Understanding virtual speakers

JENS NIRME
COGNITIVE SCIENCE | LUND UNIVERSITY

Understanding virtual speakers

Jens Nirme

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Humanities and Theology, Lund University, Sweden.
To be defended at LUX, room C126, 21 February 2020, at 10:15.

Faculty opponent
Catherine Pelachaud
Université Pierre et Marie Curie (UPMC)

Organization
LUND UNIVERSITY

Document name: Doctoral dissertation

Date of issue:

Author(s) Jens Nirme

Sponsoring organization: CCL

Title: Understanding virtual speakers

Abstract

This thesis addresses how verbal comprehension is affected by seeing the speaker and in particular when the speaker is an animated *virtual speaker*. Two people visually co-present – one talking and the other listening, trying to comprehend what is said – is a central and critical scenario whether one is interested in human cognition, communication or learning.

Papers I & II are focused on the effect on comprehension of seeing a virtual speaker displaying *visual speech cues* (lip and head movements accompanying speech). The results presented indicate a positive effect in the presence of background babble noise but no effect in its absence. The results presented in paper II also indicate that the effect of seeing the virtual speaker is at least as effective as seeing a real speaker, that the exploitation of visual speech cues by a virtual speaker may require some adaptation but is not affected by subjective perception of the virtual speakers' social traits.

Papers III & IV focus on the effect of the temporal coordination of speech and gesture on memory encoding of speech, and the feasibility of a novel methodology to address this question. The objective of the methodology is the precise manipulating of individual gestures within naturalistic speech and gesture sequences recorded by motion capture and reproduced by virtual speakers. Results in paper III indicate that such temporal manipulations can be realized without subjective perception of the animation as unnatural as long as the shifted (manipulated) gestural movements temporally overlap with some speech (not pause or hesitation). Results of paper IV were that words accompanied by associated gestures in their original synchrony or gestures arriving earlier were more likely to be recalled. This mirrors the temporal coordination patterns that are common in natural speech-gesture production.

Paper V explores how factual topics are comprehended and approached metacognitively when presented in different media, including a video of an animated virtual speaker with synthesized speech. They study made use of an interface where differences in information transience and navigation options are minimized between the media. Results indicate improved comprehension and a somewhat stronger tendency to repeat material when also seeing, compared to only listening to, the virtual speaker. Instances of navigation behaviours were, however, overall scarce and only tentative conclusions could be drawn regarding differences in metacognitive approaches between media.

Paper VI presents a virtual replication of a choice blindness experimental paradigm. The results show that the level of detail of the presentation of a virtual environment and a speaker may affect self-reported presence as well the level of trust exhibited towards the speaker.

The relevance of these findings is discussed with regards to how comprehension is affected by visible speakers in general and virtual speakers specifically, as well as possible consequences for the design and implementation of virtual speakers in educational applications and as research instruments.

Key words Verbal Comprehension, Multimodality, Audiovisual integration, Gesture, Educational Technology

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language English

ISSN 1101-8453 Lund University Cognitive Series 177

ISBN 978-91-88899-84-3

Recipient's notes

Number of pages 180

Price

Security classification

Signature

Date 2020-01-10

Understanding virtual speakers

Jens Nirme

LUND
UNIVERSITY

Coverimage by Jens Nirme

Faculty of Humanities and Theology
Department of Philosophy
Cognitive Science

ISBN 978-91-88899-84-2
ISSN 1101-8453

Printed in Sweden by Media-Tryck, Lund University
Lund 2020

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is an environmentally
certified and ISO 14001:2015 certified
provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

To Milo and Otto

Acknowledgements

To start with, I want to extend my gratitude to everyone who has had a hand in making this thesis a reality.

- To Agneta Gulz, Magnus Haake and Marianne Gullberg for their excellent supervision and infinite patience.
- To all my colleagues in the Educational Technology Group (ETG) and the rest of Lund University Cognitive Science (LUCS), the Lund University Humanities Lab and fellow members of the Cognition Communication and Learning research environment (CCL), for inspiration, support and feedback, and not least for providing a welcoming and highly enjoyable work environment!
- To all my collaborators on the papers.
- To the IT unit and the Humanities Lab for providing the infrastructure and support needed for my work.
- To the student collaborators who have assisted with data collection: Stina Christoffersson, Johanna Pihl and Oscar Borring.
- To Pernilla Kile for checking lip sync in animations and to Tobias Kastberg for help with cleaning up audio recordings.
- To Joel Parthemore for proofreading.
- To my dear parents and Cris for making it possible to finish this project while raising two stubborn little creatures.
- Finally, to all my friends and to Galina for keeping me sane.

Table of Contents

List of original papers

Paper I

Nirme, J., Haake, M., Lyberg Åhlander, V., Brännström, J., & Sahlén, B. (2019). A virtual speaker in noisy classroom conditions: supporting or disrupting children's listening comprehension?. *Logopedics Phoniatrics Vocology*, *44*(2), 79-86.

Paper II

Nirme, J., Sahlén, B., Åhlander, V. L., Brännström, J., & Haake, M. (2020). Audio-visual speech comprehension in noise with real and virtual speakers. *Speech Communication*, *116*, 44-55.

Paper III

Nirme, J., Haake, M., Gulz, A., & Gullberg, M. (2019). Motion capture-based animated characters for the study of speech–gesture integration. *Behavior Research Methods*, 1-16.

Paper IV

Nirme, J., Haake, M., Gulz, A., & Gullberg, M. (Unpublished manuscript). *Early or synchronized gestures facilitate recall of speech*.

Paper V

Nirme, J., Fredriksson, O., Haake, M., & Gulz, A. (2020) Exploring students' approach to factual texts in different presentation media. *Lund University Cognitive Studies*, *176*. Retrievable from: https://www.lucs.lu.se/publications/lund-university-cognitive-studies/

Paper VI

Lingonblad, M., Londos, L., Nilsson, A., Boman, E., Nirme, J., & Haake, M. (2015, August). Virtual blindness – A choice blindness experiment with a virtual experimenter. In: Brinkman WP., Broekens J., Heylen D. (eds.), *Intelligent Virtual Agents* (*IVA 2015): LNCS*, *vol. 9238* (pp. 442-451). Springer, Cham.

My contributions to the papers

Papers I-II

I assisted with the recording of the audio stimuli used in both studies and was alone responsible for the simultaneous capture of video and 3D-data, as well as subsequent processing and generation of the visual stimuli. The experimental design of the studies was determined in collaboration with the other authors. Data collection and hearing screening of participants was performed by students enrolled in the Master program at the Div. of Logopedics, Phoniatrics and Vocology, Dept. of Clinical Sciences, Lund University. Data management and data analyses, as well as the main part of writing and revisioning, was performed by me.

Papers III-IV

I planned, recruited speakers for and performed the recording sessions that formed the basis of the stimuli used in both studies. All stimuli creation, including processing of audio and 3D-data, animation, rendering and video editing was done by me. Selection of experimental items and experimental design was done in collaboration with the other authors. Data collection and data management (including development of a platform enabling blind coding of responses) was performed by me. Coding of responses was done by me and one other (naïve) coder. Statistical analyses, as well as the main part of writing and revisioning, was performed by me.

Paper V

The texts used for the stimuli were produced by the second author. The audio-visual stimuli based on these texts, as well as the experimental platform and navigation interface were created by me. Design of experiment and questionnaires were done in collaboration between all authors. Data collection was performed by the two first authors (including myself). Statistical analyses, as well as the main part of writing and revisioning, was performed by me.

Paper VI

Conference paper based on a student project. I had a supervisory role in formulation of research questions, experimental design, development of experimental platform and data collection. I was alone responsible for audio- and 3D-data recordings used for the stimuli, as well as post-processing and animation. The paper was written in collaboration between me, the first and the last author, and presented at the conference by me.

Other work by the author

Anikin, A., Nirme, J., Alomari, S., Bonnevier, J., & Haake, M. (2015). Compensation for a large gesture-speech asynchrony in instructional videos. In *Gesture and Speech in Interaction (GESPIN 4)* (pp. 19-23).

Ballester, B. R., Nirme, J., Duarte, E., Cuxart, A., Rodriguez, S., Verschure, P., & Duff, A. (2015). The visual amplification of goal-oriented movements counteracts acquired non-use in hemiparetic stroke patients. *Journal of neuroengineering and rehabilitation*, *12*(1), 50.

Ballester, B. R., Nirme, J., Camacho, I., Duarte, E., Rodríguez, S., Cuxart, A., ... & Verschure, P. F. (2017). Domiciliary VR-based therapy for functional recovery and cortical reorganization: randomized controlled trial in participants at the chronic stage post stroke. *JMIR serious games*, *5*(3), e15.

Nirme, J., & Garde, H. (2017). Computational camera placement optimization improves motion capture data quality. In *International Conference on Multimodal Communication: Developing New theories and Methods*. Abstract retrievable from: https://sites.google.com/a/case.edu/icmc2017/abstracts

Rudner, M., Lyberg-Åhlander, V., Brännström, J., Nirme, J., Pichora-Fuller, M. K., & Sahlén, B. (2018). Listening comprehension and listening effort in the primary school classroom. *Frontiers in psychology*, *9*.

Rudner, M., Lyberg-Åhlander, V., Brännström, J., Nirme, J., Pichora-Fuller, M. K., & Sahlén, B. (2018). Effects of background noise, talker's voice, and speechreading on speech understanding by primary school children in simulated classroom listening situations. *The Journal of the Acoustical Society of America*, *144*(3), 1976-1976.

Silvervarg, A., Kirkegaard, C., Nirme, J., Haake, M., & Gulz, A. (2014, August). Steps towards a challenging teachable agent. In *International Conference on Intelligent Virtual Agents* (pp. 410-419). Springer, Cham.

Preface

Why this and why me?

The introductory course for PhD students at the faculties of Humanities and Theology – in the incarnation of the course I happened to take – relied heavily on senior researchers talking about their own work. One of the questions they had to answer was, *in what way is your research important?* As many of my non-researcher friends said, why are you doing this?

I hope to be able – in the following chapters and included papers – to make a case why the research field to which I have contributed is important to the world. However, I would like to start with the question, why is this interesting to *me*?

In one of the early presentations of the introductory course, a senior researcher made a comment that stuck with me. She said that, though she could see several benefits and even practical applications for her research, what really drove her was curiosity: her personal interest in the research questions she was addressing. The frankness of her answer resonated with me.

So, why am I interested in understanding "virtual" speakers?

The topic was not an obvious fit for me. Like many other researchers, I have not ended up working in my area of *first choice*. That said, the best advice my mother ever gave me – sometime during my angry youth – was that life is not, and should not be, a straight line.

I had been working in a research group at another university on a more applied project: developing an interactive digital platform for stroke rehabilitation. This was the sort of work I had sought out after I returned to studies and got my master's degree in my late twenties. However, even though applied research was rewarding in that I got to see the direct impact of my work – in my case, improvement on some of the patients participating in our studies – it was missing something. I was missing something. I wanted to ask questions and test hypotheses where there was no clearly preferred outcome. Perhaps you are developing what you hope will be the best solution to a problem, testing against a control missing a specific function or a benchmark.

I was confused to say the least, wanting both to do research that would allow me to address questions I happened to find interesting and real-world problems.

This is where I found myself when I returned to my native Sweden, moving to Skåne and back into the Swedish academic life I had not been part of for the previous eight years. One of the first persons I contacted was Agneta Gulz, the director of the Educational Technology Group (ETG), who would become my supervisor. She expressed a clear awareness of the trade-offs involved and an appreciation for my desire to seize both horns of my dilemma.

We had many common interests: virtual environments, learning environments, identifying learners' needs and adapting to them…. I knew something about sensorimotor learning from my work with stroke patients. She had ideas how sensorimotor learning could be combined with the more language-dependent learning addressed by ETG's educational applications. My interest was sparked.

I am fascinated by how people understand each other: how they are able to understand each other. I am not what you would call a *people person.* I like people, but sometimes I have trouble making myself understood. Perhaps I am what Malcolm Gladwell describes (2019) as "mismatched". I am terrible at remembering people's names and faces. However, I often remember what they said and have a fairly vivid image of them saying it, along with their surroundings. The shared representation of the world in the moment of interaction, as words take on meaning, lies at the heart of my fascination.

I also love computers and all they can do. Papert writes (1980, p viii): *"The computer is the Proteus of machines. Its essence is its universality, its power to simulate. Because it can take on a thousand forms and can serve a thousand functions, it can appeal to a thousand tastes."*

I was not particularly interested in computers or technology growing up. I fell in love with computers when I started seeing them as a tool to do whatever you might want, so long as you can imagine the problem in enough detail and tell the computer exactly what steps to take. Sergi Jordà, a teacher in my master's program, captured this for me in recounting his journey from experimental musician to interactive installation artist to researcher and inventor of interactive artifacts. He described it as being like overseeing the construction of the Great Pyramids by an army of workers blindly following his orders, faster than he could think. I love computers the most when they seem to come alive or capture some vital essence of being alive.

When I finally got accepted as a PhD student after a couple of failed project proposals, it was within Lund University's Cognition, Communication and Learning (CCL) cross-faculty research environment bringing together researchers from cognitive science, psychology, linguistics, and neurophysiology, along with logopedics, phoniatrics, and audiology. In 2008, CCL was awarded ten years' funding from the Swedish Research Council, as part of its Linnaeus program.

I was welcomed into this environment with open arms. Over the years I received plenty of good advice and feedback. My work has involved collaborations with linguists, speech pathologists, audiologists, and practicing teachers. As a cognitive scientist, I have taken a deliberately multidisciplinary approach. The determining factor for planning my studies has been finding a connection to some practical application, whether that be students listening to a teacher in a classroom, developing expressive digital characters as research instruments, teachers supporting students with special needs, and so on. I have actively sought out people with deep knowledge about how to listen under adverse, classroom or classroom-like conditions. I have sought out others with knowledge about the interplay of speech and gesture, or with knowledge about supporting reading comprehension within special education. I am grateful for the welcoming research environment at Lund University generally and within CCL in particular for allowing me these opportunities. It can be humbling, sometimes even overwhelming, to collaborate across disciplinary lines with those having different knowledge, different skillsets and different points of view. In retrospect, it was exactly what I needed!

Introduction

I have studied how verbal comprehension is affected by seeing the speaker, particularly when the speaker is an animated *virtual speaker*. The situation with two people, visually co-present – one talking, the other listening, trying to comprehend what is said – is critical whether one is interested in human cognition, communication, or learning.

In the section *Definitions* I clarify what I mean by *comprehension* and *virtual speaker*, along with two kinds of visual information available to listeners observing a speaker: *visual speech cues* and *gesture*.

Verbal comprehension is complex and depends on many factors involving the speaker, the listener, and the setting. Much is known about the impact that visual information has on verbal comprehension, but there is still much to discover. It remains an active field of research implicating several disciplines, as will become apparent from the range of work referenced in this introduction and the following papers.

Why is it interesting and relevant to focus on *virtual* speakers? The answer is two-fold. First, by doing so one can address research questions related to verbal comprehension between "real" people in ways not previously possible. Second, one can gain better understanding of how human beings comprehend the virtual speakers themselves – virtual speakers that are becoming ubiquitous in daily life. It is important to figure out if and how one understands virtual speakers differently from real speakers, to better guide expectations, attitudes and strategies for designing virtual agents – not least in educational applications, where they are already widespread. The section *Motivations* makes the case for studying the comprehension of virtual speakers with respect to these two goals.

By means of virtual speakers, my research has primarily addressed the following broad questions:

> What can studies where participants listen to a virtual speaker tell us about how listeners comprehend real speakers?
>
> How does comprehending a virtual speaker differ from comprehending a real speaker?

Successfully answering these questions may also inform the design of virtual speakers in real-world applications, including educational ones.

Again, verbal comprehension is complex, drawing together a diverse class of research topics. This thesis focuses on a small but important subset. I have used virtual speakers as a tool to study possible ways that seeing a speaker affects listener comprehension. One is how the visual cues supporting audio-visual integration of speech influence comprehension under different circumstances, such as in a primarily auditory environment with several other speakers overheard in the background. Another is how seeing gestures that are coordinated with speech seems to help listeners remember what they hear, influenced by the timing of the gestures. I have also explored some factors that relate to verbal comprehension more indirectly such as how seeing a virtual speaker in a video affects the metacognitive strategies a listener uses and how trust in a virtual speaker and the listener's perception of the virtual speaker's social traits may affect comprehension.

The section *Summaries of included papers* gives a brief overview of each paper, with emphasis on what research questions I addressed, why I addressed them, and what my findings were. The papers are divided into three groups according to primary focus: *Visual speech cues, Gestures,* and *Indirect effects.* (Of course, some secondary research questions and findings overlap, meaning that the same study might well be discussed in more than one paper.)

In three sections titled *Visual speech cues*, G*estures* and M*etacognitive and social effects*, I discuss the relevance of my findings – both more obvious and more speculative interpretations – in relation to previous research. The first two of these three sections discuss speech comprehension in general as well comprehension of virtual speakers specifically.

The section titled *Design guidelines* summarizes my findings in the form of guidelines for the design of virtual speakers. Such agents are meant to include both those that promote verbal comprehension in educational and other practical software, and those intended primarily as research instruments.

The closing section *Outlook* discusses ways in which methodologies developed for my thesis work can be developed further to answer questions whose scope goes beyond that of this thesis.

Definitions

Virtual speakers

Several related terms refer to computer visualized, animated, more or less humanlike characters who are able to communicate in some way with a human being: *embodied conversational agent*, *virtual human, animated pedagogical agent*. What all have in common – their lowest common denominator – is that they involve a dynamic, visual rendering of a human or anthropomorphic agent, whose speech is presented either visually as text or vocally as audio.

I will use these related terms below to situate the notion of a *virtual speaker.*

First, there is the primary distinction between agents and avatars (even though lay persons often use *avatar* to refer to computer-based characters in general). Strictly speaking, an avatar is a character whose behavior is controlled by a human end-user – most often, the player of a game – via some input device. By contrast, virtual agents' behaviors are determined in software. In computer games, agents are often called *non-player characters* (NPCs), or *bots* if they are designed to imitate the behavior of human players. Hybrids (Chase, Chin, Oppezzo, & Schwartz, 2009) combine aspects of avatars and NPCs. Consider the characters in the game series *Sims*, where a player's actions indirectly determine agents' behavior. Consider as well the *teachable agents* (Blair, Schwartz, Biswas, & Leelawong, 2007; Silvervarg, Kirkegaard, Nirme, Haake, & Gulz, 2014) in many an educational software, where the student takes the role of teacher to instruct a student of her own (the teachable agent); the behavior of the teachable agent when taking tests or otherwise solving problems reflects what (and how well) the "real" student has taught it.

Cassell (et al*.,* 2000, p.2) defines *embodied conversational agents* as:

> … interfaces that have bodies and know how to use them for conversation, interfaces that realize conversational behaviors as a function of the demands of dialogue and also as a function of emotion, personality, and social convention.

Virtual Humans are defined by Swartout et al. (2006, p. 96) as*:*

> … software artifacts that look like, act like, and interact with humans but exist in virtual environments.

These both definitions emphasize interactivity as well as the fact that the relevant behavior is produced by algorithms modeling aspects of human behavior. Algorithms generate movements procedurally, often in real time and in response to interaction with a human user. An agent's movements may also be predetermined using manually key-framed animations, high-level scripting of general movement schemas (e.g. Kopp et al., 2006), or motion capture based on the movements of a human actor. Any one of these can be used on its own, in combination with one another, or in combination with algorithms.

In the educational field, the term *animated pedagogical agent* is often used (Johnson, Rickel & Lester, 2000). Their presentation can be more or less detailed: anything from static images and written text to life-like speech and full-body animation. Animated pedagogical agents can vary in visual appearance (e.g., naturalistic vs stylized: Haake, 2009), use of visual text vs. speech, and movement expressivity (including lip movements, facial expressions, gestures, postural animation and gaze behavior). Levels of interactivity range from largely autonomous in response to user actions, to all but fully non-autonomous, delivering predefined content in a predefined order. Other, non-visual aspects may also be emphasized, such as pedagogical role (Haake & Gulz, 2009), with consequent consideration for experts, motivators, mentors (Baylor & Kim, 205), peers, and teachable agents (Brophy et al., 1999).

The term *virtual speaker*[1] emphasizes that the agent's main purpose is to speak (as opposed to displaying text visually: e.g., in a speech bubble), as reflected in how the agent is animated. For my purposes, it was important to avoid the term "agent" as much as possible. The virtual speakers involved in the following studies of this thesis are non-interactive: i.e., their speech and movement are predetermined; they do not require and cannot respond to input from their human listener. They are thus not, properly speaking, agents at all.

Consider Figure 1, which sets out a two-dimensional space based on level of presentation detail and interactivity. The two-dimensional space can usefully be divided into four quadrants. In the top right quadrant we find virtual speakers that are both rich in presentation detail and interaction capabilities, exemplified by the *BabyX* agent created by Soul Machines, that is naturalistic both in terms of appearance and animation while being able to learn new words by responding to a

[1] Although throughout this introduction I consistently use the term "virtual speakers", in the following papers I refer to specific implementations as "virtual speakers" (papers I and II), "digitally animated speakers/characters" (papers III, IV and V) or "virtual experimenters" (papers VI) depending on the target journals and audience.

combination of visual input and natural language (Johnsson, 2017). In the bottom right quadrant we find *Anna*, IKEA's customer service agent that has a static and simplistic presentation but is capable of text-based natural language interaction, inferring user's intentions within the specialized domain of furniture shopping (Shah & Pavlika, 2005). In the bottom-left quadrant we find *Dr. Bob*, the cartoon-like figure that was simplistically animated to point out elements of presentation slides at certain points during prerecorded audio narrations in a study by Dunsworth and Atkinson (2007). The research presented in this thesis stays mostly within the top-left quadrant (minimum interactivity and maximum presentation detail), exemplified by a virtual speaker utilized in papers III and IV that reproduces speech as well as postural, gestural and facial animation recorded from a real speaker. The exception is the *virtual experimenter* in paper VI, that responded with somewhat different speech animation segments depending on some choices the participants make by voice commands (implemented by a *Wizard of Oz* method; Guindon, Shuldberg & Connor, 1987).

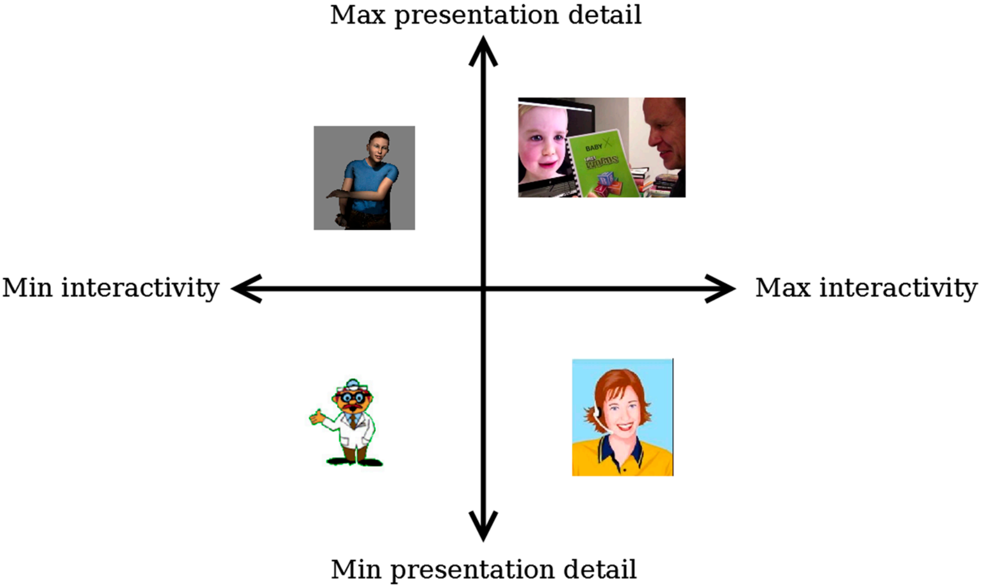

Figure 1. Examples of virtual speakers distributed within a two-dimensional space based on level of presentation detail and interactivity.

Comprehension

By *comprehension*, I mean to refer specifically to when someone extracts meaning from spoken language and applies it to a subsequent task. Papers I, II, and V use *comprehension* in an even narrower sense, as the ability to answer questions related to spoken content correctly.

I assume that comprehension involves several subprocesses. One is speech recognition: the perception and distinguishing of words in a speech signal. Another is memory encoding: representing the form and meaning of spoken content, as reflected in the recall of words in paper IV.

These are obviously not the only subprocesses that may be involved in comprehension. A full account would depend on the specific nature of the material and task at hand.

Visual information available to listeners

Typical human beings with no visual impairment rely heavily on their visual sense. Almost 50% of the brain is involved directly or indirectly in processing visual information (Marieb & Hoehn, 2007). When seeing someone speak, one has a great deal of visual information immediately available. Integration of speech and visual information takes place in many stages on many levels of processing. I chose to focus on two sources of information: visual speech cues and gestures. (I have thus excluded *inter alia* information on speakers' emotional states as seen through facial expressions and posture.) Examples of each are offered in Figure 2, using individual frames from the virtual speaker used in Paper II.

Visual speech cues

Visual speech cues include visible movements of the lips – and, to some degree, tongue and teeth (Saitoh, Morishita & Konishi, 2008) – synchronized with speech articulation, along with movements of the eyebrows and head synchronized with prosodic peaks. The former movements are referred to as *visemes,* commonly defined as by Bear and Harvey (2017): "a set of phonemes which have identical appearance". Such a definition implies a one-to-many relationships between visemes and phonemes. The exact nature of the mapping and the range of acceptable forms visemes may take vary between cultures and languages (Saitoh, Morishita & Konishi, 2008) and even among individual speakers (Cox, Harvey, Lan, Newman, & Theobald, 2008). Taking into consideration *coarticulation effects* – how something is pronounced depends on what precedes or follows it –

the mapping between visemes and phonemes might better be described as many-to-many (Mattheyses, Latacz & Verhelst, 2013). Yet out of all visual speech cues, lip movements are especially tightly linked with speech. They can be highly informative for speech recognition, particularly under noisy conditions (Grant & Seitz, 2000; Sumby & Pollack, 1954).

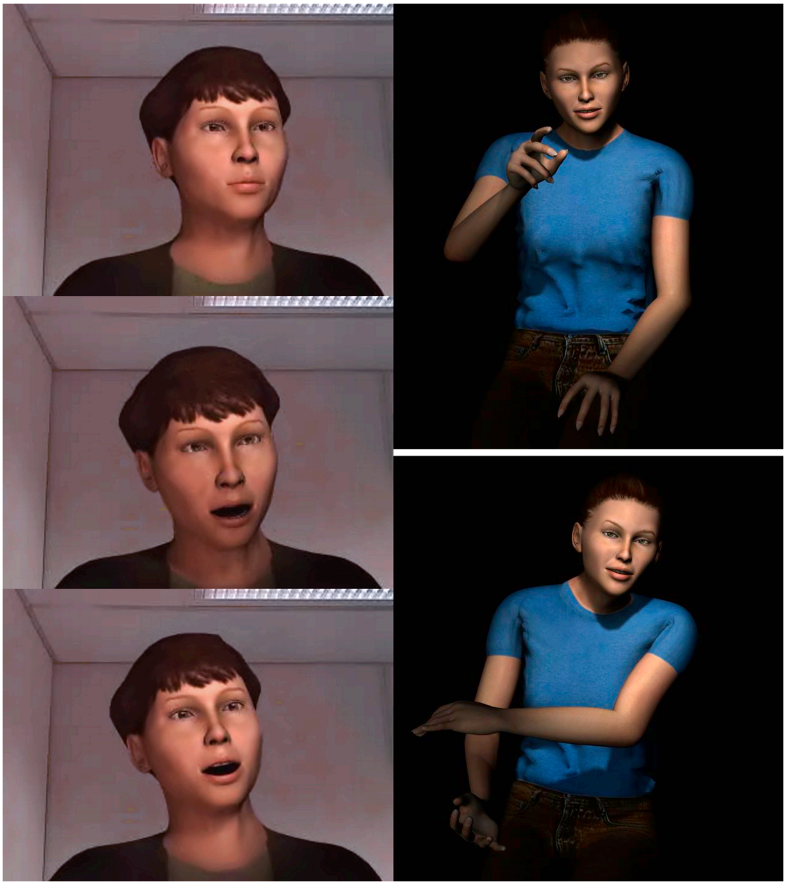

Figure 2. Virtual speakers exhibiting visual speech cues (left) and gestures (right).

Head and eyebrow movements related to speech prosody may be referred to as *visual prosody* (Munhall, Jones, Callan, Kuratate & Vatikiotis-Bateson, 2004). Studies have shown visual prosody aids speech recognition (Munhall et al., 2004). It also helps listeners perceive which words are being emphasized (Swerts & Krahmer, 2008). While speaking invariably involves visible movement of the mouth – unless one is a ventriloquist – it is possible to speak without moving the head or making any facial expressions. It just is not commonly done and, indeed, the link with speech prosody has been shown to be both strong and regular. So for example Cavé et al. (1996) found that eyebrow movements and variations in speech frequency are strongly (though not perfectly) correlated. Visual prosody interacts with expressions of emotional state (so-called *expressive modes*) to encourage a perception of stress or make informative key words more prominent (Beskow, Granström, & House, 2006).

Gesture

The other source of information available to listeners that I have focused on is gesture. More specifically, I have studied co-speech gestures: i.e., gestures that are performed together with speech, with the exception of so-called *emblems* like the conventionalized *thumbs up* and *OK* hand signs (Kendon, 2004). Gesture's relationship to speech and meaning is more complicated than that of visual speech cues. There is no established equivalent of grammatical rules, although attempts have been made (Müller, 2017; Schlenker & Chemla, 2018). McNeill and Duncan write (2000), "… the gestures we analyze are 'idiosyncratic' in the sense that they are not held to standards of good form; instead they are created locally by speakers while they are speaking".

McNeill (2008, pp. 38-42) classifies gestures into four classes according to information expressed, with awareness that the categories can and do overlap. *Iconic gestures* capture features of the concrete actions or objects they are meant to evoke. *Metaphoric gestures* typically evoke linguistic metaphors, such as rejecting a proposed dinner engagement by (literally) brushing it off. *Deictic gestures* point out entities either physically present or (in the case of *abstract deictic gestures*) imagined at a specific location in the environment. *Beat gestures* are simple rhythmic movements that do not express meaning on their own but synchronize with and can be used to emphasize speech.

It is generally difficult to classify what a speaker expresses with an individual gesture as clearly one thing or the other. Consider the gesture shown at top right in Figure 2, associated with the words *"[the cartoon cat] sees [the cartoon bird]*". It is *iconic* by virtue of the way it traces the line of sight from the bird to the cat and *deictic* by virtue of pointing out the bird's location.

Gestures that are iconic, metaphorical or deictic can all be considered *representational gestures* (McNeill, 1992). This super category excludes beats, which do not carry information the same way, along with pragmatic gestures whose function is in one way or another to regulate the conversation (Kendon, 2017a). In a similar fashion, Clark (2016, p. 342) describes what he calls *depicting* (e.g., iconic gesture) and *indicating* (e.g., deictic gesture) as *"basic methods of communication".*

The movements involved in gesture can be divided into phases: what Kendon (2004) calls *preparations*, *strokes*, *holds* and *retractions*, of which the expressive stroke is arguably the most important. It is the one that the method described in Paper III and results obtained in Paper IV revolve around. Strokes can involve one or both hands, positioned dynamically or statically (as when one hand provides a reference frame: see Figure 2, bottom right). Although most often described in terms of what is happening with the hands, strokes can and often do involve movement of the entire body!

I have limited myself to studying comprehension of co-speech gesture as one-way communication, with fixed roles for speaker and listener(s): i.e., I have not considered how listeners may interact with a speaker. In papers III and IV, the virtual speakers reported on were animated using motion capture to catch the speech and movements, including gestures, of real speakers who were instructed to address a person rather than the camera. The specific gestures included in the stimulus material could all be classified as representative.

Motivations

There is more than one good reason to study how virtual speakers are understood. This chapter will elaborate these reasons in relation to what research can say about how human listeners comprehend both human and virtual speakers.

Face-to-face: The fundamental context for comprehension

The idea that language should be studied as a multimodal phenomenon, abandoning the traditional focus on speech or written text, has grown and continues to grow in influence (see for example Perniss, 2018). Bavelas and Chovil (2000) posit that face-to-face dialog, combining audible and visible elements, is the primary context for language use. We are accustomed to see a speaker that we want to understand.Most children who are not visually impaired begin to learn to understand the world around them via face-to-face interaction with a parent or other caretaker (Yu & Smith, 2016; Yurovsky et al., 2011). The same can be said about how children come to understand language specifically. Kuhl (2007) proposes that language learning is "gated" by social interaction, including joint visual attention between children and their caregivers.

Several accounts of language evolution stress the importance of perceiving the bodily actions of others, particularly those involving the hands. Donald (1993) sees the human ability to mimic others' movement and deliberately trigger action sequences without an external cue (such as an object to act on) as an evolutionary breakthrough and a critical step on the way to developing language. Arbib (2012) proposes that the neural *mirror system* – so-called *mirror neurons* activated when performing *and* when perceiving others performing actions – supports understanding intentions and actions. This in turn supports imitation and can be extended to support pantomime. With pantomime, reduced forms of actions with hands, face, and voice provide the basis for a protolanguage: i.e., unitary utterances or *holophrases* depicting complete actions get reduced into the building blocks necessary for symbolic language. Some make the claim speech co-evolved with communicative gestures (Kendon, 2017b; Levinson & Holler, 2014) or even that gestural communication is a precursor to spoken language (Gentilucci &

Corballis, 2006; Tomasello, 2008), although the later position has lost some popularity.

Theories on the evolution of language and knowledge of the tight link between gestures and speech are difficult to reconcile with early theories in cognitive science that viewed language as a brain module working on amodal, symbolic representations independent from interactions with the world (Chomsky, 1975; Fodor, 1983). The multimodal context of language and, in particular, the role of gestures are better explained by situated and embodied theories of language, which stress how one's body and sensorimotor capabilities shape cognition (Barsalou 1999; Glenberg; 1997; Zwaan, 2004).

Virtual speakers as research tools

The idea that visual information plays an important role in verbal comprehension is not controversial and has generated plenty of excellent research. Technological progress has made available tools for generating, collecting and managing visual data – such as 3D-engines, motion capture and eye-tracking – which have become increasingly accessible and affordable. The parallel technological development and increasing research interest, is comparable to how the adaptation of video technology to studying young children's language acquisition triggered increased interest in co-speech actions and gestures (Kendon, 2007).

Virtual speakers are another tool in the research toolbox, valuable not least for their configurability. Researchers can manipulate features of interest in a systematic way while keeping other features constant, in ways they cannot do with real speakers. A real speaker cannot reproduce the same voice quality, intonation, facial *micro-expressions* (Ekman & Friesen, 1969) or involuntary eye-movements over multiple experimental sessions, or identical gestures and visual speech cues for that matter. The biases that human listeners can have towards speakers of different ethnicities or other groups can also more readily be controlled for using virtual speakers.

Listeners understand virtual speakers based on their experience with real speakers. Therefore, how virtual speakers are understood or misunderstood has plenty to say about how people understand speech in general. Speech comprehension is a complex topic; virtual speakers provide the means to start peeling the different components of speech comprehension apart by focusing in on specific aspects of behavior and presentation. Some aspects can be carefully re-created, others exaggerated, others simplified, yet others muted or left out entirely. Consider Rosenblum, Johnson and Saldaña's (1996) study where they presented a speaker's face as a set of points; or Alghamdi, Maddock, Barker and Brown's (2017) study, where they used exaggerated lip movements.

The virtual speakers used in the studies in this thesis present varying degrees of visual naturalism. Nevertheless, the general strategy has been to come as close as possible to how real speakers sound and move, controlling for subjective factors related to their animation (see paper III) or appearance (see paper II). I implemented the virtual speakers with the help of motion capture and simultaneous voice recording. Paper V constitutes an exception, as it includes a condition wherein both speaker voice and movements were synthesized – with the purpose to bring things closer to a feasible real-world application to support reading comprehension.

Of course there are differences between a virtual and a real speaker and differences between listening to a virtual and a real speaker. For all the efforts at naturalism, there is never any doubt, in the studies presented in this thesis, that what is visually presented is a digital creation and not a recording of a real speaker. Even though I make the case in the following papers that the findings have relevance for how real speakers are understood, I take care to discuss how understanding virtual and real speakers may differ. Indeed, exploring the differences is a recurring theme and a key motivation to my work. When real are replaced by virtual speakers, it is vital to know what is liable to be lost or gained, and to consider what is most important to get right. Paper II directly addresses differences in how a virtual speaker is comprehended compared to a real speaker presented in a video recording. Paper III addresses speech and gesture synchrony as one aspect of how speakers are perceived, using a task implying that the depicted movements are real or virtual to varying degrees.

Another recurring theme in this thesis – resurfacing in many forms throughout the discussion sections of the included papers – is that reductionist approaches to speech processing, focusing on sub-processes such as speech recognition in isolation to one another, risk losing ecological validity, producing results that are not generalizable to the real world. Balancing experimental control with ecological validity poses a pervasive challenge to cognitive science, as for any field studying human behavior (Araújo, Davids, & Passos, 2007; Brunswik, 1957). Paper III has a particularly strong focus on methodology, describing a distinctive approach to studying speech and gesture processing, along with its potential implications. Paradoxically, the artificial nature of the stimuli makes it possible to go beyond a narrow focus on disparate phenomena towards the study how speech is understood in the real world, as it is possible to precisely alter specific aspects of rich and naturalistic behavior.

Listeners perspective

This thesis focuses on taking the listener's perspective. One important limitation to the experimental tasks used in the following studies is that the communicative tasks were not, for the most part, interactive: i.e., there was no two-way communication. In each case, the roles of listener and virtual speaker were fixed, with no turn taking. Only in one of the studies (Paper VI) did the human participant and the virtual speaker in any way interact. Given the central role of interaction to communication (Bavelas & Chovil, 2000; Clark, 1996; De Ruiter et al., 2010), the relative lack of interaction is significant. That said, the research questions addressed by that study were not concerned with interaction but strictly with how the virtual speaker's speech was evaluated.

Human communication also relies on a social context that shapes cognition, communication and learning, for example by establishment of *rapport* between two interlocutors (Bernieri, 1988). As with interaction, discussions of social context lie outside the scope of this thesis. It is however possible to trigger social schemas by presenting learners with non-interactive and even minimally expressive characters (Lester et al., 1997; Sjödén, Tärning, Pareto, & Gulz, 2011). There are situations where learners must perceive, understand, and learn from speech without being able to interact; university lectures often take this form.

Understanding virtual speakers specifically

Technological artifacts have started talking to us. Their voices may be recorded from a human voice or synthesized. Children and adults alike learn from interacting with applications and artefacts that speak – or even respond when spoken to.

People understand speech, natural or artificial, on several levels. They are able to distinguish words whose meaning they are familiar with and sometimes infer the meaning of unfamiliar ones. This is such a basic and necessary skill that people tend to forget just how complex comprehension is. It requires distinguishing a speaker's voice from any other voices or background sounds; disambiguating words and recognizing sentence structure; extracting semantic meaning while keeping track of what was said a minute ago; relating the extracted meaning to what is already known. All this must happen in real time, in delicate synchrony, often under suboptimal conditions because people are busy, tired, or distracted. In the following papers I discuss cognitive skills that I argue are vital to successful verbal comprehension: directed attention, suppression of distracting stimuli, memory encoding, adaptation to each new speaker, and metacognitive self-regulation.

This thesis is about understanding virtual speakers in more than one sense. How do listeners understand what a virtual speaker says, and how do they understand the nature of the virtual speaker herself? With regards to the first sense, I have already argued for the position that the primary mode for producing and interpreting spoken language is in face-to-face interaction where speakers are both seen and heard. Understanding of virtual speakers will be shaped by past experience and present expectations based on interactions with real people speaking to or with us.

People tend ascribe intentions and traits to a speaker, be it real person or programmed artifact. Reeves and Nass' (1996) *media equation* ('media equals real life') captures in a pithy phrase how people approach artefactual media, even simple text-based instructions (Nass, Moon, Fogg, Reeves, & Dryer, 1995), *as if* they were interacting with a real person – despite knowing full well they are not. How people understand a message has a lot to do with how it is presented, and who they imagine is presenting it. Reeves and Nass (1996, p. 183) write:

> When participants were asked to assess the credibility of news on television, they were influenced the most by the credibility of the news anchor, even though they knew that his stories were written and researched by several other people.

This thesis contributes to the existing body of knowledge regarding how speakers are perceived and understood, with particular focus on digitally rendered (animated) speakers. When a virtual speaker is presented this way, listeners are able to make use of the available visual information - be it a minimally animated stick figure, a detailed three-dimension human model with naturalistic appearance and behavior, or anything in between.

Virtual speakers are becoming more and more prevalent in people's everyday lives. People interact with and listen to virtual speakers for entertainment (video games) and practical purposes (automated customer support, public services, learning purposes in educational and training applications). Virtual speakers act as proxies for real speakers who are not physically co-present (avatars in online virtual environments). They may be better suited than real speakers for certain situations. A Swedish company recently started providing "unbiased recruitment" by conducting all interviews via *Tengai*: an artificially intelligent robotic head with facial animation (Savage, 2019). Virtual speakers have also been used for treating social phobia (Klinger et al, 2005). They are sometimes used in educational contexts because they do not suffer from fatigue, or from the consequences of the mistakes that learners might make. For example, language learners can practice their verbal commination skills with a partner who has infinite patience (Bédi et al., 2016) and medical students can train their clinical diagnostic skills by interviewing virtual patients who will survive being misdiagnosed (Sia, Halan, Lok & Crary, 2016).

Another motivation for studying how virtual speakers are understood is more hypothetical. Clark & Chalmers' (1998) Extended Mind Hypothesis states that cognition spans not only what goes on inside our heads, but also artifacts in the environment that we interact with. If this is true, increasing exposure to virtual speakers and environments will change human cognition and, in turn, effect how people understand speech in the real world.

As virtual speakers become more prevalent, it is important to learn as much as possible about how they impact understanding and their potential broader effects on cognition, communication, and learning. These questions are relevant not only for researchers interested in the cognitive processes behind speech processing, communication, and learning; but also for application developers – particularly, with respect to this thesis, developers of educational applications. I have limited my scope to what goes on while listening, with the hope that this can provide detailed insights into how comprehension works and how information from different sensory modalities gets integrated. Although conceived as a pure research project with no direct applications, my thesis has yielded results meant to inform the design of embodied virtual speakers in educational software.

Summaries of papers

Visual speech cues

Papers I and II focus on the impact of visual speech cues on comprehension. It is well known that visual speech cues support speech *recognition* in the midst of background noise. Their role in speech *comprehension* – a more complex task that requires processing information on several levels – is less clear. Visual speech cues may actually increase perceptual load under what are already challenging listening conditions: a particular concern for virtual speakers, whose listeners are less familiar with them compared to real speakers.

Paper I. A virtual speaker in noisy classroom conditions: Supporting or disrupting children's listening comprehension?

School children listened to narratives selected from a validated verbal comprehension test with or without naturalistic background noise (babble), with or without seeing a virtual speaker using naturalistic visual speech cues. The results suggested that seeing the virtual speaker might support comprehension under noisy conditions, but a statistical test showed the effect to be non-significant. However, more careful re-analysis post publication revealed a significant effect. The effect was weak compared to the strongly significant main effect of background noise, as well as compared to previous research into the effects of visual speech cues on speech recognition. The paper discusses several possible explanations, including that the appearance of the virtual speaker or participants' unfamiliarity with her made her less effective as speech-comprehension support compared to a real speaker. Seeing the virtual speaker without noise produced no effect.

Paper II. Listening comprehension of real and virtual speakers

This paper can be regarded as a follow-up to Paper I using a larger sample. Each participant listened to one of three narratives under one of the conditions: a video of a real speaker, a video of a virtual speaker whose movements match the real speaker, and an audio-only recording. The order of the narratives was varied and balanced across conditions, so as to detect possible adaptation effects. All conditions were presented with background noise (babble). The re-designed study included measures of self-reported effort and speaker's perceived social traits. Improved comprehension from seeing a virtual speaker was confirmed and shown to be at least as strong as with the video of the real speaker. The benefit of the virtual speaker depended on some adaptation and only appeared with the second narrative to which participants listened. Interestingly, no significant adaptation effect was observed for the video of the real speaker even though its animation and viewpoint matched the video with the virtual speaker. Also, participants selected more negative words to describe the virtual speaker than the audio-only speaker. However, even if the virtual speakers were perceived more negatively, this attitude did not seem to interfere with the improvement in comprehension.

Gestures

Papers III and IV focus on gesture and how, in coordination with speech, gestures affect the way a virtual speaker is perceived and how her speech is encoded in memory. Most theoretical accounts of gesture agree on a close link between speech and gesture, evidenced by a temporal coordination whereby gesture strokes precede or coincide with the pronunciation of associated words. Comprehension of gesture is less studied, partly due to methodological challenges. How does one create precisely controlled experimental stimuli that still correspond to naturally occurring speech and gesture? One open question I addressed is whether gestures affect how listeners encode speech when not overtly engaging in a listening task. Another concerns whether listeners' processing of speech and gesture depends on the aforementioned temporal coordination: i.e., if preceding or coinciding gestures serve a communicative purpose. Yet another concerns the feasibility of using virtual speakers with configurable gesture as a research tool.

Paper III. Motion-capture-based animated characters for the study of speech-gesture integration

The researchers developed a workflow for creating stimuli with virtual speakers and manipulating gestures within naturalistic sequences based on motion-capture (MOCAP) recordings. A validation experiment revealed that participants did not recognize unnaturally timed gestures as unnatural, so long as they coincided with speech and did not occur during pauses. Introspective ratings indicated that, in the latter cases, participants might have become explicitly aware that something was wrong with the virtual speakers' hand gestures. Overall, the results demonstrated the usefulness of the method and core idea: to precisely manipulate certain parameters in naturalistic contexts to test implicit effects of speech and gesture, in this case their temporal coordination. The pros and cons of the methodology and how it can be developed for future studies are discussed.

Paper IV. Synchronized gestures facilitate recall of associated words

Expanding on the stimulus developed for Paper III, Paper IV reports a follow-up study demonstrating that temporal coordination has an effect on processing words associated with gesture. Both eliminated gestures and gestures that arrived late relative to the associated words – something that is rarely seen in natural speech – made the associated words less likely to be recalled. This finding is in marked contrast to the results reported in paper III. No difference was found between gesture that came early – often seen in natural speech – and gesture with “correct” timing. Overall, the results indicate that listeners are tuned to the natural coordination of speech and gesture, and that timing has an effect on memory encoding.

Indirect effects

Papers I-IV demonstrate that virtual speakers – like real speakers – provide direct, real-time visual cues that can be exploited for comprehension.

There can also be more indirect comprehension effects coming from how a speaker is perceived by the listener. For example, with regards to how trust in the speaker can influence the listeners’ evaluation of what the speaker says. Another regards whether and how seeing or not seeing a speaker influences the choice of metacognitive approaches in a listener.

Both virtual speakers and synthesized speech narrating simultaneous visually presented text are used to scaffold independent reading, most often for students who are poor readers. However, empirical research has yet to demonstrate any positive effects conclusively. Direct comparisons are confounded by a simple observation: the independent reader has a whole page of (printed or displayed) text continuously available, which lightens working-memory load and facilitates self-pacing, in contrast to someone listening to a speaker deliver material in a linear and time-dependent fashion. An open question remains whether students’ comprehension of information presented across various media – including by virtual speaker – is determined more by the informational content of the media or more by how the students approach the media.

Previous research has shown that, in many ways, people react to and interact with digital artifacts, in accordance with accepted social schema. Virtual speakers, whose appearance and behavior can be controlled precisely, have untapped potential as stimuli or instructors in behavioral experiments. That said, it is crucial first to understand how the quality of their implementations might affect their reception.

Paper V. Exploring different students' approach to factual texts in different presentation media

A study with secondary school children explored differences in the reception of factual texts delivered in different media, including a video of an animated virtual speaker with synthesized speech. Navigational interface, possibilities for repetition, and informational transience were designed to be as similar as possible across the different media. The virtual speaker improved comprehension compared to a disembodied synthesized voice (in this case without background noise). The effect can be explained by the synthesized voice constituting an adverse listening condition comparable to the background noise. An alternative explanation relies on metacognition, observing that participants watching a virtual speaker were somewhat likelier to go back and repeat content compared to the speech-only group. However, one cannot draw any hard conclusions regarding differences in metacognitive strategies, since there was overall (in any media) very little repetition and non-linear navigation despite the interface allowing for it.

Paper VI: Virtual blindness: A choice blindness experiment with a virtual experimenter

A study with adults investigated how the quality of presentation of a virtual speaker in the role of experimenter affected trust in the information provided, with the assumption that trust is reflected in detection rate within a choice-blindness paradigm. A *high-quality* condition had detailed 3D models and textures as well as naturalistic facial animation and speech compared to a *low-quality* condition.

Pilot-study participants completed a questionnaire that measured their *presence*: a subjective measure of a person’s experience of being in a virtual environment. Results revealed that the low- and high-quality speakers resulted in correspondingly low and high self-reported presence.

The main study used as its outcome variable participants’ detection of manipulations within a choice-blindness paradigm where the inconsistency lay in the alternative choices presented by the experimenter pre- and post-decision. The low-quality presentation resulted in faster detection, which suggests that the degree of realism affects the resulting trust in the virtual speaker. One can speculate that the effect was mediated by differences in the listeners’ expectations for behavioral consistency or social response to the virtual speaker. An alternative explanation is simply that participants were distracted by the richer visual stimuli.

Visual speech cues

Introduction

This section is focused primarily on the role of *visual speech cues* that are closely related to speech production (see *Definitions*). I discuss my experimental findings and general observations in relation to previous theoretical and empirical work about how and when visual speech cues facilitate speech comprehension. I consider both direct and more speculative interpretations of the available evidence.

To the simple question *"*does seeing the speaker have any effect on the processing and comprehension of speech?" the answer is "yes". Compared to simply hearing speech with one's eyes closed, a listener who sees a speaker will, by definition, have more information at hand. Seeing a speaker offers visual information that is clearly related to speech, even if not crucially informative. Comparative studies have shown positive effects on speech recognition from having the speech presented by a speaker who is physically present (Sumby & Pollack, 1954), delivered by video recording (Ma, Ross, Foxe & Parra, 2009), or virtual (Agelfors et al., 2009). Likewise, studies have found positive effects of speaker visibility on comprehension (Dunsworth & Atkinson, 2007; see also papers II and IV).

However, given that comprehension is complex and multifaceted, it is not always possible to pinpoint which visual cues are helpful, how listeners exploit them, or under what conditions they are effective. For example Mishra, Lunner, Stenfelt, Rönnberg & Rudner, 2013, reported improved recall scores but a greater working memory load with a visible speaker in a verbal memory task. There are too many factors at play, relating to the environment, the listener's perceptual and cognitive capacities, and the nature of the listening task. Given this complexity, it is hardly surprising no consensus exists around any model of comprehension, particularly once multimodality is taken into account. Despite a multitude of studies, it remains unclear when it is or is not beneficial toward comprehension to see the speaker.

The same uncertainty seems to apply to listener's spontaneous behavior, as evidenced by a recurrent observation of mine during the pilot studies for the experiments described in this thesis. Both school children and adults sometimes close their eyes or avert their gaze from the speaker when confronted with a task that requires careful listening. To avoid confounding results, participants in the studies in papers I-V were explicitly instructed to keep looking at the virtual speakers. One participant was excluded from the study in Paper IV for failing to comply. (The study in paper VI involved a more interactive and explicitly visual task, rendering explicit instructions unnecessary.) Previous research has shown that gaze behavior adapts strategically depending on the cognitive resources required by a task (Droll & Hayhoe, 2007). Listeners' gaze behavior adapts to the listening conditions, at least in some cases (Buchan, Paré & Munhall, 2008). Gaze also plays a role in memory encoding and retrieval (Johansson, Holsanova, Dewhurst & Holmqvist, 2012; Johansson & Johansson, 2014). In the cases where my pilot-study participants spontaneously looked away, one could argue that they adopted a suboptimal strategy given my experimental results, which show that they actually stood to benefit from seeing the virtual speaker.

Visual speech cues and real speakers

Under ideal conditions, speakers are clearly audible and demands on listeners' comprehension manageable. Visual speech cues are arguably not then required. In other scenarios, they usefully compensate for lost information. Mattys, Davis, Bradlow and Scott (2012) reviewed studies testing the impact of various factors on speech recognition, comprehension, working memory, and attentional load. They classified factors contributing to adverse conditions as stemming either from *source degradation* (e.g., a hoarse or unnatural voice), *environmental degradation* (e.g., reverb, background noise, or visual distraction), or *receiver limitations (e.g.,* hearing impairments or reduced working memory or attentional capacity). The Framework for Understanding Effortful Listening (FUEL, Pichora-Fuller et al., 2016) offers a model whereby task demands interact with listener motivation and effort to determine comprehension outcomes.

One example of an adverse condition based on an environmental factor is found in Paper I. The visibility of the virtual speaker – with naturalistic facial animation – had no effect on comprehension *unless* the speech was presented together with background noise (babble). Paper II confirmed the positive effect of seeing the virtual speaker in the background noise condition.

This result is not obvious, even though a positive effect of visual speech cues on speech recognition in noisy conditions has (as said) been well established. The problem is that comprehension is more complex than recognition and involves a greater range of listener cues: perceptual, syntactic, and semantic. Comprehension relies throughout on a combination of top-down and bottom-up effects, even though research findings indicate that visual speech cues are already integrated at early, pre-attentional and pre-lexical processing stages (Haxby, Hoffman & Gobbini, 2002; Soto-Faraco, Navarra & Alsius, 2004). Peelle and Sommers (2015) propose that audiovisual integration of speech takes place on many levels in many processing stages. Tye-Murray, Sommers and Spehar (2007) show that audio (phoneme) and visual (viseme) similarity, as well as their mutual coincidence, provide contextual cues to a word recognition task. Rosenblum (2008) compiles evidence from a few studies showing that audiovisual speech integration can be influenced top-down, by the semantic context or if a word is a real or a non-sense word. Baart, Stekelenburg and Vroomen (2014) report that visual speech cues (visemes) only affect phoneme recognition in a generated sine-wave approximation of speech when participants were expecting to hear actual speech – indicating another top-down influence on integration.

The study presented in Paper V found a positive effect for seeing the virtual speaker compared to only hearing it. There was no background noise of any kind, but another factor may have made the listening conditions challenging: the speech was presented via a synthetized voice. Voice synthesis has been linked to reduced speech recognition and comprehension (Drager, Reichle & Pinkoski, 2010; Winters & Pisoni, 2006). Another possibility is that the factual texts used in Paper V are more challenging than the narratives of papers I and II, introducing novel concepts and new vocabulary. Both these possibilities would be examples of what Mattys et al. (2012) would consider *receiver limitations*: the limited perceptual and cognitive capacities of humans in general or of different populations or individuals in particular, in relation to the demands – *perceptual* or *cognitive load* – of listening tasks. Working memory and executive functioning have been linked to how well listeners deal with such demands and how likely they are to integrate and exploit visual speech cues successfully under such conditions (Jansen, Chaparro, Downs, Palmer & Keebler, 2013; Picou, Ricketts & Hornsby, 2011).

Given that listeners have limited perceptual and cognitive capacities, the processing of richer sensory information comes at a cost. Fraser, Gagné, Alepins, and Dubois (2010) found improved and less effortful speech recognition in noisy conditions when listeners could see the (human) speaker's face (audiovisual condition) compared to when they could not (audio-only condition). However, when noise levels were adjusted so that recognition accuracy was the same with or without seeing the speaker, the researchers found greater effort and longer reaction times in the audiovisual condition. Integration of visual speech cues does not come for free and may sometimes require top-down directed attention (Talsma, Senkowski, Soto-Faraco & Woldorff, 2010), even as audiovisual integration on some levels might be best modelled as bottom-up sensory driven. The well-known ventriloquist effect whereby synchronized lip movement shifts the apparent source of speech does not require directed attention (Bertelson, Vroomen, De Gelder & Driver, 2000; Vroomen, Bertelson & De Gelder, 2001). Meanwhile, bottom-up driven attention to competing audiovisual speech can interfere with listening to a primary speaker (Senkowski, Saint-Amour, Gruber & Foxe, 2008). The well-known McGurk Effect – incongruent visemes modulate auditory perception of articulated phonemes (McGurk & MacDonald, 1976) – has been demonstrated to depend on visual attention (Andersen, Tiippana, Laarni, Kojo & Sams, 2009). The question arises whether inhibiting unattended, potentially distracting audiovisual speech is more or less demanding than inhibiting audio-only speech. One finds surprisingly little research on the topic; however some evidence (Cohen & Gordon-Salant, 2017) suggests that competing speech presented audiovisually is more distracting than audio-only in a speech recognition task.

Visual speech cues and virtual speakers

Virtual speakers whose facial animation matches speech with different degrees of accuracy and detail are common in entertainment and educational applications alike. The synchronization of speech and animation makes users attribute the speech as emanating from the visual representation of the speaker. As noted before though, visual speech cues might in fact be detrimental to comprehension under non-adverse circumstances. Sweller (2005) offers the Redundancy Principle stating that "redundant material interferes with rather than facilitates learning". Sweller, Ayres and Kalyuga (2011, p.144) clarify:

> In contrast, if the two sources of information can be understood in isolation, only one source, either the audio or the visual source should be used. If both are used, one source will be redundant and having to process both will lead to an extraneous cognitive load.

It should be noted that these researchers come from the field of so-called *multimedia learning*, where overlapping or complementary information is presented through images, animations, text or audio rather than face-to-face, (Virtual speakers are sometimes involved.) Craig, Gholson and Driscoll (2002) found that when a virtual speaker was presented in a multimedia-learning environment, redundantly presenting both text and speech impeded learning compared to speech only. That finding is line with Mayer's *modality effect* for multimedia learning: when competing visual stimuli are present, speech is better presented as audio (Ginns, 2005; Mayer, 2001; Moreno & Mayer, 2002). Although inconclusive, results from multimedia learning studies seem to indicate that visual speech cues do not have the same influence on comprehension and learning as does redundant text presentation.

Dunsworth and Atkinson (2007) directly compared conditions with or without an animated speaker narrating a presentation of the cardiovascular system. They found that the virtual speaker had a positive effect on learning, in contrast to Mayer, Dow and Mayer (2003). Clark & Choi (2005) suggest that, since animated virtual speakers exhibit visual information that is nonessential to speech processing – such as facial expressions and purely cosmetic features of visual appearance – they risk increasing the cognitive load in a learning situation. Veletsianos, Heller, Overmyer and Procter (2010) acknowledge the problem but insist that a naturalistic presentation of visual appearance, animation, and voice can circumvent it.

No consensus exists on when virtual speakers do or do not aid understanding. Valid arguments can be made that seeing virtual speakers might both distract from, or highlight, relevant information. On one hand, virtual speakers can be designed to *cut out the fat*: i.e., it is possible to remove superfluous details of visual behavior while exaggerating crucial details such as lip movement (Alghamdi, Maddock, Barker & Brown, 2017). On the other hand, virtual speakers as stimuli are less familiar than real speakers, and discrepancies in their appearance and behavior (Kätsyri, Förger, Mäkäräinen, & Takala, 2015) might be off-putting or distracting for listeners.

That said, listeners exploiting visual speech cues from a virtual speaker provides the most straightforward explanation of the positive effect on speech comprehension amid background noise reported in papers I and II. The positive effect – compared to listening without visible speaker – persisted even when the virtual speaker was perceived negatively, suggesting that integration of visual speech cues is unaffected by listeners' subjective experience of the speaker. It is however problematic to generalize from this case to other virtual speakers, since the virtual speaker's movements were particularly well matched to a real speaker, being obtained by motion capture during the voice recordings. Self-reports from Paper II's participants offer no indication that the virtual speaker was any more or less distracting than the real speaker.

Exploiting unfamiliar audiovisual speech cues may require some adaptation or "getting used to" on the listener's part, as evidenced in Paper II, where the benefit to comprehension of seeing the virtual speaker in a noisy environment only "kicked in" after the second narrative. Adaption to novel visual speech stimuli for speech recognition is well documented (Alghamdi, Maddock, Barker & Brown, 2017; Rosenblum, Johnson & Saldaña, 1996). Adaptation can be a demanding process, with capabilities differing across populations (Tye-Murray, Spehar, Myerson, Sommers & Hale, 2011). Even if listeners can interpret and adapt to a range of nonrealistic visual speech cues, there is a limit to when cues are useful! The McGurk effect is one example where irregular lip movements lead listeners astray. Paper IV reports a benefit from integrating gesture information with speech, but for that to work, the gestures should conform to the temporal patterns found in natural production.

This brings up another issue: it can be difficult and time consuming to animate virtual speakers, however it is done. That said, new technologies facilitating motion capture or automatic generation of naturalistic movements can facilitate this work (Ginosar et al., 2019). In particular, it is necessary when animating to balance any exaggerated visual speech cues against naturalism and coherence.

Gesture

Introduction

One research area that is critical for this thesis is how gesture contributes to spoken language communication. In this section I first explain why I, as a researcher, am especially interested in gestures. Next I outline the dominant view of spoken language and gesture as intimately linked, at least when produced by a speaker, and how timing is a key factor in understanding this link. Finally I briefly discuss the usefulness and issues involved with using virtual speakers who gesture as research tools and as actors in educational applications.

Gesture studies is a multi-disciplinary research field with contributions from such diverse fields as anthropology, linguistics, psychology, history, neuroscience, communication, art history, performance studies, computer science, music, theater, and even dance (ISGS, 2019). The aim is to understand how gestures are produced and understood. That most gesture researchers have some connection to linguistics is hardly surprising given the tight link gesture and spoken language – as well as the conceptual overlap with sign language.

I find gestures fascinating because they are in a way hidden in plain sight. People perform lots of gestures while speaking (Kendon, 2004), and listeners use gestures for comprehension. At the same time, it is as if neither speakers nor listeners are much aware of them. Eye-tracking studies reveal that listeners mostly look at speakers' faces and only fixate on gestures in specific situations, like when a hand is held in the end position of the gesture stroke or speakers look at their own gestures (Beattie, Webster & Ross, 2010; Gullberg & Holmqvist, 2006). The typical association people make when I tell them that I study gesture is to emblematic gesture, which is far from the most frequent (see *Definitions*). Stimulus material used in papers III and IV was based on MOCAP recordings of speakers giving spontaneous descriptions of cartoons, physical scenes or objects. At the end of these recording sessions, the volunteer speaker had explained to her the intended use of the recordings: primarily to recreate and optionally manipulate their gestures using virtual speakers. Faced with this information, the volunteer speakers – in many cases – expressed regret for not having gestured, even though their gestures (on reviewing the recordings) were found to be plentiful! Indeed, at times it has been more challenging finding individual gestures with enough temporal separation from adjacent gestures to be useful as stimulus, than it has been finding gestures in the first place. Sometimes I myself have been surprised of the prevalence of gestures in the data myself, reviewing recordings where I had not been aware that much gesturing was going on. It seems that gesture is (almost) always there, wherever and whenever one speaks: even when speaking on the phone (Bavelas, Gerwing, Sutton & Prevost, 2008).

Besides catching my imagination, this mysterious nature of gesture might be part of the reason that other researchers take interest. They take gestures as clues that speakers inadvertently exhibit, revealing details of the cognitive processes involved in language production and comprehension that are not available to introspection.

Gesture and real speakers

Gesture is sometimes talked about as part of a person's *non-verbal behavior*. This usage is however controversial when it comes to co-speech gesture, given its close link to speech (McNeill, 1985). There is general agreement that gestures and speech are tightly linked, but there are competing proposals as to the stage at which, in the preparation of an utterance, they become coupled.

De Ruiter (2007) identifies three views: (1) gestures are windows into the mind, directly expressing thought in parallel with speech (see also Beattie, 2003), (2) gesture is shaped by how information is organized in the speaker's language (see also Kita & Özyürek, 2003), and (3) that gestures are "postcards of the mind" revealing a coordinated process whereby utterances are planned for gesture and speech simultaneously, sometimes with redundant and sometimes with complementary information. De Ruiter argues for the third view.

McNeill's Growth Point Theory (2008) also argues for early-stage coupling as the starting point for an idea to be expressed in speech via the categorization of visuospatial information into linguistic forms. In his view, spoken language and gesture are linked closely from their very conception. In contrast, Hostetter & Alibali (2019) with their theoretical framework of "gestures as simulated action" do not emphasize the link as strongly. Instead they propose that gestures arise from sensorimotor simulations involved in both speech and thinking.

Empirical evidence supporting a tight link partly comes from how speech and gestures are naturally performed in synchrony. Graziano & Gullberg (2018) survey the relevant research to show that "when speech stops, gesture stops" for both children and adults, regardless of whether one is using one's first or second language. As said, gestures not only coincide with speech but do so according to a pattern. That pattern led McNeill (1992, p.26) to formulate his Phonological

Synchrony Rule: *"gesture precedes or ends at, but does not follow, the phonological peak syllable of speech"*.

Gesture theories are mostly concerned with speech and gesture production more than the role of gesture in comprehension. Researchers disagree to some extent concerning the degree to which gestures are communicative (Krauss, Morrel-Samuels & Colasante, 1991). Still, plenty of evidence suggests that they are useful for speaker (for an overview, see Goldin-Meadow & Alibali, 2013) and listeners (for an overview, see Hostetter, 2011; Kendon, 1994). Relatively little work has been devoted directly to the role of temporal coordination in comprehension, and results are consequently inconclusive (Anikin, Nirme, Alomari, Bonnevier & Haake, 2015; Pruner, Popescu & Cook, 2016; Woodall & Burgoon, 1981).

The findings presented in Paper IV suggest that the information expressed by gesture and spoken language is closely linked, in the minds of listeners as they are in the minds of speakers. Paper IV investigates the effect on how strongly spoken words are encoded, as measured by subsequent recall, of synchronous vs. asynchronous gesture strokes. The latter either adhering (advanced gesture strokes) to or violating (delayed gesture strokes) the Phonological Synchrony Rule. The results show that recall was helped by seeing a virtual speaker perform gesture strokes synchronized with speech, as they would be naturally with a real speaker. This is one example how gesture serves a purpose for listeners, making spoken words more memorable. Delayed strokes resulted in worse recall compared to the synchronized condition, while no such effect was observed for advanced strokes. To make words more memorable, gestures do indeed need to adhere to the Phonological Synchrony Rule: more precisely, listeners attend to temporal coordination patterns in the natural production of gestures. Previous findings (Obermeier & Gunter, 2014) – using EEG measurements – have shown that listeners can integrate non-redundant information in early gestures, but they do so less automatically (i.e., with more effort).

A common critique of embodied theories of language comprehension concerns how abstract concepts are grounded, given that the connection to action and perception is less obvious (Dove, 2016). Lakoff & Johnson (2008) propose that abstract concepts are understood by concrete conceptualization; language development relies on the overgeneralization of spatial terms. Gesture can play an important role in grounding abstract concepts, since they occupy the intersection where concrete, sensorimotor representations meet abstract language representations.

Paper IV reports that the temporal patterns typical in natural speech and gesture production facilitate the effect of gesture on memory encoding of speech. One possible explanation is that gesture helps listeners put themselves in the mind of the speaker, without the need for clear expression of the abstract concepts the speaker has in mind. For a virtual speaker, much may depend on how naturalistically gestures are performed. As mentioned in *Definitions*, the forms that gestures take are often idiosyncratic, complicating matters. The video stimulus material in Swedish that was used in the study included a phrase that can be translated to the fairly abstract "only one remaining" and the accompanying (synchronized) gesture placed this last remaining entity in space. This information serves no obvious purpose for the task at hand, to remember the utterance including that there was only one remaining of something. Still, seeing the synchronized gesture had a measurable effect on recall, which is difficult to pinpoint to anything more specific that it added some richness to the listeners understanding of what the speaker meant, or how much importance was given to this particular part of the utterance by the speaker. Any interpretation of this observation is, however, highly speculative.

Gesture and virtual speakers

One reason why accounts of gesture generally focus on speakers' production over listeners' comprehension is that comprehension is more difficult to study. Speakers can be observed speaking. One cannot see what goes on in the listener's mind.

Comprehension is hard to study experimentally, given the difficulty of designing stimulus material. Traditional methods for producing stimuli generally fall into one of two categories: (1) actors are asked to alter their gestures while speaking; (2) edited video recordings of a speaker where the video- and/or audio- tracks have been manipulated, to alter how gestures are combined with speech. The latter often requires masking out or otherwise concealing the speaker's face so as not to reveal manipulations. Single gestures are often presented in isolation with little else going on visually. Given that – as I noted earlier – gestures are seldom fixated upon or otherwise given explicit attention, it is questionable how well traditional methods capture the way comprehension works in the ordinary world.

Paper III briefly overviews how virtual speakers ('digitally animated speakers' in the paper) have been used in gesture studies. With the support of my supervisors, I decided that virtual speakers animated using motion-capture technology would be the best way to test for more implicit effects of gestures in a more ecologically manner, at the same time as being able to manipulate individual gestures precisely.

Paper III describes the method I developed for producing these stimuli – also used in Paper IV's study – over the course of the first year of my PhD project. Both studies demonstrated that it was possible to make the temporal manipulations I and my fellow researchers wanted, without listeners perceiving the video stimuli[2] as less natural than non-manipulated segments. The one exception was when manipulated gesture strokes coincided with pauses or hesitations in speech: a rare occurrence in natural production (Graziano & Gullberg, 2018).

Together, the findings reported by papers III and IV indicate that seeing gestures can affect speech comprehension by making spoken content more memorable. This effect does not require explicit attention to gesture: listeners in these studies appeared not to perceive gestural asynchrony even though it influenced encoding.

One area where virtual speakers commonly perform gesture is in educational applications. Gesture is often used to point at information presented graphically or emphasize key words in speech (e.g., Craig, Gholson & Driscoll, 2002; Dunsworth & Atkinson, 2007). As mentioned in *Motivations*, it is possible to reduce superfluous visual detail with virtual speakers: e.g., by simplifying their movements. This can effectively limit redundant information and what Sweller (2005) calls "extraneous load". The risk is that one throws out the baby with the bathwater. As became obvious in creating my MOCAP-animated virtual speakers, gestures are seldom performed in isolation and often involve movements of more than just the hands and arms. Gestural movements combine seamlessly with other movements such as shifting weight on a chair, averting one's gaze, establishing eye contact with a listener, or movements related to visual prosody such as rhythmic head movements. This is in stark contrast to virtual speakers gesturing as experimental stimuli. Their gestures are often simplified into discrete movements, performed with isolated arm and hand movements. Happily, there are exceptions: platforms where features of gestures, visual speech cues, and other visually observable behaviors can be parameterized and combined (e.g. Boker et al., 2009; Xu, Pelachaud, & Marsella, 2014).

The stimuli presented in papers III and IV had reduced detail in the virtual speaker's hand and finger movements compared to the real movements that were its basis (as described in Paper III), but still had the effect of strengthening encoding. Indeed, it is possible to recognize human actions from only a few animated dots representing the movement of the main joints (Johansson, 1973). That said, there are situations where precise finger movements or hand configurations of may be important to gesture (e.g. Gullberg, 2010). Reduced detail may compromise a listener's overall experience. Lifelike appearance and behavior is often considered a determining factor driving engagement with educational material, through ascription of social agency to the virtual speaker (Clark & Mayer, 2016; Mayer & DaPra, 2012). That said, an analysis not reported in papers III or IV revealed no correlation (r = .16, n = 48, p = .26) between how

[2] These showed short speech segments of speech including at least two gestures, of which one may have been manipulated depending on experimental condition.

natural a virtual speaker seemed and how well her speech was remembered. The findings of Paper VI, indicating reduced trust in a less detailed, less naturalistic virtual speaker, point toward a different way that naturalism might matter.

Metacognitive and social effects

Introduction

The preceding sections offered examples of how the information in virtual speakers' visual speech cues, including gestures, can be exploited for verbal comprehension. Seeing a virtual speaker also has indirect effects on comprehension by triggering attitudes or behavior patterns in listeners. This section will discuss my findings in relation to how listeners perceive themselves and the listening situation.

Metacognitive strategies

There is a crucial difference between reading text and listening to someone speak. When text is printed (on paper or on a screen) it has a permanence that speech, by its time-dependent nature, does not. For words and ideas to be available for readers in the process of comprehending a text, they need not be maintained in working memory. They can be accessed at any time simply by remembering where to look for them, should long-term memory fail. Donald (1993) sees the *externalization of memory* as one of three cognitive transformations that shaped the *modern mind*.

In ordinary life, listeners can interact with speakers directly by asking for clarifications, repetitions, or follow-up answers; but the conversation cannot be paused or rewinded. For virtual speakers, who often have limited interactive capabilities but whose behavior is highly configurable, the inverse may well hold.

When investigating the effects virtual speakers have on comprehension, researchers face the problem that study participants seldom have the possibility to control their own pace or revisit things they failed to understand in the way they can when reading texts. The degree of which content is only temporarily available is called *information transience*. Empirical research has shown that information transience increases the load on working memory and can thus be detrimental to comprehension. These studies often demonstrate the effect by varying the length of segments that informational content is divided into (Dowell & Shmueli, 2008; Singh, Marcus & Ayres, 2012; Wong, Leahy, Marcus & Sweller, 2012). The differences in information transience make direct comparison between comprehension of material presented via text vs. virtual speaker difficult. Nevertheless, educational applications employing virtual speakers are often discussed as viable alternatives to textbooks.

The metacognitive strategies a listener or reader adopts to comprehend material relates to information transience. If a reader or listener chooses not to revisit material, the benefit of any information permanence is less obvious. To adopt an effective strategy requires self-regulation: in particular, being able to assess one's own comprehension or memory (Nelson, 1990; Metcalfe, 2009; Zimmerman & Moylan, 2009). As the Dunning-Kruger Effect so aptly demonstrates, this kind of meta-cognitive self-assessment is inherently difficult for people whose comprehension is sufficiently incomplete (Kruger & Dunning, 1999). Another challenging metacognitive skill is selecting the optimal learning media. A study by Salomon (1984) found that more school children chose a TV-program over printed text in a learning/comprehension task, even though the printed text generally resulted in better outcomes.

In the study presented in Paper V, the researchers wanted to explore whether metacognitive strategies for comprehension (specifically the tendency to revisit material) and comprehension itself were dependent on (1) possibilities to navigate materials by repeating or skipping ahead and (2) participants' preconceptions about the media (their *action possibilities* or *affordances* in Norman's terminology; Norman, 1988). Toward this end, we constructed an interface that minimized so far as possible the differences in these aspects between presentation formats: a text on its own, simultaneous text and speech, and speech on its own with or without a virtual speaker.

Not surprisingly, the results indicate a stronger tendency for repetition when reading compared to any condition that involved listening, to a disembodied voice or virtual speaker. This suggests that listening does not facilitate self-pacing the way reading does, even when the interface allows for it. The researchers observed a slightly stronger tendency for participants to repeat material if they were listening to and seeing a virtual speaker vs. listening to a disembodied voice. Methodological issues prevent drawing any firm conclusions however. The maximum permitted time to study each topic may have been too limited. The program logs revealed few attempts at navigation; participants who did repeat anything mostly only navigated back once. A follow-up study with less strict time constraints might help clarify the matter. Self-regulation has previously been correlated with successful learning in a similar age group (Pintrich & De Groot, 1990) – but self-pacing is not always ideal. Choosing optimal strategies demands sufficient metacognitive abilities, so some metacognitively challenged learners might end up making systematically bad choices. A recent study examined learners in a common age group using an application for learning about history, learners very often declined or ignored feedback that was directly linked to incorrect answers, perpetuating their poor learning styles (Tärning et al., in revision).

Another issue that could be addressed is how various presentation media affect listeners'/readers' representation of the material. Is that representation mostly reflective of the structure of the media or the meaning of the content?

Even though paper V did not reveal poorer comprehension with a virtual speaker compared to reading, the fact that different conditions triggered different navigation behaviors despite the navigation possibilities being virtually identical across conditions points to a potential limitation of virtual speakers in educational applications. Even if listeners' comprehension is incomplete, habitual schemas for how interacting with a speaker normally works may prevent listeners from revisiting material. Also, virtual speakers may give the illusion of being responsive to listener feedback despite not having this capacity.

This brings me to the next topic: how listeners perceive virtual speakers as social agents.

Social responses

The social constructivist Lev Vygotsky (1962) laid out a theory that stressed how learning should be and is a social activity. People learn by interacting with – and communicating with – others. In a formal educational setting, this means contact with teachers and fellow students (peers). One way to promote social learning is having students work together to solve tasks. Albert Bandura's (1977) Social Learning Theory emphasizes the importance of role modeling: that is, the observation and internalization of others' behaviors for learning.

Many educational technology applications are modeled on a single student interacting solely with the learning material, solving problems and taking tests on her own. The social context risks getting lost. *Massive open online courses* (MOOCs), are free, often open-access courses that let unlimited numbers of students take university level or university preparatory courses completely online. Given the requirements of scaling, no direct contact with teachers is possible. Material is often delivered by prerecorded video lecture. Although initially seen as a potential revolution in higher education, MOOCs have faced common problems. Very few students complete the courses, while those that do don't always reach the requisite level of knowledge for further university-level study (Pappano, 2012). One factor behind this is the difficulty of creating a social context without the physical presence of teachers and students.

Social presence describes the sensation of being with other people via different communication media. A meta-analysis by Richardson, Maeda, Lv, and Caskurlu (2017) shows that social presence has a strong positive correlation to students' perceived learning and satisfaction. Can virtual speakers can help create a social context that benefits learning? Agneta Gulz (2004, p. 3) thinks so, listing *"the fulfillment of a need for personal relationships in learning"* as one potential benefit of including virtual characters in educational applications, even as she finds inconclusive results in her review of the literature.

Lester et al. (1997) describes the so called *persona effect* whereby lifelike pedagogical agents help learners view their experiences more positively. Moreno, Mayer, Spires, and Lester's (2001) *social agency effect* similarly describes higher engagement and deeper learning achieved when material is presented by a virtual speaker, in their studies a cartoon bug. Any such effect may depend on learners picking up social cues from a virtual speaker, which is more likely to occur when the virtual speaker displays lifelike appearance and behavior: what Mayer and colleagues refer to as the *embodiment principle* (Clark & Mayer, 2016; Mayer & DaPra, 2012).

On the other hand, lifelike virtual speakers might conceivably be more disruptive, constituting unfamiliar stimuli relative to real speakers. The (controversial) Uncanny Valley Hypothesis states that artifacts approaching but not quite reaching the level of human naturalism create a sense of eeriness (Mori, 1970). Kätsyri, Förger, Mäkäräinen and Takala (2015) examined the body of empirical evidence and found support for what they termed a "perceptual mismatch" formulation of the hypothesis; namely, that negative affinity arises from inconsistency in the level of human-likeness between different aspects of visual appearance and behavior (see also Garau et al., 2003). So, consistent naturalism and expressive animation might be important factors for how listeners react to a virtual speaker.

Also, not all social effects are positive. Domagk (2010) found that virtual speakers exhibiting dislikable social cues via appearance or voice had a negative effect on learning. In contrast to Domagk's (2010) results, the virtual speaker in Paper II had a positive effect on comprehension compared to the audio-only condition – despite its social traits being perceived more negatively. It seems that integration of visual speech cues is or can be relatively unaffected by how social qualities are perceived. The visual speech cues used in Paper II's study were based on MOCAP recordings of a real speaker. It is possible that they had a stronger positive effect on comprehension than those used in Domagk's study, where the focus was on social cues.

While virtual speakers may appear somewhat alien, the face-to-face encounter is very familiar. Humans are hardwired to read faces. Visemes and facial expressions already vary widely among real speakers (Cox, Harvey, Lan, Newman, &

Theobald, 2008), and perception of virtual speakers may simply represent taking this tolerance of variation one step further.

The study presented in Paper VI showed that a higher quality virtual speaker and virtual environment – including more naturalistic facial animation and voice, along with more detailed 3D models and textures – resulted in *lower* ratings of presence, including social presence, as measured by post-trial questionnaire. The explicit experimental task participants had to perform (adapted from previous choice blindness studies; Johansson, Hall, Tärning, Sikström & Chater, 2014) consisted in selecting the more attractive person from two photos. This task is not directly related to comprehension or learning, so it is questionable what relevance this measure of presence would have to an educational setting.

Another component of the presence questionnaire reported in Paper VI is *involvement*. Witmer and Singer (1998) define involvement as "a psychological state experienced as a consequence of focusing one's energy and attention on a coherent set of stimuli or meaningfully related activities and events" – making it reminiscent of what Moreno et al. (2001) call *engagement*. The measure of involvement on the questionnaire showed the greatest difference between the two conditions (high- vs. low-quality speaker and environment).

The lower quality virtual speaker resulted in a faster detection of changes that had been made in choice of answers within a choice-blindness paradigm. One possible explanation is that in the lower quality condition, participants had fewer visual stimuli to be distracted by. It is possible they also had reduced expectations of consistency and trust. Other evidence suggests that learning effects related to social agency are not that dependent on visual naturalism or natural speech. The so-called *protégé effect* – students make more effort instructing a teachable agent than learning on their own – has been amply demonstrated with agents who are minimally animated and communicate via text (Chase et al., 2009; Sjödén, Tärning, Pareto & Gulz, 2011). However, aspects of learning go beyond basic comprehension and memory. Knowledge needs to be processed and put into a wider – possibly social - context. The optimal mode of presentation may depend on the virtual speaker's intended pedagogical role (Gulz & Haake, 2006).

To conclude, it is worth noting the potential of the paradigm described in Paper VI to be further developed. Deployment of virtual speakers as experimenters in similar behavioral experiments could provide an implicit measure of attitudes towards the speaker in a highly controlled setting.

Design guidelines

In the papers included in this thesis, I present findings addressing key research questions (see *Summary of included papers*). The findings contribute to (1) understanding how human listeners comprehend real speakers, (2) understanding how human listeners comprehend virtual speakers, and (3) understanding how virtual speakers can better be designed to promote comprehension in real-world applications. While the preceding sections – *Visual speech cues*, *Gestures,* and *Metacognitive and social effects* – have focused on the first two, it is time to address the third. Altogether, the studies show that:

- Visual speech cues can support speech comprehension under adverse listening conditions, such as with background noise (papers I & II) or synthesized speech (paper V).
- Listeners do not necessarily adopt optimal listening strategies: e.g., they may spontaneously choose to ignore a visual channel despite it being potentially helpful for comprehension (papers I, II & V) or memory (Paper IV) tasks. This was a general observation from the pilot studies.
- Aversion to a speaker experienced as unnatural may interfere with the effects researchers would hope to find from the visible presence of a virtual speaker; however, such aversion need not impair integration of visual speech cues (Paper II).
- Benefitting from visual speech cues may require that listeners first get familiar with a virtual speaker (papers I and II). Adaptation – familiarization with visual speech cues – can be relatively quick (Paper II) but is nevertheless worth taking into account when designing experiments and educational applications.
- Virtual speakers can be used to study the effects of gesture on listeners' comprehension without drawing undue attention to the gesture itself (Paper III).
- Adhering to the patterns of naturally produced speech when designing the speech and gestures for a virtual speaker is important if one is to ensure that the use of gesture will serve comprehension. Virtual speakers whose gestures are naturally timed in relation to speech (arriving before or in

sync with) support comprehension by emphasizing key words and making spoken information more memorable (Paper IV).

- The more naturalistic rendering of the virtual speaker reported in Paper IV may have distracted listeners, compared to the less naturalistic rendering. The naturally timed animation supported speech-and-gesture integration better than the unnatural timing. The virtual speaker reported in Paper V was far less naturalistically rendered than the MOCAP-based virtual speakers used in the other studies, but it still had a clear positive effect on comprehension compared to a disembodied voice. The take-home message is that the benefits of naturalism depend on which aspect of learning one is addressing as well as what the desired outcome is.
- Replicating behavioral experiments by using a virtual speaker as experimenter in a virtual environment offers several advantages. It makes it possible to vary aspects of the speaker's behavior more precisely, while keeping other aspects identical in a way that would not otherwise be possible. Bear in mind that aspects of the virtual speaker's design and presentation may affect how participants respond, affecting for example their experience of social presence or trust (Paper VI). It is worth piloting with more than one style of virtual speaker so as to identify potential confounding factors.

Outlook

Introduction

So far, the thesis introduction has in different ways given background, presented and discussed implications the findings of six separate empirical studies. However, as is often the case the ambitions at the start of the thesis project exceeded what was practically achievable, and some early identified possible avenues of research were left unexplored. More importantly, during the work that was actually done, new questions arose from thinking about the results and new possibilities emerged from thinking about the methodologies and tools that were developed. In this final chapter I therefore wanted to discuss some ways that the methodologies based on virtual speakers that I have worked with can be further developed to find answers to questions that go beyond the scope of this thesis.

Controlled experiments in virtual reality

Since 2015, when a large collaborative effort to try to reproduce previously published studies in experimental psychology found that less than half of significant results were reproduced (Open Science Collaboration, 2015), the problem of replicability has been a hot topic. Performing experiments in virtual environments using immersive virtual reality technology has previously been proposed as a way to increase both the reliability and ecological validity of findings (Blascovich et al., 2002), by being able to recreate scenarios that are both rich in detail and precisely controlled. Part of this could be to eliminate experimenter bias (Bronstein, 1990) from instructions or feedback given from an experimenter. A virtual experimenter, like the one described in paper VI, could be employed for this reason, and help meet challenge of replicability. Details of its speech and behavior can be precisely controlled and experimental conditions delivered the same way, also across studies. However, it is first crucial to establish that participants understand instructions given by a virtual speaker in an equivalent way to how they would understand a human experimenter giving instructions. Also, in cases where the experimenter interacting with the participant is part of the experimental procedure itself, it is important to understand how the social traits and perceptual and cognitive capabilities that participants prescribe to a virtual speaker might change their behavior.

Spatial information in gestures

Paper III described how the natural gestures can be temporally manipulated and reproduced within naturalistic segments by a virtual speaker. The paper also discusses how the presented workflow can be expanded to be able to study how gestures affect comprehension more generally. Another interesting use would be to manipulate spatial aspects of gestures, such as location in the conversational space. If first determining a way this can be done to produce naturalistic results, having such a research instrument could be useful to study interaction between language and spatiality in memory, expanding on studies where eye-tracking has demonstrated a role of spatial locations in memory retrieval (Johansson & Johansson, 2014; Johansson, Oren & Holmqvist, 2018). Introducing mismatches between what is said in speech and what is expressed in gestures, or inconsistent gestures from one mention of a concept or object to another, can help reveal to what degree integration of speech and gesture is inevitable.

Multimodal distractions

The Visual speech cues chapter mentioned some examples of how seeing a virtual speaker may both help and distract a listener. How and what audio-visual (speech-related) stimuli are distracting is a topic that have received surprisingly little attention (with some notable exceptions; (Cohen & Gordon-Salant, 2017; Gonzalez-Franco, Maselli, Florencio, Smolyanskiy & Zhang, 2017). For example, timing has been shown to be an important factor in integration of both visual speech cues (Venezia, Thurman, Matchin, George & Hickok, 2016) and gesture (Obermeier & Gunter, 2014, paper IV) with the “target” speech one is trying to hear and comprehend. But little is known about its role in processing unattended, and potentially distracting, speech. Are gestures and visual speech cues with inconsistent or unnatural timing easier or more difficult to inhibit? Exploration of this topic could be could advantageously be studied using virtual background speakers, and be revealing with regard to the attention dependence and load involved in integration. Another motivation to study the distracting effects of audio-visual stimuli is that many educational environments are quite “busy”, both in terms of audio and visual impressions. For this purpose we have constructed a classroom in virtual reality complete with a virtual speaker (teacher), virtual peer listeners and virtual screens, all fully controllable.

References

Alghamdi, N., Maddock, S., Barker, J., & Brown, G. J. (2017). The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech. *Speech Communication, 95*, 127-136.

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K. E., & Öhman, T. (1998). Synthetic faces as a lipreading support. In *Fifth International Conference on Spoken Language Processing*.

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., & Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception. *Speech Communication, 51*(2), 184-193.

Anikin, A., Nirme, J., Alomari, S., Bonnevier, J., & Haake, M. (2015). Compensation for a large gesture-speech asynchrony in instructional videos. In *Gesture and Speech in Interaction (GESPIN 4)* (pp. 19-23).

Araujo, D., Davids, K., & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comment on. *Ecological Psychology*, *19*(1), 69-78.

Arbib, M. A. (2012). How the brain got language: The mirror system hypothesis (Vol. 16). Oxford University Press.

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*, 115-121.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577-660.

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and social Psychology*, *19*(2), 163-194.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58*(2), 495-520.

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. International Journal of Artificial Intelligence in Education, 15(2), 95-115.

Bear, H. L., & Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication, 95*, 40-67.

Beattie, G. (2013). *Visible Thought: the New Psychology of Body Language*. Hoboken: Taylor and Francis.

Beattie, G., Webster, K., & Ross, J. (2010). The fixation and processing of the iconic gestures that accompany talk. *Journal of Language and Social Psychology, 29*(2), 194-213.

Bédi, B., Arnbjörnsdóttir, B., Vilhjálmsson, H. H., Helgadóttir, H. E., Ólafsson, S., & Björgvinsson, E. (2016). Learning Icelandic language and culture in Virtual Reykjavik: starting to talk. CALL communities and culture*–short papers from EUROCALL 2016 Edited by Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny, 37.*

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. Journal of Nonverbal behavior, 12(2), 120-138.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *The American Psychologist, 37*(3), 245-257.

Bertelson, P., Vroomen, J., De Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & psychophysics, 62*(2), 321-332.

Beskow, J., Granström, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. In *Ninth International Conference on Spoken Language Processing.*

Blair, K., Schwartz, D. L., Biswas, G., & Leelawong, K. (2007). Pedagogical agents for learning by teaching: Teachable agents. *Educational Technology, 47,* 56-61.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*(2), 103-124.

Boker, S. M., Cohn, J. F., Theobald, B. J., Matthews, I., Brick, T. R., & Spies, J. R. (2009). Effects of damping head movement and facial expression in dyadic conversation using real–time facial expression tracking and synthesized avatars. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3485-3495.

Bronstein, R. F. (1990). Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality*, *5*(4), 71.

Brophy, S., Biswas, G., Katzlberger, T., Bransford, J., & Schwartz, D. (1999, July). Teachable agents: Combining insights from learning theory and computer science. In *Artificial intelligence in education, 50*, 21-28). IOS Press.

Brunswik, E. (1956). Perception and the representative design of psychological experiments. Berkeley: University of California Press.

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. Brain research, 1242, 162-171.

Cassell, J., Sullivan, J., Churchill, E., & Prevost, S. (Eds.). (2000). *Embodied conversational agents.* MIT press.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996, October). About the relationship between eyebrow movements and Fo variations. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 4, pp. 2175-2178). IEEE.

Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology, 18*(4), 334-352.

Chomsky, N. (1975). *The logical structure of linguistic theory*. New York (N.Y.): Plenum Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. John Wiley & Sons.

Clark, R. E., & Choi, S. (2005). Five design principles for experiments on the effects of animated pedagogical agents. Journal of Educational Computing Research, 32(3),209-225.

Cohen, J. I., & Gordon-Salant, S. (2017). The effect of visual distraction on auditory-visual speech perception by younger and older listeners. *The Journal of the Acoustical Society of America*, *141*(5), EL470-EL476.

Cox, S. J., Harvey, R. W., Lan, Y., Newman, J. L., & Theobald, B. J. (2008, September). The challenge of multispeaker lip-reading. In *AVSP* (pp. 179-184).

Craig, S. D., Gholson, B., & Driscoll, D. M. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features and redundancy. *Journal of educational psychology*, *94*(2), 428.

De Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture*, *7*(1), 21-38.

De Ruiter, J. P., Noordzij, M. L., Newman-Norlund, S., Newman-Norlund, R., Hagoort, P., Levinson, S. C., & Toni, I. (2010). Exploring the cognitive infrastructure of communication. *Interaction Studies*, *11*(1), 51-77.

Domagk, S. (2010). Do pedagogical agents facilitate learner motivation and learning outcomes? *Journal of Media Psychology, 22*(2), 84-97.

Donald, M. (1993). Précis of Origins of the modern mind: Three stages in the evolution of culture and cognition. *Behavioral and brain sciences*, *16*(4), 737-748.

Dove, G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic bulletin & review*, *23*(4), 1109-1121.

Dowell, J., & Shmueli, Y. (2008). Blending speech output and visual text in the multimodal interface. *Human Factors, 50*(5), 782–788.

Drager, K. D. R., Reichle, J., & Pinkoski, C. (2010). Synthesized speech output and children: a scoping review. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association, 19*(3), 259–273.

Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1352.

Dunsworth, Q., & Atkinson, R. K. (2007). Fostering multimedia learning of science: Exploring the role of an animated agent's image. *Computers & Education*, *49*(3), 677-690.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. Psychiatry, 32(1), 88-106.

Fodor, J. A. (1983). *The modularity of mind.* MIT press.

Fraser, S., Gagné, J. P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*.

Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., & Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI 2003). New York: ACM. 529–536.

Gentilucci, M., & Corballis, M. C. (2006). From manual gesture to speech: A gradual transition. *Neuroscience & Biobehavioral Reviews*, *30*(7), 949-960.

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and instruction*, *15*(4), 313-331.

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., & Malik, J. (2019). Learning Individual Styles of Conversational Gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3497-3506).

Gladwell, M. (2019). Talking to strangers: what we should know about the people we don't know. London: Allen Lane.

Glenberg, A. M. (1997). What memory is for. Behavioral and brain sciences, 20(1), 1-19.

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology, 64*, 257–283.

Gonzalez-Franco, M., Maselli, A., Florencio, D., Smolyanskiy, N., & Zhang, Z. (2017). Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific reports*, *7*(1), 3817.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197-1208.

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and cross linguistic comparisons. *Frontiers in psychology, 9*, 879.

Gullberg, M. (2010). Language-specific encoding of placement events in gestures. *Event representation in language and cognition, 11*, 166.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. Pragmatics & Cognition, 14(1), 53-82.

Gulz, A. (2004). Benefits of virtual characters in computer-based learning environments: Claims and evidence. *International Journal of Artificial Intelligence in Education, 14*(3, 4), 313-334.

Gulz, A., & Haake, M. (2006). Design of animated pedagogical agents—A look at their look. *International Journal of Human-Computer Studies, 64*(4), 322-339.

Guindon, R., Shuldberg, K. & Connor, J., (1987) Grammatical and Ungrammatical structures in *User-Adviser Dialogues: Evidence for Sufficiency of Restricted Languages in Natural Language Interfaces to Advisory Systems*, Proc, 25th ACL, Stanford, CA.

Haake,M.(2009). *Embodied pedagogical agents: From visual impact to pedagogical implications.* Department of Design Sciences, Lund University, Sweden.

Haake, M., & Gulz, A. (2008). Visual stereotypes and virtual pedagogical agents. *Journal of Educational Technology & Society, 11*(4), 1-15.

Haake, M., & Gulz, A. (2009). A look at the roles of look & roles in embodied pedagogical agents–a user preference perspective. *International Journal of Artificial Intelligence in Education, 19*(1), 39-71.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication*. Biological psychiatry, 51*(1), 59-67.

Hostetter, A.B. (2011) When Do Gestures Communicate? A Meta-Analysis. *Psychological Bulletin, 137*(2), 297–315.

Hostetter, A. B., & Alibali, M. W. (2019). Gesture as simulated action: Revisiting the framework. *Psychonomic bulletin & review*, *26*(3), 721-752.

ISGS (2019). *Who we are.* Retrieved from http://gesturestudies.com/index.php/society/who-are-we/

Jansen, S., Chaparro, A., Downs, D., Palmer, E., & Keebler, J. (2013, September). Visual and cognitive predictors of visual enhancement in noisy listening conditions. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 57, No. 1, pp. 1199-1203). Sage CA: Los Angeles, CA: SAGE Publications.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & psychophysics, 14*(2), 201-211.

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change: You will like this paper better if you (believe you) chose to read it!. *Journal of Behavioral Decision Making*, *27*(3), 281-289.

Johansson, R., Holsanova, J., Dewhurst, R., & Holmqvist, K. (2012). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1289.

Johansson, R., & Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science*, *25*(1), 236-242.

Johansson, R., Oren, F., & Holmqvist, K. (2018). Gaze patterns reveal how situation models and text representations contribute to episodic text memory. *Cognition*, *175*, 53-68.

Johnson, K. (2017, May 27). *Why Soul Machines made an AI baby.* Retrieved January 1, 2020, from https://venturebeat.com/2017/05/26/why-soul-machines-made-an-ai-baby/

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in education, 11*(1), 47-78.

Kalkstein, D. A., Kleiman, T., Wakslak, C. J., Liberman, N., & Trope, Y. (2016). Social learning across psychological distance. *Journal of Personality and Social Psychology, 110*(1), 1.

Kaplan, A. M., & Haenlein, M. (2016). Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. *Business Horizons, 59*(4), 441-450.

Kendon, A. (1994). Do gestures communicate? A review. *Research on language and social interaction*, *27*(3), 175-200.

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

Kendon, A. (2007). On the origins of modern gesture studies. In Susan Duncan, Justine Cassell, & Elena Levy (Eds.), *Gesture and the dynamic dimensions of language* (pp. 13–28). Amsterdam: John Benjamins.

Kendon, A. (2017a). Pragmatic functions of gestures. *Gesture, 16*(2), 157-175.

Kendon, A. (2017b). Reflections on the "gesture-first" hypothesis of language origins. *Psychonomic bulletin & review*, *24*(1), 163-170.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language*, *48*(1), 16-32.

Klinger, E., Bouchard, S., Légeron, P., Roy, S., Lauer, F., Chemin, I., & Nugues, P. (2005). Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *Cyberpsychology & behavior, 8*(1), 76-88.

Kopp, S., Krenn, B., Marsella, S., Marshall, A.,Pelachaud, C., Pirker, H., Thorisson, K., & Vilhjalmsson,H. (2006, August). Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents* (pp. 205-217). Springer, Berlin, Heidelberg.

Krauss, R. M., Morrel-Samuels, P. & Colasante, C. (1991). Do conversational hand gestures communicate?. *Journal of personality and social psychology, 61*(5), 743.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Kuhl, P. K. (2007). Is speech learning 'gated'by the social brain?. *Developmental science*, *10*(1), 110-120.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology, 6*, 390.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lankoski, P., & Björk, S. (2007, September). Gameplay Design Patterns for Believable Non-Player Characters. In DiGRA Conference.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997, March). The persona effect: affective impact of animated pedagogical agents. In CHI (Vol. 97, pp. 359-366).

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

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One*, *4*(3), e4638.

Marieb, E. N., & Hoehn, K. (2007). *Human anatomy & physiology*. San Francisco: Pearson Benjamin Cummings.

Mattheyses, W., Latacz, L., & Verhelst, W. (2013). Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis. *Speech Communication*, *55*(7-8), 857-876. :

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953-978.

Mayer, R. E. (2001). *Multimedia learning.* Cambridge, England: Cambridge University Press.

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied*, *18*(3), 239.

Mayer, R. E., Dow, G. T., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds?. *Journal of educational psychology, 95*(4), 806.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746.

McNeill, D. (1985). So you think gestures are nonverbal?. *Psychological review*, *92*(3), 350.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago, IL: University of Chicago Press.

McNeill, D. (2008). Gesture and thought. University of Chicago press.

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.) *Language and gesture* (pp. 141–161). Cambridge: Cambridge University Press.

Metcalfe, J. (2009). Metacognitive judgments and control of study. Current Directions in Psychological Science, 18(3), 159-163.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013). Visual information can hinder working memory processing of speech. *Journal of Speech, Language, and Hearing Research, 56*(4), 1120–1132.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of educational psychology*, *94*(1), 156.

Moreno, R., Mayer, R. E., Spires, H., & Lester, J. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, 19,* 177–213.

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy, 7*, 33-35.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science, 15*(2), 133-137.

Müller, C. (2017). How recurrent gestures mean. *Gesture, 16*(2), 277-304.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995, May). Can computer personalities be human personalities?. In *Conference companion on Human factors in computing systems* (pp. 228-229). ACM.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation, 26*, 125-173.

Norman, D. A. (1988). *The Psychology of Everyday Things*. New York: Basic Books.

Obermeier, C., & Gunter, T. C. (2014). Multisensory integration: the case of a time window of gesture–speech integration. *Journal of Cognitive Neuroscience*, *27*(2), 292-307.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Papert, S. (1980). Mindstorms: Children, computers, and powerful ideas. Basic Books, Inc..

Pappano, L. (2012). The Year of the MOOC. *The New York Times, 2*(12).

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Pichora-Fuller, M.K., Kramer, S.E., Eckert, M.A., Edwards, B., Hornsby, B.W.,Humes, L.E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C.L., Naylor, G., Phillips, N.A., Richter, M., Rudner, M., Sommers, M.S., Tremblay, K.L., Wingfield, A., (2016). Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). *Ear and Hearing, 37*, 5S-27S.

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. (2011). Visual cues and listening effort: Individual variability. *Journal of Speech, Language, and Hearing Research*.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, *82*(1), 33.

Pruner, T., Popescu, V., & Cook, S.W. (2016). *The effect of temporal coordination on learning from speech and gesture.* Oral presentation at the 7th Conf. of the International Society for Gesture Studies (ISGS 2016). Paris, France. Abstract retrieved from http://www.gesturestudies.com/files/isgsconferences/ISGS16Abstracts.pdf

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press.

Richardson, J. C., Maeda, Y., Lv, J., & Caskurlu, S. (2017). Social presence in relation to students' satisfaction and learning in the online environment: A meta-analysis. *Computers in Human Behavior, 71,* 402-417.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405-409.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. Journal of Speech and Hearing Research, 39(6), 1159-1170.

Saitoh, T., Morishita, K., & Konishi, R. (2008). Analysis of efficient lip reading method for various languages. In *2008 19th International Conference on Pattern Recognition* (pp. 1-4). IEEE.

Salomon, G. (1984). Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of educational psychology, 76*(4), 647.

Savage, M. (2019, March 12). Meet Tengai, the job interview robot who won't judge you. Retrieved from https://www.bbc.com/news/business-47442953

Schlenker, P., & Chemla, E. (2018). Gestural agreement. *Natural Language & Linguistic Theory, 36*(2), 587-625.

Senkowski, D., Saint-Amour, D., Gruber, T., & Foxe, J. J. (2008). Look who's talking: The deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions. *Neuroimage*, *43*(2), 379-387.

Shah, H., & Pavlika, V. (2005, October). Text-based dialogical e-query systems: Gimmick or convenience. In *Proceedings of the 10th international conference on speech and computers* (pp. 17-19).

Sia, I., Halan, S., Lok, B., & Crary, M. A. (2016). Virtual patient simulation training in graduate dysphagia management education—A research-led enhancement targeting development of clinical interviewing and clinical reasoning skills. *Perspectives of the ASHA Special Interest Groups*, *1*(13), 130-139.

Silvervarg, A., Kirkegaard, C., Nirme, J., Haake, M. &Gulz, A. (2014). Stepstowards a challenging teachable agent. *Proceedings of theInternational Conference on Intelligent Virtual Agents* (pp. 410-419). Berlin, Germany: Springer.

Singh, A.-M., Marcus, N., & Ayres, P. (2012). The Transient Information Effect: Investigating the Impact of Segmentation on Spoken and Written text. *Applied Cognitive Psychology, 26*(6), 848–853.

Sjödén, B., Tärning, B., Pareto, L., & Gulz, A. (2011, June). Transferring teaching to testing–an unexplored aspect of teachable agents. In International Conference on Artificial Intelligence in Education (pp. 337-344). Springer, Berlin, Heidelberg.

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, *92*(3), B13-B23.

Studdert-Kennedy, M., & Goldstein, L. (2003). Launching language: The gestural origin of discrete infinity. *Studies in the Evolution of Language, 3*, 235-254.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America, 26(2), 212-215.

Swartout, W. R., Gratch, J., Hill Jr, R. W., Hovy, E., Marsella, S., Rickel, J., & Traum, D. (2006). Toward virtual humans. *AI Magazine, 27*(2), 96-96.

Sweller, J. (2005). The redundancy principle in multimedia learning. *The Cambridge handbook of multimedia learning*, 159-167.

Sweller J., Ayres P., Kalyuga S. (2011) The Redundancy Effect. In: Cognitive Load Theory. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies, vol 1. Springer, New York, NY

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics, 36*(2), 219-238.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. Trends in cognitive sciences, 14(9), 400-410.

Tomasello M. (2008) *Origins of human communication.* Cambridge, UK: MIT Press.

Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., & Hale, S. (2011). Crossmodal enhancement of speech detection in young and older adults: Does signal content matter?. *Ear and hearing, 32*(5), 650.

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, *11*(4), 233-241.

Tärning, B., Lee, Y., Andersson, R., Månsson, K,. Gulz, A. & Haake, M. (in revision). *Entering the black box of feedback: Assessing feedback neglect in a digital educational game for elementary school students*

Perniss, P. (2018). Why we should study multimodal language. *Frontiers in psychology, 9*, 1109.

Veletsianos, G., Heller, R., Overmyer, S., & Procter, M. (2010). Conversational agents in virtual worlds: Bridging disciplines. *British Journal of Educational Technology*, *41*(1), 123-140.7

Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., & Hickok, G. (2016). Timing in audiovisual speech perception: A mini review and new psychophysical data. *Attention, Perception, & Psychophysics*, *78*(2), 583-601.

Vroomen, J., Bertelson, P., & De Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & psychophysics, 63*(4), 651-659.

Vygotsky, L.S. (1962). *Thought and Language.* Cambridge, MA: MIT Press. (Original work published in 1934).

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. Presence, 7(3), 225-240.

Winters, S., Pisoni, D. (2005). Speech synthesis: Perception and comprehension. In Brown, K.,(ed), *Encyclopedia of Language and Linguistics, 12*, 31–49.

Wong, A., Leahy, W., Marcus, N., & Sweller, J. (2012). Cognitive load theory, the transient information effect and e-learning. *Learning and Instruction, 22*(6), 449-457.

Woodall, W. G., & Burgoon, J. K. (1981). The effects of nonverbal synchrony on message comprehension and persuasiveness. *Journal of Nonverbal Behavior*, *5*(4), 207-223.

Xu, Y., Pelachaud, C., & Marsella, S. (2014). Compound gesture generation: A model based on ideational units. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Proc. of the 14th Int. Conf. on Intelligent Virtual Agents* (IVA 2014), LNCS: 8637 (pp. 477–491). Cham, Switzerland: Springer International Publishing.

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

Yurovsky, D., Wu, R., Yu, C., Kirkham, N. Z., and Smith, L. B. (2011). "Modelselection for eye movements: assessing the role of attentional cues in infantlearning," in *Connectionist Models of Neurocognition and Emergent Behavior:From Theory to Applications*, ed. E. J. Davelaar (Singapore: World Scientific),58–75.

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education* (pp. 311-328). Routledge.

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of learning and motivation*, *44*, 35-62.

Links to original papers

Paper I: https://doi.org/10.1080/14015439.2018.1455894 [Open Access]

Nirme, J., Haake, M., Lyberg Åhlander, V., Brännström, J., & Sahlén, B. (2019). A virtual speaker in noisy classroom conditions: supporting or disrupting children's listening comprehension? *Logopedics Phoniatrics Vocology*, *44*(2), 79-86.

Paper II: https://doi.org/10.1016/j.specom.2019.11.005 [Open Access]

Nirme, J., Sahlén, B., Åhlander, V. L., Brännström, J., & Haake, M. (2020). Audio-visual speech comprehension in noise with real and virtual speakers. *Speech Communication*, *116*, 44-55.

Paper III: https://doi.org/10.3758/s13428-019-01319-w [Open Access]

Nirme, J., Haake, M., Gulz, A., & Gullberg, M. (2019). Motion capture-based animated characters for the study of speech–gesture integration. *Behavior Research Methods*, 1-16.

Paper IV: [*Manuscript submitted for publication*]

Nirme, J., Haake, M., Gulz, A., & Gullberg, M. (Unpublished manuscript). *Early or synchronized gestures facilitate recall of speech*.

Paper V: https://www.lucs.lu.se/LUCS/176/LUCS_176.pdf [Free Access]

Nirme, J., Fredriksson, O., Haake, M., & Gulz, A. (2020). Exploring students' approach to factual texts in different presentation media. *Lund University Cognitive Studies*, *176.* Retrievable from: https://www.lucs.lu.se/publications/lund-university-cognitive-studies/

Paper VI: https://doi.org/10.1007/978-3-319-21996-7_47

Lingonblad, M., Londos, L., Nilsson, A., Boman, E., Nirme, J., & Haake, M. (2015, August). Virtual blindness – A choice blindness experiment with a virtual experimenter. In: Brinkman WP., Broekens J., Heylen D. (eds.), *Intelligent Virtual Agents* (*IVA 2015): LNCS*, *vol. 9238* (pp. 442-451). Springer, Cham.

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

LUND
UNIVERSITY

Faculty of Humanities and Theology
Department of Philosophy
Cognitive Science

Lund University Cognitive Studies 177
ISBN 978-91-88899-84-3
ISSN 1101-8453

9 789188 899842