface allows the user to detect the source of the problem through the "Visualize Tracking" panel. 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 from the interlocutor's face i.e., out) on top of the video recording for each frame, as shown in Figure 23: An image taken during the visualize-tracking process..

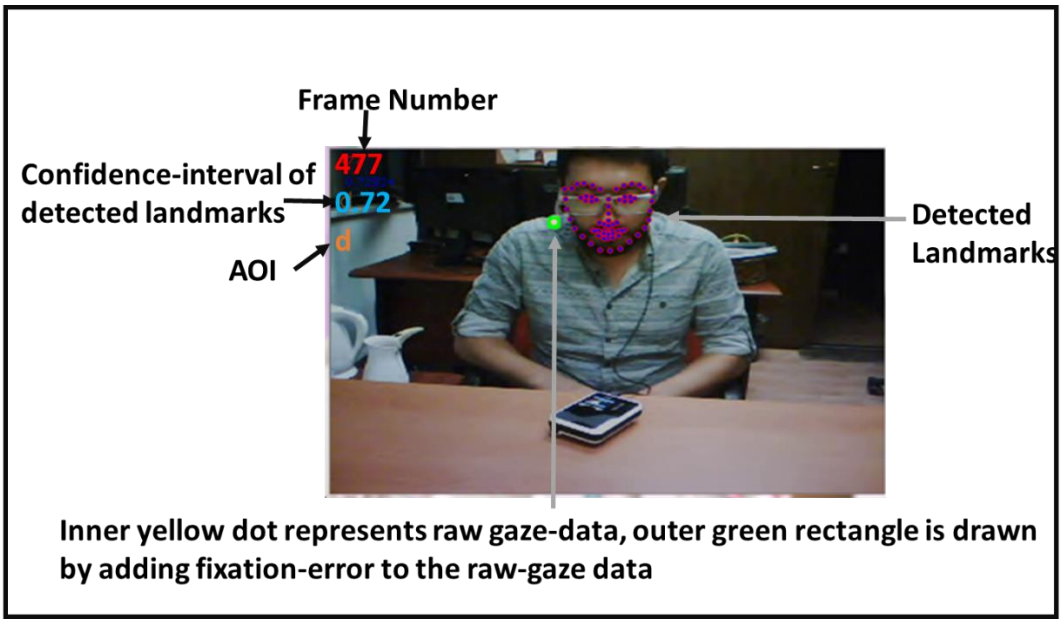


Figure 23: An image taken during the visualize-tracking process.

Our analysis of the scenes through the Visual Tracking panel revealed that the missing raw-gaze data originate from the interviewer's reading and articulation of the questions as well as the evaluating process of the interviewee's response using paper and pencil. This might be expected since the glasses lie outside the field of view of the interviewer while looking at the notebook (Figure 24). This situation exemplifies the practical difficulties that researchers face when conducting experiments in dynamical environments. To cope with these situations, the user may use MAGiC's manual-labeling function. Our manual annotation took 15 to 20 minutes per pair, depending on the length of the video and the number of missing AOI-labels.

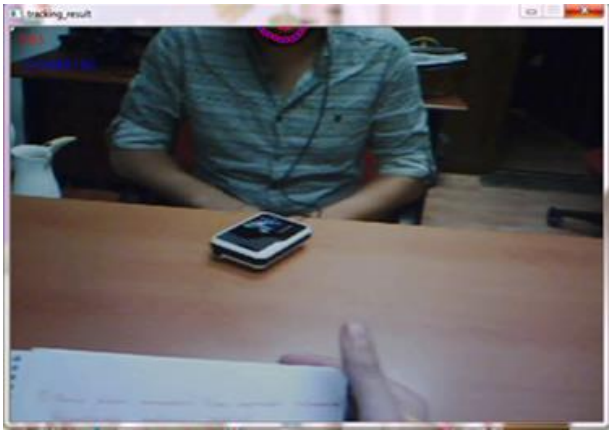


Figure 24: An image-frame captured while the interviewer was articulating a question.

The final step in the AOI-analysis was conducted by running the re-analyze and the find-detection ratio function. The re-analysis step automatically merged the detected AOIs with

the manually extracted AOI-labels. The find-detection ratio function was run to compare the detection ratio with the previous outcomes. Table 12 shows face-detection and gaze-detection accuracy for the interviewer's recordings. The results reveal an improvement of more than 30% compared to the previous analysis steps (cf. Table 10 and Table 11).

Table 12: The number and the ratio of the image-frames in which face and gaze could not be detected

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

4.5.3.3. Summary

The data obtained in speech and AOI analyses were merged into a single summary-file and using it, we calculated the percentages of gaze locations, such as in-out, AOIs, speech acts, as well as the mutual gaze behaviors of the conversation pairs. The analyses finally revealed information about 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 (Figure 25).

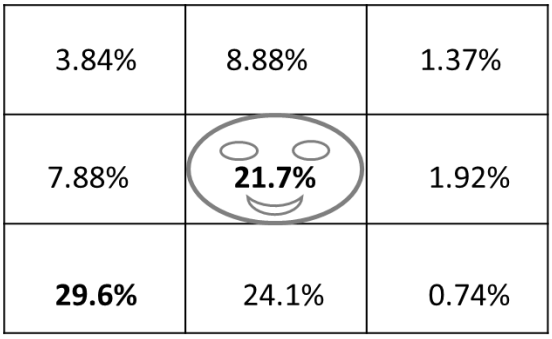


Figure 25: The distribution of gaze behavior with relative location to interlocutor's face. At 21.7% of the sessions participants looked at the interlocutor's face. The bottom left corner was the most sighted region with 29.6%.

The rightward shifts are usually associated with verbal thinking, whereas the leftward shifts are usually associated with visual imagery (Kocel, Galin, Ornstein, & Merrin, 1972). On the other hand, 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 listening to a question from their partner (see Ehrlichman & Micic, 2012, for a review).

A further investigation of the 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 the 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 a significant difference between the frequency of *Mutual Aversion* and the frequency of *Mutual Face Contact* (M=0.01, SD=0.02, $p=.02$).

4.6. 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 number of eight participants. The participants performed data analysis using publicly available sources (see Supplementary material[15]) The usability analysis was conducted in three steps, as described below:

i) Perform the analysis manually,
ii) Perform the analysis by using MAGiC,
iii) Assess the usability of MAGiC using a 7-point scale ISO 9241/10 questionnaire (see Appendix D).

We recorded the time spent to execute data analysis, and then compared it with the average duration when the participants performed the same analysis manually. When manual

[15] See the MAGiC App Channel under Youtube, https://www.youtube.com/channel/UC2gvq0OluwpdjVKGSGg-vaQ, and MAGiC App Wiki Page under Github

annotation is replaced by MAGiC, in the AOI analysis, the mean duration to annotate a single frame decreased from 29.1 seconds (SD=22.7) to an average value of 0.09 seconds (SD=0.02); and in the speech analysis, the mean duration for a single annotation decreased from 44.5 seconds (SD=8.8) to an average value of 7.1 seconds (SD=1.4). The Usability test scores are plotted in Figure 26.

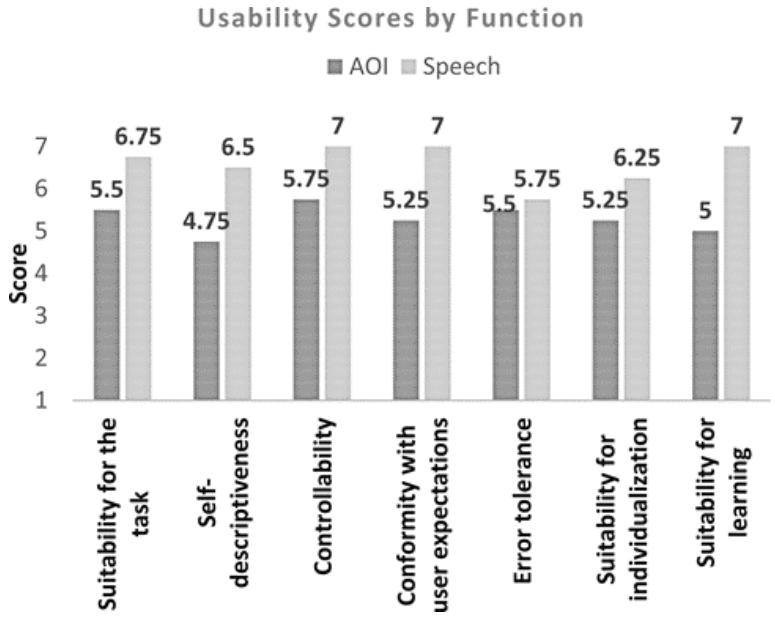


Figure 26: Usability Scores by function. All of the usability metrics were scored higher than the average.

4.7. Conclusion

In the present chapter, we introduced an environment, namely MAGiC, that allows researchers to analyze the gaze behavior of participants in a 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 the interlocutor's face in a video recording, automatically. Then it overlays gaze location data over the face to detect face 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 analysis of video data. Its capability for automated multimodal (i.e., gaze and speech) 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 manually by providing labeled images.

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 by extending their capabilities further for the purpose of detecting eye movement behavior, and for annotating speech and gaze data simultaneously. MAGiC's user interface is composed of a rich set of panels, which provides the user with an environment to conduct a guided, step-by-step analysis.

MAGiC is able to process data from a single eye tracker or 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 steps on the recorded data in the pilot study. We intentionally employed a low-frequency eye tracker, with relatively low video quality in a low-illuminated environment, as these are typical real environment challenges that affects the face tracking capabilities. Our analysis revealed that MAGiC is able to exhibit acceptable success ratio in its automatic analyses under those challenging conditions, with an average AOI labeling (i.e., face contact and gaze aversion detection) efficiency of 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. We also note that MAGiC's speech analysis component, namely CMUSphinx offers several high-quality acoustic models, although there is no pre-build acoustic model for Turkish. Despite this challenge, MAGiC returned successful results for speech analysis. The speech-act annotation also helped us to overcome speech segmentation issues by providing sub-segments for speech segment intervals.

All the data analyses were completed in approximately two hours for the three pairs of participants. Our estimations reveal that the time and effort that would have been spent on manual frame-by-frame video analysis and speech segmentation is much more, in addition to their disadvantage due to human annotator errors.

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, the face-detection ratio may be increased by employing recently-published OpenFace 2.0. Besides, in its current version, MAGiC sets an AOI-label on the interlocutor's face image. We plan to expand this labeling method so that it could 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 which are 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 synchronized recordings, in its recent version. Further development of MAGiC will address to improve its synchronization capabilities; its capability to transcribe speech into text to train speech-act annotation with pre-defined speech acts and to automate subsequent annotations. We believe that MAGiC has a chance to take advantage of being an open-source tool for behavior research, and expect further development from the community, in addition to our plans for its development.

CHAPTER 5

ANALYSIS OF GAZE AND SPEECH IN FACE-TO-FACE INTERACTION

The design and analysis of the second experiment were updated based on the experience gained from the pilot study. Similar to the previous one, in this chapter, materials, methods, and data analysis steps of the study were summarized. This time we deepened speech analysis, not only by improving tag set used for speech annotation but also by adopting an alternative annotation method, namely dialogue-act annotation.

5.1. Materials and Design

5.1.1. Participants

Seven professional interviewers, 4 female (mean age= 33.8, SD=4.72) and 3 male (mean age= 35.7, SD=0.58), with the mean age of 34.6 (SD=3.51); and 28 interviewees, 14 female (mean age=25.1, SD=2.57) and 14 male (mean age=25.4, SD=2.68), with the mean age of 25.3(SD=2.58) took part in the study. Interviewers interviewed with four participants on average. Participants in each pair did not know each other beforehand. All the participants were native Turkish speakers and had a normal or corrected-to-normal vision. (see Appendix E for detailed information about the pairs).

5.1.2. Apparatus

Both participants in a pair wore monocular Tobii eye-tracking glasses, which had a sampling rate of 30 Hz with a 56°×40° 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 threshold for the accuracy of IR-marker calibration process was 80%. Interviewers read the questions and evaluate the interviewee's response on a Wacom PL-1600 15.6 Inch Tablet which enables users to interact with the screen by using a digital pen.

5.1.3. Procedure

At the beginning of the experiment, participants were informed about the task. Additionally, in order to motivate interviewees for an interview, we asked them to think of a job-position that they were interested in. The experimental protocol was the same as the pilot study, (see Figure 3). Besides, interviewers used a digital pen and a Wacom Tablet placed vertically on the table, instead of paper and pen. Unlike the pilot study, this time, the room was uniformly illuminated and the objects all around were diminished for a less noisy background. Also, there was no time limit. Interviewers were free to ask further questions if they think it was necessary.

5.2. Data and Analysis

We excluded the data of three pairs for which the gaze data collected during a session is less than 70%. There were similarities with the analysis of the pilot study. Additional improvements were the followings: (i) We employed MAGiC in many steps of gaze and speech-tag set analysis, (ii) we extracted face contact as well as gaze aversion via a newly developed application, (iii) we reviewed and updated the proposed speech-tags, (iv) as well as speech-tag set annotation, we annotated speech with an alternative dialogue-act model and employed Praat for marking time intervals for words, (v) finally, we developed computational models by using either speech-tag set or dialogue-act annotations.

5.2.1. Speech-tag Set Analysis

"Extract and Format Audio" function of MAGiC was run to extract audio-stream from the video files obtained from each participant's recordings and to convert the audio into 16 kHz, 16 bit, mono and little-endian format (for command see Chapter 3 - Data and Analysis).

5.2.1.1. Segmentation and Synchronization

As the first step, the "Segment Audio" function of MAGiC was run to segment the audio file into smaller chunks including sub-words and pauses. Audio-segments and a text file that contained the id number and the duration of each segment were generated for each audio-stream. The mean duration of the recordings was 09:41.543 (SD=04:05.418) (in mm:ss.ms format).

Then, in order to determine session intervals, we listened to audio-segments and identified the start and the end of each session. Next, "Time Interval Estimation" function of Magic was run to provide synchronization of pair recordings. Lastly, we re-segmented synchronized pair recordings by merging segmentation information of each participant in a pair in order to improve segmentation quality. The number of segments varied depending on the length and the content of the audio (M=737.4, SD=414.1), see Appendix F

5.2.1.2. Annotation

At the final stage of Speech-tag set analysis, we firstly defined speech acts by using the related MAGiC function. The speech-tag list is given as follows.

Speech: Includes the speech itself. It is a type of commissive or declarative speech-act.

Speech While Laughing: We suppose that it might be different from regular speech, regarding gaze behavior. That is the reason for which we added this item to the list.

Asking a Question: Speaker request for information. It is a type of directive speech-act.

Confirmation: Act of verifying or making something certain. It is a type of representative speech-act.

Pre-Speech: The non-speech instance which includes the silence before the speech and the sounds for warming up the voice.

Speech Pause: Includes the pauses during the course of speech.

Micro Pause: Represents gaps up to 200 ms. We add this item because it was different from *Speech Pause*, as proposed by Heldner and Edlund (2010).

Thinking: We name the conversation segment as thinking when it included filler sounds, such as uh, er, um, eee, and drawls – the nonphonemic lengthening of syllables.

The Repetition of Question: We suppose that it might be different from the instances of *Speech* and *Thinking*, regarding gaze behavior, due to the fact that the participant who repeats the question is both thinking about the question and confirming whether understood the question correctly.

Signaling End of Speech: The conversation segments that include phrases signifying the end of the speech, such as that's all, were annotated with this item.

Questionnaire Filling: The interviewer evaluates the interviewee after each question by looking at the monitor and using a digital pen. This category is specific to interviewers.

Greeting: An action of giving a sign of welcome or to express pleasure. It is a type of expressive speech-act.

Read Question: Interviewers ask a question by reading from the monitor. We distinguish asking a question from reading it because looking at the monitor would obviously affect the gaze behavior of an interviewer. This category is specific to interviewers

Laugh: It is generally a sign of joy or positive feedback.

Later on, we annotated segments of each session by using the "Annotation" interface of MAGiC.

5.2.2. Dialogue-act Analysis

Since we conducted an experiment with two participants, dialogue act annotation might be a good alternative during the speech analysis. We employed ISO 24617-2 standard for this purpose, which was lastly updated in September 2017. Dialogue act annotation is a process that involves the following steps: (i) segmentation of dialogue into grammatical units that have a communicative function and a semantic content; (ii) assigning of communicative function labels to each segment, see Figure 2 and Table 3 for the entire list of communicative functions (Bunt, 2012; Bunt, Petukhova , & Fang, 2017).

5.2.2.1. Transcription

We transcribed text of each session into a file by listening audio-stream of both participants in a pair separately. We created a single document for each session. We first opened a Google Document and enabled speech to text feature, then started to articulate audio while listening to the interviewee's audio-stream. After that, we listened to the same recording once more so as to add non-verbal vocalizations and punctuation.

We added non-verbal vocalizations as it was proposed in the dialogue act manual (Augmented Multiparty Interaction Consortium (AMI), 2005) and related previous study (Trouvain, & Truong, 2012). Non-verbal vocalizations might be a critical clue while selecting dimension, communicative function and qualifier. Therefore, it is necessary to consider them in the segmentation and annotation phases. The list of non-verbal vocalizations that we added in the transcribed text is given below:

Unfinished Word: Depending on the context, it might be a member of either *Self-correction*, *Stalling* or *Retraction*.

Filler Sound: Such as uh, er, um, eee. It is a member of *Stalling*.

Confirmation Sound: Such as hıhı, hımm. It is a member of either *Agreement* or *Auto-Positive*.

Laughing While Speech: Sentiment qualifier might be joy.

Laugh: It might be a member of *Auto-positive***.** In addition**,** sentiment qualifier might be joy.

Drawls: The nonphonemic lengthening of syllables. It is a member of *Stalling*.

Warm-up: The sounds for warming up the voice. It is a member of *Turn Take* or *Pause*.

Breathing noise: It might be a member of *Stalling* or *Turn-Take*.

Then, while we were listening to the interviewer's audio-stream for the same pair, we completed missing words in the transcription text file of a session. Thus, we reviewed the transcription of a session twice in this phase. Lastly, we divided the transcription text file into two separate files based on the source. As a result, at the end of the Transcription phase, two files per session were created in total, one for the interviewer's transcription and other for the interviewee's. The workflow of the transcription phase is presented in Figure 27 below.

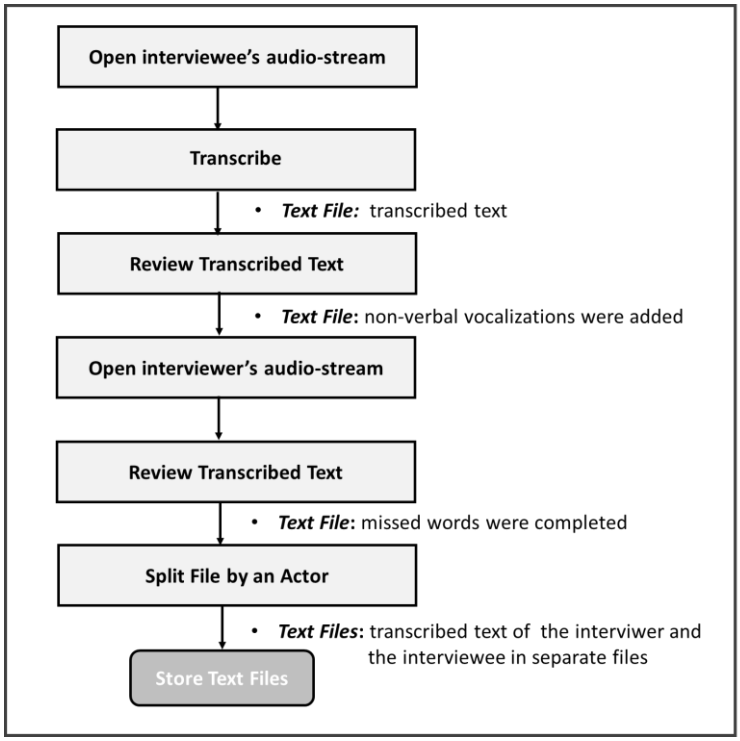


Figure 27: The workflow of transcription phase

5.2.2.2. Time Intervals and Synchronization

Secondly, time intervals of each word were marked by using Praat program[16]. Three students took part in this study and 16716 words in 15 sessions were processed in total. We selected 15 sessions by giving priority to longer ones where communication, and hence dialogue acts and RRs were more frequent.

Praat is a free application for speech analysis in phonetics. It has lots of functions for speech analysis but we employed only the "Transcribing speech with Praat function". As we have already transcribed audio-stream, the word or non-word vocalization was copied from the transcription file and pasted into the related area in an interface. Then, the time interval of a word was specified by marking the beginning and the end. As an initial process, we specified the ending time of the beeping sound and marked it as beginning of a session. We did the same for both interviewer's and interviewee's audio-stream. Then, time offset to provide synchronization was calculated based on the time difference between the starting moments of the audio streams of interviewer and interviewee in the same pair.

Even though we reviewed the transcript text twice in the previous phase, there would still be some missed words or non-word vocalizations. In such cases, the transcription file was updated with the missing word and/or non-word vocalization. In addition to that, after each word was processed, a controller checked if it was necessary to update the time-intervals of words and transcribed texts. Thus, the transcription file was reviewed four times in total since its creation and word-intervals were checked twice.

Lastly, we merged the transcription files of the interviewer and interviewee in a session. As a result, we are left with a single transcription file per session at the end of this phase. The workflow is presented in Figure 28 below:

[16] For detailed information about Praat: http://www.fon.hum.uva.nl/praat/

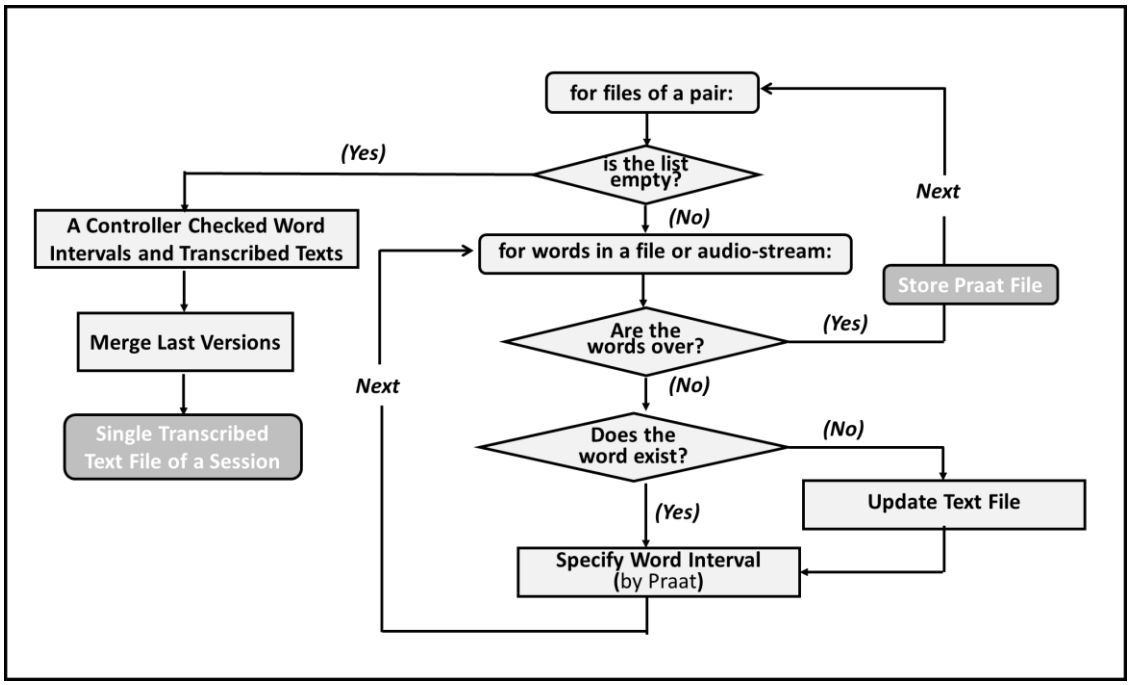


Figure 28: The workflow for generating final version of transcriptions

5.2.2.3. Segmentation and Annotation

We segmented speech utterances into dialogue act units. As proposed by Prasad and Bunt, (2015), dialogue act units were determined based on the meaning rather than the syntactic features. Since we were investigating the relation between dialogue act units and gaze behavior which was able to change quite fast, we specified dialogue-act units in smaller intervals that differed from the previous and the subsequent dialogue-act units in terms of communicative function, qualifiers and RRs.

Before starting the annotation, we studied the ISO 24617-2 standard, its revised version and the manual of annotation (Augmented Multiparty Interaction Consortium (AMI)., 2005; Bunt, Kipp, & Petukhova, 2012; Bunt et al., 2017; ISO, 2012). Furthermore, we reviewed the sample annotations given in DialogBank (Bunt et al., 2018). DialogBank was developed by Tilburg University and contains a collection of dialogues following the ISO 24617-2 standard. After that, based on the information obtained from these sources, we created three yes-no decision trees[17] to provide a tool for assigning the (i) Dimension and Communicative Function, (ii) Certainty Qualifier and, (iii) Conditionality Qualifier. They were prepared in Turkish as they were created for the annotations of dialogues in Turkish.

[17] Available under https://github.com/ulkursln

A decision tree is a flowchart like diagram which represents the various outcomes of a series of possible decisions. A primary advantage of using a yes-no decision tree is that it is easy to follow. A decision tree has three main parts: root node, leaf nodes, and branches. The root is the starting point of a tree. The root and leaf nodes prompt the user to answer to a yes-no question. Branches consist of arrows connecting nodes from questions to answers.

Even though ISO 24617-2 supports RR annotation, it does not specify any particular set for RR. Thus, we employed another standard recommended by ISO 24617-2 for the annotation of discourse relation. ISO 24617-8, also known as ISO DR-Core, was proposed as an international standard for the annotation of discourse relations (ISO, 2016; Bunt & Prasad, 2016; Prasad & Bunt, 2015). It provides a mapping of the relations among the existing annotation frameworks, including PDTB to the ISO DR-Core. This mapping enabled us to study previously annotated discourses following the PDTB framework. For instance, we benefited from the TED-Multilingual Discourse Bank, (Zeyrek et al., 2019). They followed the principles of PDTB and annotated TED-Talks in six languages including Turkish. In order to get more insight into discourse annotations in Turkish, we reviewed their annotations with PDTB Annotator, which is a tool for annotating discourse relations (Lee, Prasad, Webber, & Joshi, 2016).

ISO DR-Core proposed the markup language DiAML (Dialogue Act Markup Language) with the representation format using XML. Instead of DiAML-XML format, which is computer-friendly, we adopted an alternative one, human-friendly tabular representation, namely DiAML-MultiTab representation. According to DiAML-Multitab representation, an annotator has to assign the unique ID to each dialogue act. Moreover, if there is a functional or feedback dependence between two dialogue acts, intending to represent this relation, the ID of the preceding dialogue-act should be referenced by the succeeding one. Similarly, in case there is a RR between two dialogue-acts, the dialogue act which is the first argument of the RR, should be referenced by the other one. We developed an excel macro[18] to automatize the process of assigning unique ID's and updating references. Automation minimizes the error rate and improves performance. Suppose that you have made an update on the DiAML-MultiTab excel, for instance, you have added a missing dialog-act unit. In such a case, you would have to update all dialogue-act IDs and references accordingly, right after the line inserted. Since the excel macro took care of this process we did not perform any manual update. As a result, a single excel file in DiAML-MultiTab format was created for each session, at the end of this phase. The workflow is presented in Figure 29.

[18] It will be available under https://gist.github.com/ulkursln

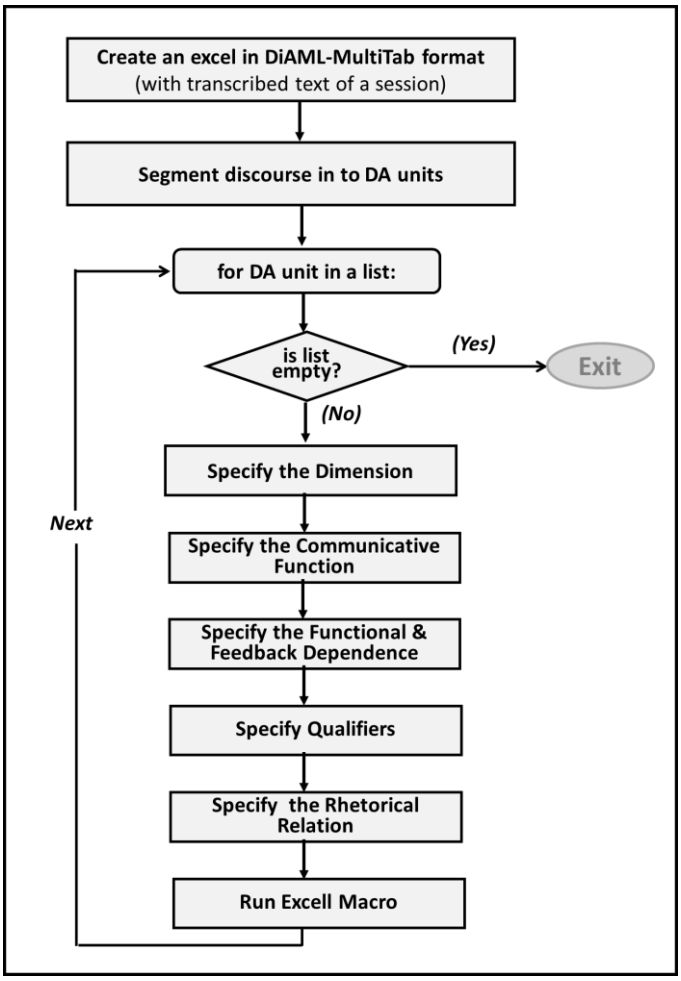


Figure 29: The workflow of segmentation and annotation

5.2.3. Gaze Analysis

We first exported the video from recordings by running the corresponding function of Tobii Studio. We obtained 56 video-streams for 28 pairs of participants. Tobii Studio supports AVI file format for movies. We converted AVI files to WMV before continuing analysis.

5.2.3.1. Face Detection

We run "Face Tracking with the default detector" function of MAGiC. Then, we extracted Area of Interest (AOI) labels corresponding to the frame-image along with the input parameter: (i) 2D landmarks of faces and (ii) linearly interpolated raw gaze data. We set the following parameters for AOI extraction function.

▪ The size of the captured images during face tracking was 720 × 480 pixels, while the eye tracker image-frame resolution was 640 × 480.

▪ The eye-tracking glasses had a reported degree of accuracy of half a degree of visual angle. The recording angles of the built-in scene camera of the eye-tracking glasses were 56 degrees horizontal and 40 degrees vertical. The seating distance between the participants was approximately 100 cm. The corresponding eye tracker accuracy was 4.84 pixels in the horizontal direction and 5.34 pixels in the vertical direction.

After that, in order to interpolate missing gaze data, first the scaling factor was calculated via Equation 4, then the location of the first sample after gap was multiplied by the scaling factor, and lastly the result was added to the location of the last sample before the gap. The max gap length that would be filled with interpolation was chosen to be shorter than a normal blink which was 75 ms as proposed by previous studies. (Benedetto et al., 2011; Komogortsev, Gobert, Jayarathna, Koh, & Gowda, 2010; Ingre, Akerstedt, Peters, Anund, & Kecklund)

$$s_{scaling\ factor} = \frac{t_{timestamp\ of\ sample\ to\ be\ replaced} - t_{timestamp\ of\ first\ sample\ after\ gap}}{t_{timestamp\ of\ last\ sample\ before\ gap} - t_{timesamp\ of\ first\ sample\ after\ gap}},$$

(Equation 4, taken from Olsen, 2012)

Then, we monitored the efficiency of face detection through the "Review Outcomes" panel. The "Find AOIs Detection Ratio" function gave the number and percentage of extracted AOI labels in frame-images. Furthermore, we trained a custom detector for the video-stream in case more than 30% of frame-images could not be assigned to an AOI-label. Then, we re-run the face-tracking function but this time with the trained detector. Again, we monitored the performance of face tracking and continued the analysis with the AOI labels which got the higher percentage rate, see Table 13.

Table 13: Performance of face-tracking with a trained custom detector

	#Face Detection <70%	#Trained detector better	#Default detector better
Interviewer	2	1	1
Interviewee	11	11	0

The detection of AOI-labels failed due to undetected faces and/or the missing gaze data. Eventually, we run the "Assign AOIs Label" function of MAGiC which enabled us to assign AOI labels to the frame-images manually. We assigned or updated the AOI labels for the following cases:

- The face of the interlocutor was on frame-image, yet it could not be detected automatically.

- The face of the interlocutor was on frame-image, but it was not detected correctly.

- The face of the interlocutor did not exist for that particular frame-image. This happens especially when an interviewer was looking at the monitor while evaluating the interviewee or reading the question. In such cases, if we already knew the relative position of an interviewee with respect to the monitor, we easily inferred AOI –label.

Lastly, after reviewing and updating the extracted AOI labels manually, we re-run "Find AOIs Detection Ratio" function and eliminated three pairs that had less than 70% of the assigned labels (the workflow is illustrated in Figure 30). Hence, we continued speech-tag set analysis with the remaining 25 pairs.

5.2.3.2. Gaze Behavior Detection

In the previous phase, we ended up with AOI labels corresponding to each frame-image. As mentioned in the MAGiC chapter, 1 of the 9 characters from *a* to *i* was assigned as an AOI-label. If the participant was looking at the face of the conversation partner, AOI-label would be *e*, otherwise it would be one of the eight remaining characters. At the beginning of this phase, we matched the label *e* with face contact and the other labels with gaze aversion. Currently, there is no commonly accepted method for fixation identification in dynamic scenes (Munn et al., 2008; Srinivasan et al., 2014).

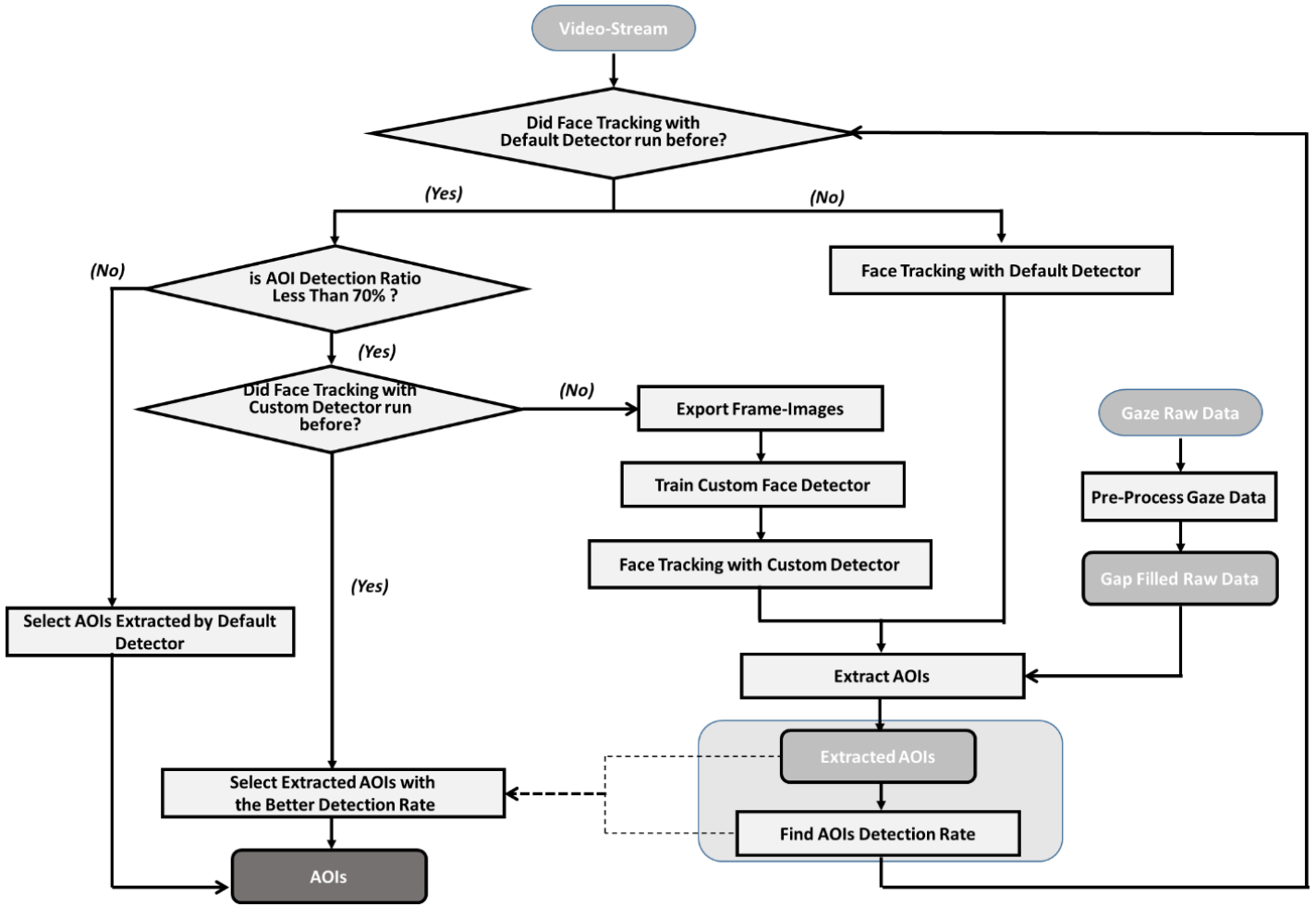


Figure 30: The workflow for selecting extracted AOIs with the better detection rate

In the present study, in line with the literature, we followed the consequent steps, as illustrated in Figure 32. In case the duration threshold was set too high, actual fixations might be missed. On the other hand, if it was set too low, false fixations might be present (Camilli, Nacchia, Terenzi, & Di Nocera, 2008). As recommended by Manor and Gordon (2003), we determined the minimum fixation duration as 100 ms. Before applying that, first, we merged adjacent aversions between which there were at most two consecutive non-aversion frames, see Figure 31. Then, we eliminated short aversions that are less than 100 ms.

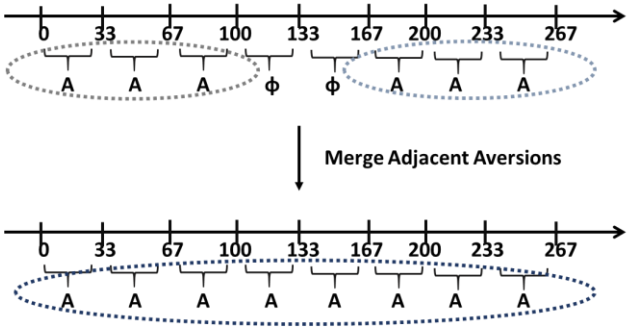


Figure 31: Merging Adjacent Aversions

Lastly, we re-run Merge and Discard functions with the same parameters, this time for face contact. We have developed a C# application and automatized the process described above. By doing so, we also reduced the error rate which would have been higher otherwise.

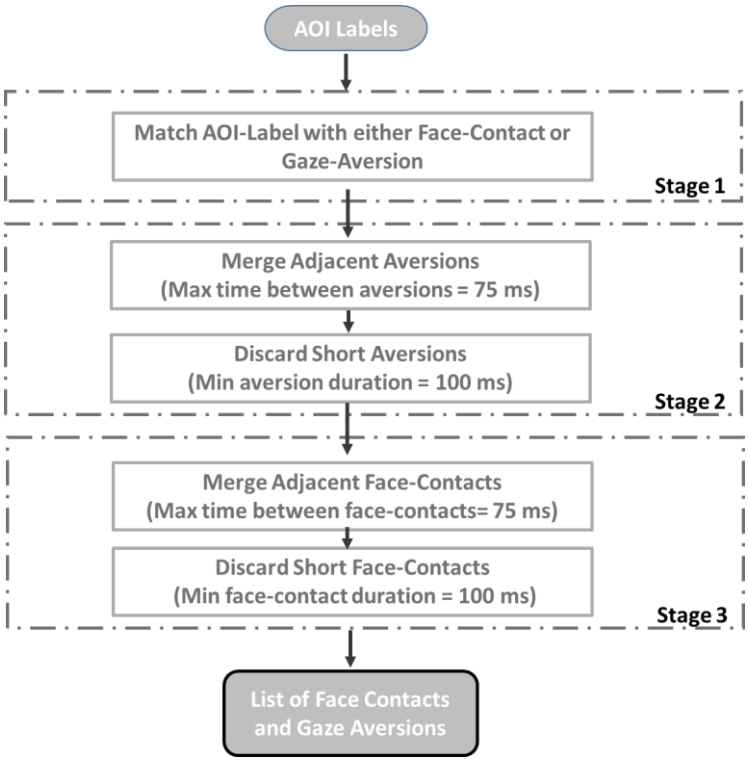


Figure 32: Process flow for detection of gaze behavior

5.2.4. Multimodal Data

5.2.4.1. Gaze and Speech-tag set

The data obtained in speech and gaze analyses were merged into a single summary-file. Each line of the generated text file corresponded to the particular frame-image. The columns of the file were the speech-tag, sender, gaze behavior of sender and of an interlocutor, raw gaze data and coordinates of the minimum bounding rectangle of the detected face on that particular frame-image. As a result, we obtained a series of gaze behavior and related features taken at successive intervals of 33 ms.

The last two columns allowed researchers to investigate gaze-aversion behavior and generate computational models predicting coordinates in the presence of gaze-aversion. On the other hand, in the present study, we only focused on the gaze-behavior, whether it was gaze aversion or face-contact.

5.2.4.2. Gaze and Dialogue-act

We first found the time interval of a particular dialogue unit by concatenating the time intervals of each word that produced a dialogue unit together. To this aim, we have developed a C# application. This application processed the text grid file generated via Praat analysis and excel file in DiAML-Multitab representation. The user is warned in case there is a mismatched text or an absent word.

In the summary file, each line represented the gaze behaviors of a sender and an interlocutor on that particular time with the corresponding communicative function(s), dimension(s), sender information, and if exist; RR(s), functional dependence(s), feedback dependence(s), certainty and sentiment qualifier.

5.2.5. Statistical Analysis

All analyses were carried out in R programming language and environment (R Core Team, 2016). We first screened data and removed outliers. After that, we checked assumptions and decided whether we should transform data or run either the parametric test or the non-parametric one. In addition, we handled individual differences by employing mixed models. We modeled individual variation on the dependent variable by assuming different intercepts for each subject. For instance, for a linear regression model with a single explanatory variable X, the single response variable Y is given by

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ i=1...n} \qquad \text{(Equation 5)}$$

where i=1…n stands for the i^{th} observation (Xi, Yi) with n being the number of observations. Here β_0 is the intercept, β_1 is the slope , and ε_i is the uncorrelated random error. In a linear model, the explanatory variable which is also known as fixed effect concerns a single individual in a row. On the other hand, as in the case of the present study, the researcher might have collected data from the same subject several times. Since the linear model requires independence of data points assuming that each row in the dataset comes from a different subject, we could not run the linear model when we elicit multiple responses from the same subject causing non-independent responses. Therefore, we have to handle these non-independencies before performing the linear model. One option is taking the average over items for subject analysis. However, the previous studies discussed on the pros and cons of averaging and a general conclusion was that even though it is legitimate in principle, a mixed model enables researchers to take full data into account and gives them much more flexibility (Clark, 1973; Raaijmakers, Schrijnemakers, & Gremmen, 1999; Locker, Hoffman, & Bovaird, 2007)

If we add a random effect for the subject in order to handle individual differences, we will get a mixed model with both fixed and random effects. The updated version of Equation 5, where j represents the individual, is written as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij}, \text{ i=1...n.} \qquad \text{(Equation 6)}$$

This time there are two random terms both of which are specific to an individual: u_j and ε_{ij} .

The linearity of the models was checked via a residual plot. If there was a non-linear or a kind of curvy pattern, then it would indicate a violation of the linearity assumption. In such cases, the transformation of data might resolve the issue. In this study, if the linearity assumption was violated, we applied log and square-root transformations, and re-plotted

residuals. We continued the analysis with the transformed data only if the transformation resolved the linearity problem. Homoskedasticity was also checked by observing the residual plot. The residual value is a measure of how much a regression line misses a data point. If individual residuals of the model did not have a similar amount of deviation from the predicted value, the violation of Homoskedasticity occurred. To overcome this problem, we used similar methods that we applied in the case of the violation of linearity.

Furthermore, we checked the normality of residuals and collinearity. In order to make sure that there would be no collinearity, we chose explanatory factors in such a way that they are not correlated with each other. After we checked all the assumptions, we decided on the analysis method. If we could not achieve to handle violations, then we used the proper non-linear analysis

We decided on the explanatory factors by comparing the likelihood ratios among models with or without that particular factor. If there was a significant difference among the models, then, we concluded that it was due to that particular factor, thus we included it in the model. Lastly, we performed post-hoc tests to investigate the relationship or patterns between subgroups that would otherwise have remained undetected.

5.3. Results

In this chapter, the results of the statistical analysis and computation models are presented. The statistical results are organized under four main categories as frequency, duration, discourse annotation schemes, and evaluation scores. We generally investigated the relationship between annotated speech and gaze behaviors. Finally, we presented the computational models with their architecture and accuracy scores.

5.3.1. Frequency

The frequency of gaze aversion and face contact were examined. We calculated the normalized frequency by dividing the count of gaze behavior of a particular session by the duration of that session. The frequency of gaze behavior per minute was calculated using the following equation, where p represents the participant in session s:

$$f(s,p) = \frac{N_{s,p}}{T_s/60.000} \text{ ,.} \qquad \text{(Equation 7)}$$

Here $N_{s,p}$is the number of gaze behaviors of the participant p in session s and T_s represents the duration of that session which was divided by 60.000 to convert milliseconds to minutes.

The paired sample t-test was performed to compare the frequencies of gaze aversion and face contact per role. The analysis revealed that there was no significant difference between the frequencies of gaze aversion (M=20.8, SE=2.62) and face contact (M=23.2, SE=1.86) for interviewers, t(22)=-1.82, p=0.08. On the other hand, interviewees' gaze

aversion frequency (M=44.7, SE=3.6) was significantly higher than their face contact frequency (M=35, SE=3.13), t(24)=2.49, p=0.02.

5.3.1.1. Gaze Aversion

The paired sample t-test was performed to compare the gaze aversion frequency of interviewers with those of interviewees. The analysis revealed that interviewees performed gaze aversions more frequently (M = 44.7, SE = 3.60) compared to the interviewers (M = 20.8, SE = 2.62) and the difference was significant t(23)=-5.03, p<.000.

We also performed a multivariate analysis of variance to test the gender effect. The aversion frequencies of interviewer and interviewee were dependent variables. There was no statistically significant effect of gender in gaze aversion frequency, for both the role of an interviewer, F(2,19)=0.13, p=0.26 and an interviewee F(2,19) =0.08, p=0.45.

5.3.1.2. Face Contact

The paired sample t-test was performed to compare the face contact frequency of interviewers with those of interviewees. The analysis revealed that interviewees performed face contact more frequently (M = 35, SE = 3.13) compared to the interviewers (M = 23.2, SE = 1.86) and the difference was significant t(22)=-3.28, p=0.003

We also performed a multivariate analysis of variance to test the gender effect. The face contact frequencies of interviewer and interviewee were dependent variables. There was no statistically significant effect of gender in face contact frequency of interviewer, F(2,18)=0.16, p=0.2 and interviewee F(2,18) =0.14, p=0.26.

5.3.1.3. Mutual Gaze Behavior

The previous analyses were performed by considering the gaze behavior of a single participant. The dyadic experimental design also allows us to investigate mutual gaze behavior. There were four possible pairs of gaze behaviors as presented in Figure 33 below.

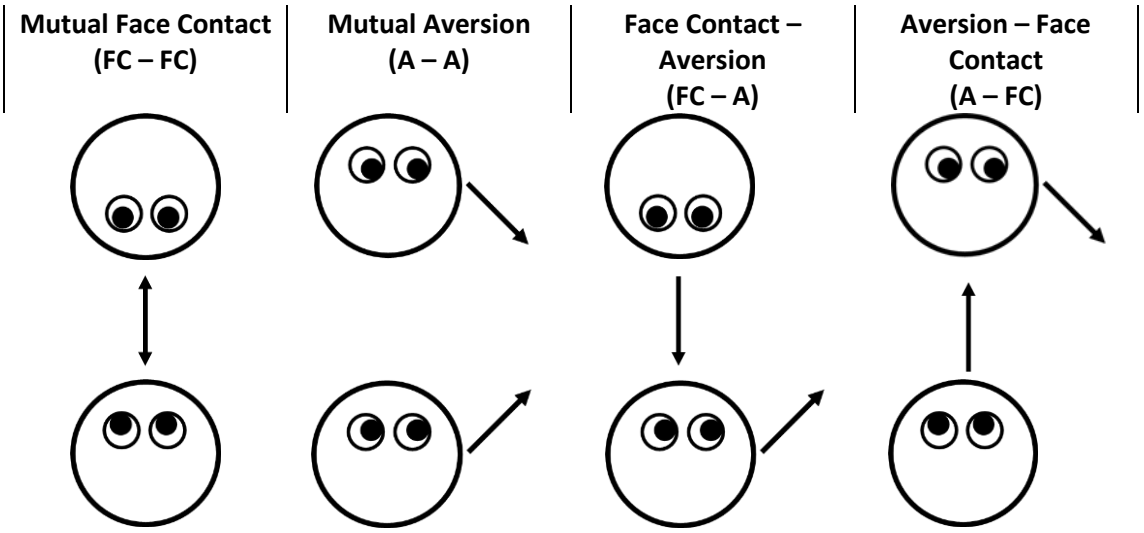


Figure 33: A visualization of dyadic gaze behaviors

We designed a linear mixed model. We compared the potential models by ANOVA test to find out which one fits best. The best performing statistical model's parameters are given in Equation 8. Fixed effects were mutual gaze behavior and the gender of the interviewer and interviewee. In addition to that, the mixed effect term was added for varying intercepts by interviewers, and by interviewees that are nested within interviewers' groups.

Fixed effects = Mutual Gaze Behavior × Interviewee Gender × Interviewer Gender,

Random effects = 1 | InterviewerID /IntervieweeID.

(Equation 8)

The results revealed that when the interviewer was female and the interviewee was male, the frequency of mutual gaze aversion (M=28.6, SE=3.23) was higher than the frequency of mutual face contact (M=9.97, SE=3.6) and this difference is significant, t(88.6)=2.81, p=0.031. Furthermore, when the interviewee was male, the frequency of mutual face contact was higher for male interviewers (M=29.3, SE=6.99) compared to female-interviewer (M=9.97, SE=3.6), and this difference is significant, t(83.4)=-2.63, p=0.01. We also analyzed the ratio of mutual face contact by taking into account the cumulative raw gaze data. It comprised 34% of the conversations with the average duration of 546.1 ms.

5.3.2. Duration

We also examined the duration of gaze aversion and face contact. In the pilot study, the average duration of gaze behavior was calculated by Equation 9, where p represents the participant in the session s:

$$\overline{T}(s,p) = \frac{1}{n}\sum_{i=1}^{n} \Delta\, t_i, \qquad \text{(Equation 9)}$$

In the pilot study, we calculated the sum of durations and then divided it by the number of related gaze behaviors. We implemented those processes to guarantee the independence of observations which required that all participants in a sample are only counted once. However, even though averaging over items is legitimate in principle, it causes disregarding of subject variation. Therefore, this time, we designed a mixed model that takes the full data into account.

We first screened data and removed outliers, then tested the assumptions of the linear mixed model. When we checked the plot where the y-axis represented observations and the x-axis represented quantiles modeled by the distribution, we realized that the data were best fitted to a gamma distribution. Moreover, the residual plot indicated some kind of a pattern. On this plot, better-fitted values have smaller residuals indicating that the model is more "on" with higher predicted means. Therefore, the variance is not homoscedastic: it's smaller in the higher range and vice versa (see appendix G for both residual and the probability distribution plots of gaze aversion).

As a result, since the data was non-normal and violated the homogeneity assumption we performed penalized quasi-likelihood (PQL) instead of linearity test. PQL is a flexible model that can deal with unbalanced design, non-linear data, and random effects.

5.3.2.1. Gaze Aversion

The statistical model is given in Equation 10 below. Fixed effects were gender, partner-gender, role and their two-way and three-way interactions. In addition to that, the mixed effect term was added for varying intercepts by interviewers, and by interviewees that are nested within interviewers' groups. Lastly, we considered varying the slope of the interaction between gender and partner-gender differing across interviewers' groups.

$$\text{Fixed effects} = \text{Role} \times \text{Gender} \times \text{PartnerGender},$$

$$\text{Random effects} = 1 + \text{Gender} \times \text{PartnerGender} \mid \text{InterviewerID} / \text{IntervieweeID}.$$

(Equation 10)

There was a significant effect of the role, i.e., being an interviewer or interviewee, on the duration of gaze aversion. The post hoc tests revealed that a significant difference between the gaze aversion-durations of interviewers (M=258.2 ms, SE =5.25) and interviewees (M=313.2 ms, SE=3.43) was observed when the partner gender was female, $t(9760)=5.75$, $p<.0001$, see Figure 34.

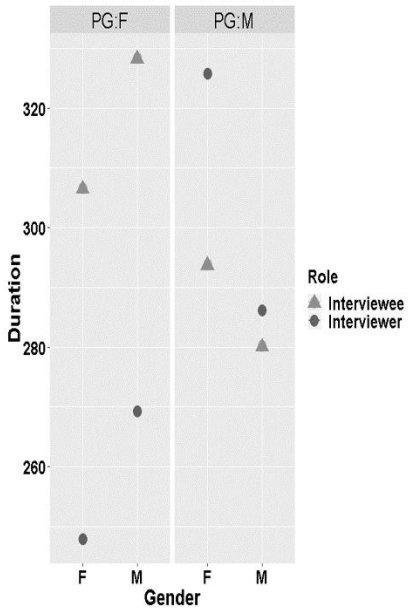


Figure 34: Gaze aversion durations per gender, partner gender and role

Furthermore, there was no statistically significant effect of gender in gaze aversion duration t(9760) = 0.92, p=0.36. We also examined the effect of gender in pairs. We found that there was a significant difference in gaze aversion duration of an interviewer when the interviewer was female while the interviewee was male (M=328.4 ms, SE=8.68) compared to the case where both the interviewer and interviewee were female (M=247.8 ms, SE=7.2), t(9760)=-3.33, p=0.005, see Figure 35.

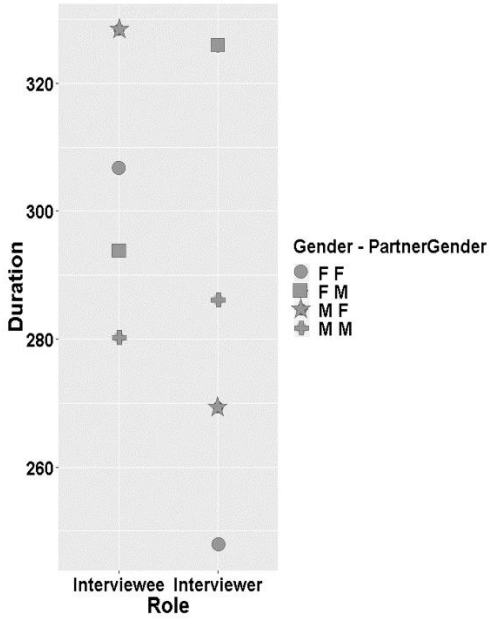


Figure 35: Gaze aversion durations per role and the pair of gender-partner gender

5.3.2.2. Face Contact

The statistical model is given in Equation 11. Fixed effects were gender, partner gender, role and their interactions. In addition to that, the mixed effect term was added for varying intercepts by interviewers, and by interviewees that are nested within interviewers' groups.

$$\text{Fixed effects} = \text{Role} \times \text{Gender} \times \text{PartnerGender},$$

$$\text{Random effects} = 1|\ \text{InterviewerID}\ /\text{IntervieweeID}.$$

(Equation 11)

There was a significant effect of the role, i.e., being an interviewer or interviewee, on the duration of face contact. The interviewer's face contact duration (M=648.9 ms, SE=7.06) was significantly higher than the interviewee's face contact duration (M=585.8 ms, SE=6.06), t(10434)=-1.977, p=0.048.

Furthermore, we found that there was a significant difference in face contact duration of an interviewee when the interviewee was male while the interviewer was female (M=645.8 ms, SE=15.3) compared to the case where both the interviewer and interviewee were female (M=555 ms, SE=9.47), t(10435)=-3.19 p=0.008. In addition to that, we also found a significant difference in face contact duration of an interviewer when the interviewer was female while the interviewee was male (M=472.4 ms, SE=16.9) compared to the case where both the interviewer and interviewee were female (M=873.7 ms, SE=16), t(10435)=1.5, p<.0001, see Figure 36.

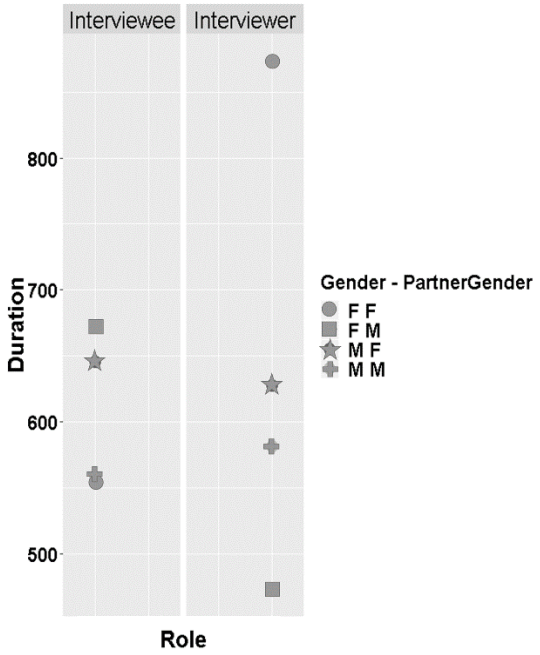


Figure 36: Face contact durations per role and the pair of gender-partner gender

5.3.3. Multimodal Analysis

We also examined the relation between gaze behavior and speech-tag or gaze behavior and dialogue-act. In this section, we will describe the analysis steps via speech-tag set. Similar calculations were also performed for dialogue-act analysis.

Primarily, we extracted the ratio of gaze behavior observed during an instance of speech-tag set. Each instance of speech-tag set might be assigned several times during a session. In Equation 12, let B is a set including percentages of gaze aversion and face contact during occurrences of speech-tags, for session x and participant p where i is the element of F which is a set of frames labeled with speech-tags. D function gets the gaze behavior type and frame numbers as input parameters and returns the duration of that specified gaze behavior among those frames.

$$B_{x,p}(S,A)=\{\, i \in F_s:\ D(i,A)/\,(\,D(i,A)+D(i,FC)\,)\,\}, \qquad \text{(Equation 12)}$$

The process details are given in Table 14. We intentionally skip the frames between 10 and 25 to simulate realistic data. During the analysis, we excluded the frames in which there was no extracted gaze behavior for the interviewer or interviewee.

Table 14: Illustration of calculating the ratio of gaze behavior (GB) to the particular speech-tags, S_1 and S_2. Only the interviewer's gaze behavior is considered. A similar calculation is also performed for interviewees.

Frame No	Speech-Tag	GB	Ratio of GBDuration
1		A	
2		A	
3		A	$\|A\| / \|S_{1,1}\| = 6/9$
4		A	
5	$S_{1,1}$	A	
6		A	
7		FC	
8		FC	$\|FC\| / \|S_{1,1}\| = 3/9$
9		FC	
26		FC	
27		FC	$\|FC\| / \|S_{2,1}\| = 10/20$
28-35	$S_{2,1}$	FC	
36-44		A	$\|A\| / \|S_{2,1}\| = 10/20$
45		A	
46		A	
47		A	$\|A\| / \|S_{1,2}\| = 25/50$
48-70	$S_{1,2}$	A	
71		FC	$\|FC\| / \|S_{1,2}\| = 11/50$
72-81		FC	
82		A	$\|A\| / \|S_{1,2}\| = 14/50$
83-95		A	

A sample implementation of Equation 12 for Table 14 is given as follows:
Frame Set:
F_{s1} ={[1-9], [46-95]}

Gaze Behavior Counts:
$D([1-9], A)$ ={6}, $D([46-95], A)$ ={25, 14}
$D([1-9], FC)$ ={3}, $D([46-95], FC)$ ={11}

Set of Aversion Percentages, During S_1:
$B_{1,interviewer}(S_1, A)$={i ϵ {[1-9], [46-95]}: $D(i, A)/\big(D(i, A) + D(i, FC)\big)$ }
$B_{1,interviewer}(S_1, A)$= {6/9, 25/50, 14/50}

Set of Face Contact Percentages, During S_1:
$B_{1,interviewer}(S_1, FC)$={i ϵ {[1-9], [46-95]}: $D(i, FC)/\big(D(i, A) + D(i, FC)\big)$ }
$B_{1,interviewer}(S_1, FC)$= {3/9, 11/50}

As well as the duration, we also calculated the frequency of fixations of gaze behavior during a particular speech-tag. This time, we just consider the fixation counts of related gaze behavior. For instance, in Table 14, the frequency of face contact was 1 for $S_{1,2}$,

whereas the frequency of gaze aversion was 2. Thus the percentages were 1/3 and 2/3 respectively.

5.3.3.1. Speech-tag Set Annotation

The statistical analysis was conducted on the top 5 speech-tags, namely *Speech*, *Micro Pause*, Speech *Pause*, *Thinking*, *Pre-Speech*, which in total cover 80.3% of the whole data. (see Appendix H).

The data was non-normal and violated the homogeneity assumption, thus we performed penalized quasi-likelihood (PQL). The statistical model is described by Equation 13. Fixed effects were role, speech-tag set, their mutual interaction, interviewer gender, interviewee gender and their mutual interaction. Besides, the mixed effect term was added for varying intercepts by interviewers, and by interviewees that are nested within interviewers' groups. Lastly, we added the speech-tag ID which was a unique identifier for each occurrence of speech-tag set, as a mixed effect term.

$$\text{Fixed effects } = \text{Role} \times \text{SpeechTagSet} + \text{ Interviewer Gender} \times \text{Interviewee Gender},$$

$$\text{Random effects } = 1|\ \text{InterviewerID}/\text{IntervieweeID } + 1|\text{Speech tag ID}.$$

(Equation 13)

There was a significant difference in frequency of gaze behavior between the interviewers and interviewees when the speech tag was *Thinking* ($t(6840)=13$, $p<.0001$), *Speech* ($t(6840)=12.9$, $p<.0001$), *Speech Pause* ($t(6840)=10.8$, $p<.0001$) or *Micro Pause* ($t(6840)=7.23$, $p<.0001$), see Figure 37. Moreover, we conducted pairwise comparisons between speech-tags. There were significant differences between *Thinking* and *Speech* ($t(6840)= 4.28$, $p=0.0002$), *Thinking* and *Micro Pause* ($t(6840)=6.64$, $p<.0001$), *Speech* and *Pre-Speech* ($t(6480)= -3.66$, $p=0.0024$), *Speech* and *Micro Pause* ($t(6840)=3.62$, $p=0.0027$), *Speech Pause* and *Micro Pause* ($t(6840)=5.34$, $p=<.0001$), and *Pre-Speech* and *Micro Pause* ($t(6840)=5.77$, $p=<.0001$).

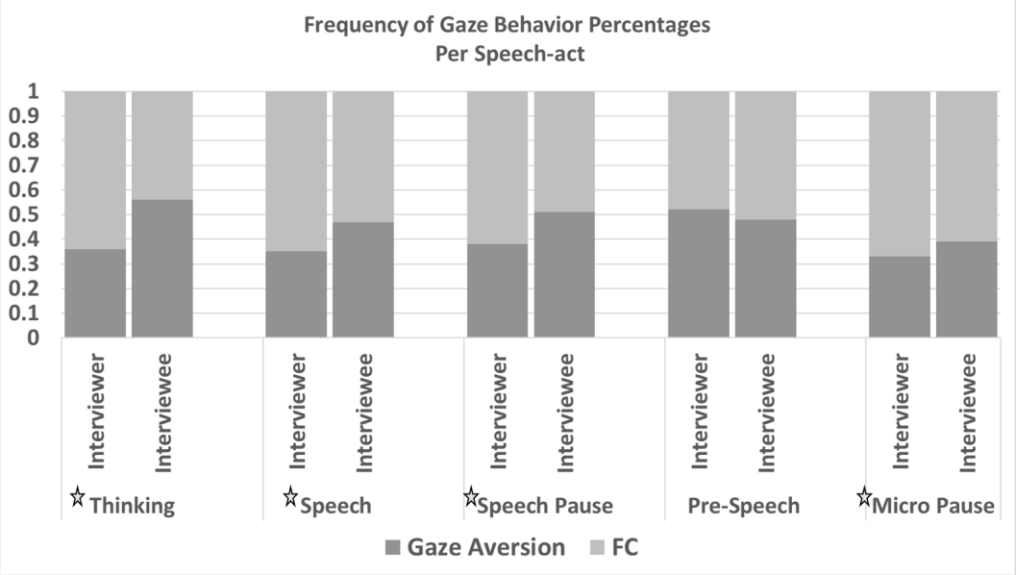


Figure 37: Frequency of gaze behavior percentages for speech-tag set. Significant differences are presented with * character.

We also examined the difference in duration of gaze behavior between the interviewers and interviewees. Similarly, results revealed that when the speech-tag was *Thinking* ($t(6840)=13.3$, $p<.0001$), *Speech* ($t(6840)=12.9$, $p<.0001$), *Speech Pause* ($t(6840)=10.7$, $p<.0001$) or *Micro Pause* ($t(6840)=7.8$, $p<.0001$), interviewee's gaze aversion duration was significantly longer than the interviewer's.

5.3.3.2. Dialogue Act Annotation

The statistical analysis was conducted on the top 5 speech- acts, namely *Stalling*, *Answer*, Auto *Positive*, *Inform*, *Turn Take*, which in total cover 80.3% of the whole data. (see Appendix H).

The data was non-normal and violated the homogeneity assumption, thus we performed PQL. The statistical model is described by Equation 14. Fixed effects were role, dialogue act, their mutual interaction, interviewer gender, interviewee gender and their mutual interaction. In addition, the mixed effect term was added for varying intercepts by interviewers, and by interviewees that are nested within interviewers' groups. Lastly, we also added the dialogue act ID which was a unique identifier for each occurrence of dialogue acts, as a mixed effect term.

$$\text{Fixed effects } = \text{Role} \times \text{Speech Act} + \text{ Interviewer Gender} \times \text{Interviewee Gender},$$

$$\text{Random effects } = 1|\,\text{InterviewerID/IntervieweeID } + 1|\text{Dialogue Act ID}.$$

(Equation 14)

There was a significant difference in percentage frequency of gaze behaviors between the interviewers and interviewees when the dialogue act was *Answer* (t(5334)=13.1, p<.0001), *Stalling* (t(5334)=19.9, p<.0001), or *Turn Take* (t(5334)=5.69, p<.0001), see Figure 38. Moreover, we conducted pairwise comparisons between communicative functions of dialogue-act. There were significant differences in the frequency of gaze behavior between *Answer* and *Inform* (t(5320)= -3.31, p=0.0085), *Answer* and *Stalling* (t(5320)=-3.97, p=0.0007), *Answer* and *TurnTake* (t(5320)= -7.20, p<.0001), *AutoPositive* and *TurnTake* (t(5320)=-4.66, p=<.0001), *Inform* and *TurnTake* (t(5320)=-2.77, p=0.0444), and *Stalling* and *TurnTake* (t(5320)=-4.57, p=<.0001).

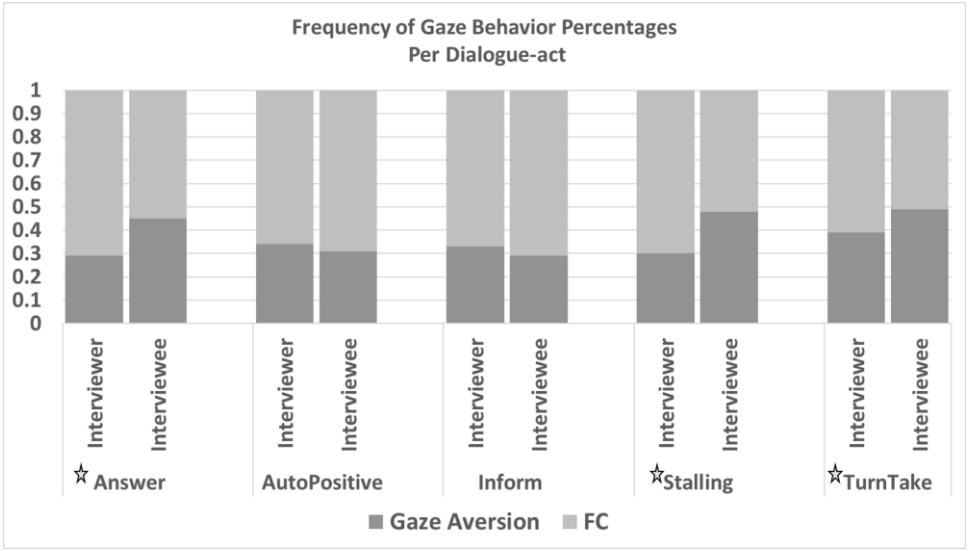


Figure 38: Frequency of gaze behavior percentages per dialogue act

We also examined the difference in duration of gaze behavior between the interviewers and interviewees. Similarly, results revealed that when the dialogue act was *Answer* (t(5334)=14.2, p<.0001), *Stalling* (t(5334)=19.8, p<.0001) or *Turn Take* (t(5334)=5.58, p<.0001), interviewee's gaze aversion duration was significantly longer than the interviewer's.

5.3.4. Average Scores of Evaluation Questionnaire

We investigated the relation between gaze behavior and the average scores of evaluation criteria. We made use of linear mixed models to predict the average score of each of the 4 questions by considering the frequency or duration of the gaze behavior. The only significant effect was found in the model in which fixed effects were interviewer gender, interviewee gender, their mutual interaction, aversion frequency of interviewer, aversion frequency of interviewee and their mutual interaction. We also added interviewer ID as a mixed effect term (see Equation 15)

$$\text{Fixed effects} = \text{Interviewer's Frequency} \times \text{Interviewee's Frequency} + \text{Interviewer Gender} \times \text{Interviewee Gender},$$

$$\text{Random effects} = 1|\text{Interviewer ID}.$$

(Equation 15)

We found that the average score of the first question significantly decreased as the aversion frequency of an interviewee increased (χ^2 (1)= 4.78, p=0.29) , see Figure 39.

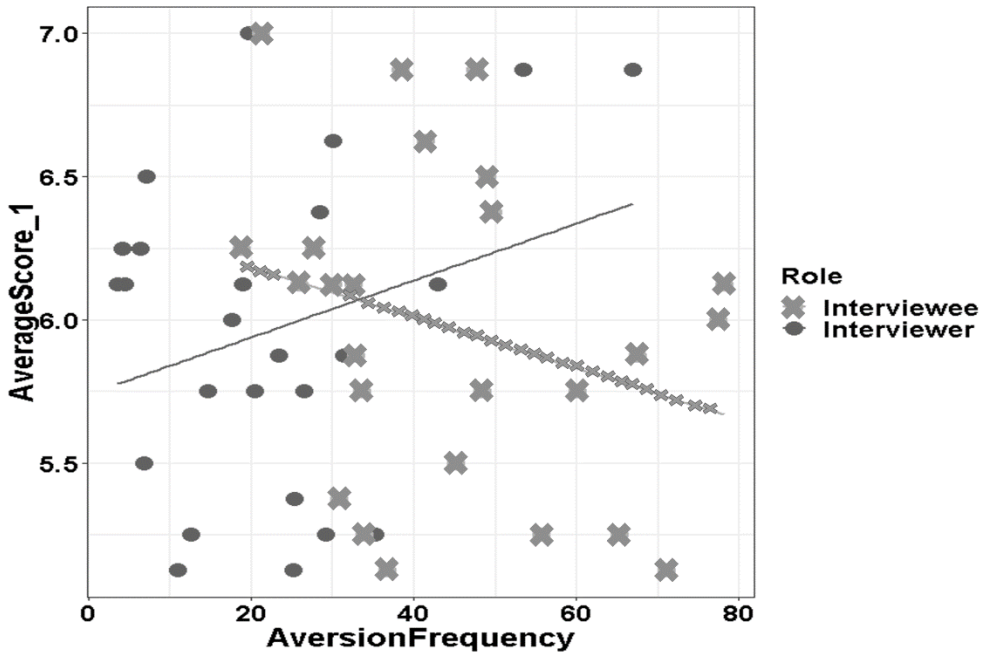


Figure 39: Average score of the first question as a function of aversion frequency. Considering both the interviewers and interviewees, there are 25 points in total that correspond to a single session. Both the aversion frequency and the score of the first question were averaged over a session. The lines represent the general direction followed by either an interviewer or an interviewee.

CHAPTER 6

COMPUTATIONAL MODEL OF SPEECH DRIVEN GAZE IN FACE-TO-FACE INTERACTION

The secondary research questions of this dissertation are making computational model of gaze behavior with the high-level features of speech and the difference in the performance of computational models when the input features are extracted from either Dialogue-act or speech-tag set analysis. For this aim, we trained simplified versions of ResNets (He, Zhang, Ren, & Sun, 2016) and VGGNet (Simonyan & Zisserman, 2015) which are CNN (Convolution Neural Network) architectures, with the data from the 28 pair experiment (for details see Chapter 5). In this chapter, firstly the structure of CNN architecture is presented. Afterwards, the list of input data, their representations and the way how we align them as a multimodal time series are explained. Lastly, we report the performances of computational models and visualize filters of the model with the best performance to get insight on why and how a particular prediction was made.

6.1. Introduction

Convolutional Neural Networks (CNNs) are a particular type of Deep Neural Networks (DNNs). In the following subsections, the brief history and basic building block of DNNs, components of a basic CNN architecture, and the topology of VVGNet (Simonyan & Zisserman, 2015) and ResNet (He, Zhang, Ren, & Sun, 2016) are summarized. Moreover, in the present study we use speech driven-gaze data as a time series. Accordingly, the shape of time series data and the basic method for calculating the convoluted features using this data are presented.

6.1.1. Deep Neural Networks

The deep learning approach has greatly improved many artificial intelligence tasks including machine translation, object detection and speech recognition. In addition to classical AI tasks, researchers have adapted deep learning to various areas. Osako, Singh, and Raj (2015) tried to eliminate noise from speech signals by using a particular type of DNN, namely RNNs (Recurrent Neural Networks), Wang, Meghawat, Morency, and

Xing (2017) performed sentiment analysis with data from multiple modalities, and Gatys, Ecker, and Bethge (2016) utilized neural models to produce images in different styles.

The basic building block of neural networks are neurons which are inspired by their biological counterparts, yet they still differ in several ways. The idea behind neural networks is based on the assumption that several parts of neurons like dendrites, cell bodies, axons and their inner workings can be imitated by simple mathematical models. McCulloch and Pitts (1943) produced the first mathematical model of neural networks. Afterwards, at the end of the 1950s, Rosenblatt (1958) proposed Perceptron as a simplified mathematical model representing the operations of neurons in our brains. Accordingly, a neuron takes binary inputs from a group of neighborhood neurons. These data are then multiplied by the weight values corresponding to the synapse strength between the neurons, and in parallel with the all-or-none principle, if the sum of these weighted inputs is above a certain threshold value, one is produced as the output otherwise zero is produced, see Figure 40.

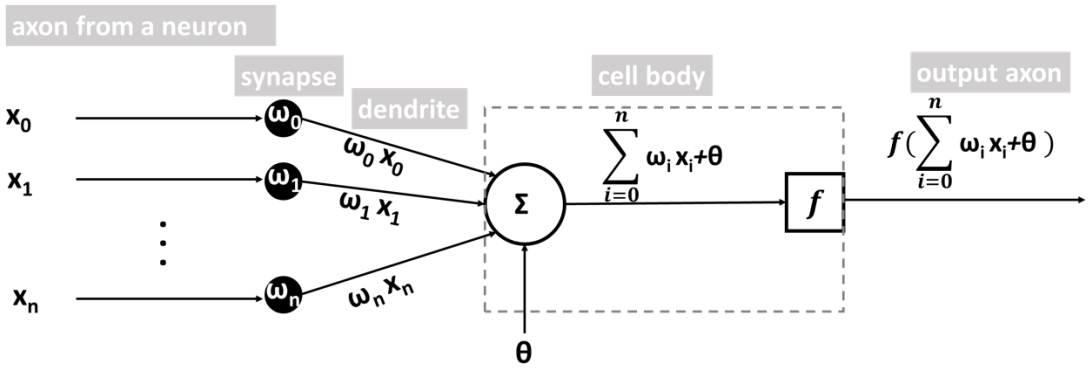


Figure 40: Biological inspiration for neural networks. Here x_i represents *i*th input, n represents the number of elements in neuron sets which is greater than zero. ω_i represents the weights associated with the *i*th connection, θ is the bias term. The output is computed by applying the activation function *f* to the weighted sum of the input signals. The gray background labels represent the parts of biological neurons corresponding to the related structure of the mathematical model presented below them.

In order to solve more complex tasks than a single neuron can do, networks containing multiple layers of neurons are produced. Networks up to three layers are generally called shallow neural networks, while networks that have more than three layers are called deep neural networks. In a DNN, instead of a single output layer, the input data is directed to hidden layers, which consist of a set of neurons. The output of a hidden layer is used as the input for the next layer both of which cannot be directly observed from outside. In other words, data in between these layers are hidden until the very end of the network, where one gets the output, i.e. the apparent response of the network. Intermediate calculations done by hidden layers allow tackling very complex problems by focusing on input data characteristics instead of noisy raw data. An activation function defines whether a node in the network will be active based on the sum of weighted inputs. This function is generally selected to be non-linear to allow learning about complex non-linear

transformations on the input signal (Goodfellow, Warde-Farley, Mirza, Courville, & Bengio, 2013), whereas bias values help for better data fitting with left or right shifts from the activation function (Goodfellow, Bengio, & Courville, 2017).

The main mechanism of learning in neural networks is backpropagation. The network propagates the signals of the input data forward through its parameters, and then propagates the information about the errors backwards through the network so that the parameters can be updated (for more detailed coverage, see Goodfellow et al., 2017), see Figure 41.

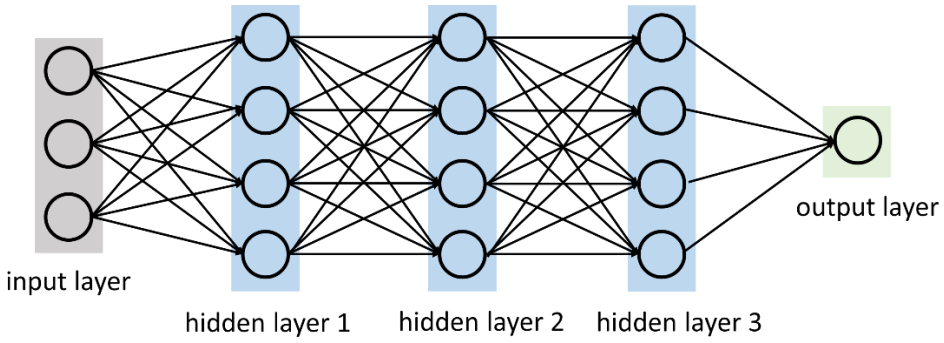


Figure 41: DNN structure with three hidden layers

In the present study, simplified VVGNet (Simonyan & Zisserman, 2015) and ResNet (He, Zhang, Ren, & Sun, 2016) which are CNN (Convolution Neural Network) architectures are trained. The brief history and operations of CNN, as well as the topology of VGGNet and ResNet are presented in the following subsection.

6.1.2. Convolution Neural Networks

Convolution Neural Networks (CNN) are regularized versions of fully connected networks. It takes the name CNN from an important mathematical operation, namely convolution, which integrates the product of two functions. It is useful for calculating the derivatives, finding patterns in signals, detecting edges, applying blurs and so on. CNN dates back to the studies in the neural basis of visual perception performed by Hubbel and Wiesel (1959). They recorded the activity of neurons in the visual cortex when a certain pattern of light was projected on a screen in front of a cat. Depending on the angle at which light is incident, they observed activation in different neuron groups. They also noticed that some neurons, which they called simple cells, showed different activation levels when exposed to light and some other neurons, which they named complex cells, played a crucial role in edge detection. This study has demonstrated that the visual system generates complex visual information from simple features of a stimulus. Parallel to the studies on brain, in the 1980s, Fukushima proposed hierarchical network models inspired by the theory of simple and complex cells (Fukushima, 1980, 1988). There have been neural

network models that mimic humanly cognitive faculty at the behavioral level. As one of the first example of such studies, LeCun, Bottou, Bengio, and Haffner (1998) developed LeNet-5 which was specialized in recognition of handwriting characters. The basic comprehension of LeNet-5 was that the features of the image were scattered throughout the entire image and the similar features in different parts of the image could be effectively revealed out by a few learnable parameters. The basic features of LeNet-5 can be listed as follows: (i) sparse matrix between layers to reduce computational load (ii) use of multilayer neural networks as a classifier in the last stage, (iii) use of nonlinear activation functions such as sigmoid and hyperbolic tangent, (iv) spatial averaging.

A basic CNN architecture includes four fundamental operations respectively: (i) convolution, (ii) nonlinearity (ReLU), (iii) pooling or subsampling and (iv) classification (Fully Connected), see Figure 42.

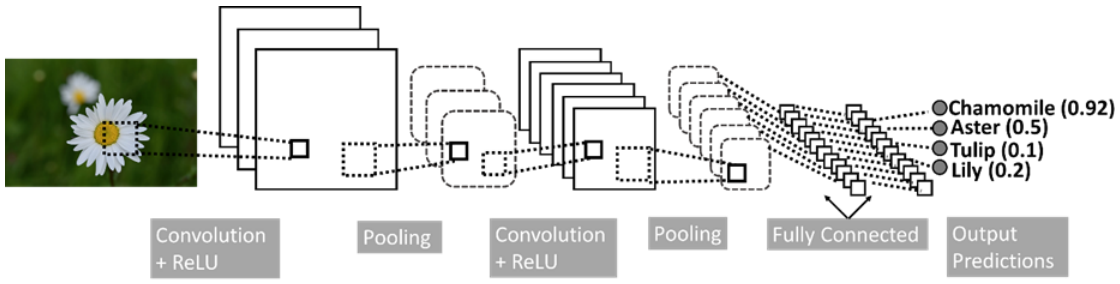


Figure 42: A simple CNN model. The output is one of the four flowering plants, image adapted from LeCun et al., 1998

In the first step, the purpose of the convolution layer is to extract the features of an input image. Convolution learns features over the small square areas on the image, and this method enables maintaining spatial relations. Suppose that we have a 5 x 5 input image as matrix and another 3 x 3 matrix which is called “filter” or “kernel” or “feature detector”. In the convolution step, the element-wise multiplication between those two matrices is computed and then, the outputs of multiplication are added in order to get a single integer, see Figure 43. This process would be repeated until all pixels of an input image are covered by sliding kernel over the input matrix via a predefined number of pixels, named stride, and forming a new matrix which is termed “Convolved Feature” or “Feature Map” or “Activation Map”. In practice, CNN determines the content values of the kernel itself in the learning phase, but of course some parameters like the number of filters and the filter size must be specified beforehand. The more the number of filters, the more features of the image will be extracted, which also would improve the performance at recognizing patterns in the images that we have never encountered.

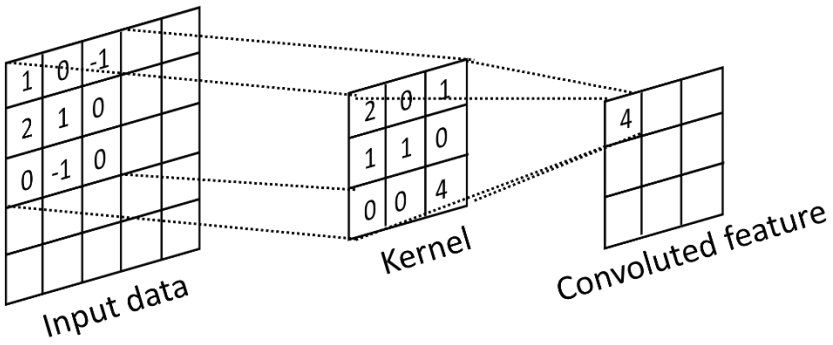


Figure 43: Illustration of convolution operation for 2D input data.

As the second step, after each convolution process, ReLU (Rectified Linear Unit) (He, Zhang, Ren, & Sun, 2016) is applied. It is a non-linear operation and updates all negative pixel values on the feature map to zero. The purpose of the ReLU layer is to introduce non-linearity to the model in consistency with the generally non-linear real-world data. ReLU is not the single alternative, sigmoid and hyperbolic tangent would be applied as well. The output is referred to as the "Rectified feature map".

In the third step, the rectified feature map is reduced in size without losing valuable information. In this reduction process, three different methods can be applied; Max, Average and Sum. Suppose that we adopted the average operation. First, we select a window to determine the amount of spatial neighborhood, for instance 2 x 2. Then, we slide this window until we pass through all elements of the rectified feature map, and each time take the average of the four values within that window. Since pooling is applied to each of the input maps separately, the number of output maps will be the same with the input one. Pooling brings important benefits. Reducing the size of the feature dimension makes the model easier to manage. Furthermore, it enables to handle the overfitting problem by reducing the number of parameters and hence calculations to be performed. In addition, it prevents small distortions on the input from adversely affecting the model performance.

The reason why the last step is called Fully Connected Layer (FCL) is that it is a classic multilayer perceptron where all of the input and output neurons are connected. In case the model is trained for classifying the input image as a type of flowering plant, until the FCL step, the high-level features of the input image are extracted. The FCL consumes these features to calculate the probability of classes. The probabilities are normalized to be between 0 and 1 with the sum of all being 1, as the `softmax` function is applied.

These four steps we have described so far are members of forward propagation. Just before running them, the parameters and weights in the network are initialized randomly. In the very first training phase, the output class probabilities are likely to diverge from expected values, as the weights are randomly assigned. The output probabilities are optimized by updating the weights with individual gradient descent values calculated in the

backpropagation phase. Forward and back propagations are repeated for the predefined number of times of complete passes through the training dataset, i.e., epochs.

So far, we discussed the architecture of LeNet-5, one of the first CNNs. Until the early 2010s, there were not many developments in the field of CNN. After 2010, with the increase in the amount of data collected and computational power, the number of issues that CNN had the opportunity to work on increased. In 2012, a more complex version of LeNet-5, called AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a large margin, re-accelerating researches on CNN. In the present study, we used simplified versions of the ResNets (He, Zhang, Ren, & Sun, 2016) and VGGNet (Simonyan & Zisserman, 2015) which are also the winners of ILSVRC 2014 and 2015 challenges respectively. In contrary to large convolutions, the VGG network demonstrated that using multiple 3 x 3 convolutions sequentially can emulate a similar effect to represent complex features, see Table 15.

The reason behind our selection is that VGG has simple architecture, thus it is easy to implement and ResNet, in recent years, is one of the networks with the highest performance (Canziani, Paszke, & Culurciello, 2016).

Table 15: VGG16 architecture. VGG16 to classify input image [224 X 224 X 3] into one of 16 categories.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000
dense_1 (Dense)	(None, 16)	16016

Convolutional Neural Network (CNN) models produce successful results, especially in the areas of image classification and recognition. Although they are mostly used for image processing, they have also been used for time series in recent years (Fawaz, Forestier, Weber, Idoumghar, & Muller, 2019). In this study, we collected the gaze data in the form of a time series and trained 1D CNN networks. In this way, the location of the feature within the input segment is not of high relevance when compared to 2D inputs. The illustration of the convolution operation for time series data is presented in Figure 44.

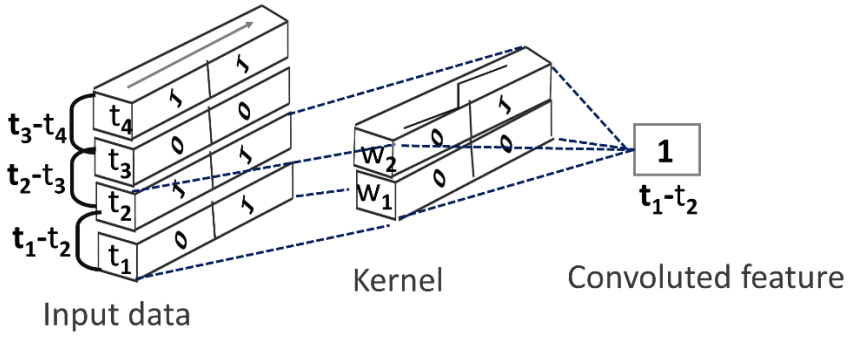


Figure 44: Illustration of the convolution operation for time series data. Input shape is (1,4,2), i.e., one sample of four time points, each item having two channels. The kernel shape is (2,2).

1D CNN is mostly used in NLP studies. Data points in time series are generally introduced to the network as a group of instances, rather than one by one. The number of instances in a group is referred to as *timestamps* and distance between consecutive groups is called *step-distance*. For instance, for a discourse consisting of eight words, each having two channels, the updated shape of an input that will be introduced to CNN is (3,4,2) where the corresponding timestamp is four and step-distance is two, see Figure 45.

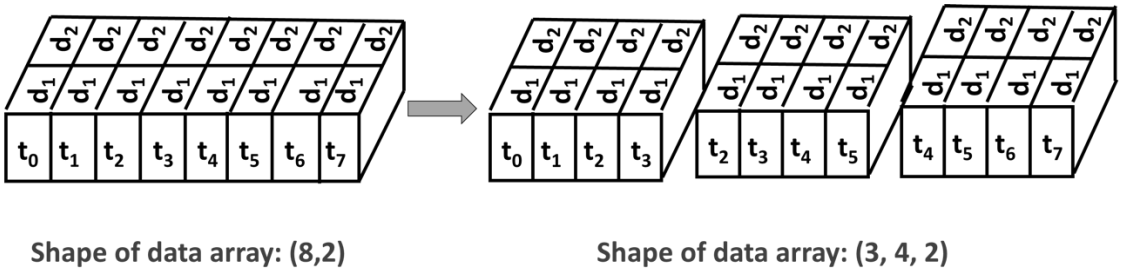


Figure 45: Shape of the time series array. On the left, the shape of data is (8,2) where the number of data points is eight and the number of dimensions is two. On the right-hand side, the shape has changed to (3,4,2) where the timestamp is four and the step distance is two.

In this study, we obtained a series of gaze behavior and related features taken at successive intervals of 33 ms. According to the data obtained from the human-human experiment, the average gaze behavior duration is 300 ms. Therefore, we assigned nine to timestamps

as each frame was 33 ms and since the minimum fixation duration was 100 ms, we assigned three to step distance.

6.2. Data and Analysis

One of the research questions of this dissertation was the difference in the performance of computational models when the input features were extracted from either Dialogue-act or speech-tag set analysis. With this aim, we trained CNN models and then compared the best performances obtained from the test results. 2D CNN models were generally developed for the classification of images. It is called a 2D convolution since the movement of the filter across the image takes place in two dimensions: width and height. 2D CNN enables one to derive interesting features as well as local spatial features of each pixel. On the other hand, the location of the feature does not have much significance in time series. Since we produced time series of gaze behavior with corresponding conversation tags in the previous phases, we constructed and trained 1D CNN models.

We worked on 2 well-known CNN models. First, we constructed a smaller and more compact 1D variant of the VGGNet which was the runner-up at the ILSVRC 2014. We preferred this model because it is quite appealing with its uniform architecture. In addition to that, we constructed an alternative Residual Neural Network architecture (ResNet). Fawaz et al. (2019) showed that ResNet performs with a high accuracy when applied to the time series.

In our experimental design, while the interviewees participated in a single session, interviewers took part in multiple interviews. Thus, we collected more data from interviewers than we acquired from interviewees. Because of that, we designed computational models for predicting gaze behavior of interviewers rather than the interviewees.

At the beginning of the training, we split data set into two as training and testing. The test set comprised 20% of the whole dataset and was not used in the training. Moreover, for parameter tuning, we performed 5-fold cross validation on the training data set and split it into training and validation data sets. K-fold cross validation assumes that each observation is independent. Therefore, when evaluating a model for time series forecasting, classic k-fold cross validation cannot be directly used. Instead, we performed back-testing in which data was split by respecting the temporal order, in contrast to random splitting, see Figure 46.

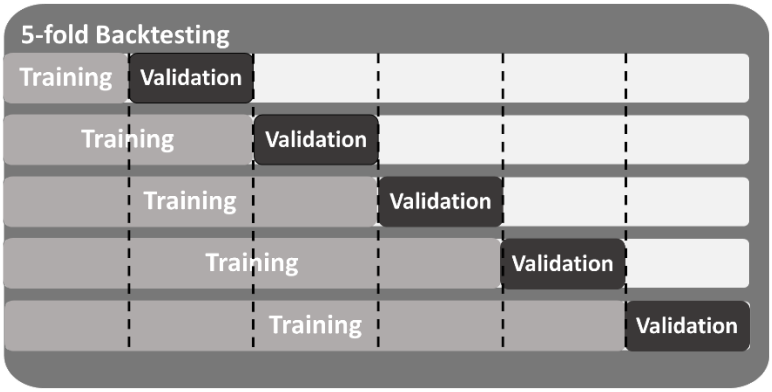


Figure 46: Illustration of 5-fold Backtesting.

The validation accuracy fluctuated between splits in the course of 5-fold back-testing (for details see result under section 6.3). We first applied pooling, weight and dropout regularizations to the networks in order to handle underfitting (i.e., the network cannot capture the underlying trend of the data) and overfitting (i.e., the network captures the noise in the data and have poor generalizability). Since the problem persisted, then, we examined the data and back-testing method more carefully. In the n-fold back-testing, the ratio of data provided for the training and validation is different at each split. It is five for the fifth split and one for the first split, see Figure 46. When the training data is not big enough, the network might not quite learn about the underlying trend of the data. Moreover, as presented in Table 16, even in different networks, there were similar fluctuations in the validation accuracy of splits. In the case presented in Table 16, validation accuracy of the second and the fourth splits are lower than the others.

Table 16: Validation accuracy in each split of 5-fold back-testing. The tagging scheme of input data was Speech-tag set and feature map size of the first blocks was 16.

	gazeResNet	gazeVGG
1.	75.9	75.9
2.	62.5	61.7
3.	78.5	80.1
4.	61.8	57.1
5.	66.9	66.6

We examined the reason for such similarities. In the second split training performed with the data that involved the first and the second interviewer, along with a half data of the third one, whereas the test was performed on the remaining data of the third interviewer and approximately the half data of the fourth interviewer. Similarly, on the fourth split, training was performed with the data that involved the first five interviewers, whereas it was tested with the sixth interviewer. The difference in the frequency of gaze behaviors

between the interviewers changed as presented in Figure 47. The second and the third interviewer had a greater tendency to aversion whereas the sixth one had a tendency in the opposite direction. Hence, especially for the second and the fourth splits, the distribution of data for training and testing was different which resulted in validation fluctuations.

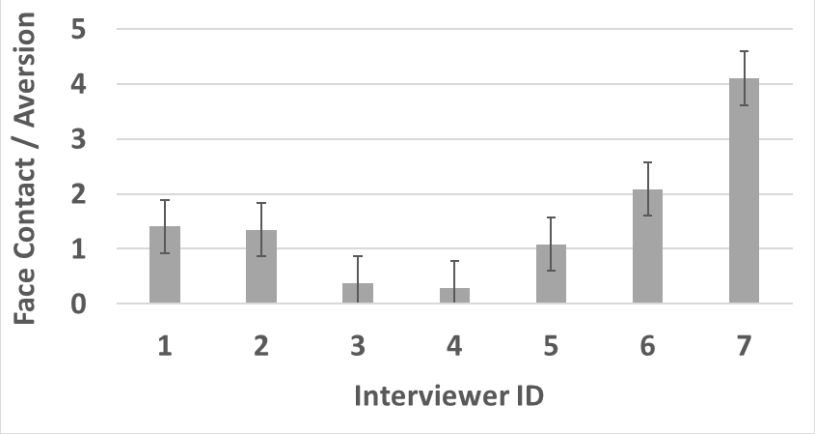


Figure 47: The ratio of the frequency of face contact to aversion per interviewer.

In light of the information we discussed above, one of the possible reasons for the fluctuation we experienced in validation accuracy was the lack of a sufficient amount of training data for the first splits. This issue is caused by the n-fold back-testing that is adopted for cross validation of time series data. Another reason was the order of interviewers in the data presented to the network. Particular orders of interviewers in the input data result in specific orders of interviewers in splits used for training and validation. This might cause testing the network with a different distribution than the one used in training. The classical cross validation method enables to handle such distribution issues by randomly dividing the set of input data into training and test sets. However, time-series data have temporal relations that prevent randomized division. Thus, in order to handle fluctuations in the validation accuracy, we created five input data by shuffling the order of interviewers, and then in each data set, we separated the initial 80% of the data for training and the remaining 20% for testing. Since we used fewer data when tagging with the dialogue-act scheme, we created a distinct order for both tagging schemes. To make predictions over almost all of the data, we shuffle the order of interviewers in five data sets so that the pair of last two interviewers in the order was different from the ones in the remaining four data sets, see Table 17.

Table 17: The list of data sets created by shuffling the orders of interviewers in the input data. Numbers represent interviewer ID's.

	Orders of Interviewers	
	Speech-Tag Set	Dialogue Act
1.	1-2-3-4-5-6-7	1-2-3-5-6-7
2.	2-7-3-5-1-4-6	2-3-5-6-7-1
3.	3-1-4-7-6-2-5	3-6-5-7-1-2
4.	4-6-5-2-7-1-3	5-3-1-7-2-6
5.	6-3-2-1-7-5-4	6-7-1-2-3-5

We trained mini VGGNet and mini ResNet models with input data either annotated with speech-tag set or dialogue acts. As discussed above, the 5-fold back-testing is used for tuning feature map size which was either 16 or 32, and the one that minimized the errors on the validation set was chosen. Finally, we trained the models with the chosen feature map size on the data sets in which orders of interviewers differ from each other (see Table 17). Then we evaluate the model with test data that corresponds to the last 20% of data. Later we compared the test accuracies of VGGNet and ResNet to determine the best accuracy rate specific to the tagging scheme.

In the speech-tag set model, we provided the computational model with the following 20 features as input variables:

Sender: It can be either an interviewer or an interviewee.

Speech Instance: It can be one of the following items: *Speech, Speech while Laughing, Asking a Question, Confirmation, Pre-Speech, Speech Pause, Micro Pause, Thinking, The Repetition of Question, Signaling End of Speech, Questionnaire Filling, Greeting, Read-Question, Laugh*

Gender: Gender information was held for both the interviewer and the interviewee participant in a separate column.

Is the Same Person: It is a polar question to specify whether the current individual is the same person with the one in the previous line. This value was stored separately for the interviewer and the interviewee participants.

Gaze Behavior: If raw gaze data of the participant could not be extracted and/or there was a problem in face detection, we omitted those cases. Thus, the value for gaze behavior can be either *Aversion* or *Face Contact*. This value was stored separately for the interviewer and the interviewee participants. We treated the interviewee's gaze behavior as an input feature when constructing a model for

predicting the interviewer's gaze behavior, i.e., the target variable was the interviewer's gaze behavior, and vice versa.

On the other hand, in the dialogue-act tagging model, a total of 137 features involving the following input variables in addition to *Sender*, *Gender*, *Is the Same Person* and *Gaze Behavior* features:

Communicative Function: We annotated conversations of 15 sessions with Dialogue-act tagging. We encountered 43 out of 56 communicative functions, except the following 13 functions: *Correction, Accept Offer, Decline Offer, Decline Request, Decline Suggestion, Auto Negative, Allo Negative, Feedback Elicitation, Return Self Introduction, Question, Address Offer, Address Request, Address Suggest* (see Table 3 for the whole list) One or more functions might be assigned per utterance.

Dimension: It represents the type of semantic content. ISO 24617-2 proposed nine dimensions and we encountered all of them: *Task, Turn Management, Time Management, Auto Feedback, Own Communication Management, Discourse Structuring, Social Obligation Management, Allo Feedback, Partner Communication Management.* One or more dimensions might be assigned per utterance.

Certainty: It can be one of the following items: *Certain, Uncertain, Empty*

Sentiment: It can be one of the following items: *Joy, Surprised, Empty*

Functional Dependence: It represents whether there is a functional dependence or not.

Feedback Dependence: It represents whether there is a feedback dependence or not.

Rhetorical Relation: It represents the coherence of text and discourse. There are four options: (i) an utterance might not have a RR relation (characterized by *Empty* label), (ii) an utterance is associated with a single RR (characterized by one of the RR labels), (iii) an utterance is associated with the same RR multiple times (characterized by multiple option of the related RR), (iv) an utterance is associated with multiple RR (characterized by the related multiple RR labels). As a result, a total of 37 labels are available: 18 labels for the RR category, 18 labels for multiple choice of each RR category and a single label for the *Empty* category.

Argument Number of Rhetorical Relations: Argument number represents the argument-order of an utterance in a related RR. Each RR has two arguments, a first and a second. Since the same RR might be associated with an utterance multiple times, an utterance might be both the first and the second argument of a certain RR. This feature characterizes the state of being the first and second argument for a

particular RR. Eventually, a total of 36 labels are available for this category, 18 labels for being the first argument and the remaining for the second one.

Lastly, we preprocessed categorical data by applying One-Hot-Encoding, in which a feature with n category is represented by n variable, instead of a single one. The new variables were coded as numbers: 1 represents the presence of category and 0 otherwise. Such a transformation also enabled us to handle the multi-dimensional aspect of Dialogue-act tagging. Consequently, instead of assigning a value to a single column holding multiple categories, we assigned the value to each category individually.

6.3. Results

We trained models on Google Colab which is a free Jupyter notebook environment provided by Google. Colab offers Tesla K80 GPU, using Keras and Tensorflow. We implemented training codes in Python 3.0 by Keras libraries with Tensorflows being its backend[19]. First, we have trained the simplified version of the VGG network with either 16 or 32 filters. We trained the models by applying 5-fold backtesting on the separated training, validation and test sets. We also applied generally accepted methods for the following problems: (i) validation loss was much higher than the training loss, i.e., overfitting, (ii) training took too long, and (iii) model performance was poor on the test data set, i.e underfitting. Batch normalization, pooling, weight and dropout regularization were applied to the proposed networks. In Figure 48, the updated gazeVGG architecture is presented.

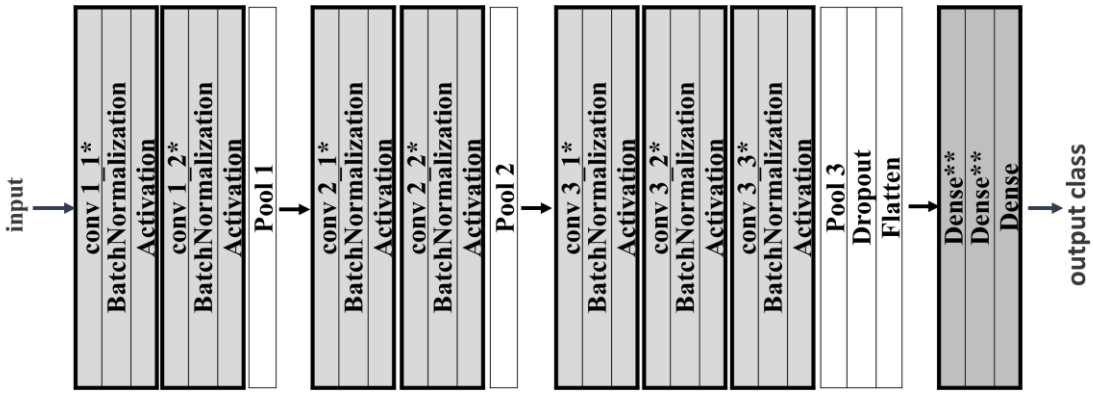


Figure 48: GazeVGG architecture with batch normalization, regularization and pooling. * L2 kernel and L2 bias regularizers were applied. **L2 kernel regularizer was applied.

As stated in the second research question, we are trying to investigate the performance differences among computational models based on the data annotated either with

[19] Colaboratory Notebooks including training codes will be publicly available under https://github.com/ulkursln

dialogue-act or speech-tag set. Since the number of input features depends on the speech annotation scheme, dialogue-act annotation resulted in more features. Taking that difference in input numbers into account, we tested the performances on multiple models with varying filter-lengths. In addition to VGG, the other model trained was the simplified version of ResNet. To apply batch normalization and regularization techniques, we made further improvements on the source code of ResNet mentioned in Fawaz et al. (2019), see Figure 49 for the architecture.

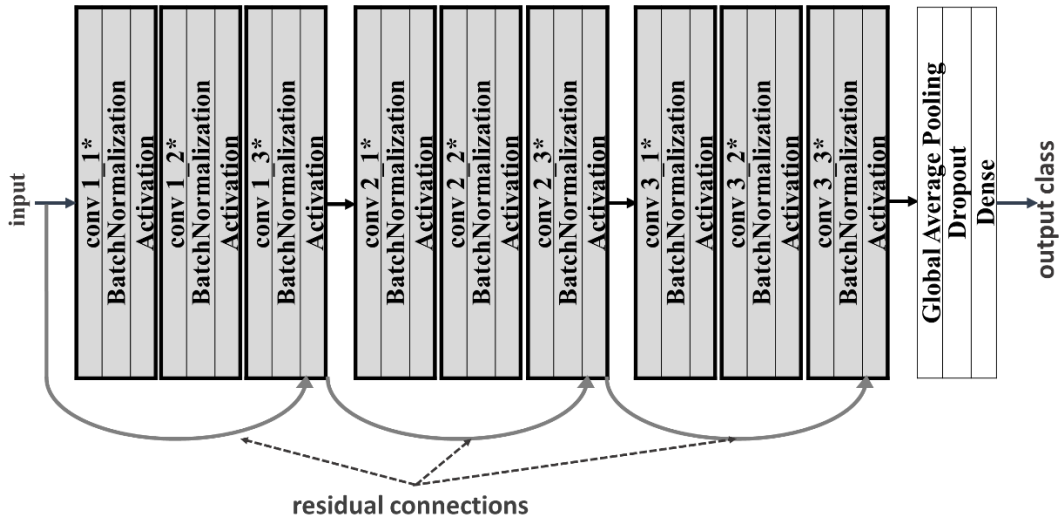


Figure 49: Simplified ResNet architecture with batch normalization, regularization, and pooling. There is a convolution between blocks and a residual connection between the last item of the previous block and the current one, which means the training in between layers is omitted. * L2 kernel and L2 bias regularizers are applied.

As a result of 5-fold back-testing, 16 is the optimal filter number for training VGG network with input data involving dialogue act annotation, and 32 for the remaining (see, appendix I). We performed 5-fold cross-validation on the models created with optimal filter numbers and input data created by shuffling the orders of interviewers. In Table 18, we summarized the performances of computational models. According to the results, computational models running on the data annotated with Speech-tag set generally perform better than the ones running on the data annotated with Dialogue-acts.

Table 18: Performances of computational models with 5-fold cross-validation. The highest test accuracy was obtained from the ResNet model that had 32 filters and received data annotated with the Speech-tag set method.* represents the filter numbers in the first block.

Tagging Scheme	CNN Architecture	Filter-Number*	Avg. Training Accuracy (%)	Avg. Test Accuracy (%)
Dialogue-Act	*VGG*	16	83.2 (SD: 1.20)	69.6 (SD: 11.3)
	ResNet	32	83.1 (SD: 0.88)	70.7 (SD: 12.3)
Speech-Tag Set	*VGG*	32	81.1 (SD: 0.18)	76.9 (SD: 5.82)
	ResNet	32	81.1 (SD: 0.14)	**78.8** (SD: 5.94)

As can be seen in Table 18, there is a 10% difference between training and test accuracy performances in the models that receive the data annotated by ISO 24617-2 standard. In order to get a more robust estimation about how accurately models make predictions on unseen data, we then performed 10-fold cross-validation on those data by splitting the last 10% of data for testing in each iteration., see Appendix J for the interviewers' orders. We obtained accuracy performances similar to the 5-fold validation, see Table 19.

Table 19: Performances with 10-fold cross-validation. The highest test accuracy was obtained from the ResNet model with 32 filters. .* represents the filter numbers in the first block

Tagging Scheme	CNN Architecture	Filter-Number*	Avg. Training Accuracy (%)	Avg. Test Accuracy (%)
Dialogue-Act	*VGG*	16	82.9 (SD: 1.1)	69.9 (SD: 12.9)
	ResNet	32	82.8 (SD: 0.8)	70.3 (SD: 12.8)

In order to examine the quantitative differences between models classification accuracy, we also created confusion matrices that contain the ratio of false and correct estimations, see Table 20.

Table 20: Confusion Matrix of the models with the highest performances for each tagging scheme. It represents the percentages of true and false predictions made on actual classes, i.e., aversion and face contact. The percentage of true aversion predictions is 76.3% for the Speech-tag set scheme, while it is 54% for the dialogue act scheme.

		Predicted Class			
		Speech Tag Set		Dialogue Act	
		Face Contact	Aversion	Face Contact	Aversion
Actual Class	Face Contact	85.1%	14.9%	94.8%	5.2%
	Aversion	23.7%	76.3%	46%	54%

CHAPTER 7

GENERAL DISCUSSION AND CONCLUSION

7.1. Discussion

Gaze provides an effective way to receive and send information in a face to face interaction, similar to other non-verbal communication channels accompanying speech. When studying gaze and speech, it is necessary to decide from which level both models will be addressed. We can examine the gaze through low-level eye movements such as *saccadic movements* which occur while focusing on small parts of an object to get higher resolution in resolving the grand picture of it, *vestibulo ocular reflexes* necessary for fixing the position of the moving object in the retina, or *smooth pursuit movements* that take a role in the perception of depth and tracking moving objects. As an alternative, features underlying biologically realistic eye movements such as the physiology of the eye and eyelid, or motion kinematics of eye, eyelid and combination of eye-head movements can be employed. Low-level eye movements, anatomic features of the eye and kinematics of eye movements have been extensively studied by physiologists. As a result, a wealth of information has been obtained to enable biologically realistic eye animations in virtual agents or robots. However, although there exist studies in the related fields, eye movements have some other high-level characteristics still waiting to be resolved, like when they occur, how long they last, what their roles are in communication and what their relations with cognitive status are (Ruhland et al., 2015).

As in the gaze studies, researchers have dealt with the speech at different levels for modeling non-verbal communication components driven by speech. In the Zoric, Forchheimer, and Pandzic (2011) study, they produced facial expressions in real-time with prosodic information obtained from speech signals. In the modeling of gaze behavior, Marsella et al. (2013) went one step further. In addition to the prosodic information of speech signals, they also employed superficial semantic information. Moreover, to identify roles of a conversation, Cassell, Torres, and Prevost (1999) proposed information structures, particularly the *theme* that associate the previous discourse with the current discourse and *rheme* that presents the new information regarding the theme. In the present study, we investigated the roles of the high-level characteristic of eye movements driven by high-level features of speech in face-to-face interaction.

7.1.1. Gaze in Relation with Speech

The first research questions of the present study are as follows: "What are the underlying features of gaze behavior among humans" and "What is the relation between gaze and speech to achieve conversational goals in a specified face-to-face interaction environment?" To examine this question, we first conducted a pilot study with three pairs. It was a mock job interview task comprising the Turkish translations of eight common job interview questions adapted from Villani et al. (2012). Then, we performed the next study with the experience gained from the pilot study and by addressing the constraints encountered in experimental procedures, methodology and analysis of previous studies. Twenty-eight pairs consisted of seven professional interviewers and 28 interviewees took part in the study. The participants in the pairs had not met before. They wore Tobii glasses throughout the study. Tobii glasses record eye-tracking data, sound and scene camera video that has a frame rate of 30 fps, which means the duration of each frame is 33 ms. The Interviewers read questions and evaluated the interviewees' responses on the Wacom PL-1600 15.6 Inch Tablet which enabled users to interact with the screen by using a digital pen. Prior to the experiment, we allowed the interviewer to test the interview interface on the screen and ensured that the interviewer got familiar with interview questions and evaluation criteria, and gained enough experience with the digital pen. Moreover, to increase the motivation for participation, we asked the interviewee to think about the position and the company that he or she wanted to work for and to answer the questions posed by the interviewer by aiming for this position. We also shared this information with the interviewer at the beginning of the experiment. To conduct a more realistic interview, in addition to the given eight questions, interviewers were allowed to ask further ones, if they thought it was necessary. There was no time limit for the study. The participants stayed alone in the room throughout the session. In addition, we adjusted the lighting of the room accordingly so as to improve the quality of the collected eye data.

We first analyzed gaze behavior. Since both participants were wearing Tobii eyeglasses during experiments, a total of six video streams were extracted from the recordings of the pilot study, and a total of 56 video-streams were extracted from the recordings, 28 of them obtained from interviewers' recordings and the remaining from interviewees. Gaze behavior is identified as either face contact or gaze aversion by detecting whether the participant is looking at the other person's face or not. We investigated gaze aversion in the pilot study, whereas, in the next study, we expanded the analysis and examined both face contact and gaze aversion. The gaze analysis was carried out in three steps: (i) determining the boundaries of the face, i.e., face detection, (ii) deciding whether the partner's gaze was within those boundaries, i.e., identification of gaze behavior, (iii) fixation detection of related gaze behavior. In the second step, gaze behavior was identified for a single raw data, while in the third step, a fixation of gaze behavior was defined, which is the process of handling a group of raw gaze data rather than a single one.

We, first, used the Viole-Jones (Viola & Jones, 2001) method for face detection, then, in order to minimize the number of undetected faces, we used face detection and face

tracking methods together. In cases where the Viole-Jones method could not detect the face, Camshift face-tracking was executed with the coordinates of the last detected face. Besides, we added the Kalman filter, since Camshift method might not produce very robust results against background noise. As a result, although we achieved to detect faces with better performances comparing to using Viole-Jones method alone, the face boundaries were constructed by a rectangle shape. The assumption that the faces are rectangular can lead to wrong estimation of gaze behavior, especially when the raw gaze data was near the corners of the rectangle, see Figure 7. Therefore, to specify more realistic face boundaries, we adopted OpenFace framework which defines face boundaries over 68 facial landmarks, see Figure 16. We used the information we gained during the gaze analysis of the pilot study in the development of MAGiC application, which we discussed in the next section.

We performed the gaze behavior analysis of 28 pairs via MAGiC application. We monitored the ratio of unidentified gaze behavior of each recording and observed that the identification rate of 11 interviewees and two interviewers were less than 70%. The visualization function of MAGiC enables us to reveal the underlying reason for such a difference in the identification rate between interviewers and interviewees. Even there was a raw data of interviewees, the interlocutors' (i.e., interviewers') face might not be detected while they were reading a question or evaluating the responses of an interviewee by turning their head and face to the screen. For such cases, we trained a custom face detector instead of using Haar-Cascade classifiers which were provided by OpenFace as the default detector. Face tracking with the custom detectors improved the gaze analysis on the recordings for one of two interviewers and all 11 interviewees.

The reason for the failure of identifying gaze behavior is either the undetected partner's face or the person's missing gaze data. In order to minimize data loss, we manually determined the related gaze behavior on frame-images with an interface provided by MAGiC. We manually assigned or updated the identity of raw gaze behavior for the following situations: (i) the partner's face was on the frame, however, the face could not be detected automatically or was detected incorrectly. (ii) the partner's face does not exist for a specific frame-image, thus even the raw gaze data of the participant for that particular frame-image cannot be obtained, aversion can be assigned as gaze behavior. After all these processes, we re-monitored the ratio of undefined gaze behaviors, and then we excluded the data of the 3 pairs that still have an identification rate below 70%. In this way, we completed the face detection and identification of gaze behavior which are the first two steps of gaze analysis. As a result, using gaze raw data we ended up with whether the gaze behavior on each frame-image was gaze aversion or gaze on a face.

Raw gaze data includes noise and saccadic movements which are rapid and designed to direct the fovea to the vision of interest. They are identified as the jump from one fixation to another. Saccadic behavior might be important for particular research questions like searching for visual targets, but in the present study, since we focused on maintaining gaze on the interlocutor's face or out of the face, we should eliminate jumping behaviors as

well as noise from the data. Fixation identification algorithms are employed to group the POR data within a specified neighborhood or velocity.

Classical eye-tracking systems (i.e., remote eye tracking) require users to sit against the screen without moving their heads. On the other hand, wearable or mobile eye tracking offers the user the freedom to move around and interact with the real dynamic world. Hessels, Niehorster, Nyström, Andersson, & Hooge (2018). pointed out that while classical eye tracking allows identifying head-centered fixations, mobile eye tracking should consider head movements and larger world space, i.e., world centered approach. This is important for preventing to present the eye movement metrics that cannot be compared with each other. The fixation detection algorithms in the software packages of some eye tracking device manufacturers generally offer black box solutions. However, we thought it would be useful to state the fixation algorithm explicitly in order to represent comparable results.

In the present study, we detected fixations in a few steps decided in line with the findings in the literature and the information contained in the manual of the eye-tracking device that we utilized during the study. We firstly interpolate missing data. The maximum gap length that would be filled with interpolation was chosen to be shorter than an ordinary blink which was 75 ms as proposed by previous studies (Benedetto et al., 2011; Ingre et al., 2006; Komogortsev et al., 2010) Then, we merged adjacent gaze behaviors of the same type (i.e. both aversion or face contact) between which there were again a gap shorter than an ordinary blink, i.e. 75 ms. Finally, gaze behaviors under 100 ms were discarded as recommended by Manor and Gordon (2003). Thus, we completed preparation for gaze analysis and then we passed to the speech analysis.

We handled speech annotation in two ways: (i) discourse and rhetorical relations with ISO 24617-2 and ISO 24617-8 respectively, (ii) an alternative set of speech tags that we produce based on the roles attributed specifically to the gaze in social communication studies. Our aim of annotating speech with two different methods is to investigate which characteristics of speech will produce better performance in modeling social gaze. In the pilot study, we only used the speech tags for annotating speech (viz. speech-tag set). We updated the tag set proposed in the pilot study for the speech annotation performed in the experiment. The updated set we proposed for speech annotation was determined based on studies on the role of eye movements in social communication as well as our observations on the data we collected. We considered the followings while creating and updating the tag set:

- We identified separate labels for communication, requesting information and providing feedback, which are functions of communication. (*Speech*, *Asking a Question*, *Confirmation*)

- We assigned labels to the pauses. We classified pauses by their duration and role as proposed by Heldner and Edlund (2010). (*Pre-Speech*, *Speech Pause*, *Micro Pause*)

- In parallel with the turn management role of speech, we labeled statements implying turn change and indicating that the speaker will release the turn at the end of the speech. (*Signaling End of Speech).*

- We named the conversation segment as thinking when it included filler sounds, such as uh, er, um, eee, and drawls. We defined a separate label for thinking, because, we assumed that gaze behavior might be effected in the course of thinking with the aim at reducing the burden of what is being looked upon (*Thinking)*.

- As the interviewer reads the questions from the screen, the interviewer's gaze would evidently be directed towards the screen, so we tagged this case separately (*Read Question*).

- A separate label for repeating the question is identified instead of annotating related speech segments as *Speech* or *Thinking* since repeating the question is both a sort of speech and it might be used by participants to save time to think (*Repetition of the Question*).

- We assumed that gaze behavior would be affected by laughter, thus, separate labels are defined for them. (*Laugh, Speech While Laughing*)

- The interviewers evaluated the interviewee's answer before proceeding to the next question. This evaluation process performed by looking at the screen. We categorized it separately from *Asking a Question* because in the meantime there is generally a no verbal or non-verbal exchange of information (*Questionnaire Filling*).

- We handled greeting apart from *Speech* because we assumed that the sender would aim to signal intimacy while greeting and this might have an effect on gaze behavior (*Greeting*).

In speech tag set analysis, we use MAGiC for segmentation and synchronization of pair recordings. Later on, we annotated segments of each session by using the “Annotation” interface of MAGiC.

As well as speech-tag set annotation, we annotated speech with an alternative dialogue-act model. For the analysis of the dialogue act, we first transcribed the conversations by listening to the audio streams of both the interviewer and the interviewee in each session. After that, we listened to the audio streams once more to add non-verbal vocalizations such as *Unfinished Word, Filler Sound , Laugh, Drawl, Warm-up and so on.* Adding non-verbal vocalizations is recommended by the standard depending on they have an effect on the choice of communicative function, or qualifiers. At this point, we also updated the transcribed text in case there were missing words.

As the last step of the transcription phase, we separated the transcription file of a session according to the sender. Thus, we ended up with two transcribed text files, one involved

the transcribed text of the interviewer and the other of the interviewee. After that, using the Praat program, three students marked the time interval of a total of 16716 words in 15 out of 25 sessions. When selecting these 15 sessions, we have given priority to long sessions in which dialogue act and RR tagging might be more frequent. Each marker file was checked by another person. In addition, transcribed text files were once more updated if non-verbal vocalizations or words that were not included before and caught while listening sessions on Praat. Thus, transcribed text files were reviewed four times in total since its creation and the word intervals were checked by two people. We performed dialogue act annotation on those transcribed texts of each session.

Dialogue act represents the communicative function that serves in a dialogue to change the state of mind of an addressee by means of its semantic content. We employed ISO 24617-2 which is a semantically based standard for dialogue annotation. ISO 24617-2 proposed nine dimensions based on the type of semantic content: *Task, Turn Management, Time Management, Auto Feedback, Own Communication Management, Discourse Structuring, Social Obligation Management, Allo Feedback, Partner Communication Management* and 56 communicative functions. In the present study, we encountered 43 out of 56 communicative functions, except the following ones: *Correction, Accept Offer, Decline Offer, Decline Request, Decline Suggestion, Auto Negative, Allo Negative, Feedback Elicitation, Return Self Introduction, Question, Address Offer, Address Request, Address Suggest*. The multi-dimensional approach of this standard allows multiple functions or dimensions to be assigned to a single utterance. When the speaker repeats the question posed to him, he or she might signal the following simultaneously: taking the turn, understanding the question and the need for time to answer. This standard allows handling such cases. Another strength of the standard is that many frameworks of the dialog act annotation neglect some minor nuances that the speaker wants to convey to an addressee. The utterance that the speaker is uttering for giving information should be labeled with *Inform*. However, the assignment of *Inform* alone cannot indicate whether the speaker is sure about the information that he or she is giving, or this information makes him or her happy, or this information is conditional. For such cases, ISO 24617-2 recommends three qualifiers: (i) *Certainty*, (ii) *Sentiment*, (iii) *Conditionality*. Moreover, this standard recommends ISO 24617-8 or better known as ISO DR-Core for Rhetorical Relation annotation. To understand the discourse as a whole, the relation between the sentences or clauses in the discourse (i.e., Rhetorical Relations) should be considered. In this standard, 18 labels are recommended for RR. In the present study, all 18 labels were included.

In order to ensure that the communicative functions are assigned as accurately as possible, as proposed in the annotation guideline, we have identified segments as the minimal stretch of utterances having a communicative function. While assigning functions to the identified segments, we tried to understand what the speaker meant by imagining ourselves in the place of an addressee. As suggested in the annotation guideline, whatever the way the speaker expressed himself, we considered following questions during annotation:(i) why did the speaker say it, (ii) what is the purpose of the speaker in using this utterance, and (iii) what are the speaker's assumptions about the person he was

addressing. ISO 24617-2 indicates that labeling should be based on the speaker's intention, instead of what he or she says literally. Therefore, this standard proposes to think functionally rather than relying on Linguistic cues, which are useful, but focusing only on them could make us miss what the speaker really wants to say and that would cause false labeling. In English, question sentences often contain words starting with "wh", such as "what can I know?", "which rule is violated?". They are labeled as *Set Question* according to the ISO 24617-2 scheme. But we can't label all the sentences in this form as *Set Question.* For instance, "why don't you go tomorrow" might be a suggestion rather than a question, depending on the context. Another issue to consider is to pick more specific communicative functions while annotating the functional segment. For instance, *Check Question* is a *Propositional Question*, but it additionally expects that the answer will be positive. In the present study, for segmentation and annotation of dialogue act, as well as considering the all criteria mentioned above, we studied the ISO 24617-2 standard, its revised version, the manual of annotation (Augmented Multiparty Interaction Consortium (AMI)., 2005; Bunt et al., 2012, 2017; ISO, 2012) the sample annotations given in DialogBank (Bunt et al., 2018) and we benefited from the TED-Multilingual Discourse Bank in order to get more insight into discourse annotations in Turkish(Zeyrek et al., 2019). We adopted DiAML (Dialogue Act Markup Language)-Multitab representation because it is human-friendly and easy to understand from the perspective of a third person. In DiAML-Multitab excel, we automatize the process of assigning unique ID's and updating references by using the developed macro.

We performed statistical analysis by merging the gaze and speech analysis data into a single summary file. Each line of the generated text file corresponded to a particular frame-image and the columns are gaze behavior of an interviewer and an interviewee on that particular frame-image and either the features related with speech-tag set analysis, such as sender, speech-tag label or the features related with dialogue act analysis such as, communicative function, RR label, qualifier and so on. As a result, we obtained two summary files, one involves the features of speech-tag set and the other dialogue-act, a series of gaze behavior and related features taken at successive intervals of 33 ms.

All statistical analyses were carried out in the R programming language (R Core Team; 2016). Instead of averaging the data points of subjects, we performed mixed models in order to provide independence of data points in a linear model. We represented individual differences by adding both fixed and random effects to a model.

We observed that interviewees performed face contact and gaze aversion more frequently when compared to interviewers. Moreover, the gaze aversion-durations of interviewers were longer than that of interviewees. On the other hand, face contact durations of interviewees were longer than that of interviewers. When we examined gaze behavior per role, there was no difference between the frequencies of gaze aversion and face contact for interviewers, while a difference was observed for interviewees. Interviewees avert their gaze more frequently compared to performed face contact.

The findings are in line with the conclusions summarized by Kendon (1967) in his detailed study investigating the function of gaze in face-to-face conversation. Kendon (1967) stated that individuals tend to look at others more frequently when listening compared to speaking and the glances of speakers would be shorter than the listeners. He had grouped the roles in the conversation as speakers and listeners. In the present study, due to the role of interviewees, they spoke more frequently than the interviewers. Comparing interviewers and interviewees, the gaze behavior of the latter was more similar to that of the speakers mentioned in Kendon (1967).

Broz et al. (2012) studied mutual gaze in a face-to-face conversation with participants wearing ASL eye-tracking devices. They observed a mutual face gaze occurring for about 46% of a conversation. Rogers et al. (2018) also conducted a dual eye-tracking study and reported that the mutual face gaze comprised 60% of the conversation with 2.2 seconds duration on average. On the other hand, when cumulative data of all sessions are taken into account, we found a lower ratio in the present study, which was 34% and the average duration was 546.1 ms. There are two crucial steps in determining mutual face gaze: (i) deciding whether the gaze of an individual was inside the face boundaries of an interlocutor, and (ii) synchronization of recordings exported from eye-trackers. Broz et al. (2012) and Rogers et al. (2018) manually annotated gaze behavior in each frame. However, in the present study interlocutor's face boundaries were detected based on 68 facial landmark points and gaze behavior was generally decided automatically via MAGiC. For synchronization, in Broz et al. (2012) study, the experimenter produced handclaps at the start and end of an experiment. Then, to synchronize the pair's recordings, they watched the video files exported from the eye tracker's scene camera and manually determined the beginning and end of a conversation. On the other hand, Rogers et al. (2018) utilize behavioral annotation software, namely Mangold INTERACT[20] for manual and synchronous coding of pair's audiovisual files. In the present study, a beeping sound was generated to indicate the beginning of a session. We used the semi-automatic synchronization function of MAGiC, in which audio files of pairs were automatically divided into segments involving time interval information and then the audio-segments containing beeping sound was determined by listening to audio segments. The starting time of the following segment was set as the initial time of that participant's recording in a pair. Then MAGiC calculated the time offset between the pair's recordings which is necessary for synchronization. So we tried to ensure synchronization of pairs' recordings as precisely as possible. This is crucial because, for instance, even a 33 ms shift in recordings exported from a 30 Hz eye tracker will result in incorrect synchronization and consequently incorrect analysis. One of the reasons for the differences in the ratio of mutual face gaze may be the method employed for synchronization.

[20] For detailed information about Mangold INTERACT, see https://www.mangold-international.com/en/software/interact

Manual coding of gaze behavior might be the other reason since it is open to human-related errors. Manual coding involves the process of detecting the face boundaries. Comparing to the previous studies, we employed state of the art technologies for face boundary detection. Moreover, because of the hardware or operational constraints, eye-tracking devices might estimate gaze positions with deviations. Eye tracker manufactures provide the estimated error that is specific to device in degrees for the visual angle. In the present study, we utilized the MAGiC application which considers such error margins to estimate gaze behavior automatically, to visualize gaze and face boundaries overlaid on a frame-image for enabling manual annotation. It is not possible to code gaze behavior manually when this margin of error is taken into account. For instance, Rogers et al. (2018) used 15 pixels for the size of the circle that represents the gaze position. They decided on a size of 15 pixels to achieve a balance between comfort in the coding process while providing distinguishable regions. In addition, Broz et al. (2012) studied with raw gaze data, while Rogers et al. (2018) employed a fixation extraction function provided by the eye-tracking manufacturer. Working on fixations rather than PORs not only decreases the amount of data to be analyzed but also eliminates the noise and saccadic movements. We adopted a similar approach with Rogers et al. (2018) study and worked on fixations instead of raw data. However, there was a crucial difference. They extracted fixations with a block-box solution that we could not get an insight into the inner processing of the provided function so we cannot make inferences about its suitability for dynamic scenes. On the other hand, in the present study, the fixation extraction algorithm that is suitable for dynamic scenes was explained step by step to provide repeatability of the study.

Lastly, differences in eye-tracking equipment, cultures, spoken language and experimental procedures might be the underlying reasons for the variety of findings. For instance, we performed a mock job interview task whereas other studies conducted conversations without a predetermined topic and in Broz et al. (2012) study, the experimenter stayed in the room throughout the data collection, even though he stood out of the participant's sight.

Throughout this study, we tried to automate the analysis as much as possible by utilizing the state of the art methods. Thus, we aimed to overcome some methodological problems in the pilot study and to reduce the amount of human-related errors and the time necessary for annotation. In parallel with this aim, we developed MAGiC for the analysis of gaze which involves face detection, gaze behavior identification and speech analysis including segmentation, annotation and synchronization of pair's recordings.

7.1.2. MAGiC

MAGiC was developed to allow researchers from various disciplines to work on it without a technical background. It is a desktop application written in C# programming language and an open-source software application which is publicly available for non-profit use.

Comparing to remote eye tracking, mobile eye-tracking analysis has technical difficulties in recognizing and tracking objects in the dynamic scenes. MAGiC focuses on face

recognition, which is one of the most studied subdomains of object recognition. It automatically detects whether the extracted raw gaze data is gaze on the face of an interlocutor or aversion. In addition, it provides interfaces for speech analysis involving segmentation, synchronization of pair recordings and annotation of segments. MAGiC significantly reduces the time and effort required for manual annotation of eye and audio recording data. For instance, a 10-minute video recording extracted from a 60 Hz eye tracker contains 36,000 frame-images. It takes more than 10 hours to manually annotate the entire recording assuming that 1 second is required for annotation of a single frame. MAGiC automatically performs this analysis, and Depending on the capabilities of the computer used, MAGiC automatically completes this annotation around 10 minutes. It also significantly reduces effort and time required for the segmentation and annotation of audio recordings by segmenting audio recordings in a couple of seconds and by providing interfaces for the synchronization of pair's recordings and annotation of segments.

MAGiC employs OpenFace frameworks for gaze behavior analysis, CMUSphinx for audio recording analysis and dlib for machine learning of face tracking. Through the capabilities of those frameworks, it provides researchers with information that cannot be obtained in manual annotations, such as extracting the coordinates of some facial features like eyes or mouth and creating separate segments for pauses in a millisecond duration which is virtually impossible for human annotators to detect at this level of temporal granularity.

Facial recognition with the default detector provided by OpenFace may lead to poor face detection in some video recordings and consequently, give low gaze behavior detection ratio. MAGiC provides interfaces to monitor problematic recordings, to train custom detectors for face detection on those problematic recordings and to perform face detection with the trained detectors. Moreover, it is developed based on the Separation of Concerns (SoC) design principle. Accordingly, each module in MAGiC can be used in isolation, for instance, it is possible to use MAGiC only for speech segmentation, face tracking, or annotation of speech segments.

We represented MAGiC's capabilities with the data gathered from the pilot study. The segmentation of the audio recordings was 1-2 seconds, and the annotation of segments took 10-20 minutes depending on the duration of the session. The face detection took 4-10 minutes depending on the duration of the video. Our analysis revealed that MAGiC identified gaze behaviors with a success ratio of over 80%.

Furthermore, we conducted a usability analysis of MAGiC. A total of eight participants took part in the usability study. They firstly installed the MAGiC application on their personal computer by using publicly available sources. Then, they performed the randomly assigned Gaze or Speech analysis both manually and using related interface of MAGiC, respectively. Finally, they assessed the usability of MAGiC using a 7-point scale ISO 9241/10 questionnaire. We sent them an excel file at the beginning of the study. They filled the related sheets and columns of excel with the annotations they made manually, the elapsed time they allocated for both manual and automatic analysis, and the notes they

responded with for the questionnaire. We compared the elapsed time of manual and automatic analysis. The mean duration to annotate a single frame-image decreased from 29.1 seconds (SD=22.7) to an average value of 0.09 seconds (SD=0.02); and in the speech analysis, the mean duration for annotation of a single segment decreased from 44.5 seconds (SD=8.8) to an average value of 7.1 seconds (SD=1.4). Furthermore, all of the usability metrics were scored higher than the average, namely 5.29 out of 7 for gaze analysis and 6.61 out of 7 for speech analysis. In order to investigate the reason for the differences in usability scores in speech and gaze analysis, we asked participants about their ratings. Even though they do not work specifically in this field, people are showing more tendency to comprehend the speech analysis and the motivation behind it when compared to the gaze analysis. For this reason, we think that the usability of gaze analysis would be even higher for the researchers working in this field. Since MAGiC is an open-source project for behavior research, it is open to further developments from the community. Some of the features that can be added later are as follows: automatic detection of facial expressions by using Facial Action Coding System (FACS) which is already provided by MAGiC, detection of gaze on specified objects or regions or face, and automating speech annotation.

7.1.3. Computational Models

The secondary research questions of the present study are as follows: "How can we computationally model gaze behavior with the high-level features of speech" and "How appropriate is employing discourse analysis scheme, namely ISO 24617-2 standard, in a computational model of gaze behavior?" To this aim, we trained two common Convolutional Neural Network (CNN) architectures, namely VGGNet and ResNet.

A Convolution Neural Network is a particular type of deep neural network and it is most commonly applied to the processing of 2D images. On the other hand, gaze data is in the form of a time series (i.e., gaze on face or an aversion) and one requires to employ 1D CNN, which is mostly used in NLP studies and gets the input data as a group of *timestamps* with a pre-defined *step-distance* between consecutive groups of *timestamps*. In the present study, we found that the average duration of gaze behavior including aversion and gaze on face was around 300 ms. Therefore, we assigned nine to timestamps which corresponds to 300 ms for the 30 Hz eye tracker. Moreover, we assigned three to step distance which corresponds to the minimum fixation duration, namely 100 ms.

We, first, preprocessed categorical data by applying One-Hot-Encoding, in which a feature with n category is represented by n variable, instead of a single one. For the input data including speech annotation with the speech-tag set, we provided a total of 20 features including *Sender*, *Speech Instance*, *Gender*, *Is the Same Person* and *Interviewee's Gaze Behavior*. On the other hand, we provided a total of 137 channels involving *Sender*, *Gender*, *Is the Same Person*, *Interviewee's Gaze Behavior, Communicative Function*, *Dimension*, *Certainty*, *Sentiment*, *Functional Dependence*, *Feedback Dependence*, *Rhetorical Relation* and *Argument Number of Rhetorical Relations*. We conducted 28 sessions with seven professional interviewers and 28 interviewees, where each

interviewee took part in a single session and an interviewer attended more than one session. Therefore, we collected more data for a single interviewer compared to an interviewee. We trained computational models to predict the gaze behavior of interviewers.

We trained VGG and Resnet models with 16 or 32 filters in the first block and taking an input data annotated either with a dialogue act or speech-tag set. In the parameter tuning phase, we used backtesting that is specific to the time series as a cross-validation method. For the data annotated with dialogue acts, ResNet with 32 filters and VGG with16 filters, and for the data annotated with a speech-tag set, 16 filters for both VGG and ResNet achieved better accuracies. After we decided the filter size in the first block, we trained models with the decided parameters and evaluated the models' performances by building 5-fold cross-validation with data sets created by shuffling the orders of interviewers in the input data. We observed that ResNet models achieved better accuracies for both annotation methods due to VGG bottleneck which causes loss of generalization capability after some depth whereas ResNet handles this vanishing gradient problem by using residual connections. Moreover, we found that the speech tag set gave rise to better performances compared to dialogue-act annotations. To get further insight into the differences in performances, we draw a confusion matrix representing the percentages of true and false predictions made on the actual gaze behavior. Although both ResNet models predicted face contact with higher accuracies, the dialogue-act method was not good at predicting aversions. The probable reasons might be the differences in the number of features and the number of input sessions. In addition, speech-tag set involves *Pre-Speech*, *Speech Pause* and *Micro Pause* for annotation of pauses whereas dialogue act annotation does not handle pauses.

7.2. Concluding Remarks and Future Directions

We investigated gaze accompanying speech in a face-to-face interaction. Firstly, we studied the characteristics of gaze and its relations with speech with an experimental research conducted via mobile eye tracking devices. The results indicate that the frequency and duration of gaze differ significantly depending on the role. We showed these differences could not be observed in the analysis performed with raw gaze data instead of detected fixations. As in some of the previous studies, performing gaze analysis with raw gaze data or with detected fixations by using black box solutions are inadequate to obtain comparable results. Moreover, in multimodal analysis, it is important to automate annotations with the state of the art methods. Manual annotation is vulnerable to human-related errors and in addition, automatic annotation with the state of the art methods provide further information that may not be extracted manually such as, detecting the coordinates of facial landmarks, taking into account the error margins while annotating the gaze behavior or segmentation of the speech at milliseconds precision. MAGiC offers an analysis environment for researchers working in the field without requiring a technical background. It is also open to future developments as it is an open source project.

Secondly, we developed CNN models of gaze behavior in a face-to-face interaction. The widely used VGGNet and ResNet architectures were adopted for this aim. In the first model, we annotated the speech data with the speech-tag set that we created by benefiting from the founding of previous social gaze studies and also by examining the data we have collected. In the second model, we used a particular semantic annotation framework proposed for dialogue act annotation, namely ISO 24617-2. The performance of the first model was higher than the second one. Our goal here was not to suggest an alternative scheme for discourse annotation. Due to the increasing number of Embodied Conversational Agents (ECAs) in recent years, studied in face-to-face interaction gain more importance, and it requires to handle interaction in a multimodal manner instead of speech in isolation. From this perspective, results showed that, in the computational model of gaze, comparing to performance of one of the discourse annotation scheme with a great effort behind it, annotation with a simple tag set performs better. Thus, multimodality should be taken into account when proposing an automatic speech annotation schemes. In addition, results showed that CNN allows us to predict high level features of eye movement with high level features of speech.

As future work, other non-verbal cues accompanying speech might be experimentally investigated to examine their characteristics, roles and relations in social communication. In addition, in order to see the effect of language, culture and personal differences, similar experimental investigations might be performed. Moreover, instead of mock job interview task in which the role of participants causes a kind of asymmetry between the participants, an experimental investigation of open conversation might be another future study. As an outcome of the present study, we also provided a Turkish corpus that involves the time intervals of each words. This corpus might be used in different studies for different aims, such as investigating the effects of conjunctions in the prediction of gaze behavior. In the future studies, it is important to obtain comparable results to facilitate the flow of interdisciplinary knowledge on a face-to-face interaction that is studied by many different disciplines including, linguistics, computer engineering, AI and psychology. Thus, it will be useful to support open source environments in the field. Because they enable researchers from different backgrounds to work together, and also allow progressive information flow between different disciplines.

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

Adams, R. B., & Kleck, R. E. (2003). Perceived Gaze direction and the processing of facial displays of emotion. *Psychological Science*, *14*(6), 644–647. https://doi.org/10.1046/j.0956-7976.2003.psci_1479.x

Adams, R. B., & Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, *5*(1), 3–11. https://doi.org/10.1037/1528-3542.5.1.3

Alexandersson, J., B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, & M. Siegel (1998). Dialogue acts in VERBMOBIL-2. Verbmobil Report 226. Saarbrucken: DFKI

Allen, J. & M. Core (1997) DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1). http://www.cs.rochester.edu/research/ cisd/resources/damsl/RevisedManual/

Andrist, S., Mutlu, B., & Gleicher, M. (2013). Conversational gaze aversion for virtual agents. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8108 LNAI, pp. 249–262). https://doi.org/10.1007/978-3-642-40415-3_22

Archer, D., & Akert, R. M. (1977). Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of Personality and Social Psychology*, *35*(6), 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

Argyle, M., Lefebvre, L., & Cook, M. (1974). The meaning of five patterns of gaze. *European Journal of Social Psychology*, *4*(2), 125–136. https://doi.org/10.1002/ejsp.2420040202

Augmented Multiparty Interaction Consortium (AMI). (2005). Guidelines for dialogue act and addressee annotation version 1.0. Unpublished script

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. F., Strodtbeck, F. L., Mills, T. M., & Roseborough, M. E. (1951). Channels of Communication in Small Groups. *American Sociological Review*, *16*(4), 461. https://doi.org/10.2307/2088276

Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019, February 1). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE Computer Society. https://doi.org/10.1109/TPAMI.2018.2798607

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

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

Baron-Cohen, S. (1994). How to build a baby that can read minds: Cognitive mechanisms in mind reading. *Curr. Psychol. Cogn.*, *13*(5), 513–552.

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

Basu, S., Choudhury, T., Clarkson, B., Pentland, A., & others. (2001). Learning human interactions with the influence model. *Proc NIPS Vancouver, British Columbia, Canada*. Retrieved from

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.768&rep=rep1&type=pdf

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566–580. https://doi.org/10.1093/joc/52.3.566

Beck, R. S., Daughtridge, R., & Sloane, P. D. (2002, January). Physician-patient communication in the primary care office: A systematic review. *Journal of the American Board of Family Practice*.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(3), 199–208. https://doi.org/10.1016/j.trf.2010.12.001

Binetti, N., Harrison, C., Coutrot, A., Johnston, A., & Mareschal, I. (2016). Pupil dilation as an index of preferred mutual gaze duration. *Royal Society Open Science*, *3*(7). https://doi.org/10.1098/rsos.160086

Broz, F., Lehmann, H., Nehaniv, C. L., & Dautenhahn, K. (2012). Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (pp. 858–864). https://doi.org/10.1109/ROMAN.2012.6343859

Bunt, H. (1994). Context and dialogue control. *Think Quarterly*, *3*(1), 19–31. Retrieved from http://www.cs.uu.nl/docs/vakken/uem/bunt.pdf

Bunt, H. (2006). Dimensions in dialogue act annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006* (pp. 919–924). European Language Resources Association (ELRA).

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. *Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009)*, 13–24. https://doi.org/10.1038/ncomms2383

Bunt, H. (2012). Data categories for dialogue acts. Unpublished manuscript, Tilburg University.

Bunt, H. (2019). Guidelines for using ISO standard 24617-2. S.l.: [s.n.].

Bunt, H., Kipp, M., & Petukhova, V. (2012). Using DiAML and ANVIL for multimodal dialogue annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012* (pp. 1301–1308). European Language Resources Association (ELRA).

Bunt, H., Petukhova, V., & Fang, A. C. (2017). Revisiting the ISO standard for dialogue act annotation. In *Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. Retrieved from http://www.iso.org/diaml https://aclanthology.info/papers/W17-7404/w17-7404

Bunt, H., & Prasad, R. (2016). ISO-DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12) (pp. 45–54).

Bunt, H., Volha P., Andrei M., Alex C. F., and Kars W. (2018) The DialogBank: Dialogues with Interoperable Annotations'. *Language Resources and Evaluation*, pp. 1-37. DOI 10.1007/s10579-018-9436-9.

Camilli, M., Nacchia, R., Terenzi, M., & Di Nocera, F. (2008). ASTEF: A simple tool for examining fixations. *Behavior Research Methods*, *40*(2), 373–382. https://doi.org/10.3758/BRM.40.2.373

Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *In IEEE International Symposium on Circuits & Systems*

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory (pp. 1–10). Association for Computational Linguistics (ACL). https://doi.org/10.3115/1118078.1118083

Cassell, J. (2000). Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents. In J Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied Conversational Agents* (pp. 1–27). MIT Press.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjailmsson, H., & Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Conference on Human Factors in Computing Systems – Proceedings* (pp. 520–527). https://doi.org/10.1145/302979.303150

Cassell, J., Torres, O. E., & Prevost, S. (1999). Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. *Machine Conversations*, 143–154. https://doi.org/10.1.1.52.2297

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.

Choudhury, T., & Pentland, A. (2004). Characterizing social interactions using the sociometer. In *Proceedings of NAACOS* (pp. 1–4). Retrieved from

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.7399&rep=rep1&type=pdf

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Demberg, V., Scholman, M. C. J., & Asr, F. T. (2019). How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue and Discourse*, *10*(1), 87–135. https://doi.org/10.5087/dad.2019.104

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

Dunham, P. J., Dunham, F., & Curwin, A. (1993). Joint-Attentional States and Lexical Acquisition at 18 Months. *Developmental Psychology*, *29*(5), 827–831. https://doi.org/10.1037/0012-1649.29.5.827

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

Ekman, P., & Davidson, R. (1994). The nature of emotion: Fundamental questions. Series in affective science. In *The Nature of Emotion: Fundamental Questions* (pp. 56–58). Oxford University Press.

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.

Emery, N. J. (2000, August). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/S0149-7634(00)00025-7

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth.*Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9602–9605. https://doi.org/10.1073/pnas.152159999

Fasola, J., & Matarić, M. J. (2012). Using socially assistive human-robot interaction to motivate physical exercise for older adults. In *Proceedings of the IEEE* (Vol. 100, pp. 2512–2526). https://doi.org/10.1109/JPROC.2012.2200539

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. doi: 10.1007/s10618-019-00619-1

Fillmore, C. J. (1968). The case for case. In E. Bach, & R. T. Harms (Eds.), Universals in linguistic theory (pp. 1-88). New York, NY: Holt, Rinehart, and Winston.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. In *Robotics and Autonomous Systems* (Vol. 42, pp. 143–166). https://doi.org/10.1016/S0921-8890(02)00372-X

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

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920–1931. https://doi.org/10.1016/j.visres.2011.07.002

Fragopanagos, N., & Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural Networks*, *18*(4), 389–405. https://doi.org/10.1016/j.neunet.2005.03.006

Freire, A., Eskritt, M., & Lee, K. (2004). Are eyes windows to a deceiver's soul? Children's use of another's eye gaze cues in a deceptive situation. *Developmental Psychology*, *40*(6), 1093–1104. https://doi.org/10.1037/0012-1649.40.6.1093

Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., & Hagita, N. (2002). Messages embedded in gaze of interface agents - Impression management with agent's gaze. In *Conference on Human Factors in Computing Systems – Proceedings* (Vol. 4, pp. 41–48). https://doi.org/10.1145/503384.503385

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119–130. https://doi.org/10.1016/0893-6080(88)90014-7

Gabbott, M., & Hogg, G. (2000). An empirical investigation of the impact of non-verbal communication on service evaluation. *European Journal of Marketing*, *34*(3/4), 384–398. https://doi.org/10.1108/03090560010311911

Gallup, A. C., Chong, A., & Couzin, I. D. (2012). The directional flow of visual information transfer between pedestrians. *Biology Letters*, *8*(4), 520–522. https://doi.org/10.1098/rsbl.2012.0160

Gatys, L., Ecker, A., & Bethge, M. (2016). A Neural Algorithm of Artistic Style. *Journal of Vision*, *16*(12), 326. https://doi.org/10.1167/16.12.326

Gerwing, J., & Allison, M. (2009). The relationship between verbal and gestural contributions in conversation: A comparison of three methods. *Gesture*, *9*, 312–336.

Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, *136*, 359–364. https://doi.org/10.1016/j.cognition.2014.11.040

Goldman-Eisler, F. (1968). *Psycho-linguistics: Experiments in spontaneous speech*. New York, NY: Academic Press.

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning. Cambridge, MA*: MIT Press.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). *Maxout networks. In 30th International Conference on Machine Learning, ICML* 2013 (pp. 2356–2364). International Machine Learning Society (IMLS).

Goodwin, C. (1980). Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry*, *50*(3–4), 272–302. https://doi.org/10.1111/j.1475-682X.1980.tb00023.x

Goodwin, C. (1981). Conversational Organization: Interaction Between Speakers and Hearers. *Conversational Organization: Interaction Between Speakers and Hearers*, (September), 173.

Gorman, W. G., & Hall, G. D. (1964). Dielectric constant correlations with solubility and solubility parameters. *Journal of Pharmaceutical Sciences*, *53*(9), 1017–1020. https://doi.org/10.1002/jps.2600530905

Gredebäck, G., Johnson, S., & Von Hofsten, C. (2010, January). Eye tracking in infancy research. *Developmental Neuropsychology*. https://doi.org/10.1080/87565640903325758

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. (1986). Attention, Intention and the Structure of Discourse. *Computational Linguistics*, *12*(3), 175–204.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2016-December, pp. 770–778). IEEE Computer Society. https://doi.org/10.1109/CVPR.2016.90

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555-568. http://dx.doi.org/10.1016/j.wocn.2010.08.002

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. C. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science, 5*(8). https://doi.org/10.1098/rsos.180502

Hieke, A. E., Kowal, S., & O'Connell, D. C. (1983). The trouble with "Articulatory" pauses. *Language and Speech, 26*(3), 203–214. https://doi.org/10.1177/002383098302600302.

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

Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PLoS ONE, 10*(8). https://doi.org/10.1371/journal.pone.0136905

Holler, J. & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica, 146*, 81–116.

Holmqvist, K., Nystrom, 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.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 148*(3), 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Ide, N. (2017). Introduction: The Handbook of Linguistic Annotation. 10.1007/978-94-024-0881-2_1.

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., & Kecklund, G. (2006). Subjective sleepiness, simulated driving performance and blink duration: Examining individual differences. *Journal of Sleep Research, 15*(1), 47–53. https://doi.org/10.1111/j.1365-2869.2006.00504.x

ISO DIS 24617-2 (2010) Language resource management – Semantic annotation framework (SemAF), Part 2: Dialogue acts. ISO, Geneva, January 2010.

ISO DIS 24617-8 (2016) Language resource management – Semantic annotation framework (SemAF), Part 8: Semantic Relations in discourse, core annotation schema (DR-Core)

Izard, C. E. (1991). The Psychology of Emotions. New York, NY: Plenum Press

Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, *108*(1–2), 116–134. https://doi.org/10.1016/j.cviu.2006.10.019

Kanan, C., Bseiso, D. N. F., Ray, N. A., Hsiao, J. H., & Cottrell, G. W. (2015). Humans have idiosyncratic and task-specific scanpaths for judging faces. *Vision Research*, *108*, 67–76. https://doi.org/10.1016/j.visres.2015.01.013

Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., & Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4451 LNAI, pp. 91–112). https://doi.org/10.1007/978-3-540-72348-6_5

Kelly, S. D., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture and action during language comprehension. *Psychonomic Bulletin & Review*, *22*, 517–523.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*(C), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press

Kendrick, K. H., & Holler, J. (2017). Gaze Direction Signals Response Preference in Conversation. *Research on Language and Social Interaction*, *50*(1), 12–32. https://doi.org/10.1080/08351813.2017.1262120

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

Kim,G.W., & Kang, D.S. (2015). Improved Camshift algorithm based on Kalman filter, Adv. Sci. Technol. Lett., 98, pp. 135–137.

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.

Kleinke, C. L. (1986, July). Gaze and Eye Contact. A Research Review. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.100.1.78

Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye.*Nature*. Nature Publishing Group. https://doi.org/10.1038/42842

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

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11), 2635–2645. https://doi.org/10.1109/TBME.2010.2057429

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (Vol. 2, pp. 1097–1105).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2323. https://doi.org/10.1109/5.726791

Lee, A., Prasad, R., Webber, B., & Joshi, A. (2016). Annotating discourse relations with the PDTB annotator. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: System Demonstrations* (pp. 121–125). Association for Computational Linguistics, ACL Anthology.

Lees, R. B., & Chomsky, N. (1957). Syntactic Structures. *Language*, *33*(3), 375. https://doi.org/10.2307/411160

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B, 369 (1651)* 20130302, doi: 10.1098/rstb.2013030

LIRICS (2006a) D4.2, Preliminary set of semantic data categories. http://lirics.loria.fr.

LIRICS (2006b) D4.3, Documented compilation of semantic data categories. http://lirics.loria.fr.

Locker, L., Huffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, *39*(4), 723–730. https://doi.org/10.3758/BF03192962

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Manor, B. R., & Gordon, E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of Neuroscience Methods*, *128*(1–2), 85–93. https://doi.org/10.1016/S0165-0270(03)00151-1

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., & Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings - SCA 2013: 12th ACM SIGGRAPH / Eurographics Symposium on Computer Animation* (pp. 25–36). https://doi.org/10.1145/2485895.2485900

Mason, M. F., Tatkow, E. P., & Macrae, C. N. (2005). The look of love: Gaze shifts and person perception. *Psychological Science*, *16*(3), 236–239. https://doi.org/10.1111/j.0956-7976.2005.00809.x

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Mehrabian, A., & Wiener, M. (1967). Decoding Of Inconsistent Communications. *Journal of Personality and Social Psychology*, *6*(1), 109–114. https://doi.org/10.1037/h0024532

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

Meyer, T., & Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra). Association for Computational Linguistics*, *003*, 129–138. Retrieved from http://dl.acm.org/citation.cfm?id=2387973

Miltsakaki, E., Robaldo, L., Lee, A., & Joshi, A. (2008). Sense annotation in the penn discourse treebank. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4919 LNCS, pp. 275–286). https://doi.org/10.1007/978-3-540-78135-6_23

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, *20*, 336–366

Munn, S. M., Stefano, L., & Pelz, J. B. (2008). Fixation-identification in dynamic scenes (p. 33). Association for Computing Machinery (ACM). https://doi.org/10.1145/1394281.1394287

Myllyneva, A., & Hietanen, J. K. (2016). The dual nature of eye contact: To see and to be seen. *Social Cognitive and Affective Neuroscience*, *11*(7), 1089–1095. https://doi.org/10.1093/scan/nsv075

Olsen, A. (2012). The Tobii I-VT fixation filter. Copyright © Tobii Technology AB

Osako, K., Singh, R., & Raj, B. (2015). Complex recurrent neural networks for denoising speech signals. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/WASPAA.2015.7336896

Otsuka, K., Sawada, H., & Yamato, J. (2007). Automatic inference of cross-modal nonverbal interactions in multiparty conversations (p. 255). Association for Computing Machinery (ACM). https://doi.org/10.1145/1322192.1322237

Otteson, J. P., & Otteson, C. R. (1980). Effect of teacher's gaze on children's story recall. *Perceptual and Motor Skills*, *50*(1), 35–42. https://doi.org/10.2466/pms.1980.50.1.35

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., & Joshi, A. (2009). The hindi discourse relation bank. In *ACL-IJCNLP 2009 - LAW 2009: 3rd Linguistic Annotation Workshop, Proceedings* (pp. 158–161). https://doi.org/10.3115/1698381.1698410

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). Human computing and machine understanding of human behavior: A survey. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4451 LNAI, pp. 47–71). https://doi.org/10.1007/978-3-540-72348-6_3

Perrett, D. I., & Emery, N. J. (1994). Understanding the intentions of others from visual signals: Neurophysiological evidence. *Current Psychology of Cognition*, *13*(5), 683–694. Retrieved from http://doi.apa.org/psycinfo/1995-24608-001

Peter, H. W., & Chomsky, N. (1968). Aspects of the Theory of Syntax. *The Modern Language Review*, *63*(1), 132. https://doi.org/10.2307/3722650

Petukhova, V., & Bunt, H. (2012). The coding and annotation of multimodal dialogue acts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012* (pp. 1293–1300). European Language Resources Association (ELRA).

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*, *6*(11). https://doi.org/10.1371/journal.pone.0027591

Pfeiffer, U. J., Vogeley, K., & Schilbach, L. (2013, December). From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2013.07.017

Picard, R. W. (1999). Affective Computing for HCI. In H. J. Bullinger & J. Ziegler (Eds.), *Proceedings of the 8th HCI International on Human-Computer Interaction: Ergonomics and User Interfaces* (pp. 829–833). Munich: Lawrence Erlbaum Associates. Retrieved from http://dl.acm.org/citation.cfm?id=647943.742338

Popescu-Belis, A. (2016). Manual and automatic labeling of discourse connectives for machine translation. In TextLink–Structuring Discourse in Multilingual Europe Second Action Conference Karoli G´asp´ar University of the Reformed Church in Hungary Budapest, 11–14 April, 2016 ´(p. 16).

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

Prasad, R., & Bunt, H. (2015). Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings 11th joint ACL-ISO workshop on interoperable semantic annotation (ISA-11)* (pp. 80–92).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008* (pp. 2961–2968). European Language Resources Association (ELRA).

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Prasad, R. ;, … Webber, B. L. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. *IRCS Technical Reports Series*. https://doi.org/10.1136/bmj.331.7518.689

Prasad, R., Webber, B., & Lee, A. (2018). Discourse Annotation in the PDTB : The Next Generation. *Proceedings 14th Joint ACL - ISO Workshop on Interoperable*

Semantic Annotation, 87–97. Retrieved from https://aclweb.org/anthology/papers/W/W18/W18-4710/ http://aclweb.org/anthology/W18-4710

Prasov, Z., & Chai, J. Y. (2008). What's in a Gaze? The role of eye-gaze in reference resolution in multimodal conversational interfaces. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (pp. 20–29). https://doi.org/10.1145/1378773.1378777

R Team Core. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions. *Journal of Memory and Language*, *41*(3), 416–426. https://doi.org/10.1006/jmla.1999.2650

Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., … McDonnell, R. (2015). A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum*, *34*(6), 299–326. https://doi.org/10.1111/cgf.12603

Qu, S., & Chai, J. Y. (2009). The role of interactivity in human-machine conversation for automatic word acquisition. In *Proceedings of the SIGDIAL 2009 Conference: 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 188–195). https://doi.org/10.3115/1708376.1708404

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., … Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on 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. E., … Ansari, R. (2000). Gesture, speech, and gaze cues for discourse segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2*, 247–254. https://doi.org/10.1109/CVPR.2000.854800

Reade, C. (1989). Elements of Functional Programming. Boston, MA, USA: Addison-Wesley Longman.

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the Fourth Wall of Cognitive Science: Real-World Social Attention and the Dual Function of Gaze. *Current Directions in Psychological Science*, *25*(1), 70–74. https://doi.org/10.1177/0963721415617806

Rogers, S. L., Speelman, C. P., Guidetti, O., & Longmuir, M. (2018). Using dual eye tracking to uncover personal gaze patterns during social interaction. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-22726-7

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning, and culture. In *Conversation Analysis: Comparative Perspectives* (pp. 187–229). Cambridge University Press. https://doi.org/10.1017/CBO9780511635670.008

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696. https://doi.org/10.2307/412243

Sanders, T. J. M., Demberg, V., Hoek, J., Scholman, M. C. J., Asr, F. T., Zufferey, S., & Evers-Vermeul, J. (2018). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2016-0078

Schank, R. C., & Tesler, L. (1969). A conceptual dependency parser for natural language (pp. 1–3). Association for Computational Linguistics (ACL). https://doi.org/10.3115/990403.990405

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.

Sharp, R., Jansen, P., Surdeanu, M., & Clark, P. (2015). Spinning straw into gold: Using free text to train monolingual alignment models for non-factoid question answering. In *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* (pp. 231–237). Association for Computational Linguistics (ACL). https://doi.org/10.3115/v1/n15-1025

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the Classroom: Reciprocal Effects of Teacher Behavior and Student Engagement Across the School Year. *Journal of*

Educational Psychology, *85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

Srinivasan, V., Bethel, C. L., & Murphy, R. R. (2014). Evaluation of head gaze loosely synchronized with real-time synthetic speech for social robots. *IEEE Transactions on Human-Machine Systems*, *44*(6), 767–778. https://doi.org/10.1109/THMS.2014.2342035

Stern, H., & Efros, B. (2002). Adaptive color space switching for face tracking in multi-colored lighting environments. In *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002* (pp. 249–254). IEEE Computer Society. https://doi.org/10.1109/AFGR.2002.1004162

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 *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014* (pp. 5739–5742). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/EMBC.2014.6944931

Sundaram, D. S., & Webster, C. (2000, September 1). The role of nonverbal communication in service encounters. *Journal of Services Marketing*. https://doi.org/10.1108/08876040010341008

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, *57*(6), 1454–1463. https://doi.org/10.1111/j.1467-8624.1986.tb00470.x

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution*, *52*(3), 314–320. https://doi.org/10.1016/j.jhevol.2006.10.001

Trouvain, J., & Truong, K. P. (2012). Comparing non-verbal vocalisations in conversational speech corpora. In L. Devillers, B. Schuller, A. Batliner, P. Rosso, E. Douglas-Cowie, R. Cowie, & C. Pelachaud (Eds.), *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012)* (pp. 36–39). Paris, France: European Language Resources Association (ELRA). Retrieved from http://doc.utwente.nl/80906/

Turing A.M. (1950) Computing Machinery and Intelligence. *Mind* 49: 433-460.

Van Dijk, T. A. (1979). Pragmatic connectives.*Journal of Pragmatics*, *3*(5), 447–456. https://doi.org/10.1016/0378-2166(79)90019-5

Vertegaal, R., Slagter, R., Van Der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Conference on Human Factors in Computing Systems – Proceedings* (pp. 301–308).

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

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007

Viola, P., & Jones, M. (2001). Robust Real-time Object Detection. *In Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling* (p. 25).

Wall, D. P., Kosmicki, J., Deluca, T. F., Harstad, E., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry*, *2*. https://doi.org/10.1038/tp.2012.10

Wang, H., Meghawat, A., Morency, L. P., & Xing, E. P. (2017). Select-additive learning: Improving generalization in multimodal sentiment analysis. In *Proceedings - IEEE International Conference on Multimedia and Expo* (pp. 949–954). IEEE Computer Society. https://doi.org/10.1109/ICME.2017.8019301

Wang, J., & Yagi, Y. (2008). Integrating color and shape-texture features for adaptive real-time object tracking. *IEEE Transactions on Image Processing*, *17*(2), 235–240. https://doi.org/10.1109/TIP.2007.914150

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*, 2322–2333.

Woods, W. A. (1970). Transition Network Grammars for Natural Language Analysis. *Communications of the ACM*, *13*(10), 591–606. https://doi.org/10.1145/355598.362773

Zeyrek, D., Demirşahin, I., & Bozşahin, C. (2018). Turkish Discourse Bank: Connectives and Their Configurations (pp. 337–356). https://doi.org/10.1007/978-3-319-90165-7_16

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2019). TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-019-09445-9

Zitoune, F. B., & Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics, *SEM 2015* (pp. 147–152). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/s15-1016

Zoric, G., Forchheimer, R., & Pandzic, I. S. (2011). On creating multimodal virtual humans-real time speech driven facial gesturing. *Multimedia Tools and Applications*, *54*(1), 165–179. https://doi.org/10.1007/s11042-010-0526-y

APPENDICES

APPENDIX A

QUESTIONS

The Turkish translations of eight common job interview questions adapted from Villani et al. (2012) study:

1) Biraz sizden bahsedelim. Neler yapmaktan hoşlanırsınız? Beklentileriniz hedefleriniz nelerdir?
2) Önerdiğimiz bu pozisyonla nedne ilgileniyorsunuz?
3) Bu pozisyonun hangi açılardan size uygun olduğunu düşünüyorsunuz?
4) Önümüzdeki 5 yıl içinde kendinizi nerede görüyorsunuz?
5) Biraz da kişisel özelliklerinizden konuşalım. Sizi tanımlayan 3 en belirgin özelliğiniz nelerdir?
6) Liderlik yetenekleriniz hakkında ne düşünüyorsunuz? Nasıl bir lider olurdunuz?
7) Liderlik özelliklerinizi ön plana çıkaran yaşadığınız bir deneyimi paylaşır mısınız?
8) Başka bir yerden transfer edilme teklifi alsanız bu size nasıl hissettirir?

APPENDIX B

EVALUATION QUESTIONNAIRE

The interviewers evaluated the interviewees' responses on a 7-point Likert Scale.

Bilinçli cevap verdi:	1	2	3	4	5	6	7
Doğal hareket etti:	1	2	3	4	5	6	7
Yaratıcıydı:	1	2	3	4	5	6	7
Göz hareketleri doğaldı:	1	2	3	4	5	6	7

APPENDIX C

THE OUTPUT CONTENT OF THE AOI MODULE

The output file options of "Face Tracking" panel.

***2d landmark**:* The first line of an output file generated for 2d landmark is header and is written as follows: frame, timestamp, confidence, detection_success and landmark coordinates, namely x_0, x_1 ... x_67, y_0, y_1 ... y_67.[21] Confidence is a measure between 0 and 1 which represents the confidence level of tracking. For each of the 68 landmark points, first the x coordinate and then the y coordinate is recorded. This option is non-editable and is selected by default.

***3d landmark*:** The first line of an output file is header and is specified as follows: X_0... X_67, Y_0...Y_67, Z_0...Z_67. For each of the 68 landmark points, first the x coordinate, then the y and the z coordinates are recorded respectively. Every 3d point is written in millimeters and represents the facial landmark location with respect to the camera. A focal length and an optical center (by default, center of an image is assigned) are required to calculate the camera-related position. The Focal length is estimated using Algorithm 1.

Algorithm 1. Estimating the focal length.

Input: w_{c_i}: width of the captured image
h_{c_i}: height of the captured image

Output: (f_x, f_y): focal length

begin
$f_x \leftarrow 500 \times (w_{c_i}/640)$
$f_y \leftarrow 500 \times (h_{c_i}/480)$
$f_x \leftarrow (f_x + f_y)/2$
$f_y \leftarrow f_x$
end

[21]The naming convention is specified by the content module developers. We keep the naming convention intact in the present study.

Head Pose: The first line of an output file is header and is written as: `pose_Tx, pose_Ty, pose_Tz, pose_Rx, pose_Ry, pose_Rz`. The first three are translations and they represent the location of the head with respect to the camera in millimeters. Others are rotations in radians around X, Y, Z axes.

Action Units (AU). AUs are proposed to represent facial muscular activity (Ekman & Friesen, 1978). They are employed to extract facial expressions from facial appearance changes. The Facial Action Coding (FAC) system characterizes a spontaneous facial behavior among a group of items in a set of 46 AUs. OpenFace is able to recognize intensity and/or presence of following AUs: `1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28,` and `45`. The first line of an output file is header and is written as: `AU01_c, AU02_c… AU45_c, AU01_r, AU02_r… AU45_r.` The presence of AU is represented with the c suffixed column name. It is encoded as 0 for the absence and 1 for the presence. Intensity, on the other hand, is specified with r suffixed column name. It is encoded as a continuous value in the range from 0 to 5 where 0 represents absence, 1 represents presence at minimum intensity and 5 represents presence at maximum intensity.

Tracking Framework Image-Size. The last option is to create an output file holding the captured-image size. This file is necessary for further analysis. The raw gaze data depends on the image resolution of the eye tracker and it might have a different size than the image captured during face tracking has. In such cases, MAGiC automatically scale coordinates to match each other's dimensions. This option is non-editable and selected by default.

APPENDIX D

USABILITY METRICS

The 7-point scale ISO 9241/10 questionnaire.

Suitability for the task	**1:** The software inappropriately meets the demands of the tasks. **7:**Software is suitable, if it supports the user to realize his tasks effectively and efficiently.
Self-descriptiveness	**1:** The software offers insufficient information regarding the inputs which are allowed or necessary **7:** Software is self-descriptive, if every step is understandable in an intuitive way, or, in case of mistakes supported by immediate feedback.
Controllability	1:The software forces an unnecessary inflexible sequence of commands. 7:Software is controllable, if the user is able to start the sequence and influence its direction as well as speed till he reaches his aim.
Conformity with user expectations	1:The software makes more difficult the orientation because of a non-conforming design. 7:Software conforms with the user's expectations, if it is consistent, complying with the characteristics of the user, e.g. taking into account the knowledge of the user in that special working area, accounting education and experience as well as general acknowledged conventions.
Error tolerance	1:The software gives unspecific information regarding error correction and management. 7:Software is error tolerant, if it requires no or just minimal additional effort despite obvious faulty steering or wrong input.
Suitability for individualization	1: The software is difficult for the user to expand if new tasks arise. 7: Software is suitable for individualization, if the system allows customizing according to the task as well as regarding the individual capabilities and preferences of a user.
Suitability for learning	1:The software is difficult to learn without outside direction or handbooks. 7:Software supports the suitability of learning, if the user is accompanied through different states of his learning process and the effort for learning is as little as possible.

APPENDIX E

THE DETAILED INFORMATION OF PAIRS

Table 21: The detailed information of participants

Session	InterviewerID	Interviewer Gender	Age	Interviewee Gender	Age
1	Interviewer-1	Female	35	Female	27
2	Interviewer-1	Female	35	Male	30
3	Interviewer-1	Female	35	Female	30
4	Interviewer-2	Male	35	Male	25
5	Interviewer-2	Male	35	Male	22
6	Interviewer-2	Male	35	Female	22
7	Interviewer-2	Male	35	Male	21
8	Interviewer-2	Male	35	Male	25
9	Interviewer-3	Female	38	Female	22
10	Interviewer-3	Female	38	Female	22
11	Interviewer-4	Female	27	Female	24
12	Interviewer-4	Female	27	Female	28
13	Interviewer-4	Female	27	Female	24
14	Interviewer-4	Female	27	Male	26
15	Interviewer-4	Female	27	Male	26
16	Interviewer-3	Female	38	Male	25
17	Interviewer-3	Female	38	Male	25
18	Interviewer-3	Female	38	Male	30
19	Interviewer-5	Male	36	Male	22
20	Interviewer-5	Male	36	Female	25
21	Interviewer-5	Male	36	Female	24
22	Interviewer-6	Male	36	Female	24
23	Interviewer-6	Male	36	Male	24
24	Interviewer-6	Male	36	Male	27
25	Interviewer-6	Male	36	Male	27
26	Interviewer-7	Female	35	Female	26
27	Interviewer-7	Female	35	Female	25
28	Interviewer-7	Female	35	Female	29

APPENDIX F

THE NUMBER OF SEGMENTS

Table 22: The number of segments of each session

SessionID	#Segments
1.	683
2.	383
3.	356
4.	608
5.	526
6.	378
7.	403
8.	379
9.	611
10.	715
11.	623
12.	394
13.	455
14.	466
15.	935
16.	664
17.	721
18.	515
19.	893
20.	342
21.	362
22.	1219
23.	1452
24.	1497
25.	2002
26.	672
27.	752
28.	850
Total	**19856**

APPENDIX G

RESIDUAL AND THE PROBABILITY DISTRIBUTION PLOTS

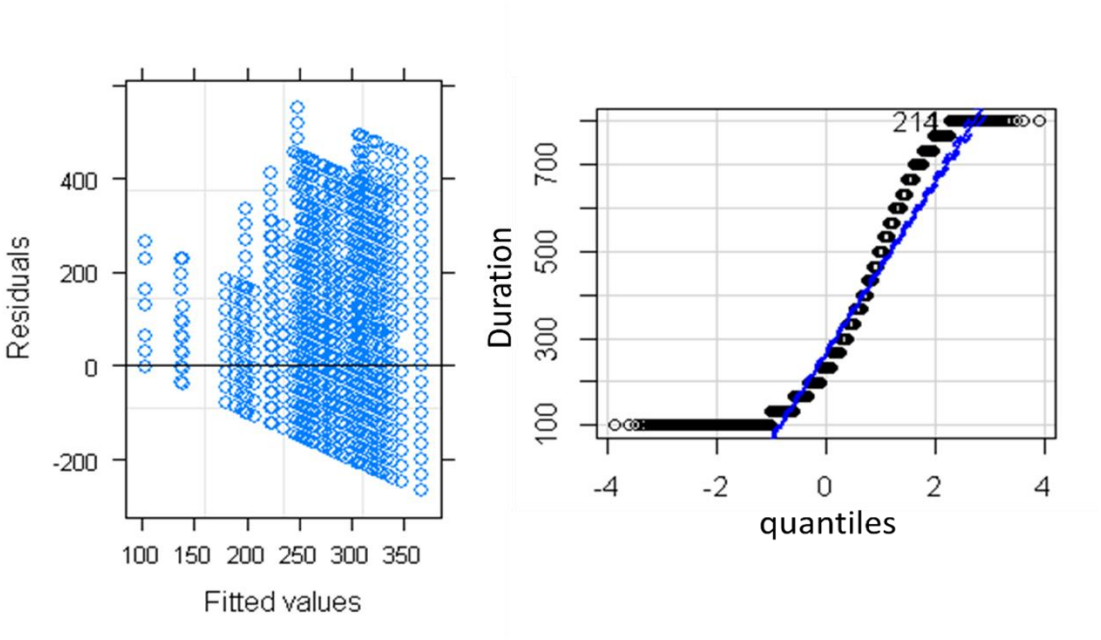


Figure 50: The residual and the probability distribution plots. They were used to test the assumptions of the linear mixed model for the duration of gaze aversion. In Residual plot, higher fitted values have larger residuals indicating that the model is more "off" with larger predicted means. So, the variance is not homoscedastic: it's smaller in the lower range and larger in the higher range.

APPENDIX H

NUMBER OF LABELS

Table 23: Number of Speech-tags

Tag	Frequency
Speech	2867
Micro Pause	1387
Speech Pause	1262
Thinking	927
Pre-Speech	402
Sub-Total	**6845** (covers 80.3%)
Asking A Question	363
Confirmation	337
Read-Question	325
Questionnaire Filling	239
Signaling End of Speech	174
Speech While Laughing	92
Greeting	52
Laugh	50
The Repetition of Question	44
Total	**8521**

Table 24: Number of Dialogue-acts

Dialogue-act	Frequency
Stalling	1948
Answer	1430
Auto Positive	801
Inform	753
Turn Take	407
Sub-Total	**5339** (covers 79.6%)
Turn Keep	340
Set Question	230
Self Correction	140
Interaction Structuring	117
Agreement	80
Request	69
Retraction	52
Propositional Question	35
Check Question	32
Turn Release	31
Confirm	28
Thanking	24
Turngrab	23
Accept Request	16
Initial Greeting	16
Return Greeting	16
Opening	13
Allo Positive	10
Completion	10
Choice Question	8
Initial Self Introduction	8
Pausing	8
Turn Accept	8
Accept Thanking	7
Apology	7
Accept Apology	6
Suggest	6
Instruct	5
Turn Assign	4
Initial-Goodbye	3
Promise	3
Accept Suggest	2
Disagreement	2
Offer	2
Return-Goodbye	2
Self Error	2
Correct Misspeaking	1
Disconfirm	1
Total	**6706**

Table 25: Number of Dialogue-act dimensions

Dimension	Frequency
Task	2658
Turn Management	1948
Time Management	820
Auto Feedback	363
Own Communication Management	337
Discourse Structuring	325
Social Obligation Management	239
Allo Feedback	174
Partner Communication Management	92
Total	**8521**

Table 26: Number of Rhetorical-Relation

Rhetorical Relation	Frequency
Elaboration	1589
Conjunction	1549
Cause	743
Expansion	491
Exemplification	347
Contrast	314
Concession	312
Restatement	303
Condition	280
Similarity	172
Substitution	151
Disjunction	121
Asynchrony	116
Manner	108
Purpose	42
Synchrony	22
Negative Condition	18
Exception	6
Total	**6684**

APPENDIX I

VALIDATION AND TRAINING ACCURACIES OF BACKTEST

Table 27: The 5-fold backtesting results of Speech Tag Set. The highest validation accuracies of each architecture is written in bold. *) It represents the filter numbers in the first block.

	16*		**32***	
	Training	*Validation*	*Training*	*Validation*
ResNet	84.7 (SD:0.28)	69.1 (SD:7.71)	84.6 (SD:0.25)	**71.9 (SD:7.44)**
VGG	84.8 (SD:0.21)	68.3(SD:9.61)	84.9(SD:0.21)	**69.1 (SD:7.51)**

Table 28: The 5-fold backtesting results of Dialogue act. The highest validation accuracies of each architecture is written in bold. *) It represents the filter numbers in the first block.

	16*		**32***	
	Training	*Validation*	*Training*	*Validation*
ResNet	84.1(SD:1.53)	64.9(SD:6.71)	83.7(SD:1.25)	**65.9 (SD:7.99)**
VGG	85.8(SD:1.03)	**64.5 (SD:7.51)**	86.7(SD:1)	64.4 (SD:7.88)

APPENDIX J

ORDERS OF INTERVIEWERS

10- fold cross validation for the models trained with the input data involving Dialogue-act annotation.

Table 29: The orders of interviewers for 10-fold cross validation of dialogue-act annotation.

Orders	Interviewer IDs
1.	1-2-3-5-6-7
2.	2-3-5-6-7-1
3.	3-6-5-7-1-2
4.	5-3-1-7-2-6
5.	6-7-1-2-3-5
6.	1-5-3-6-2-7
7.	6-2-3-7-5-1
8.	2-1-7-3-6-5
9.	5-7-6-1-3-2
10.	3-1-7-2-5-6

CURRICULUM VITAE

Ülkü ARSLAN AYDIN
Address: Am Tierpark 51, 10319 Berlin
Mobile: +90 5066070512, +49 1514 2453256
Email: ulku.arslan@gmail.com

Education:

- Middle East Technical University, Ankara, Turkey
  Ph.D. in Cognitive Science, 2012- (GPA: 4.0/4.0)

- Middle East Technical University, Ankara, Turkey
  M.Sc. in Cognitive Science, 2009-2012 (GPA: 3.7/4.0)

- Hacettepe University, Ankara, Turkey
  B.S. in Computer Engineering, 2001-2005 (GPA: 2.9/4.0)

Experience:

- System Analyst at Capital Markets Board of Turkey (currently in maternity leave)
  March 2009 -

- Software Engineer at STM A.Ş
  July 2006- February 2009
  Worked in TBS (Turkish Armed Forces Information System) project.

- Software Engineer at CyberSoft
  June 2005 - July 2006
  Worked in AVIS (Full Automation of Azerbaijan Tax Offices) project.

- Internship at The Scientific and Technological Research Council of Turkey
  June 2004 - July 2004

Publications:

Arslan Aydin, Ü., Kalkan, S., & Acarturk, C. (2018). MAGiC: A multimodal framework for analysing gaze in dyadic communication. *Journal of Eye Movement Research*, 11(6). https://doi.org/10.16910/jemr.11.6.2

Arslan Aydin, Ü., Kalkan, S., Acartürk, C. (in prep.) Computational Models of Speech Driven Gaze in Face-to-Face Interaction

Arslan Aydin, Ü., Kalkan, S., Acartürk, C. (2017). Dinamik Gözbakışı Analizi: Yüzyüze İletişim için bir Uygulama Ortamı [Dynamic Gaze Analysis: An Application Enviroment for Face-to-Face Communication]. *International Artificial Intelligence and Data Processing Symposium IDAP* (pp.1-6). Malatya, Turkey.

Arslan Aydin, Ü., Kalkan, S., Acartürk, C. (2017). A Gaze-Centered Multimodal Approach to Human-Human Social Interaction.*3rd IEEE International Conference on Cybernetics CYBCONF*(pp. 1-6). June 21-23, 2017. Exeter, UK. doi: 10.1109/CYBConf.2017.7985753

Arslan-Aydin, U., Acartürk, C., & Cagiltay, K. (2013). The role of visual coherence in graphical passwords. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 1774-1779). Austin, TX: Cognitive Science Society.

Arslan-Aydin, U., Acartürk C. (2011). Kullanılabilir Güvenlik ve Grafik Parolalar. Proceedings of the16th Annual Conference of the "Türkiye'de İnternet" (pp. 173-182). Izmir, Turkey

Award and Summer School:

Student Travel Award, EUCog - European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics (2013)

International Summer School and Workshop on Brain Dynamics: Connectivity & Cognition (2012)