ROYAL INSTITUTE
OF TECHNOLOGY

Recognition and Generation of Communicative Signals

Modeling of Hand Gestures, Speech Activity and Eye-Gaze in Human-Machine Interaction

KALIN STEFANOV

Doctoral Thesis
Stockholm, Sweden 2018

KTH Royal Institute of Technology
EECS Electrical Engineering and Computer Science
SE-100 44 Stockholm
SWEDEN

TRITA-EECS-AVL-2018:46
ISBN 978-91-7729-810-6

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i datalogi Torsdag den 7 juni 2018 klockan 14.00 i K2, Kungliga Tekniska Högskolan, Teknikringen 28, Stockholm.

Tryck: Universitetsservice US-AB

In loving memory of my mother
Rumiana Stefanova

Abstract

Nonverbal communication is essential for natural and effective face-to-face human-human interaction. It is the process of communicating through sending and receiving wordless (mostly visual, but also auditory) signals between people. Consequently, a natural and effective face-to-face human-machine interaction requires machines (*e.g.*, robots) to understand and produce such human-like signals. There are many types of nonverbal signals used in this form of communication including, body postures, hand gestures, facial expressions, eye movements, touches and uses of space. This thesis investigates two of these nonverbal signals: hand gestures and eye-gaze. The main goal of the thesis is to propose computational methods for real-time recognition and generation of these two signals in order to facilitate natural and effective human-machine interaction.

The first topic addressed in the thesis is the real-time recognition of hand gestures and its application to recognition of isolated sign language signs. Hand gestures can also provide important cues during human-robot interaction, for example, emblems are type of hand gestures with specific meaning used to substitute spoken words. The thesis has two main contributions with respect to the recognition of hand gestures: 1) a newly collected dataset of isolated Swedish Sign Language signs, and 2) a real-time hand gestures recognition method.

The second topic addressed in the thesis is the general problem of real-time speech activity detection in noisy and dynamic environments and its application to socially-aware language acquisition. Speech activity can also provide important information during human-robot interaction, for example, the current active speaker's hand gestures and eye-gaze direction or head orientation can play an important role in understanding the state of the interaction. The thesis has one main contribution with respect to speech activity detection: a real-time vision-based speech activity detection method.

The third topic addressed in the thesis is the real-time generation of eye-gaze direction or head orientation and its application to human-robot interaction. Eye-gaze direction or head orientation can provide important cues during human-robot interaction, for example, it can regulate who is allowed to speak when and coordinate the changes in the roles on the conversational floor (*e.g.*, speaker, addressee, and bystander). The thesis has two main contributions with respect to the generation of eye-gaze direction or head orientation: 1) a newly collected dataset of face-to-face interactions, and 2) a real-time eye-gaze direction or head orientation generation method.

Sammanfattning

Naturlig och effektiv interaktion människor emellan kräver icke-verbal kommunikation, dvs sändande och mottagande av ordlösa (ofta visuella) signaler mellan människor. Följaktligen kräver en naturlig och effektiv interaktion mellan människor och maskiner (t.ex. robotar) att även maskinerna kan förstå och producera sådana människolika signaler. Det finns många typer av icke-verbala signaler som används i denna form av kommunikation, inklusive kroppsställningar, handgester, ansiktsuttryck, ögonrörelser, beröring och spatiala referenser. Denna avhandling undersöker två av dessa icke-verbala signaler: handgester och blickbeteende. Huvudmålet med avhandlingen är att föreslå beräkningsmetoder för realtidsigenkänning och generering av dessa två signaler för att underlätta naturlig och effektiv interaktion mellan människor och maskiner.

Det första ämnet som tas upp i avhandlingen är realtidsigenkänning av handgester och dess tillämpning på igenkänning av isolerade teckenspråks tecken. Handgester kan också innehålla viktig information i människa-robot interaktion, till exempel är emblem typ av handgester med specifik betydelse som används för att ersätta talade ord. Avhandlingen har två huvudbidrag avseende igenkänning av handgester: 1) en nyinspelat dataset med isolerade svenska teckenspråkstecken, och 2) en i realtid fungerande metod för igenkänning av handgester.

Det andra ämnet som tas upp i avhandlingen är det allmänna problemet med detektering av talaktivitet i bullriga och dynamiska miljöer utifrån visuell information, och dess tillämpning på tillägnande av språk i ett socialt sammanhang. Talaktivitet kan också ge viktig information under människa-robot interaktion, till exempel kan den aktuella aktiva talarens handgester och ögonrörelser eller huvudriktning spela en viktig roll för att förstå interaktionstillståndet. Avhandlingens huvudbidrag med avseende på talaktivitetsdetektering: en metod för att detektera talaktivitet från video i realtids.

Det tredje ämnet som behandlas i avhandlingen är realtidsgenerering av blickriktning eller huvudorientering och dess tillämpning på människa-robot-interaktion. Ögonblickriktning eller huvudorientering kan ge viktiga signaler under människa-robot-interaktion, till exempel kan den styra vem som ska prata när, och koordinera förändringar i rollerna i konversationen. Avhandlingen har två huvudbidrag med avseende på generering av blickriktning eller huvudorientering: 1) ett nyinspelat dataset med flerpartsinteraktioner ansikte mot ansikte, och 2) en metod för att generera blickriktning eller huvudorientering i realtid.

Acknowledgements

This thesis has been a challenging and rewarding experience and here I will take the opportunity to express my sincere gratitude to the people that turned this experience into a pleasurable one.

I thank my supervisor, Jonas Beskow, for the invaluable guidance that ensured I do not lose focus on the way. I thank my co-supervisor, Hedvig Kjellström, for the persistent encouragement to pursue my ideas. I also want to thank my undercover supervisor, Giampiero Salvi, for the countless insightful discussions and friendship.

I thank Iolanda Leite for the quality review of the thesis.

I thank Akihiro Sugimoto, with whom I had the privilege to collaborate during my visit at the National Institute of Informatics. I thank Stefan Scherer for the productive collaboration during my visit at the Institute for Creative Technologies.

I thank my collaborators in the TIVOLI project for introducing me to the fascinating world of sign language communication: Britt Claesson, Sandra Derbring, and Krister Schönström.

I thank the professors at TMH for the leadership: Anders Askenfelt, Jonas Beskow, Johan Boye, Rolf Carlson, Jens Edlund, Olov Engwall, Anders Friberg, Björn Granström, Joakim Gustafson, David House, Peter Nordqvist, Giampiero Salvi, Bo Schenkman, Gabriel Skantze, Johan Sundberg, and Sten Ternström.

I thank my colleagues at TMH for the great environment: Samer Al Moubayed, Simon Alexanderson, Bajibabu Bollepalli, Mattias Bystedt, Saeed Dabbaghchian, Gaël Dubus, Anders Elowsson, Laura Enflo, Per Fallgren, Morgan Fredriksson, Jana Götze, Anna Hjalmarsson, Martin Johansson, Patrik Jonell, Christos Koniaris, Dimosthenis Kontogiorgos, José David Lopes, Zofia Malisz, Raveesh Meena, Joe Mendelson, Daniel Neiberg, Catharine Oertel, Glaucia Laís Salomão, Andreas Selamtzis, Todd Shore, Sofia Strömbergsson, Eva Szekely, Luca Turchet, Niklas Vanhainen, Preben Wik, and Meg Zellers.

I thank my parents, who always emphasized the importance of good education and who always encouraged me to learn more. I thank my father, Momchil Stefanov, for teaching me to pursue my ambition by example. I thank my mother, Rumiana Stefanova, for the immense support and encouragement to explore and to be curious about the world. I miss you!

I thank my brother, Svetlin Stefanov, for the much helpful advice on the way.

Last but not least, I thank my girlfriend, Ayumi Maruyama, for being such an amazing person and for being there in the hardest moments.

心より感謝申し上げます

Included Publications

The publications included in the thesis are the result of collaborative efforts. The contributions of the individual authors are outlined below.

Paper A
Stefanov, K. and Beskow, J. (2013). 'A Kinect Corpus of Swedish Sign Language Signs'. In *Proceedings of the Multimodal Corpora: Beyond Audio and Video*, Edinburgh, UK.

Kalin and Jonas discussed and designed the dataset. Kalin developed the recording application and performed the data collection. Kalin also performed the data post-processing and initial analysis. Kalin wrote the paper with contributions from Jonas.

Paper B
Stefanov, K. and Beskow, J. (2017). 'A Real-time Gesture Recognition System for Isolated Swedish Sign Language Signs'. In *Proceedings of the Symposium on Multimodal Communication*, Copenhagen, Denmark.

Kalin and Jonas discussed the methods. Kalin proposed and developed the methods and performed the experiments. Kalin wrote the paper with contributions from Jonas.

Paper C
Stefanov, K., Beskow, J., and Salvi, G. (2017). 'Vision-based Active Speaker Detection in Multiparty Interactions'. In *Proceedings of the Grounding Language Understanding*, Stockholm, Sweden.

Kalin, Giampiero and Jonas discussed the methods. Kalin proposed and developed the methods and performed the experiments. Kalin wrote the paper with contributions from Giampiero and Jonas.

Paper D
Stefanov, K., Beskow, J., and Salvi, G. (2017) 'Self-Supervised Vision-Based Detection of the Active Speaker as a Prerequisite for Socially-Aware Language Acquisition'. *Submitted to the IEEE Transactions on Cognitive and Developmental Systems Special Issue on Language Learning in Humans and Robots.*

Kalin, Giampiero and Jonas discussed the methods. Kalin proposed and developed the methods and performed the experiments. Kalin wrote the paper with contributions from Giampiero and Jonas.

Paper E
Stefanov, K. and Beskow, J. (2016) 'A Multi-party Multi-modal Dataset for Focus of Visual Attention in Human-human and Human-robot Interaction'. In *Proceedings of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Kalin and Jonas discussed and designed the dataset. Kalin developed the recording application and performed the data collection. Kalin also performed the data post-processing and initial analysis. Kalin wrote the paper with contributions from Jonas.

Paper F
Stefanov, K., Salvi, G., Kontogiorgos, D., Kjellström, H., and Beskow, J. (2018) 'Analysis and Generation of Candidate Gaze Targets in Multiparty Open-World Dialogues'. *Submitted to the ACM Transactions on Human-Robot Interaction Special Issue on Artificial Intelligence for Human-Robot Interaction.*

Kalin, Giampiero, Hedvig and Jonas discussed the methods. Dimosthenis provided one of the datasets. Kalin proposed and developed the methods and performed the experiments. Kalin wrote the paper with contributions from Giampiero, Dimosthenis, Hedvig and Jonas.

Additional Publications

In addition to the publications included in the thesis, the author has contributed to the following published papers.

Stefanov, K., Sugimoto, A., and Beskow, J. (2016). 'Look Who's Talking: Visual Identification of the Active Speaker in Multi-party Human-robot Interaction'. In *Proceedings of the Advancements in Social Signal Processing for Multimodal Interaction*, Tokyo, Japan.

Stefanov, K. and Beskow, J. (2016) 'Gesture Recognition System for Isolated Sign Language Signs'. In *Proceedings of the Symposium on Multimodal Communication*, Copenhagen, Denmark.

Chollet, M., Stefanov, K., Prendinger, H., and Scherer, S. (2015). 'Public Speaking Training With a Multimodal Interactive Virtual Audience Framework'. In *Proceedings of the International Conference on Multimodal Interaction*, Seattle, WA.

Al Moubayed, S., Beskow, J., Bollepalli, B., Hussen-Abdelaziz, A., Johansson, M., Koutsombogera, M., Lopes, J., Novikova, J., Oertel, C., Skantze, G., Stefanov, K., and Varol, G. (2014). 'Tutoring Robots'. *Innovative and Creative Developments in Multimodal Interaction Systems*. Springer Berlin Heidelberg.

Al Moubayed, S., Beskow, J., Bollepalli, B., Gustafson, J., Hussen-Abdelaziz, A., Johansson, M., Koutsombogera, M., Lopes, J., Novikova, J., Oertel, C., Skantze, G., Stefanov, K., and Varol, G. (2014). 'Human-robot Collaborative Tutoring Using Multiparty Multimodal Spoken Dialogue'. In *Proceedings of the International Conference on Human-Robot Interaction*, Bielefeld, Germany.

Al Moubayed, S., Beskow, J., Bollepalli, B., Hussen-Abdelaziz, A., Johansson, M., Koutsombogera, M., Lopes, J., Novikova, J., Oertel, C., Skantze, G., Stefanov, K., and Varol, G. (2014). 'Tutoring Robots: Multiparty Multimodal Social Dialogue With an Embodied Tutor'. In *Proceedings of the International Summer Workshop on Multimodal Interfaces*, Lisbon, Portugal.

Koutsombogera, M., Al Moubayed, S., Bollepalli, B., Hussen-Abdelaziz, A., Johansson, M., Lopes, J., Novikova, J., Oertel, C., Stefanov, K., and Varol, G. (2014). 'The Tutorbot Corpus – A Corpus for Studying Tutoring Behaviour in Multiparty Face-to-Face Spoken Dialogue'. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

Meena, R., Dabbaghchian, S., and Stefanov, K. (2014) 'A Data-driven Approach to Detection of Interruptions in Human-Human Conversations'. In *Proceedings of the FONETIK*, Stockholm, Sweden.

Beskow, J., Alexanderson, S., Stefanov, K., Claesson, B., Derbring, S., Fredriksson, M., Starck, J., and Axelsson, E. (2014) 'Tivoli – Learning Signs Through Games and Interaction for Children with Communicative Disorders'. In *Proceedings of the Conference of the International Society for Augmentative and Alternative Communication*, Lisbon, Portugal.

Beskow, J., Alexanderson, S., Stefanov, K., Claesson, B., Derbring, S., and Fredriksson, M. (2013). 'The Tivoli System – A Sign-driven Game for Children with Communicative Disorders'. In *Proceedings of the Symposium on Multimodal Communication*, Msida, Malta.

Beskow, J. and Stefanov, K. (2013) 'Web-enabled 3D Talking Avatars Based on WebGL and HTML5'. In *Proceedings of the Intelligent Virtual Agents*, Edinburgh, UK.

Eyben, F., Gilmartin, E., Joder, C., Marchi, E., Munier, C., Stefanov, K., Weninger, F., and Schuller, B. (2012) 'Socially Aware Many-to-Machine Communication'. In *Proceedings of the International Summer Workshop on Multimodal Interfaces*, Metz, France.

Al Moubayed, S., Skantze, G., Beskow, J., Stefanov, K., and Gustafson, J. (2012). 'Multimodal Multiparty Social Interaction With the Furhat Head'. In *Proceedings of the International Conference on Multimodal Interaction*, Santa Monica, CA.

Contents

Tables

Figures

Part I

Thesis Background

Chapter 1

Communicative Nonverbal Signals

This chapter presents a broad introduction to the categories and functions of different nonverbal behaviors that can be employed as communicative signals in face-to-face interactions. The abundance of literature studying the importance of nonverbal behavior in face-to-face interaction is the broad motivation for this thesis. The broad goal of the thesis can be defined as: to propose and develop computational methods that could enable machines to recognize and generate similar communicative signals.

Nonverbal behavior is a well-studied area of human behavior, with roots leading back to the 19th century, most notably to "The Expression of Emotions in Man and Animals" (Darwin, 1873). In its narrow sense, nonverbal behavior refers to actions as distinct from speech. It includes facial expressions, hand and arm gestures, postures, positions, and movements of the body, the legs and the feet. In the broader sense, nonverbal behavior also includes vocal phenomena, such as fundamental frequency range and intensity range, speech errors, pauses, speech rate, and speech duration (Mehrabian, 1972).

Human communication is the process of one person inducing an interpretation in the mind of another person by means of verbal and/or nonverbal signals. Consequently, nonverbal human communication is the process of one person inducing an interpretation in the mind of another person by means of nonverbal signals (Richmond et al., 2012). Nonverbal behavior is any of a wide variety of human behaviors that also have the potential for forming communicative messages. Such nonverbal behavior becomes nonverbal communication if another person interprets the behavior as a message and attributes meaning to it. We can engage in nonverbal behavior whether we are alone or someone else is present. We can engage in nonverbal communication only in the presence of one or more people who interpret our behavior as messages and assign meaning to those messages. Therefore, for human communication to exist, whether verbal or nonverbal, a source must send a message and a receiver must receive and interpret that message. Sometimes receivers perceive our verbal and nonverbal behavior as messages, and sometimes they do not. Table 1.1

presents the four distinct possibilities: the source's behavior is intended to send/not send a message and the receiver interprets the behavior either as a message or not.

		Source	
		Message	No message
Receiver	Message	Communication	Communication
	No message	Behavior	Behavior

Table 1.1: Nonverbal Behavior as Opposed to Nonverbal Communication

A communication channel is the sensory route on which a signal travels. Verbal communication relies mostly on one channel, because spoken language is transmitted through sound and picked up by the ears. Nonverbal communication, on the other hand, can be taken in by all five senses. Verbal and nonverbal communication include both vocal and nonvocal signals. Table 1.2 presents the relationship between verbal/nonverbal communication and vocal/nonvocal signals.

	Verbal	**Nonverbal**
Vocal	Spoken language	Paralanguage
Nonvocal	Sign language	Body language

Table 1.2: Verbal Communication as Opposed to Nonverbal Communication

1.1 Categories of Communicative Nonverbal Signals

Numerous categories of nonverbal behavior can be selected from the following realms (Mehrabian, 1972),

- **Signals:** facial expressions, verbalizations, movements, and postures.
- **Referents:** feelings and attitudes.
- **Attributes:** personality, age, gender, and status.
- **Media:** face-to-face, telephone, and video.

Books on the topic of nonverbal behavior, (Knapp et al., 2013, Moore et al., 2013), mostly agree on the different categories of nonverbal behaviors including, body movements and gestures, managing space and territory, touch, tone of voice, and appearance. Similar categorization is used in this text and is presented next,

- **Appearance:** Appearance involves physical characteristics and artifacts. There are many aspects of physical appearance that can potentially produce messages including, attractiveness, body size, body shape, facial features, hair, skin color, height, weight, clothing, watches, and necklaces.

- **Gestures and movement:** This type of behavior is often called body language (Table 1.2). The study of the communicative aspects of all gestures, eye behaviors, facial expressions, posture, and movements of the hands, arms, body, head, legs, feet, and fingers is called *kinesics* (Birdwhistell, 1970).

- **Face and eyes:** We also communicate through eye behaviors, primarily eye contact and face behaviors, primarily facial expressions. While face and eye behaviors are often studied under the category of kinesics, communicative aspects of eye behaviors have their own branch of studies called *oculesics*.

- **Voice:** Paralanguage refers to the vocalized but nonverbal part of the communication (Table 1.2). The study of the communicative aspects of voice including, pitch, volume, rate, vocal quality, and verbal fillers is called *vocalics* (Andersen, 1999).

- **Space:** The study of the communicative aspects of space and distance is called *proxemics*. Proxemic distances can be grouped into several categories including, public, social, personal, and intimate distance (Hall, 1990). The concept of territoriality groups space into several categories including, primary, secondary, and public space (Hargie, 2011).

- **Touch:** The study of the communicative aspects of touch is called *haptics*. Touch is important for human social development, and it can be grouped into several categories including, welcoming, threatening, and persuasive touch.

- **Environment:** Environmental factors include architecture, interior spatial arrangements, music, color, lighting, temperature, scent, and smell. The study of the communicative aspects of scent and smell is called *olfactics*.

- **Time:** The study of the communicative aspects of time is called *chronemics*. Time can be grouped into several categories including, biological, personal, physical, and cultural time (Andersen, 1999).

1.2 Functions of Communicative Nonverbal Signals

Nonverbal signals play six major functions in relation to the verbal signal. These functions are presented next,

- **Complementing:** Some nonverbal signals add to, reinforce, clarify, and elaborate the intended meaning of the verbal. For example, a shaking fist reinforces an accompanying threatening utterance.

- **Contradicting:** Some nonverbal signals contradict, dispute, and are in conflict with the verbal. For example, sarcasm is used to make a point by utilizing nonverbal signals which contradict the verbal signal.
- **Accenting:** Some nonverbal signals accent, enhance, emphasize, and highlight the verbal. For example, pausing before speaking or speaking louder than usual, highlights the verbal signal.
- **Repeating:** Some nonverbal signals repeat, reiterate, and restate the verbal. For example, emblems (Section 2.1) are gestures that can be used to repeat the verbal signal.
- **Regulating:** Some nonverbal signals regulate the verbal. For example, looking at or away from the other person is one way to regulate who is allowed to speak when.
- **Substituting:** Some nonverbal signals substitute the verbal. For example, emblems (Section 2.1) are gestures that can be used to substitute for the verbal signal.

1.3 Summary

In summary, often nonverbal behavior cannot be translated into definitions because meanings are in people's minds, not nonverbal behaviors (Table 1.1). The meaning attributed to nonverbal behaviors is influenced by the context in which these behaviors occur. Besides the context, culture, ethnic and geographic origins, gender, social status, and educational background, all contribute to the meaning attributed to these signals. Furthermore, although single nonverbal behaviors can stimulate meanings, more typically a meaning is composed of groups of nonverbal behaviors that interact to create communicative impact; to understand all of these behaviors, it seems necessary to look at the individual categories of behaviors one by one.

This chapter presented a broad introduction to the categories and functions of different nonverbal behaviors that can be employed as communicative signals in face-to-face interactions. The importance of nonverbal behavior in face-to-face interaction is the broad motivation for the work described in this thesis. The broad goal of the thesis can be defined as: to propose and develop computational methods that could enable machines to recognize and generate similar communicative signals. The chapter discussed the requirements for nonverbal communication to exist and grouped different nonverbal signals into categories. The chapter also discussed the major functions nonverbal signals play in relation to the verbal signal. Evidently, the domain of human nonverbal behavior is wide and complex. Therefore, the next chapter is dedicated to narrowing down the scope of the thesis to two distinct communicative nonverbal signals: hand gestures and eye-gaze.

Chapter 2

Scope and Delimitations

This chapter narrows down the scope of the thesis to two distinct communicative nonverbal signals: hand gestures and eye-gaze, and places the recognition and generation of certain categories of these two signals in the context of human-machine interaction. The multitude of functions these two signals serve in face-to-face interaction is the narrowed down motivation for this thesis. The narrowed down goal of the thesis can be defined as: to propose and develop computational methods that could enable machines to recognize communicative hand gesture signals and generate communicative eye-gaze signals.

On one hand, the computational methods developed in this thesis can be classified as *Social signal processing* (SSP) (Pantic et al., 2011, Vinciarelli et al., 2009). SSP defines several challenges for machine analysis of social signals including, the recording of the scene, detection of the people in the scene, extraction of multimodal signals, and multimodal signal analysis and classification in a given context. On the other hand, the computational methods developed in this thesis have direct application in *Human-robot interaction* (HRI). Robots displaying human-like behaviors are expected to take an increasing role in society, for example, by taking care of elderly (Feil-Seifer and Matarić, 2005, Tapus et al., 2007) or helping children in their learning (Castellano et al., 2013). To support natural and effective social interaction, robots should be able to recognize and generate all categories of communicative nonverbal signals discussed in Section 1.1. As a result, nonverbal behavior has been an active area of research in the field of HRI, with work mostly concentrated around kinesics, and additional emphasis within kinesics on eye-gaze, hand gestures, and facial expressions (Thomaz et al., 2016).

2.1 Communicative Hand Gesture Signals

Several categorizations of hand gestures have been proposed in the literature. One of the most adopted, (McNeill, 1992, 2005), categorizes hand gestures into four groups. These categories are presented next,

- **Deictic:** This type of gestures are related to words, including pointing gestures used for direct spatial or abstract reference. For example, talking about someone across the room and pointing them out.
- **Iconic:** This type of gestures are related to an event or an object. For example, using the hands to describe a high mountain or a wide river.
- **Metaphorical:** This type of gestures are related to abstract concepts. For example, using the fingers to create a heart-like shape and place it on the chest.
- **Beat:** This type of gestures are related to the "music" of the utterance. For example, up-and-down hand movements that coincide with spoken clauses.

Hand gestures can be placed along a continuum where their co-occurrence with speech is more and more optional (Kendon, 1980). Moving along the continuum, gestures become increasingly language-like and may take over more of the communicative functions of speech. On one side of the continuum are co-speech gestures, which are unconsciously produced in conjunction with speech. On the other side of the continuum are sign languages, which are full languages with their own morphology, phonology, and syntax (Stokoe, 1980). This continuum is presented next,

- **Gesticulations:** This type of gestures are co-occurring with speech.
- **Speech-framed:** This type of gestures are filling in a slot in speech.
- **Emblems:** This type of gestures can replace words.
- **Pantomimes:** This type of gestures are produced without speech.
- **Sign languages:** Sign languages have lexical words and full grammars.

2.1.1 Categories of Hand Gesture Signals

There are five major categories of kinesics (Ekman, 1976, Ekman and Friesen, 1969*a*,*b*, 1972, 1974). Similar categorization is used in this text and is presented next,

- **Emblems:** This type gestures and movements are often referred to as speech-independent. Besides having direct verbal translations and usually being used intentionally, emblems are socially learned in much the same manner as language. The meanings assigned to emblems are arbitrary, and the way meanings are associated with actions is highly similar to the way meanings are associated with words. Emblems are different from the signs used by the deaf people who communicate using sign languages; even though emblems have a generally agreed-on meaning, they are not part of a formal sign system like a sign language (Andersen, 1999). The primary function of emblems is to

substitute spoken words; they can be used to induce specific meanings in the minds of others in place of the verbal signal.

- **Illustrators:** This type of gestures are often referred to as speech-linked. Like emblems, illustrators are usually intentional. Unlike emblems, illustrators cannot stand alone and induce meaning; illustrators generate little or no meaning when they are not accompanying speech (Andersen, 1999). The primary function of illustrators is to clarify or complement the verbal signal. Illustrators are a large group of nonverbal behaviors, which can be further divided into three categories,

  - Gestures that are related to the speech referent or explanation. For example, using the hands to describe a high mountain or a wide river.
  - Gestures that highlight or emphasize an utterance. For example, rising a different finger to highlight different points in an argument.
  - Gestures that help the speaker in organizing the conversation. For example, hand movements that punctuate the speech.

- **Regulators:** This type of gestures and body movements, along with eye and vocal signals, maintain and regulate the back-and-forth interaction between speakers and listeners during spoken dialogue. Regulators are not as intentional as emblems and illustrators. The primary function of regulators is to manage the communication turn-taking. Regulators can be divided into four categories (Duncan, 1972, 1974),

  - Turn-yielding regulators are used by speakers who wish to discontinue talking and give the listener the opportunity to take the speaking role.
  - Turn-maintaining regulators are used by speakers who want to continue talking.
  - Turn-requesting regulators are used by listeners to signal the speaker that they want to talk.
  - Turn-denying regulators are used by listeners to signal that they decline the turn to talk.

- **Affect displays:** This type of gestures primary involve facial expressions but also include a person's posture, the way a person walks, limb movements, and other behaviors that provide information about the emotional state and mood. The primary function of affect displays is to reveal the emotional state and these signals are usually unintentional.

- **Adaptors:** This type of gestures result from uneasiness, anxiety, or a general sense that one is not in control (Andersen, 1999). Adaptors are falling into three categories,

- Self-adaptors are acts in which an individual manipulates ones own body. For example, scratching and hair twisting.
- Alter-adaptors are acts in which an individual protects from others. For example, folding the arms and placing them on the chest.
- Object-focused adaptors are acts in which an individual manipulates an object. For example, tapping a pen or twisting a ring around the finger.

2.1.2 Computational Modeling of Hand Gesture Signals

Many studies investigate aspects of robots using pointing gestures to communicate with humans. These are shown to increase people's information recall (Huang and Mutlu, 2013) as well as task performance and perceived workload (Lohse et al., 2014). Eye-gaze is shown to significantly assist the recognition of pointing gestures (Häring et al., 2012, Iio et al., 2010). Liu et al. (2013) presented a study that shows that people do not usually point to refer to a person, but instead use eye-gaze. Sauppé and Mutlu (2014) compared several variations deictic gestures on a small humanoid robot including, pointing, presenting, touching, and sweeping, and found that their effectiveness is strongly related to the context. Yamazaki et al. (2008) presented a system for a museum robot that moves its head at significant points of an explanation, such as at transition points, together with deictic words, in response to a question, upon keywords, or with unfamiliar words.

Some studies investigate the recognition of pointing gestures. In (Brooks and Breazeal, 2006) the authors presented a framework for recognition of deictic gestures of a human. Van den Bergh et al. (2011) presented a real-time pointing detection system for a robot giving directions. Similarly, Quintero et al. (2013), achieved pointing recognition for object selection. Burger et al. (2012) used computational models to recognize gestures for robot commands.

This thesis contributes to the literature on SSP and computational HRI with computational methods for recognition of communicative hand gesture signals. More specifically, one of the outcomes of the thesis is a newly collected dataset of isolated Swedish Sign Language signs. Another outcome is an investigation of computational methods for recognition of hand gestures and the development of a real-time hand gesture recognition method. The method is applied and evaluated on the task of recognizing isolated sign language signs which fall under the category of emblems.

2.2 Communicative Eye-Gaze Signals

There are three main categories of communicative eye-gaze signals: mutual gaze, one-side look, and gaze aversion (Argyle and Cook, 1976, Argyle and Dean, 1965, Argyle and Ingham, 1972, Argyle et al., 1973). Mutual gaze refers to two people looking in the direction of one another's face. Staring occurs when one person focuses in on another person and gives a long and often invasive look, which in some

cultures, is considered unacceptable and rude (Argyle and Ingham, 1972). One-side look refers to gaze of one individual in the direction of another person's face, but the gaze is not reciprocated. Gaze aversion can signal that one is not interested in what the other person has to say or is used as a regulator when wishing to stop the communication. Gaze aversion is also a natural part of an ongoing conversation to avoid staring at the other person and it is increased by speakers who are using turn-maintaining signals; speakers who want to continue talking often signal their intention by dramatically reducing their gaze toward the listener.

2.2.1 Functions of Eye-Gaze Signals

Eye-gaze signals serve four primary functions (Kendon, 1967). These functions are presented next,

- **Scanning:** The eyes scan, focus, and collect information about the world. Humans use scanning to monitor the environment and to protect from harm.
- **Establishing and defining relationships:** Eye contact is often the first stage in the initial encounter phase of a relationship. When a person catches the eye of another person, and if the receiver looks at the source, a relationship begins. If the receiver looks away from the source, a relationship is not started. Eye contact can decrease the physical distance between people and can be used to close others out of a conversation.
- **Expressing emotions:** While many areas of the face can be controlled, the eye area is one of the least controllable. As a result, the eyes and the area surrounding them reveals more accurate information about the emotional state than other areas of the face.
- **Controlling and regulating the interaction:** Bavelas et al. (2002) found that there is what they call a "gaze window" in a conversation. They suggested that speakers have minor breaks in their narratives to allow for short responses. At those points, there is a mutual gaze. This short gaze is then used by listeners to respond with "mmhm" or a nod to indicate microlevel understanding and/or agreement.

2.2.2 Computational Modeling of Eye-Gaze Signals

Eye-gaze can be used by robots to manage the conversational floor. For example, looking away can signal cognitive effort. Based on this insight, Andrist et al. (2014), presented a three-function gaze control system to control a robot. The robot uses face-tracking to engage in mutual gaze, idle head motion to increase lifelikeness, and purposeful gaze aversions to achieve regulatory conversational functions. In contrast, to achieve more natural and engaging gaze behavior, Sorostinean et al. (2014), presented a social attention system that tracks a person but attends to

strong motion when detected in its visual field. In order to generate a realistic robotic eye-gaze behavior, Kuno et al. (2006), analyzed human head orientation data in a museum setting. Based on this, they presented a system for a guide robot; the robot alternates gaze between exhibition items and human audiences while explaining the exhibits.

Joint attention can be useful during robot-to-human object handover. Grigore et al. (2013) studied a handover task where they compared a model based on only physical features of the action versus one that uses information about the human's engagement in an interaction: eye-gaze and head orientation as a sign of a human's focus of attention and engagement. Admoni et al. (2014) found that gaze signals influence people's compliance with the direction indicated by the gaze in ambiguous handover situations. Moon et al. (2014) showed that people reach for an offered object earlier when a robot signals via eye-gaze to the handover target location. Huang and Thomaz (2011) outlined a three-part model of joint attention capabilities for social robots: responding to joint attention, initiating joint attention, and ensuring joint attention.

Studies show humans to be influenced in a variety of ways by the robot's gaze. Staudte and Crocker (2009) demonstrated that a human's own gaze behavior and understanding of the robot's speech content is modulated by the coordination of that robot's speech and gaze. Admoni et al. (2013) found that people are more accurate at recognizing shorter, more frequent fixations than longer, less frequent ones. In a collaborative task, people are also found to take spatial and contextual cues from brief robotic glances (Mutlu et al., 2009). A humanoid robot's gaze has a positive impact on trust for difficult human decisions, (Stanton and Stevens, 2014), and robots are found to be more persuasive when they use gaze (Ham et al., 2015).

This thesis contributes to the literature on SSP and computational HRI with computational methods for analysis and generation of communicative eye-gaze signals in multiparty interactions. More specifically, one of the outcomes of the thesis is a newly collected multimodal dataset of multiparty face-to-face interactions. Another outcome is an investigation of computational methods for analysis and generation of eye-gaze direction or head orientation and the development of a real-time method for generation of candidate gaze targets.

2.3 Summary

In summary, most of the computational HRI research is concerned with recognizing pointing gestures and with generating regulators. While there is a considerable amount of research on the importance and effects of eye-gaze, there is little computational research on the production of appropriate gaze behaviors.

This chapter presented an overview of different categories of communicative hand gesture signals and communicative eye-gaze signals. The chapter also presented the functions these signals serve in face-to-face interaction. The multitude of functions these two signals serve in face-to-face interaction is the motivation

for this thesis. The goal of the thesis is defined as: to propose and develop computational methods that could enable machines to recognize communicative hand gesture signals and generate communicative eye-gaze signals. The next chapter is dedicated to the description of the research methodology employed in this work and concludes the first part of the thesis. The second part of the thesis is dedicated to outlining the contributions of the included publications. Each chapter in the second part follows the same structure: 1) a background section which includes the specific motivation and context of the study, and 2) a contributions section which includes discussion on the contributions and results of the study.

Chapter 3

Research Methodology

This chapter presents the main steps taken during the research described in the thesis. This research follows a data-driven approach that can be roughly divided into three main steps: data collection, signal modeling, and model evaluation.

3.1 Dataset Collection

A prerequisite for developing data-driven computational models of a certain communicative signal is a dataset that captures the signal in a given context. There are two alternatives for acquiring representative data; either to use a readily available dataset or to collect one. Using a prerecorded dataset is an appealing alternative since this considerably shortens the time needed to develop and test different methods. However, datasets are usually recorded for the purpose of investigating certain phenomena. Consequently, finding a dataset that addresses the specific needs of a novel study is difficult or such dataset might not exist. Therefore, the research in this thesis also includes the collection of datasets designed with a particular study in mind; the datasets capture the behavioral signal under consideration in a certain context. The following list describes some of the desired characteristics of the considered and collected datasets,

- **Multimodal:** The datasets are rich in perceptual signals. For example, the collected datasets include in one case (**Paper A**), streams of color, depth and body signals, and in another (**Paper E**), streams of audio, color, depth, infrared, body, face, eyes, touch and robot signals.
- **Multiparty:** The datasets involve interactions of more than two participants at a time (**Paper E**).
- **Multiuser:** The datasets involve recordings of many different participants. For example, the collected datasets include in one case (**Paper A**), 17 different participants, and in another (**Paper E**), 24 different participants.

- **Spontaneous:** The datasets include non-scripted interactions and natural behaviors (**Paper E**).

- **Dynamic:** The datasets are rich in conversational dynamics. For example, one of the collected datasets (**Paper E**) includes three types of interactions: 1) human-human-robot task-based interactions, 2) human-human-human task-based interactions, and 3) human-human-human open-world dialogues.

- **Automated:** The datasets take into consideration easy data post-processing (**Paper A** and **Paper E**). For example, during the data collection, signals and methods are used to ensure automated temporal and spatial alignment of the collected data streams.

- **Large:** The datasets capture representative data of many users and interactions (**Paper A** and **Paper E**).

3.2 Signal Modeling

The methods used for behavioral signal modeling are established methods from the field of *Machine learning* (Bishop, 2006, Mitchell, 1997). The research in this thesis relies on problem-specific modifications of different supervised, unsupervised, probabilistic (Murphy, 2012) and deep learning methods (Goodfellow et al., 2016). Some of the work also investigates application of transfer learning. The developed methods are either used for recognition/estimation of a communicative signal or for generation/synthesis of a communicative signal. It is important to outline the main difference between recognition and generation. While the main goal of recognition is to achieve average high performance for as many as possible subjects, the main goal of generation is to achieve average high performance for one subject.

Supervised machine learning methods rely on data labeling/annotation in order to perform some sort of optimization that "learns" a function which maps the input (data representations/features) to the output (target classes/labels). When using large datasets as the ones used in this work, manual data labeling is time consuming and erroneous process. Furthermore, manual data labeling is a subjective process and usually requires several annotators in order to minimize this subjective bias. These observations are the main motivation to use semi-automated data labeling techniques. This means that the labels are usually generated by an automated procedure and then they are manually inspected and corrected when necessary.

The main goal of the thesis is to propose and develop methods that can be used in real-time and real-world human-machine interactions. This puts some limitations on the type of assumptions that can be made for the environment in which the machine (*e.g.*, robot) will engage in interactions. The following list outlines some of the desired characteristics of the proposed methods. The proposed methods are designed in a way so most of the desired characteristics are met. However,

there are certain limitations and assumptions which are discussed in the included publications,

- **Real-time:** The proposed methods should be able to operate in real-time (**Paper B**, **Paper C**, **Paper D**, and **Paper F**). For example, the methods should not use future information for the perceptual signals. This means that the methods should predict/estimate the target signal based only on the history of the perceptual signals.
- **Non-intrusive:** The proposed methods should use non-intrusive sensors as perceptual signal generators (**Paper B**, **Paper C**, **Paper D**, and **Paper F**). For example, ideally the goal would be to limit the multi-sensory input to human-like perceptual abilities (*e.g.*, 2 color cameras or one RGB-D camera and 2 microphones).
- **Automatic:** The proposed methods should use data representations which are automatically generated (**Paper B**, **Paper C**, **Paper D**, and **Paper F**). For example, data representations that cannot be automatically generated in real-time or require manual annotation, should be avoided in order to ensure the real-time characteristic of the methods.
- **General:** The proposed methods should extend to unseen subjects. For example, a method that recognizes/estimates a certain signal produced by a group of subjects should be able to generalize beyond that group without significant decrease in performance.
- **Robust:** The proposed methods should not assume a specific interaction environment (**Paper B**, **Paper C**, **Paper D**, and **Paper F**). For example, assumptions like noise-free environment or known spatial configuration should be avoided.

3.3 Model Evaluation

Model evaluation involves many steps including, data partitioning, experimental setup, and choice of an evaluation metric. Data partitioning is an important process which ensures that the models are trained on a fraction of the dataset (train set) and tested on unseen data (test set). Furthermore, in order to select the final set of model parameters, a validation set is reserved for validating and comparing different model instances. The studies included in the thesis follow the same approach to data partitioning; all available data is randomly partitioned without replacement, where $\sim$80% of the data is used for training, $\sim$15% is used for testing, and $\sim$5% is used for validation (**Paper C**, **Paper D**, and **Paper F**). This type of data partitioning is reasonable when the amount of available data is large. However, this is not always the case which motivates other approaches to splitting the available data. One such approach used in this thesis is cross-validation, and more specifically, the

leave-one-out cross-validation technique (**Paper B**). For example, let us assume that the data for a certain hand gesture consists of only 5 examples. The leave-one-out cross-validation technique builds a model using 4 examples and tests the model on the 5th left-out example. This process is repeated 5 times (in this example) where every time a new model is built using different combinations of 4 examples and tested on the left-out example. The final model performance is reported in terms of average performance of the 5 different models (in this example).

The experiments conducted in the included studies can be classified as quantitative evaluation methods and are generally divided into three groups: subject dependent, multi-subject dependent, and subject independent. The goal of the subject dependent experiments is to test the proposed methods on data for only one subject at a time. This type of experiments is the first indication of the appropriateness of the methods for modeling of a certain behavioral signal. The goal of the multi-subject dependent experiments is to test the proposed methods on data for several subjects at a time. This type of experiments tests the scalability of the methods to more than one subject. Finally, the goal of the subject independent experiments is to test the proposed methods on data for unseen subjects. This type of experiments tests the transferability of the methods.

Given a model of a behavioral signal and an experimental setup, evaluation is generally reported in terms of accuracy (classification) and error (regression). In the broad sense, the model's accuracy in a classification task, is the amount of correctly predicted classes/labels as a fraction of the target classes/labels (**Paper B**, **Paper C** and **Paper F**). Specifically, counting the number of true positives (tp), false positives (fp), true negatives (tn), and false negatives (fp), the accuracy is calculated,

$$\mathrm{ACC} = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{tn} + \mathrm{fp} + \mathrm{fn}} \tag{3.1}$$

Accuracy can be an unreliable metric when the dataset is unbalanced (the number of observations in different classes vary greatly). The baseline chance performance using this metric can vary greatly in these cases. This observation motivated the use of weighted accuracy in one of the studies where the dataset is unbalanced (**Paper D**). As a consequence, regardless of the actual class distribution in the dataset, the baseline chance performance using weighted accuracy was always 50%. In that study, for a two-class classification task, the weighted accuracy is calculated,

$$\mathrm{WACC} = \frac{\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}} + \frac{\mathrm{tn}}{\mathrm{fp}+\mathrm{tn}}}{2} \tag{3.2}$$

In the broad sense, the model's error in a regression task, is the magnitude of error the model commits when estimating some continuous target value. The study described in **Paper F** uses mean absolute error for model evaluation. Denoting the number of samples with n, the estimated value with $\hat{y}$ and the target value with y, the model's mean absolute error is calculated,

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.3}$$

Part II

Thesis Contributions

Chapter 4

Hand Gesture Recognition

This chapter presents a study on the first topic addressed in the thesis: the real-time recognition of hand gestures and its application to recognition of isolated sign language signs. Hand gestures can also provide important cues during human-robot interaction, for example, emblems are type of hand gestures with specific meaning used as substitutes of words (Section 2.1).

4.1 Background

There has been a substantial research effort to develop assistive technology for deaf people. This group of people is at a disadvantage when it comes to communicating with society or access to information, such as in education and social services (Efthimiou and Fotinea, 2007, Steinberg et al., 1998). Outside the deaf community, there are groups of people who use signing complementary to speech in order to enhance communication. This group includes people with various disabilities such as developmental disorder, language disorder, cerebral palsy and autism. Within this group, *key word signing*, is a well established method of augmented and alternative communication. Key word signing borrows individual signs from a sign language to support and enhance the verbal signal (Windsor and Fristoe, 1991). Furthermore, key word signing discards the grammatical constructs in the sign language. Therefore, this type of communication support can be seen as communication based on emblematic gestures (Section 2.1).

While many deaf children have sign language as their first language and are able to acquire it in a natural way from the environment, children that need signs for other reasons do not have the same opportunities to be introduced to signs and signing. This observation motivates the development of a learning environment where children can learn signs in a game-like setting. During the game-play an avatar can present different signs and give the child certain tasks to accomplish, and by doing so the child practices signing. The learning environment is thus required to interpret the signs produced by the child.

4.1.1 Related Work

The study described in this chapter presents a hand gesture recognition method that models and recognizes manual components (hand shape/orientation and movement trajectories) of isolated sign language signs. A comprehensive review of the research on sign language recognition (SLR) and the main challenges is provided in (Cooper et al., 2011), while Mitra and Acharya (2007) and Rautaray and Agrawal (2015) surveyed the literature on gesture recognition. Several long-term research projects have been funded to develop sign language technology, such as ViSiCAST (2000), eSign (2008), SignCom (2011), SIGNSPEAK (2012), and Dicta-Sign (2012, 2012). In addition, there are several projects creating sign language datasets; in Sweden, the SSL Corpus Project (2009), in Germany, the DGS-Korpus dictionary project (2010), and in the UK, the BSL Corpus Project (2010).

4.1.2 Related Methods

Early work on SLR applied Artificial Neural Networks (ANN) for modeling isolated signs. The idea of one of the first papers on SLR, (Murakami and Taguchi, 1991), was to train an ANN given the features from a DataGlove, (Kadous, 1996), and recognize isolated signs. In (Kim et al., 1996), the researchers used DataGloves and Fuzzy Min-Max ANN to recognize 25 isolated gestures. The work in (Waldron and Kim, 1995) presented an isolated SLR system using ANN, and Huang and Huang (1998) presented an SLR system using a Hopfield ANN. Hidden Markov Models (HMM), (Rabiner, 1989, Yamato et al., 1992), are modeling techniques well suited for the problem of SLR (Starner et al., 1998). Grobel and Assan (1997) presented an isolated sign recognition system based on HMMs and Vogler and Metaxas (1997) showed that word-level HMMs are SLR suitable. In their following work, (Vogler and Metaxas, 1999), they demonstrated that Parallel HMMs are superior to regular HMMs, Factorial HMMs and Coupled HMMs for recognition of sign language. A more detailed review of the literature on SLR is presented in **Paper B**. The method developed in this study is based on HMMs because they are effective techniques for modeling spatio-temporal information and achieved state-of-the-art results at the time of the study. A short description of HMMs is presented next.

A time-domain process demonstrates a Markov property if the conditional probability density of the current event, given all present and past events, depends only on the jth most recent event. If the current event depends solely on the most recent past event, then the process is termed a first order Markov process. This assumption is reasonable to make when considering the position/orientation of the hands of a person through time. More specifically, an HMM is a double stochastic process governed by,

- An underlying Markov chain with a finite number of states.
- A set of random functions associated with each state.

In discrete time instants, the process is in one of the states and generates an observation symbol according to a random function corresponding to that state. Each transition between the states has a pair of probabilities defined as follows,

- Transition probability is the probability for undergoing a transition from one state to another.
- Output probability is the conditional probability of emitting an output symbol from a finite alphabet when the process is in a certain state.

The model is termed hidden because all that can be seen is a sequence of observations. An HMM is expressed as $\lambda = (A, B, \Pi)$, where A is state transition probability, B is observation symbol probability and Π is initial state probability. The generalized topology of an HMM is a fully connected structure, known as the ergodic model, where any state can be reached from any other state. For the recognition of isolated gestures, the goal is to predict the unknown class of an observation sequence O into one of C classes. If we denote C different models (*i.e.*, one per class) by λ_c, $1 \leq c \leq C$, then an observation sequence O is classified to class $\hat{c}$ using,

$$\hat{c} = \text{argmax}_{c \in C} P(O|\lambda_c) \tag{4.1}$$

4.2 Contributions

Previous Swedish Sign Language resources include the Swedish Sign Language Dictionary with approximately 8000 video recorded signs (Mesch and Wallin, 2012, Mesch et al., 2012). The main reason for recording a dataset instead of using the existing resources is that the study needed the depth information to be included in the recordings (RGB-D). This requirement stems from the real-time recognition goal of the learning environment. RGB-D recordings further help with resolving known issues related to hand tracking in video, such as different lighting, motion blur, and cluttered backgrounds. In addition, in order to capture enough variability for signer dependent as well as signer independent recognition, the study needed each sign to be repeated many times by different signers.

The dataset described in **Paper A** is an essential part of the developed method. The dataset captures 51 signs from the Swedish Sign Language (SSL) and the vocabulary is composed of four sets: objects, colors, animals, and attributes. The dataset has 2 parts: the first part captures experienced signers while the second part, captures inexperienced signers. The first part is composed of 7 signers that are experts in SSL (sign language researchers and/or teachers) and for six of them, SSL is the first language. The second part of the dataset is composed of 10 signers that have no prior experience in SSL. The recordings took place in a dedicated space but the recording environment was not explicitly controlled, the only requirement was that the upper body of the signer falls in the field of view of the RGB-D sensor. The

recording setup is illustrated in Figure 4.1. The signers were recorded performing all 51 signs in one session; 5 sessions were recorded, resulting in 5 instances of each sign per signer. In total, the dataset consists of 1785 sign instances for the experienced signers and 2550 sign instances for the inexperienced signers.

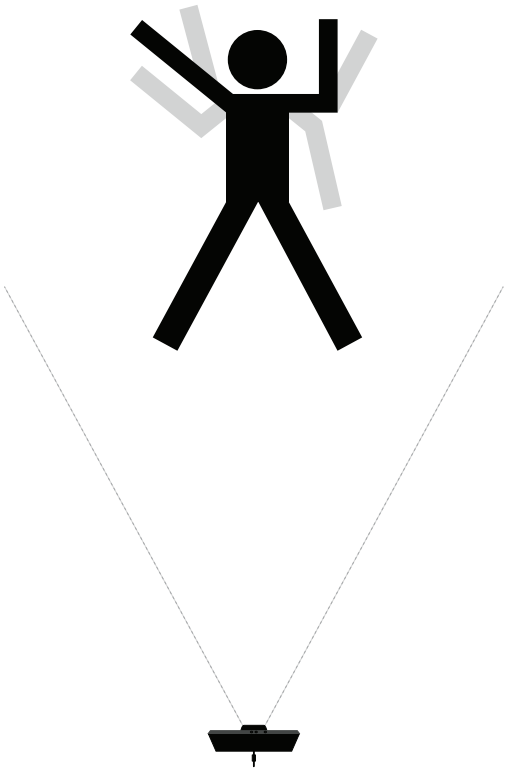


Figure 4.1: Spatial configuration of the setup in the Swedish Sign Language dataset. The dataset includes color, depth and body streams that are automatically segmented and aligned during the recording.

The method developed in this study is based on HMMs and is described in **Paper B**. In the developed method, the state index of the models transits from left to right with time, as illustrated in Figure 4.2, and the first and the last states, x_s and x_e, are always non-emitting. Here the state transition probabilities $a_{ij} = 0$ if $j < i$ and $\sum_{j=1}^{N} a_{ij} = 1$. Having the models' topology and the dataset (*i.e.*, the observation symbols are spatial representations of the trajectories of both hands), we train an HMM for each sign (λ_c). When an unknown observation sequence O is presented to the models, we calculate the likelihood $P(O|\lambda_c)$ for all $1 \leq c \leq 51$. Following Equation 4.1, the model which yields the maximum likelihood of observing the unknown sequence is selected by the method.

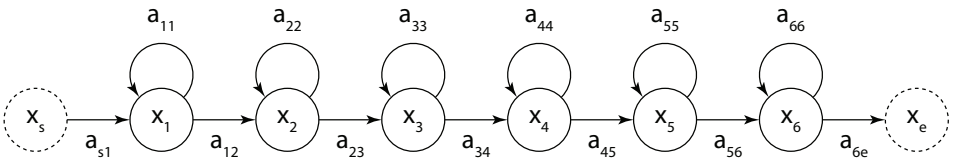


Figure 4.2: Topology of an 8-state left-to-right (no state skipping) HMM developed for modeling and recognition of isolated Swedish Sign Language signs.

The method is evaluated in two experiments: in a signer dependent experiment using leave-*one sign*-out cross-validation procedure, and in a signer independent experiment using leave-*one signer*-out cross-validation procedure (Section 3.3). The mean accuracy of the method in signer independent mode is significantly lower than in signer dependent: 57.9% (experienced signers) and 68.9% (inexperienced signers) compared to 87.6% (experienced signers) and 90.3% (inexperienced signers). These accuracy rates are however for the full set of 51 signs. In the learning environment, the method needs to recognize one out of a small number of signs (*e.g.*, one out of five animals). For this type of limited recognition tasks, accuracy will increase drastically since we can control the signs which are presented and make sure that signs that are confused by the method never appear together.

The applicability of the developed environment as a learning tool has been tested on a group of children (ages $10-11$) with no prior sign language skills. This study involved 38 children divided in two equal groups. Both groups played the same sign language related computer games. The first group accomplished tasks given by an avatar by performing isolate sign language signs, while the second group accomplished the same tasks by mouse clicking. A week after the children in both groups were asked to perform the signs they learned during the game-play. The main hypothesis of the study was that the group of children who interacted with the learning environment through signing will outperform the other group in terms of recalling the signs and carrying them out. Statistical test on the collected results confirmed the hypothesis (Potrus, 2017). This result showed that computer games that employ isolated signs as an interaction medium are a successful learning environment that can support the acquisition of sign language skills.

In summary, the thesis has two main contributions with respect to hand gesture recognition: 1) a newly collected dataset of isolated Swedish Sign Language signs, and 2) a real-time hand gesture recognition method. This chapter presented the importance of key word signing as method of augmented and alternative communication which motivated the development of a learning environment where children can learn signs in a game-like setting. The method presented here is an important component of that learning environment.

Chapter 5

Speech Activity Detection

This chapter presents a study on the second topic addressed in the thesis: the general problem of real-time speech activity detection in noisy and dynamic environments and its application to socially-aware language acquisition. Speech activity can also provide important information during human-robot interaction, for example, the current active speaker's hand gestures (Chapter 4) and eye-gaze direction or head orientation (Chapter 6) can play an important role in understanding the state of the interaction.

5.1 Background

The literature on language acquisition offers several theories of how infants learn their first words. One of the main problems which researchers face in this field is the problem of *referential ambiguity* as discussed in (Clerkin et al., 2016, Pereira et al., 2014, Yurovsky et al., 2013). Referential ambiguity stems from the idea that infants must acquire language by linking heard words with perceived visual scenes, in order to form word-referent mappings. In everyday life however, these visual scenes are highly cluttered which results in many possible referents for any heard word (Bloom, 2000, Quine et al., 2013). Many computational methods of language acquisition are rooted in finding statistical associations between verbal descriptions and the visual scene (Clerkin et al., 2016, Räsänen and Rasilo, 2015, Roy and Pentland, 2002, Yu and Ballard, 2004), or in more interactive robotic manipulation experiments (Salvi et al., 2012).

The literature on language acquisition does not consider how infants might know which caregiver is talking and therefore requires attention. This observation motivates the development of methods for inferring the active speaker in noisy and dynamic environments. Such methods could support an artificial cognitive system that attempts at acquiring language in similar manner as infants. Therefore, these methods should be plausible from a developmental perspective: one of the main implications is that the methods should not require manual annotations.

5.1.1 Related Work

The study described in this chapter presents methods for speech activity detection in social interactions. One of the mechanisms to cope with the problem of referential ambiguity is by using social signals related to the caregivers' intent. Although a word is heard in the context of many objects, infants may not treat the objects as equally likely referents. Instead, infants can use other social signals to rule out contenders to the named object. Yu and Smith (2016) used eye-tracking to record gaze data from both caregivers and infants and found that when the caregiver visually attended to the object to which infants' attention was directed, infants extended their duration of visual attention to that object, thus increasing the probability for successful word-referent mapping. Furthermore, infants do not learn only from interactions they are directly involved in, but also observe and attend to interactions between their caregivers. Handl et al. (2013) and Meng et al. (2017) performed studies to examine how the body orientation can influence the infants' gaze shifts. The main conclusion was that static body orientation alone can function as a signal for infants' observations and guides their attention.

5.1.2 Related Methods

Speech activity detection is important for many applications and each area imposes different constraints on the methods. Generally, there are two main approaches to speech activity detection: audio-only and audio-visual.

Audio-only active speaker detection is the process of finding segments in the input audio signal associated with different speakers. This type of detection is known as *speaker diarization*. Speaker diarization has been studied extensively and Anguera et al. (2012) offered a comprehensive review of the research in this field. Audio-visual speaker detection combines information from both the audio and the video signals. The application of audio-visual synchronization to speaker detection in broadcast videos was explored by (Nock et al., 2003). Unsupervised audio-visual detection of the speaker in meetings was proposed in (Friedland et al., 2009) and Zhang et al. (2008) presented a boosting-based multi-modal speaker detection algorithm applied to distributed meetings. In more recent studies, a multi-modal Long Short-Term Memory model that learns shared weights between modalities was proposed in (Ren et al., 2016) and Hu et al. (2015) proposed a Convolutional Neural Network model that learns the fusion function of face and audio information. A more detailed review of the literature on language acquisition and speech activity detection is presented in **Paper D**. The methods developed in this study are based on Perceptrons, Feedforward Artificial Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks. A short description of these methods is presented next.

A single artificial neuron (also called Perceptron) has the following mode of operation: it computes a weighted sum of all of its inputs X, using a learnable

weight vector W along with a learnable additive bias term, b. Then it potentially applies a non-linearity σ to the result,

$$h = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{5.1}$$

In other words, an artificial neuron performs a dot product with the input and its weights, adds a bias and applies a non-linearity, called an *activation function*. Some choices for activation functions include,

- Logistic function: $f(x) = \frac{1}{1+e^{-x}}$
- Hyperbolic tangent: $f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
- Rectifier: $f(x) = \max(0, x)$

It is possible to connect the outputs of artificial neurons to the inputs of other artificial neurons, giving rise to Artificial Neural Networks (ANN). In a feedforward ANN the neurons are typically organized in *fully-connected layers*, such that the neurons in adjacent layers have full pair-wise connections, but neurons within a layer are not connected. Feedforward ANNs receive an input through a fully-connected *input layer*, and transform it through series of fully-connected *hidden layers*. The last fully-connected layer is called the *output layer* and in classification mode it represents the target class/label scores. One drawback of feedforward ANNs is that they do not scale well to inputs like images.

Convolutional Neural Networks (CNN) make the explicit assumption that the inputs are images. CNNs are very similar to feedforward ANNs: they are made up of neurons that have learnable weights and biases and each neuron receives some inputs, performs a dot product and optionally applies a non-linearity. There are three main types of layers in CNNs: *convolutional layer*, *pooling layer*, and fully-connected layer. Using these layers, CNNs transform the original image from pixel values to the target class/label scores. CNNs take advantage of the fact that the input consists of images to constrain the architecture; the main difference between fully-connected and convolutional layers is that the neurons in a convolutional layer are connected only to a local region in the input, and that many of them share the learnable parameters. The main function of the pooling layer (periodically inserted between successive convolutional layers) is to reduce the spatial size of the representation.

Traditional artificial neural networks are memoryless; it is difficult for these networks to use previous events to inform later ones. Recurrent Neural Networks (RNN) is one approach (of many) to addresses this issue. This type of networks have loops in them, allowing information to persist; the network looks at some input x_t at time t and outputs a value h_t. The loops allow information to be passed from one step of the network to the next. Long Short-Term Memory networks (LSTM) are RNNs capable of learning long-term dependencies. RNNs have the form of a

chain of repeating modules such as a single fully-connected layer. LSTMs also have this chain like structure, but the repeating module consists of four fully-connected layers that interact in a specific way. The key to LSTMs is a fully-connected layer called the *cell state*; LSTMs have the ability to remove or add information to the cell state, via process regulated by structures called *gates*, which are fully-connected layers as well. An LSTM has three of these gates: *forget*, *input*, and *output* gate.

5.2 Contributions

For automatic language acquisition, the goal is to infer the possible objects the active speaker is focusing attention on. In this context, assumptions such as known sensor arrangement or participants' position and number in the environment are unrealistic, and should be avoided (these assumptions should be avoided in the context of HRI as well). Therefore, the methods developed in this study have several desirable characteristics for such types of scenarios, 1) they work in real-time, 2) they do not assume specific spatial configuration (sensors and participants), 3) the number of possible (simultaneously) active speakers is free to change during the interaction, and 4) no externally produced labels are required, but rather the acoustic inputs are used as reference to the visually based learning.

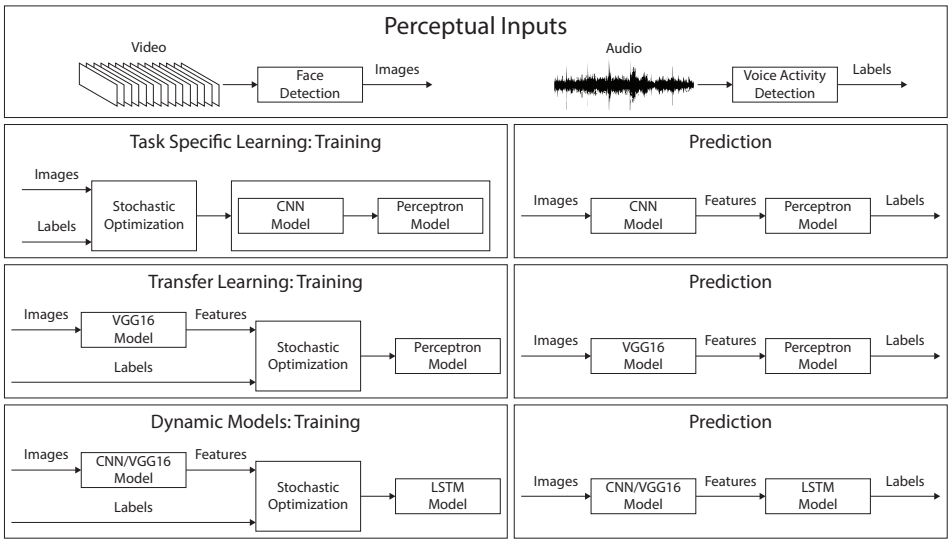


Figure 5.1: Overview of the approaches to active speaker detection. In the top row are the perceptual inputs and the way they are automatically modified before being passed to the static (second row), transfer learning (third row) and dynamic (forth row) methods.

The dataset described in **Paper E** is an essential part of the developed methods. The dataset is described in Section 6.2 where it is also used for analysis and generation of eye-gaze direction or head orientation. The study discussed here investigates a self-supervised learning approach to construct an active speaker detector: the machine learning methods are supervised, but the labels are obtained automatically from the auditory modality to learn detectors in the visual modality. All detectors, illustrated in Figure 5.1, make use of data representations based on a CNN and a following classifier. The study considers two types of classifiers: static (Perceptron) and dynamic (LSTM). Additionally, two approaches to training are considered: transfer learning that employs a pretrained CNN for data representation and only the classifier is trained, and task specific learning that employs simultaneous training of the CNN for data representation and the classifier. The incremental development of these ideas and methods is described in details in **Paper C** and **Paper D**.

The methods are evaluated in three experiments: speaker dependent, multi-speaker dependent, and speaker independent. The mean weighted accuracy (Section 3.3) of the methods in speaker independent mode is significantly lower than in speaker dependent and multi-speaker dependent: 60.3% compared to 75.9% and 80.2%. The results of the multi-speaker dependent experiment show that the proposed methods can scale beyond a single subject without decrease in performance. Combining this observation with the shown applicability of transfer learning to the task suggests that, the proposed methods can generalize to unseen perceptual inputs by incorporating a model adaptation step for each new speaker.

In summary, the thesis has one main contribution with respect to speech activity detection: a real-time vision-based speech activity detection method. This chapter presented the importance of speech activity for socially-aware language acquisition which motivated the development of the active speaker detection methods. The methods presented here are a prerequisite for socially-aware language acquisition and they can be seen as mechanisms for constraining the visual input thus helping to resolve the referential ambiguity in dynamic visual scenes.

In face-to-face interaction people not only hear when someone is talking, but they also see that person talking. This observation and the inherited difficulties of audio-only speech activity detection in noisy and dynamic environments, is another motivation for the developed methods. From an HRI perspective, the methods could enable a robot to see which of its interlocutors is talking. Furthermore, there has been little work on the question of whom (or what) to follow during face-to-face human-robot interaction. This study proposes one way to addresses this question. Our hypothesis was that speech activity is one of the predictors of a perceiver worth following during interactions. This hypothesis was confirmed in an experiment described in **Paper F** and discussed in Section 6.2.

Chapter 6

Eye-Gaze Analysis and Generation

The third topic addressed in the thesis is the real-time generation of eye-gaze direction or head orientation and its application to human-robot interaction. Eye-gaze direction or head orientation can provide important cues during human-robot interaction, for example, it can regulate who is allowed to speak when and coordinate the changes in the roles on the conversational floor (Section 2.2).

6.1 Background

Gaze patterns related to regulating the interaction are generally unconscious (Argyle and Graham, 1976, Kendon, 1967, Vertegaal et al., 2001). It is also important to consider environmental factors when trying to model gaze under varying contexts, as it is otherwise impossible to make any universal judgments (Peters et al., 2010). While there are some patterns that hold more globally, it has been found that many gaze patterns vary substantially from one pair to the next (Cummins, 2012). It has also been found that, gaze coordination is related to factors such as established common ground and mutual knowledge (Shockley et al., 2009). Another interesting finding is that speakers and listeners are different with regards to their gaze patterns. While listeners gaze at speakers for long periods of time, the speakers gaze at listeners in short but frequent periods (Argyle and Cook, 1976).

Since, clear conversational roles in face-to-face communication are vital for smooth and effective interaction, a robot which is aware of the established roles could avoid misunderstandings or talking over its interlocutors. On one hand, most of the research on gaze in conversational settings has been carried out on *dyadic* interactions. On the other, research on gaze in *multiparty* settings usually develops rule-based methods for gaze generation. These observations motivate the development of data-driven methods for generation of gaze for a robot involved in multiparty interactions.

6.1.1 Related Work

The study described in this chapter presents methods for generating candidate gaze targets in multiparty interactions. There are three main approaches to generating social gaze: biologically-inspired, data-driven, and heuristic (Admoni and Scassellati, 2017). Biologically inspired approaches are either mimicking *bottom-up* neurological responses to the visual input from the environment or are *top-down* cognitive architectures that derive context to detect visual saliency. Humans have the ability to rapidly orient their attention to visually salient locations and only a small fraction of humans' visual input is registered and processed (Itti and Koch, 2000). Several researchers have computationally described and modeled aspects of visual attention and saliency (Borji and Itti, 2013, Frintrop et al., 2010, Itti and Koch, 2001).

Heuristic approaches design rule-based methods that use observations from human interactions in order to mimic visual attention behavior. Some studies have tested heuristic methods in dyadic human interactions. The study in (Zhang et al., 2017) described an interactive gaze method implemented on a humanoid robot. The system's usability for establishing mutual gaze with a user was also tested. Peters et al. (2005) presented a method for an embodied conversational agent able to establish, maintain and end the conversation based on its perception of the level of interest of its interlocutor. In this work the *speaker* and the *listener* were modeled separately. Heuristic methods have also been proposed in multiparty interactions. The study in (Bennewitz et al., 2005) presented a humanoid museum guide robot. The proposed system was able to interact with people in multiparty scenarios using attention shifts among other modalities.

6.1.2 Related Methods

This chapter presents a study on data-driven methods for generation of candidate gaze targets in multiparty interactions. Data-driven approaches model behavioral aspects of gaze in human conversations with empirical measurements such as gaze timings, frequencies and locations. The model parameters are typically extracted by analyzing video data of human dyadic interactions. The studies in (Andrist et al., 2013) and (Andrist et al., 2014) proposed computational methods for generation of gaze aversions in relation to conversational functions and speech. Admoni and Scassellati (2014) presented a computational method for generation of robot nonverbal behavior. The method can be both predictive (by recognizing the context of new nonverbal behaviors) and generative (by creating new nonverbal behavior based on the desired context). A more detailed review of the literature on approaches to generating gaze is presented in **Paper F**. The methods developed in this study are based on Feedforward Artificial Neural Networks and Recurrent Neural Networks. A short description of these methods is presented Section 5.1.

6.2 Contributions

The methods developed in this study model the eye-gaze direction and head orientation of a person in three-party open-world dialogues, (Bohus and Horvitz, 2010), as a function of low-level multimodal signals generated by the interlocutors. These signals include, speech activity, eye-gaze direction and head orientation which can be automatically estimated during the interaction.

The dataset described in **Paper E** is an essential part of the development of the methods. The dataset involves multiparty human-robot interactions based around objects, multiparty human-human interactions based around the same objects, and multiparty human-human open-world dialogues. The primary purpose of the dataset is to serve as a source for modeling visual attention patterns for robots interacting with humans, but the richness of the dataset also makes it useful for other studies (**Paper C** and **Paper D**). In total 15, ∼30-minute sessions were recorded, resulting in ∼7.5 hours of data. Three participants took part in each recording session where a pair of participants was new in every session, and one participant took the role of moderator for all sessions. All interactions were in English and all data streams were spatially and temporally synchronized and aligned. All interactions occurred around a round interactive surface and the participants were seated. There were 23 unique participants and 1 moderator. Figure 6.1 illustrates the spatial configuration of the setup.

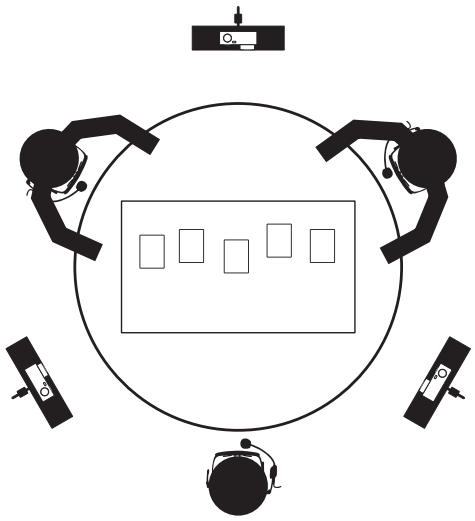
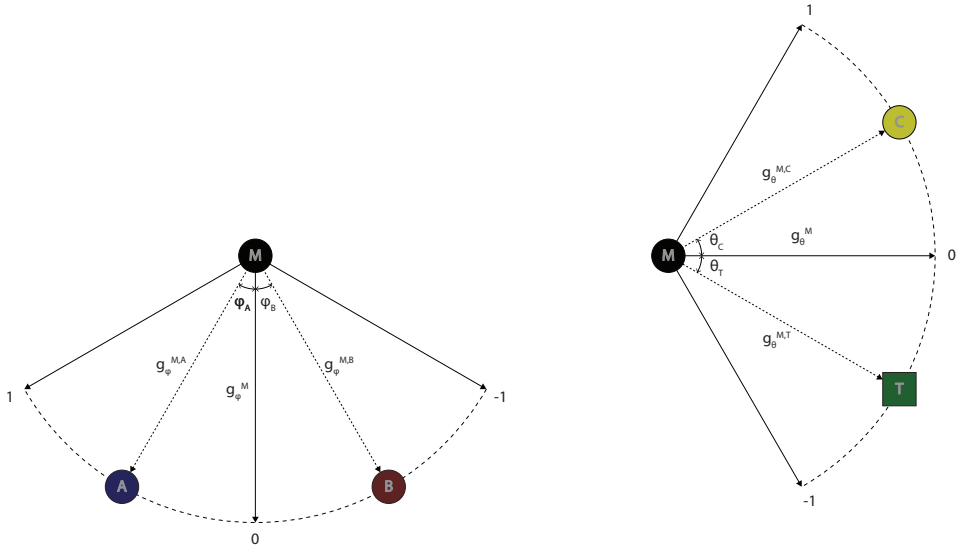


Figure 6.1: Spatial configuration of the setup in the multiparty interaction dataset. The dataset includes audio, color, depth, infrared, body, face, eyes, touch and robot streams and is semi-automatically aligned and synchronized during the recording.

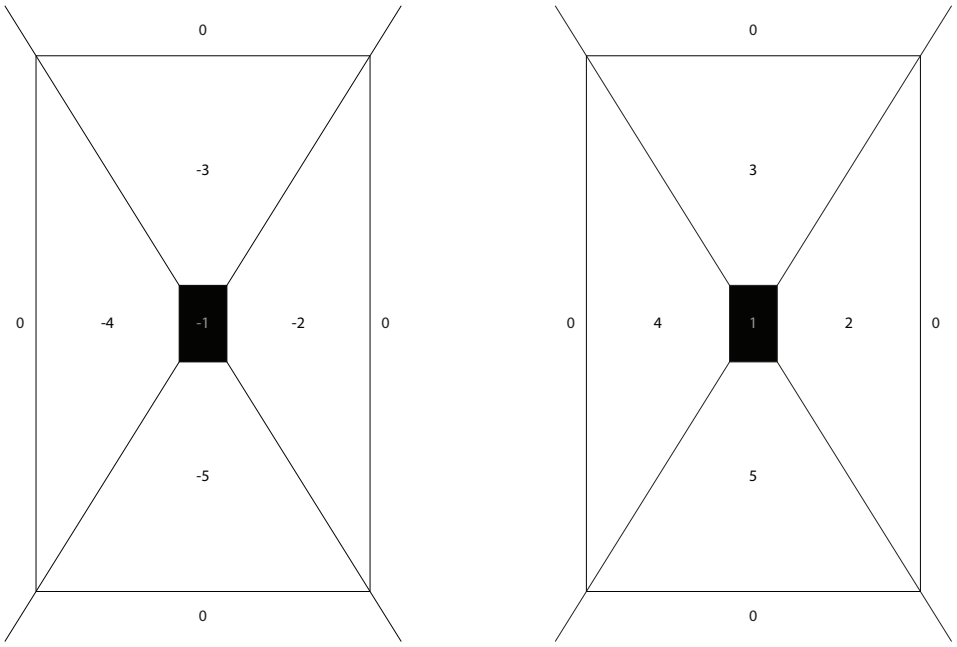
A raw three-dimensional representation of the eye-gaze direction or head orientation does not capture the dynamic relation between the candidate gaze targets and the current spatial state of the interaction. One of the contributions of this study (**Paper F**) is data representations suitable for modeling spatial relations in the context of multiparty open-world dialogues. The study proposes two types of data representation: continuous and discrete.



(a) Horizontal. (b) Vertical.

Figure 6.2: Continuous and active data representation (for simplicity, only the eye-gaze directions are drawn). In the figures M is the person being modeled. The azimuthal angles of the eye-gaze directions or head orientations of M that pass through the current position of A (interlocutor A) and B (interlocutor B) are used to create a data representation interval $[-0.5, 0.5]$ for the azimuthal angles. The polar angles of the eye-gaze directions or head orientations of M that pass through C (the current mean position of A and B) and the position of T (a static object) are used to create a data representation interval $[-0.5, 0.5]$ for the polar angles. Both data representation intervals are dynamically extended to $[-1, 1]$ based on the current position of A and B. Then, the current state of the interaction from the perspective of M is expressed in terms of: M's eye-gaze direction or head orientation. This process of encoding is used to encode the current state of the interaction from the perspective of all participants.

Within the continuous data representation the study considers two cases: passive and active. The passive case uses the human visual field as a reference frame for data representation (Walker et al., 1990). The active case uses the current position of the two interlocutors as a reference frame for data representation. This later case is illustrated in Figure 6.2. A more detailed description of the continuous approaches to data representation is presented in **Paper F**.



(a) Interlocutor A. (b) Interlocutor B.

Figure 6.3: Discrete and simple data representation. As seen from the perspective of M, given the current position of interlocutor A, a grid centered at interlocutor A is defined. The grid is defined in such a way that it partitions the space around the position of interlocutor A into 5 regions. The same type of partitioning is applied to interlocutor B. The regions (one for interlocutor A and one for interlocutor B) intersected by the current eye-gaze direction or head orientation of M are used as encoding. This process of encoding is used to encode the current state of the interaction from the perspective of all participants.

Within the discrete data representation the study considers two cases: complex and simple. The complex case partitions the space around the two interlocutors into 42 (+1) regions while the simple case partitions the space around the two interlocutors into 10 (+1) regions. This later case is illustrated in Figure 6.3.

A more detailed description of the discrete approaches to data representation is presented in **Paper F**.

The study discussed here investigates a supervised learning approach for generation of candidate targets for eye-gaze direction or head orientation. The study considers two types of methods: static (ANN, Section 5.1) and dynamic (LSTM, Section 5.1) performing both classification and regression tasks. In the case of discrete data representation, the goal of the classifier (ANN or LSTM) is to predict the most likely regions (one per interlocutor) for eye-gaze direction or head orientation. In the case of continuous data representation, the goal of the regressor (ANN or LSTM) is to estimate the azimuthal and polar angles for eye-gaze direction or head orientation. The methods are evaluated in one experiment: subject dependent (Section 3.2). This study also uses a second dataset described in (Kontogiorgos et al., 2018), which is similar to the dataset described in **Paper E**. Therefore, the study presents a subject dependent experiment for both moderators, for short, moderator A (**Paper E**) and moderator B (Kontogiorgos et al., 2018).

For the continuous and passive representations and for both moderators, the methods reach the lowest mean absolute error (the best is 0°, Section 3.3) when estimating the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input: $\sim$9° and $\sim$13° for the azimuthal angle, and $\sim$3° and $\sim$7° for the polar angle, for moderator A and B, respectively. For the continuous and active representations and for both moderators the methods reach the lowest error when estimating the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input: $\sim$9° and $\sim$13° for the azimuthal angle, and $\sim$5° and $\sim$7° for the polar angle, for moderator A and B, respectively.

For the discrete and complex representations and for both moderators the methods reach the highest accuracy (the best is 1, Section 3.3) when estimating the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input: 0.71 and 0.75 for interlocutor A, and 0.40 and 0.61 for interlocutor B, for moderator A and B, respectively. For the discrete and simple representations and for moderator A the methods reach the highest accuracy when estimating the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input: 0.98 and 0.84 for interlocutor A and B, respectively. For moderator B the methods reach the highest accuracy when estimating the moderator's eye-gaze direction using the eye-gaze direction of the interlocutors and the speech activity as input: 0.97 and 0.95 for interlocutor A and B, respectively.

In summary, the thesis has two main contributions with respect to generation of eye-gaze direction or head orientation: 1) a newly collected dataset of face-to-face interactions, and 2) a real-time eye-gaze direction or head orientation generation method. The main finding of the study is that the used descriptors are good predictors for eye-gaze direction or head orientation when the moderators are in a listening state. When the moderators are in a speaking state, the used descriptors are not sufficient since they do not encode the moderators' intentions. In addition,

the results from the study clearly show that a candidate gaze targets generation method that takes into account the speech activity significantly outperforms one which does not use this information. This result further motivates the methods for speech activity detection described in Chapter 5.

Chapter 7

Conclusions

This thesis comes in the pursuit of the ultimate goal of building autonomous socially intelligent artificial systems that are able to interact with humans in a natural and effective way. Such systems need to recognize and generate the subtle, rich and multimodal communicative signals that complement the stream of words – the communicative signals humans typically use when interacting with each other. The studies included in this thesis, propose, develop and evaluate methods for real-time recognition and generation of such communicative signals.

The work in **Paper A** and **Paper B** addresses the problem of real-time recognition of hand gestures and its application to the recognition of isolated sign language signs. This work includes the collection of a new dataset of isolated Swedish Sign Language signs and development of new a recognition method. The developed method is an important component of a learning environment that employs isolated sign language signs as an interaction modality. The conducted user study investigates the applicability of the learning environment as a learning tool for a group of children with no prior sign language skills. The result of the study shows that the learning environment can support the acquisition of sign language skills.

The work in **Paper C** and **Paper D** addresses the problem of real-time speech activity detection in noisy and dynamic environments and its application to socially-aware language acquisition. This work includes the incremental development of new speech activity detection methods. The developed methods attempt to limit the assumptions about the interaction environment to a minimum by not assuming specific spatial configuration and specific number of possible (simultaneously) active speakers. Furthermore, the methods are plausible from a developmental perspective because they do not require manual annotations. These methods could be an important component of any artificial system that engages in social interactions with humans.

The work in **Paper E** and **Paper F** addresses the problem of real-time generation of eye-gaze direction or head orientation and its application to human-robot interaction. This work includes the collection of a new dataset of face-to-face

human-human and human-robot interactions and the development of new a eye-gaze direction and head orientation generation methods. Since eye-gaze direction and head orientation have important functions in human-human interactions such as regulating the changes in the roles on the conversational floor, the developed methods could also play an important role in any artificial system that engages in social interactions with humans.

There are numerous directions for future work. Although single communicative signal can play an important function in face-to-face interactions, more typically this function is composed of groups of communicative signals that interact to create communicative impact. Therefore, an obvious direction for future work is unifying the methods described in this thesis into the perception and generation modules of an artificial system and let them interact. In turn, such system can have a wide range of applicability, including social robotics for healthcare, education and entertainment. Alternatively, the methods presented here are only one attempt at modeling very complex communicative signals that are influenced by the context, among others, in which they occur. This motivates further investigation of these signals under different context and using different modeling approaches.

Bibliography

Admoni, H., Dragan, A., Srinivasa, S. and Scassellati, B. (2014), Deliberate delays during robot-to-human handovers improve compliance with gaze communication, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 49–56.

Admoni, H., Hayes, B., Feil-Seifer, D., Ullman, D. and Scassellati, B. (2013), Are you looking at me? perception of robot attention is mediated by gaze type and group size, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 389–396.

Admoni, H. and Scassellati, B. (2014), Data-driven model of nonverbal behavior for socially assistive human-robot interactions, *in* 'Proceedings of the ACM International Conference on Multimodal Interaction', pp. 196–199.

Admoni, H. and Scassellati, B. (2017), 'Social eye gaze in human-robot interaction: A review', *Journal of Human-Robot Interaction* **6**(1), 25–63.

Andersen, P. (1999), *Nonverbal Communication: Forms and Functions*, Mayfield Publishing.

Andrist, S., Mutlu, B. and Gleicher, M. (2013), Conversational gaze aversion for virtual agents, *in* 'Proceedings of the Intelligent Virtual Agents', pp. 249–262.

Andrist, S., Tan, X., Gleicher, M. and Mutlu, B. (2014), Conversational gaze aversion for humanlike robots, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 25–32.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. and Vinyals, O. (2012), 'Speaker diarization: A review of recent research', *IEEE Transactions on Audio, Speech, and Language Processing* **20**(2), 356–370.

Argyle, M. and Cook, M. (1976), *Gaze and Mutual Gaze*, Cambridge University Press.

Argyle, M. and Dean, J. (1965), 'Eye-contact, distance and affiliation', *Sociometry* **28**(3), 289–304.

Argyle, M. and Graham, J. (1976), 'The central europe experiment: Looking at persons and looking at objects', *Environmental Psychology and Nonverbal Behavior* **1**(1), 6–16.

Argyle, M. and Ingham, R. (1972), 'Gaze, mutual gaze, and proximity', *Semiotica* **6**.

Argyle, M., Ingham, R., Alkema, F. and McCallin, M. (1973), 'The different functions of gaze', *Semiotica* **7**, 19–32.

Bavelas, J., Coates, L. and Johnson, T. (2002), 'Listener responses as a collaborative process: The role of gaze', *Journal of Communication* **52**(3), 566–580.

Bavelas, J. and Gerwing, J. (2011), 'The listener as addressee in face-to-face dialogue', *International Journal of Listening* **25**(3), 178–198.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M. and Behnke, S. (2005), Towards a humanoid museum guide robot that interacts with multiple persons, *in* 'Proceedings of the IEEE-RAS International Conference on Humanoid Robots', pp. 418–423.

Birdwhistell, R. (1970), *Kinesics and Context: Essays on Body Motion Communication*, University of Pennsylvania Press.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Springer-Verlag.

Bloom, P. (2000), *How Children Learn the Meanings of Words*, MIT press.

Bohus, D. and Horvitz, E. (2010), Facilitating multiparty dialog with gaze, gesture, and speech, *in* 'Proceedings of the International Conference on Multimodal Interfaces'.

Borji, A. and Itti, L. (2013), 'State-of-the-art in visual attention modeling', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(1), 185–207.

Brooks, A. and Breazeal, C. (2006), Working with robots and objects: Revisiting deictic reference for achieving spatial common ground, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 297–304.

BSL (2010), 'http://www.bslcorpusproject.org/'.

Burger, B., Ferrané, I., Lerasle, F. and Infantes, G. (2012), 'Two-handed gesture recognition and fusion with speech to command a robot', *Autonomous Robots* **32**(2), 129–147.

Castellano, G., Paiva, A., Kappas, A., Aylett, R., Hastie, H., Barendregt, W., Nabais, F. and Bull, S. (2013), Towards empathic virtual and robotic tutors, *in* 'Proceedings of the Artificial Intelligence in Education', pp. 733–736.

Clerkin, E., Hart, E., Rehg, J., Yu, C. and Smith, L. (2016), 'Real-world visual statistics and infants' first-learned object names', *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **372**(1711).

Cooper, H., Holt, B. and Bowden, R. (2011), *Sign language recognition*, Springer London, pp. 539–562.

Cummins, F. (2012), 'Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals', *Language and Cognitive Processes* **27**(10), 1525–1549.

Darwin, C. (1873), *The Expression of the Emotions in Man and Animals*, John Murray.

DGS (2010), 'http://www.sign-lang.uni-hamburg.de/dgs-korpus/'.

DICTA (2012), 'http://www.dictasign.eu'.

Droeschel, D., Stückler, J. and Behnke, S. (2011), Learning to interpret pointing gestures with a time-of-flight camera, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 481–488.

Droeschel, D., Stückler, J., Holz, D. and Behnke, S. (2011), Towards joint attention for a domestic service robot — person awareness and gesture recognition using time-of-flight cameras, *in* 'Proceedings of the IEEE International Conference on Robotics and Automation', pp. 1205–1210.

Duncan, S. (1972), 'Some signals and rules for taking speaking turns in conversations', *Journal of Personality and Social Psychology* **23**, 283–292.

Duncan, S. (1974), 'On the structure of speaker-auditor interaction during speaking turns', *Language in Society* **3**(2), 161–180.

Efthimiou, E. and Fotinea, S. (2007), An environment for deaf accessibility to educational content, *in* 'Proceedings of the International Conference on Information and Communication Technology and Accessibility', pp. 125–130.

Efthimiou, E., Fotinea, S., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P. and Lefebvre-Albaret, F. (2012), Sign language technologies and resources of the dicta-sign project, *in* 'Proceedings of the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon', pp. 37–44.

Ekman, P. (1976), 'Movements with precise meanings', *Journal of Communication* **26**(3), 14–26.

Ekman, P. and Friesen, W. (1969*a*), 'Nonverbal leakage and clues to deception', *Psychiatry* **32**(1), 88–106.

Ekman, P. and Friesen, W. (1969*b*), 'The repertoire of nonverbal behavior: Categories, origins, usage, and coding', *Semiotica* **1**(1), 49–98.

Ekman, P. and Friesen, W. (1972), 'Hand movements', *Journal of Communication* **22**(4), 353–374.

Ekman, P. and Friesen, W. (1974), 'Detecting deception from the body or face', *Journal of Personality and Social Psychology* **29**, 288–298.

Elliott, R., Glauert, J., Kennaway, J. and Marshall, I. (2000), The development of language processing support for the visicast project, *in* 'Proceedings of the ACM International Conference on Assistive Technologies', pp. 101–108.

Feil-Seifer, D. and Matarić, M. (2005), Defining socially assistive robotics, *in* 'Proceedings of the IEEE International Conference on Rehabilitation Robotics', pp. 465–468.

Friedland, G., Yeo, C. and Hung, H. (2009), Visual speaker localization aided by acoustic models, *in* 'Proceedings of the ACM International Conference on Multimedia', pp. 195–202.

Frintrop, S., Rome, E. and Christensen, H. (2010), 'Computational visual attention systems and their cognitive foundations: A survey', *ACM Transactions on Applied Perception* **7**(1), 1–39.

Gibet, S., Courty, N., Duarte, K. and Naour, T. (2011), 'The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation', *ACM Transactions on Interactive Intelligent Systems* **1**(1), 1–23.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep Learning*, MIT Press.

Grigore, E., Eder, K., Pipe, A., Melhuish, C. and Leonards, U. (2013), Joint action understanding improves robot-to-human object handover, *in* 'Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 4622–4629.

Grobel, K. and Assan, M. (1997), Isolated sign language recognition using hidden markov models, *in* 'Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics', pp. 162–167.

Hall, E. (1990), *The Hidden Dimension*, Anchor.

Ham, J., Cuijpers, R. and Cabibihan, J. (2015), 'Combining robotic persuasive strategies: The persuasive power of a storytelling robot that uses gazing and gestures', *International Journal of Social Robotics* **7**(4), 479–487.

Handl, A., Mahlberg, T., Norling, S. and Gredebäck, G. (2013), 'Facing still faces: What visual cues affect infants' observations of others?', *Infant Behavior and Development* **36**(4), 583–586.

Hargie, O. (2011), *Skilled Interpersonal Communication: Research, Theory and Practice*, Routledge.

Häring, M., Eichberg, J. and André, E. (2012), Studies on grounding with gaze and pointing gestures in human-robot-interaction, *in* 'Proceedings of the International Conference on Social Robotics', pp. 378–387.

Hu, Y., Ren, J., Dai, J., Yuan, C., Xu, L. and Wang, W. (2015), 'Deep multimodal speaker naming', *Computing Research Repository* **abs/1507.04831**.

Huang, C. and Huang, W. (1998), 'Sign language recognition using model-based tracking and a 3d hopfield neural network', *Machine Vision and Applications* **10**(5), 292–307.

Huang, C. and Mutlu, B. (2013), Modeling and evaluating narrative gestures for humanlike robots, *in* 'Proceedings of the Robotics: Science and Systems', pp. 57–64.

Huang, C. and Thomaz, A. (2011), Effects of responding to, initiating and ensuring joint attention in human-robot interaction, *in* 'Proceedings of the IEEE International Conference on Robot and Human Interactive Communication', pp. 65–71.

Iio, T., Shiomi, M., Shinozawa, K., Akimoto, T., Shimohara, K. and Hagita, N. (2010), Entrainment of pointing gestures by robot motion, *in* 'Proceedings of the International Conference on Social Robotics', pp. 372–381.

Itti, L. and Koch, C. (2000), 'A saliency-based search mechanism for overt and covert shifts of visual attention', *Vision Research* **40**(10), 1489–1506.

Itti, L. and Koch, C. (2001), 'Computational modelling of visual attention', *Nature Reviews Neuroscience* **2**(3).

Kadous, M. (1996), Machine recognition of auslan signs using powergloves: Towards large-lexicon recognition of sign language, *in* 'Proceedings of the Integration of Gesture in Language and Speech', pp. 165–174.

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

Kendon, A. (1980), *The Relationship of Verbal and Nonverbal Communication*, Mouton Publishers.

Kim, J., Jang, W. and Bien, Z. (1996), 'A dynamic gesture recognition system for the korean sign language (ksl)', *IEEE Transactions on Systems, Man, and Cybernetics* **26**(2), 354–359.

Knapp, M., Hall, J. and Horgan, T. (2013), *Nonverbal Communication in Human Interaction*, Cengage Learning.

Kontogiorgos, D., Avramova, V., Alexandersson, S., Jonell, P., Oertel, C., Beskow, J., Skantze, G. and Gustafson, J. (2018), A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction, *in* 'Proceedings of the International Conference on Language Resources and Evaluation'.

Kuno, Y., Sekiguchi, H., Tsubota, T., Moriyama, S., Yamazaki, K. and Yamazaki, A. (2006), Museum guide robot with communicative head motion, *in* 'Proceedings of the IEEE International Conference on Robot and Human Interactive Communication', pp. 33–38.

Liu, P., Glas, D., Kanda, T., Ishiguro, H. and Hagita, N. (2013), It's not polite to point: Generating socially-appropriate deictic behaviors towards people, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 267–274.

Lohse, M., Rothuis, R., Pérez, J., Karreman, D. and Evers, V. (2014), Robot gestures make difficult tasks easier: The impact of gestures on perceived workload and task performance, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', pp. 1459–1466.

McNeill, D. (1992), *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press.

McNeill, D. (2005), *Gesture and Thought*, University of Chicago Press.

Mehrabian, A. (1972), *Nonverbal Communication*, Transaction Publishers.

Meng, X., Uto, Y. and Hashiya, K. (2017), 'Observing third-party attentional relationships affects infants' gaze following: An eye-tracking study', *Frontiers in Psychology* **7**.

Mesch, J. and Wallin, L. (2012), From meaning to signs and back: Lexicography and the swedish sign language corpus, *in* 'Proceedings of the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon', pp. 123–126.

Mesch, J., Wallin, L. and Björkstrand, T. (2012), Sign language resources in sweden: Dictionary and corpus, *in* 'Proceedings of the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon', pp. 127–130.

Mitchell, T. (1997), *Machine Learning*, McGraw Hill.

Mitra, S. and Acharya, T. (2007), 'Gesture recognition: A survey', *IEEE Transactions on Systems, Man, and Cybernetics* **37**(3), 311–324.

Moon, A., Troniak, D., Gleeson, B., Pan, M., Zheng, M., Blumer, B., MacLean, K. and Croft, E. (2014), Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 334–341.

Moore, N., Hickson, M. and Stacks, D. (2013), *Nonverbal Communication: Studies and Applications*, Oxford University Press.

Murakami, K. and Taguchi, H. (1991), Gesture recognition using recurrent neural networks, *in* 'Proceedings of the Conference on Human Factors in Computing Systems', pp. 237–242.

Murphy, K. (2012), *Machine Learning: A Probabilistic Perspective*, MIT Press.

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H. and Hagita, N. (2009), Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 69–76.

Nagai, Y. (2005), The role of motion information in learning human-robot joint attention, *in* 'Proceedings of the IEEE International Conference on Robotics and Automation', pp. 2069–2074.

Nock, H., Iyengar, G. and Neti, C. (2003), Speaker localisation using audio-visual synchrony: An empirical study, *in* 'Proceedings of the International Conference on Image and Video Retrieval', pp. 488–499.

Ou, S. and Grupen, R. (2010), From manipulation to communicative gesture, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 325–332.

Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroeder, M. and Vinciarelli, A. (2011), *Social signal processing: The research agenda*, Springer, pp. 511–538.

Pereira, A., Smith, L. and Yu, C. (2014), 'A bottom-up view of toddler word learning', *Psychonomic Bulletin & Review* **21**(1), 178–185.

Peters, C., Asteriadis, S. and Karpouzis, K. (2010), 'Investigating shared attention with a virtual agent using a gaze-based interface', *Journal on Multimodal User Interfaces* **3**(1), 119–130.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M. and Poggi, I. (2005), A model of attention and interest using gaze behavior, *in* 'Proceedings of the Intelligent Virtual Agents', pp. 229–240.

Potrus, D. (2017), Swedish sign language skills training and assessment, Master's thesis, KTH Royal Institute of Technology.

Quine, W., Churchland, P. and Føllesdal, D. (2013), *Word and Object*, MIT press.

Quintero, C., Fomena, R., Shademan, A., Wolleb, N., Dick, T. and Jagersand, M. (2013), Sepo: Selecting by pointing as an intuitive human-robot command interface, *in* 'Proceedings of the IEEE International Conference on Robotics and Automation', pp. 1166–1171.

Rabiner, L. (1989), 'A tutorial on hidden markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**(2), 257–286.

Rabiner, L. and Juang, B. (1993), *Fundamentals of Speech Recognition*, Prentice-Hall.

Räsänen, O. and Rasilo, H. (2015), 'A joint model of word segmentation and meaning acquisition through cross-situational learning', *Psychological Review* **122**, 792–829.

Rautaray, S. and Agrawal, A. (2015), 'Vision based hand gesture recognition for human computer interaction: A survey', *Artificial Intelligence Review* **43**(1), 1–54.

Ren, J., Hu, Y., Tai, Y., Wang, C., Xu, L., Sun, W. and Yan, Q. (2016), Look, listen and learn — a multimodal LSTM for speaker identification, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', pp. 3581–3587.

Richmond, V., McCroskey, J. and Hickson, M. (2012), *Nonverbal Behavior in Interpersonal Relations*, Pearson.

Roy, D. and Pentland, A. (2002), 'Learning words from sights and sounds: A computational model', *Cognitive Science* **26**(1), 113–146.

Salvi, G., Montesano, L., Bernardino, A. and Santos-Victor, J. (2012), 'Language bootstrapping: Learning word meanings from perception-action association', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **42**(3), 660–671.

San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L., Fernández, F., Ferreiros, J., Lucas, J., Macías-Guarasa, J., Montero, J. and Pardo, J. (2008), 'Speech to sign language translation system for spanish', *Speech Communication* **50**(11), 1009–1020.

Sauppé, A. and Mutlu, B. (2014), Robot deictics: How gesture and context shape referential communication, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 342–349.

Shockley, K., Richardson, D. and Dale, R. (2009), 'Conversation and coordinative structures', *Topics in Cognitive Science* **1**(2), 305–319.

SIGNSPEAK (2012), 'http://www.signspeak.eu'.

Sorostinean, M., Ferland, F., Dang, T. and Tapus, A. (2014), Motion-oriented attention for a social gaze robot behavior, *in* 'Proceedings of the International Conference on Social Robotics', pp. 310–319.

SSL (2009), 'http://www.ling.su.se/english/research/research-projects/sign-language'.

Stanton, C. and Stevens, C. (2014), Robot pressure: The impact of robot eye gaze and lifelike bodily movements upon decision-making and trust, *in* 'Proceedings of the International Conference on Social Robotics', pp. 330–339.

Starner, T., Weaver, J. and Pentland, A. (1998), 'Real-time american sign language recognition using desk and wearable computer based video', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(12), 1371–1375.

Staudte, M. and Crocker, M. (2009), Visual attention in spoken human-robot interaction, *in* 'Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction', pp. 77–84.

Steinberg, A., Sullivan, V. and Loew, R. (1998), 'Cultural and linguistic barriers to mental health service access: The deaf consumer's perspective', *American Journal of Psychiatry* **155**(7), 982–984.

Stokoe, W. (1980), 'Sign language structure', *Annual Review of Anthropology* **9**(1), 365–390.

Tapus, A., Matarić, M. and Scassellati, B. (2007), 'Socially assistive robotics', *IEEE Robotics Automation Magazine* **14**(1), 35–42.

Thomaz, A., Hoffman, G. and Cakmak, M. (2016), 'Computational human-robot interaction', *Foundations and Trends® in Robotics* **4**(2-3), 105–223.

Van den Bergh, M., Carton, D., de Nijs, R., Mitsou, N., Landsiedel, C., Kühnlenz, K., Wollherr, D., van Gool, L. and Buss, M. (2011), Real-time 3d hand gesture interaction with a robot for understanding directions from humans, *in* 'Proceedings of the IEEE International Conference on Robot and Human Interactive Communication', pp. 357–362.

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

Vinciarelli, A., Pantic, M. and Bourlard, H. (2009), 'Social signal processing: Survey of an emerging domain', *Image and Vision Computing* **27**(12), 1743–1759.

Vogler, C. and Metaxas, D. (1997), Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods, *in* 'Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics', pp. 156–161.

Vogler, C. and Metaxas, D. (1999), Parallel hidden markov models for american sign language recognition, *in* 'Proceedings of the International Conference on Computer Vision', pp. 116–122.

Waldron, M. and Kim, S. (1995), 'Isolated asl sign recognition system for deaf persons', *IEEE Transactions on Rehabilitation Engineering* **3**(3), 261–271.

Walker, H., Hall, W. and Hurst, J. (1990), *Clinical Methods: The History, Physical, and Laboratory Examinations*, Butterworth Publishers.

Windsor, J. and Fristoe, M. (1991), 'Key word signing: Perceived and acoustic differences between signed and spoken narratives', *Journal of Speech, Language, and Hearing Research* **34**(2), 260–268.

Yamato, J., Ohya, J. and Ishii, K. (1992), Recognizing human action in time-sequential images using hidden markov model, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 379–385.

Yamazaki, A., Yamazaki, K., Kuno, Y., Burdelski, M., Kawashima, M. and Kuzuoka, H. (2008), Precision timing in human-robot interaction: Coordination of head movement and utterance, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', pp. 131–140.

Yang, M., Ahuja, N. and Tabb, M. (2002), 'Extraction of 2d motion trajectories and its application to hand gesture recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(8), 1061–1074.

Yu, C. and Ballard, D. (2004), 'A multimodal learning interface for grounding spoken language in sensory perceptions', *ACM Transactions on Applied Perception* **1**(1), 57–80.

Yu, C. and Smith, L. (2016), 'The social origins of sustained attention in one-year-old human infants', *Current Biology* **26**(9), 1235–1240.

Yurovsky, D., Smith, L. and Yu, C. (2013), 'Statistical word learning at scale: The baby's view is better', *Developmental Science* **16**(6), 959–966.

Zhang, C., Yin, P., Rui, Y., Cutler, R., Viola, P., Sun, X., Pinto, N. and Zhang, Z. (2008), 'Boosting-based multimodal speaker detection for distributed meetings', *IEEE Transactions on Multimedia* **10**(8), 1541–1552.

Zhang, Y., Beskow, J. and Kjellström, H. (2017), Look but don't stare: Mutual gaze interaction in social robots, *in* 'Proceedings of the International Conference on Social Robotics', pp. 556–566.