Åbo Akademi

OPEN ACCESS

This is an electronic reprint of the original article. This reprint may differ from the original in pagination and typographic detail.

Audio-visual speech comprehension in noise with real and virtual speakers

Nirme, Jens; Sahlén, Birgitta; Lyberg Åhlander, Viveka; Brännström, Jonas; Haake, Magnus

Published in:
Speech Communication

DOI:
10.1016/j.specom.2019.11.005

Published: 01/01/2020

Document Version
(Peer reviewed version when applicable)

Document License
Publisher rights policy

Link to publication

Please cite the original version:
Nirme, J., Sahlén, B., Lyberg Åhlander, V., Brännström, J., & Haake, M. (2020). Audio-visual speech comprehension in noise with real and virtual speakers. *Speech Communication*, *116*, 44–55. https://doi.org/10.1016/j.specom.2019.11.005

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

This document is downloaded from the Research Information Portal of ÅAU: 17. May. 2021

Audio-visual speech comprehension in noise with real and virtual speakers.

Jens Nirme (corresponding author)
Div. of Cognitive Science, Lund University, Lund, Sweden
Lund university, Box 192, 221 00, Lund
e-mail: jens.nirme@lucs.lu.se

Birgitta Sahlén
Logopedics, Phoniatrics and Vocology, Department of Clinical Sciences, Lund University, Lund, Sweden
Lund University Hospital, s-221 85 Lund, Sweden
e-mail: birgitta.sahlen@med.lu.se

Viveka Lyberg Åhlander
Logopedics, Phoniatrics and Vocology, Department of Clinical Sciences, Lund University, Lund, Sweden
Lund University Hospital, s-221 85 Lund, Sweden
Speech Language Pathology, Faculty of Arts, Psychology and Theology, Åbo Akademi University, Turku, Finland
e-mail: Viveka.Lyberg_ahlander@med.lu.se

Jonas Brännström
Logopedics, Phoniatrics and Vocology, Department of Clinical Sciences, Lund University, Lund, Sweden
Lund University Hospital, s-221 85 Lund, Sweden

Magnus Haake
Div. of Cognitive Science, Lund University, Lund, Sweden
Lund university, Box 192, 221 00, Lund
e-mail: magnus.haake@lucs.lu.se

Abstract

This paper presents a study where a 3D motion-capture animated ‘virtual speaker’ is compared to a video of a real speaker with regards to how it facilitates children’s speech comprehension of narratives in background multitalker babble noise. As secondary measures, children self-assess the listening- and attentional effort demanded by the task, and associates words describing positive or negative social traits to the speaker. The results show that the virtual speaker, despite being associated with more negative social traits, facilitates speech comprehension in babble noise compared to a voice-only presentation but that the effect requires some adaptation. We also found the virtual speaker to be at least as facilitating as the video. We interpret these results to suggest that audiovisual integration supports speech comprehension independently of children’s social perception of the speaker, and discuss virtual speakers’ potential in research and pedagogical applications.

1 Introduction

Education at all levels has elements of verbal instruction or lectures. Students have to recognize, process, and comprehend material conveyed through speech. *Speech recognition* refers to the ability to identify words and sentences in a speech signal, and *speech comprehension* refers to the more advanced and complex ability to understand the meaning conveyed in the speech signal. Classrooms are, however, often noisy environments – not least in primary school – making both speech recognition and comprehension challenging (Bradley & Sato, 2008; Lyberg-Åhlander, Brännström, & Sahlén, 2015). We also have, by now, converging evidence indicating detrimental effects of exposure to noise, of the type that might occur in classroom environments. For example, controlled experimental studies have shown direct effects of noise on speech recognition (Grant & Seitz, 2000; Hagerman, 1982; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2006) and children have been found to be more strongly affected by noise when it comes to speech recognition compared to adults (Neuman et al., 2010). Furthermore, longitudinal studies have shown detrimental long-term effects of noisy environments on academic performance (Shield and Dockrell, 2008).

Real life listening is however typically not based on only the auditory channel, but rather an integration of speech and visual cues. It is long since established that facial visual cues such as lip movements influence speech recognition. With congruent audiovisual stimuli, speech recognition can be aided (Hagerman, 1982; Ma, Zhou, Ross, Foxe, & Parra, 2009; Neuman, Wroblewski, Hajicek, & Rubinstein, 2010; Sumby & Pollack, 1954) and with incongruent audiovisual stimuli it is modulated (McGurk & MacDonald, 1976). Head- and eyebrow movements coordinated with speech inform the perception of prominent words within sentences (Al Moubayed, Beskow, & Granström, 2009) and speech recognition (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004).

2 Background

2.1 Speech Comprehension

While speech comprehension requires successful speech recognition, it also depends on other cognitive resources. To comprehend speech, one must semantically process words and sentences, keep track of what was said a minute ago, relate it to previous knowledge, and encode new information for later retrieval. It is a complex task that has to be performed in real time. Processing on different levels of abstraction and in different sensory channels has to be coordinated, with both bottom-up (McGurk & MacDonald, 1976) and top-down (Tuomainen, Andersen, Tiippana, & Sams, 2005) processes potentially in effect. Also, speech recognition alone requires allocation of a range of perceptual and cognitive resources when done under suboptimal conditions, as summarized in a review by Mattys, Bradlow, Davis, and Scott (2013). Listening effort refers to a deliberate allocation of mental resources to meet these and other demands associated to listening and comprehension, including direction of attention to one specific

speaker (Pichora-Fuller et al., 2016). Different varieties of noise (including competing speech) have been shown to increase listening effort associated with speech recognition (Koelewijn, Zekveld, Festen, & Kramer, 2012).

Previous studies have found ensuing effects of noise on listening comprehension. Ljung, Sörqvist, Kjellberg, and Green (2011) found that adults are worse at answering questions about content they had heard in noise and Valente, Plevinsky, Franco, Heinrichs-Graham, and Lewis (2012) found that the negative effect is stronger for children. Furthermore, children have also been found to have a harder time following instructions in the presence of a competing speaker (Klatte, Lachmann, & Meis, 2010). Brännström, von Lochow, Åhlander, and Sahlén (2018) found that multitalker babble noise had a detrimental effect on children's ability to answer comprehension questions both immediately after listening to narrative passages and after 5-8 days.

The role of visual input in speech comprehension is less clear than its role in speech recognition. As mentioned, comprehension depends on a range of cognitive and perceptual capacities, and the same can be said about audiovisual integration of speech. Jansen, Chaparro, Downs, Palmer, and Keebler (2013) found that 'visual enhancement' of speech recognition performance correlated with baseline measures of contrast sensitivity and executive functioning. The integration of visual cues thus competes for the same limited cognitive resources as other processes crucial to speech comprehension. The relation between listening conditions and the demands of parallel multimodal processes involved in speech comprehension is not yet fully understood, though recently some progress has been made in the field. In a study by Picou, Ricketts, and Hornsby (2011), participants with larger working memory capacities demonstrated and reported reduced listening effort in noise when exposed to audiovisual conditions. Similarly, Mishra, Lunner, Stenfelt, Rönnberg, and Rudner (2013a) showed more 'cognitive spare capacity' (operationalized as performance on updating or inhibiting tasks) when seeing a face speaking two-digit numbers compared to only hearing – but only in steady state noise and not in speech-like babble. On the other hand, another study by the same group reported decreased cognitive spare capacity by visual presentation of the speaker in the absence of noise (Mishra, Lunner, Stenfelt, Rönnberg, and Rudner, 2013b). Fraser, Gagné, Alepins, and Dubois (2010) found reduced reported listening effort as well as improved speech recognition in noise when seeing the speakers face compared to only listening. However, when noise levels were normalized to equalize speech recognition accuracy between conditions, results indicate that accuracy came at an increased cost in reported listening effort and reaction times.

Overall, results from studies of how listening conditions influence audiovisual speech comprehension are inconclusive. Further research is warranted, although it is challenging due to the composite nature of speech comprehension.

2.2 Digital Animated Characters and Speech

Digitally animated characters are generally representations of humans or anthropomorphized animals, imaginary creatures, or even inanimate objects that move and speak (Haake, 2009). The animations, voices, and movements are either scripted, procedurally generated, captured from real human actors, or a combination thereof. These digitally animated characters are regularly referred to as *Virtual Humans* (Garau, Slater, Pertaub, & Razzaque, 2005),

Embodied Conversational Agents (Cassell, 2000), or *Intelligent Virtual Agents* (Johnson & Lester, 2015). They are, furthermore, often implemented in pedagogical settings and then referred to as *Animated Pedagogical Agents* (Clark & Choi, 2005) – 'pedagogical' denoting some kind of pedagogical role.

Due to the potential implementations of controlled behaviors that are at the same time rich and flexible, digitally animated characters have been implemented for studying a range of phenomena (examples include social influence: Blascovich et al., 2002; choice blindness: Lingonblad et al., 2015; expressivity: Pelachaud, 2009). Regarding speech recognition and comprehension, digitally rendered characters have been used in a number of studies, e.g. adults' speech recognition has been found to be facilitated with digitally rendered facial animations (Agelfors et al., 1998; Möttönen, Olivés, Kulju, & Sams, 2000) – but usually to a lesser degree than video recordings with a natural speaker. Backing this idea of using digital animated characters as research instruments is the body of research with Animated Pedagogical Agents within educational settings comprising a broad variety of interactive agent-human environments (Gulz, 2005; Johnson and Lester, 2015).

In this study, we will focus on the viability of digitally animated characters as tools to study speech comprehension under audio-visual, noisy conditions. The lecture-like experimental scenario involves, furthermore, no actual interaction or turn-taking between speaker and listeners. To clarify this distinction, we use the term 'virtual speaker' to denote the digitally animated character.

2.3 The Virtual Speaker

In a previous study (Nirme, Haake, Lyberg-Åhlander, Brännström, & Sahlén, 2018), we examined how presentation of a 3D animated 'virtual speaker' (rendered with facial expressions and lips movements recorded in parallel with an audio recording) affected speech comprehension, as measured by questions following spoken narratives. The study had a crossed 2 × 2 factor within-subject design, with the factors 1) presence or absence of background multitalker babble noise, and 2) the presentation of the narratives by the virtual speaker or by audio-only. Results showed a strong significant negative effect of noise on comprehension – while the interaction between noise and the presentational format (virtual speaker vs. audio-only) only approached significance ($p = .08$), with a stronger positive effect on speech comprehension when presented with the virtual speaker in the noise condition compared to the quiet setting (where there was virtually no effect). A reanalysis of the data, including a random slope factor representing individual variance with regard to the sensitivity to noise, found the interaction effect to be weakly significant ($p = .047$). While not confirming that the virtual speaker helped the participating children (age 8-9) to overcome the effect of babble noise, the results indicated a potential positive effect of the virtual speaker during some circumstances. In the paper we discussed possible explanations for the inconclusive result that can be summarized as follows: (1) the study's within-subject design and limited number of narratives caused a type II error – masking a true effect in that the virtual speaker was only presented in two narratives out of 4 and only in one narrative together with babble noise; (2) the facial animation was not of sufficient fidelity or realism; thus the virtual speaker was a relatively poor support for audiovisual speech integration, or even unappealing (uncanny) to listeners and distracting from the listening task.

Related to the second explanation is the question whether audiovisual speech integration relies on perceptual adaptation in order to be effective, similarly to how previous exposure to different accents leads to improved recognition (Bradlow & Bent, 2008). Hagerman (1982) observed learning effects testing speech recognition in noise, which were partly attributable to repeated presentations of sentences including similar lexical items. Rosenblum, Johnson and Saldaña (1996) found that the speech recognition improved during the second half of the experiment for all audiovisual conditions, including video showing the full face as well as point light representations of the face – while no such improvement was observed for the audio-only group. Adaptation effects are possibly more pronounced in the context of more unfamiliar audiovisual stimuli as suggested by a study by Alghamdi, Maddock, Barker, and Brown (2017) in which strong adaptation effects were observed for stimuli with exaggerated lip movements over repeated (training) sessions in a sentence recognition task. Furthermore, perceptual adaptation can even influence performance of subsequent similar but not identical perceptual tasks, like the 'McGurk aftereffect' as demonstrated by Bertelson, Vroomen, and De Gelder (2003).

Moreover, explanation 2 also relates to how much the perception of distracting or awkward visual features influence audiovisual speech processing. Behavioral and neurological findings indicate a separation between a face recognition system based on invariant features and a system for perception of facial expressions by variant features – as well as a specific separate subsystem for audiovisual speech perception integrating information from auditory cortex (Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2002). The degree of independence of these systems has, however, been debated (Calder & Young, 2005). The superior temporal sulcus (STS) has been identified as the neural locus of audiovisual integration (Beauchamp, Nath, & Pasalar, 2010) while it also seems to be involved in processing of different types of facial expressions (Haxby, Hoffman, & Gobbini, 2002).

2.4 Indirect Social Effects and Distraction

If not directly disrupting speech recognition, visual features of a virtual speaker might have more indirect effects on comprehension. Social effects related to virtual speakers are since long established. Cassell and Bickmore (2003) demonstrated the strong social effects of so-called relational agents, i.e. virtual agents designed to establish social relationships with human users. In a pedagogical context, virtual speakers are prevalent in educational software in different socially relational roles such as teachers, peers or teachable agents – as well as other roles related to the learning content such as historical figures an anthropomorphized animals acting like learning companions (Gulz, 2005; Haake, 2009; Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010; Graesser, 2016). Lester, Converse, Kahler, Barlow, Stone, & Bhogal (1997) described the so called 'persona effect', that a lifelike character makes learners view their experience more positively. Moreno, Mayer, Spires, & Lester (2001) later established the concept of 'social agency effects', i.e. virtual agents with (apparent) social communicative abilities make learners more committed to engage and make sense of instructional content. Subsequent studies on learning of educational material in different (multimedia) applications have found evidence both supporting (Mayer, 2014; Lusk & Atkinson, 2007) and contradicting (Choi & Clark, 2006) these claims. A review made in 2011 argued for the importance of including control conditions where material is presented without an agent and concluded that the

majority of studies fulfilling this criterion showed no effects on learning outcome or motivation (Heidig & Clarebout, 2011). However, in the case of 'teachable agents', i.e. virtual pedagogical agents implemented within a 'learning by teaching' paradigm, a social responsibility effect (*the protégé effect*) has proven robust effects on motivation and engagement (Chase, Chin, Oppezzo, & Schwartz, 2009).

A proposed mechanism enabling 'social agency' depends on listeners picking up 'social cues' from a character and that this is more likely to happen when it displays lifelike appearance and behavior, referred to as 'the embodiment principle' (Clark & Mayer, 2016; Mayer & DaPra, 2012). Of course, people interacting with most non-human conversational systems are at some level aware of its conversational limitations and constrained movements but might still act as if interacting with another human. In fact, social responses towards digital artifacts (Reeves & Nass, 1996) – as well as ascribing social traits to them (Nass, Moon, Fogg, Reeves, & Dryer, 1995) – can be observed even with the most basic form of text-based communication with computers. This behavior, also known as suspension of disbelief, is however volatile and may break down by mismatches between observed and expected behavior (Thomas & Johnston, 1995). As Norman (2013) has suggested, people tend to react with aversion and frustration to interactive artifacts or systems that do not match our preconceived mental models.

Even if we accept the assumption that a social context can promote learning, not all social characteristics of virtual characters are necessarily positive. Animated agents exhibiting dislikable (negative) 'social cues' of appearance and voice (determined by pretesting) have been shown to be detrimental to learning (retention and transfer) compared to both a 'likable' agent and an onscreen text-only control condition (Domagk, 2010). The 'uncanny valley hypothesis', formulated in the early days of robotics by Mori (1970), suggests that artifacts approaching but not quite reaching the level of human realism creates a sense of eeriness and diminished affinity in comparison to both perfect/total human-likeness and more clearly artificial (or cartoonish) forms. Mori also suggested the effect would be enhanced if the artifact was animated. While on one hand the uncanny valley hypothesis has been generalized to also cover non-physical artifacts such as digitally animated characters, its validity has also been disputed. The original formulation does not properly define what specific aspects of realism are crucial, or how to operationalize 'eeriness'. Kätsyri, Förger, Mäkäräinen, & Takala (2015) examined the body of empirical evidence and found some evidence for the 'perceptual mismatch' formulation of the hypothesis; that negative affinity arises from inconsistency in the level of human-likeness between different aspects of visual appearance (MacDorman, Green, Ho, & Koch, 2009) or between appearance and behavior (Garau et al., 2003). There is also some evidence that previous exposure can mediate the effect (Burleigh & Schoenherr, 2014), i.e. a kind of adaptation effect.

Even if a virtual speaker is perceived positively, the addition of non-task relevant stimuli in the visual channel might distract when listening in noise. People focusing on a spoken content sometimes spontaneously choose to focus on the auditory signal and close their eyes – in the real world as well as in virtual settings. A few children included in a pilot test for a previous study (Nirme et al., 2018) had to be reminded to look at the screen with a virtual speaker as they spontaneously closed their eyes as a strategy to focus on the spoken content. Clark and Choi (2005), warn against designing too 'complex and noisy' agents that might distract or increase the 'cognitive load' in educational

applications. At the same time, animated virtual agents can be designed to cue attention to relevant information in a multimedia environment, reducing extraneous cognitive load (Yung & Paas, 2015).

2.5 Research Questions

Based on the inconclusive result of our previous study (Nirme et al., 2018) and the possible explanations discussed in the paper, we formulated a new set of research questions to be investigated in this present study.

The purpose of this study is to investigate how a virtual speaker (based on audio and motion capture recordings of a real speaker) supports speech comprehension in babble noise compared to an audio-only presentation and a video recording of the aforementioned real speaker.

The negative main effect of babble noise on speech comprehension has already been demonstrated in previous studies (Klatte, Lachmann, & Meis, 2010; Ljung et al., 2011; Nirme et al., 2018; Valente et al., 2012), and is not addressed in this study adopting babble noise in all conditions.

The first two research questions (RQ1 and RQ2) are of primary importance to the design of the study presented below and are addressed using a validated measurement tool for speech comprehension (CELF-4, see section *Materials and Methods* below), while research questions three (RQ3) and four (RQ4) are more exploratory and addressed using self-report measures (see section *Materials and Methods* below).

RQ1: Virtual speakers and comprehension in noise

a) *Does visual presentation of a virtual speaker facilitate speech comprehension in background multitalker babble noise compared to an audio-only presentation?*

If the inconclusive result of the previous study (Nirme et al, 2018) can be explained by insufficient statistical power and thus strengthened by increasing the number of administered tasks (texts with corresponding answers) from one to three, we expect to find a significant difference as to improved results of speech comprehension after being presented with the virtual speaker compared to audio-only presentation.

b) *How does visual presentation of a virtual speaker compare to a video of a real speaker with regard to supporting speech comprehension in background multitalker babble noise?*

If the differences in static and dynamic appearance between the virtual speaker and the real speaker affect audiovisual speech processing, we expect to find impaired speech comprehension after seeing the video with the virtual speaker compared to the video with the real speaker.

RQ2: Adaptation effects

Is there an adaptation effect involved in exploiting visual speech cues from a virtual speaker resulting in improved speech comprehension in background multitalker babble noise?

If adaptation effects observed in studies of speech recognition (Alghamdi et al., 2017; Rosenblum et al., 1996) can be generalized to speech comprehension, we predict that there will be an increasing positive effect on speech comprehension (in background multitalker babble noise) by repeated exposure to the virtual speaker.

RQ3: Self-assessed listening effort and attentional effort

Will speech comprehension and sustained attention to the speaker in the presence of background multitalker babble be perceived as more effortful when watching a virtual speaker compared to only hearing an audio recording or watching a video recording of a real speaker?

If the virtual speaker's static and dynamic appearance interferes with or distracts from speech comprehension (in background multitalker babble noise), we expect the participating children to report higher levels of effort in attending to the verbal content after listening to and seeing the virtual speaker.

RQ4: Perceived social traits

a) *Will watching a virtual speaker elicit more* negative *or positive perceived social traits compared to listening to an audio recording or watching a video recording of a real speaker?*

If there is a perceived mismatch between different aspects of the virtual speaker's static appearance, voice, or movement and the children's expectations, we expect the participants/children to attribute more negative social traits to the virtual speaker.

b) *How will perceived social traits be related to speech comprehension outcome?*

Moreover, if they interfere with audio-visual speech processing, we expect to find a negative correlation between perceived negative social traits and speech comprehension. Conversely, if social agency promotes speech comprehension, we expect to find a positive correlation between perceived positive social traits and outcome.

3 Materials and Methods

3.1 Participants

To examine our research questions, we performed a study with 3rd- and 4th-graders in 5 classes from 5 different schools in the Scania region in southern Sweden. For participant data to be included in the study, the children (apart from obtaining informed consent from legal guardians) had to fulfill the following criteria: (i) they had to be between 8- to 10-years-old (corresponding to grade three to four in the Swedish educational system) with a minimum two years in Swedish schools) and (ii) they had to pass a hearing screening performed at the end of the experiment (see below). If any child (with consent from legal guardians) did not fulfil the two criteria, they were still permitted to perform the experiment, but their data excluded from the analyses.

3.2 Materials

3.2.1 Speech comprehension

Testing of speech comprehension was done using the 'Understanding Spoken Paragraphs' subtest of the Swedish version of the *Clinical Evaluation of Language Fundamentals* (*CELF-4*), a battery of tests assessing language skill that has been validated for English speaking children (Semel, Wiig, & Secord, 2004). For our study we selected three narratives (named A, B, and C) corresponding to three of the texts validated for the participants' age group.

The 'Understanding Spoken Paragraphs' subtest in CELF-4 consists of short narratives (texts) followed by five related questions: three 'content' questions about explicitly mentioned details, e.g. "*What did the students receive more than pizza and soft drinks?*"), one 'inference' question requiring inference based on the content, e.g. "*Why do you think Pricken [a dog] ran in the opposite direction*?", and one 'summary' question (a basic question about the theme).

In the current study we have only included answers to the content questions in our analyses. The rationale for this is twofold: (i) the results of the 'inference' question may be confounded by students' previous knowledge and experiences, and (ii) in the previously mentioned study (Nirme et al., 2018), we found no significant effect of babble noise on inference questions. Answers to summary questions were also excluded due to non-specific answer definitions (e.g. any answer that is 'logical' is accepted) and ceiling effects observed in previous studies.

Three versions of each narrative (CELF-4 narrative: A, B, and C) were prepared, one version for each of the three experimental conditions. The three versions were all based on a recording session of a real speaker capturing audio, video, and motion capture data, presenting: (i) a rendered video of a virtual (digitally animated) speaker generated from the audio and motion capture recordings (VIRTUAL); (ii) a video of the real speaker generated from the audio and video recordings (VIDEO); and (iii) an audio-only version without any visual presentation of the real speaker (AUDIO).

3.2.2 Stimuli for the AUDIO condition

The narrator's voice (used in all three experimental conditions) was recorded using a head mounted microphone (*Lectret HE-747*) and sampled in 16 bit, 44.1 kHz. The narrator was the same as in the previous study (Nirme et al., 2018). Each narrative was recorded two times and the recording with the best overall quality was then selected. The duration of the three selected narratives ranged from 42 to 45 s with a speech rate ranging from 145 to 187 words (267-288 syllables) per minute. The sound pressure levels of the recordings were normalized to equalize their root mean square (RMS).

Background multitalker babble noise was added to each audio track. The background multitalker babble noise was constructed from separate recordings of four girls aged nine to eleven years reading different separate chapters from a children's story book. Selected parts of the recordings of each girl were combined after being normalized to have an equal average RMS. Pauses were removed for the normalization and afterwards reinserted. For further details about the process of constructing the babble noise, see von Lochow, Lyberg-Åhlander, Sahlén, Kastberg, &

Brännström (2018a). The resulting babble noise track was added to the speech signal at -10 dB resulting in a +10 dB signal-to-noise ratio (SNR). To preserve ecological validity and account for the *Lombard effect* (Junqua, 1993), the narrator heard the same type of babble noise (at the same levels) as was later added to the recordings. Prior to recordings, the narrator went through a 'vocal loading' procedure to induce hoarseness (Whitling, Rydell, & Åhlander, 2015). This procedure reproduce the strain teachers put on their voices (resulting in hoarseness) to compensating for noisy classroom environments (Lyberg-Åhlander, Rydell, & Löfqvist, 2011; Kristiansen et al., 2014). The resulting audio files both constituted the complete stimulus material for the AUDIO condition and were integrated as the audio component in the two audio-visual conditions (VIDEO and VIRTUAL). The same recorded narrations and babble noise tracks have been used in previous studies (Brännström et al., 2018; Nirme et al., 2019; von Lochow et al., 2018a,b).

3.2.3 Stimuli for the audiovisual (VIRTUAL and VIDEO) conditions

During the audio recordings, an *ASUS Xtion Pro Live* sensor was placed on a table in front of the narrator at around 40 cm distance from and 20 cm below the level of the narrator's face. The ASUS Xtion Pro Live sensor has both a video camera (RGB; 640 × 480 pixels) and an active infrared 3D depth map sensor (640 × 480 pixels). For the study, we recorded the narrators head and upper torso at 30 frames per second.

The real speaker presentation (VIDEO) was created by combining the recorded video files and corresponding audio tracks (with added babble noise) to generate AVI video files (640 × 480 pixels, 30 fps, *Xvid* video compression, uncompressed PCM audio), using *AviDemux 2.5.6* (AviDemux, 2011). The video and sound files were synchronized by means of a clapperboard (visible next to the narrators head a few seconds before the narrative started). The resulting video files constituted the stimulus material for the VIDEO condition (fig. 1; left).

A digital 3D-model matching the apparent age and gender of the narrator was created using the 3D character modeling software *Autodesk Character Generator* (Autodesk, 2014). Both the 3D-model and the animation process were identical to those used in the study (Nirme et al., 2018). Lip movements, facial expressions, eye blinks, head and torso postures, and gaze direction were extracted from the recorded image and depth map data generated by the ASUS Xtion sensor (see above) using *FaceShift Studio* (software specialized for facial motion capture; FaceShift Studio, 2015). The software was trained on a set of predefined 'target facial poses' performed by the narrator before the recordings started. For each frame of the recorded sequences, the FaceShift algorithm finds a weighted combination of the target poses that best approximates the recorded depth map data and recognition of eyebrow, eye, and lip movements. The weighted combinations of target poses were then further informed by the recorded 2D video image data.

The resulting facial animation was transferred onto the 3D-model by mapping the weighted combination of the target poses to its weighted combination of *blendshapes* (target mesh deformations representing idealized expressions such as a smile or a visemes associated with specific phonemes; Lewis et al., 2014), and orientations of head, neck and torso to its skeletal articulation (realized by combined rotations of three 3-dimensional 'skeletal' joints). The resulting angular movement (representing the corresponding movement of the real speaker visible in the

recorded videos) was quantified as the standard deviation (SD) and range of motion (ROM) around three axes: pitch/nodding 2.78°/15.4° SD/ROM; yaw/turning 3.27°/15.8° SD/ROM, and roll/tilting 3.84°/17.3° SD/ROM. There was a change in direction of rotation around one or more axis approximately once every 1.0 s.

The 3D model and animation data were then imported into the 3D graphics software *Autodesk Maya 2014* (Autodesk, 2014). In *Autodesk Maya*, a backdrop matching that of the VIDEO condition was added and lighting, camera perspective, and orientation of the speaker relative the camera was adjusted to reproduce the scenery of the real video presentation in the VIDEO condition. Light sources were placed both from above and in front of the virtual speaker.

Each frame (30 frames per second) was rendered as a 640 × 480 image file to match the VIDEO condition (generated from the original recording of 640 × 480 pixels, see above). The rendered images were then combined with the corresponding audio tracks using Avidemux 2.5, creating videos showing the virtual speaker (VIRTUAL condition) by a process identical to the VIDEO condition (see above).

The resulting videos with the virtual speaker (fig. 1, right) were reviewed by an expert lip reader that deemed them realistic – apart from a minor issue with a few articulations of /f/. Since the aim of the current study was partly to examine how deviations of 'a virtual speaker' from a real speaker affects speech comprehension, we opted not to correct such minor issues.

Fig. 1. Left: Nonconsecutive frames from the real video condition (VIDEO) with the narrator presenting the narratives used for the CELF-4 passage comprehension test; Right: Nonconsecutive frames from the virtual speaker condition (VIRTUAL) with the virtual speaker performing the same task as the narrator in the real video (to the left).

3.2.4 Self-assessed listening effort and attentional effort

To obtain a subjective measure of how effortful the children found listening and sustaining attention to the different versions of the speaker, we created a short questionnaire that was delivered after the speech comprehension test. In the questionnaire, the children were instructed (by means of written instructions as well as verbally) to rate how difficult they perceived three different aspects of the speech comprehension test they had just performed. The rating was accomplished using a pencil on a 100 mm wide visual analog scale (VAS) with the labeled endpoints "difficult" (to the left) and "easy" (to the right); no marked 'center point' was labeled. The the children were asked to rate three different aspects addressing how easy or difficult it was: "*[...] to hear what the speaker was saying?*" (VAS.1), "*[...] to concentrate on listening to the narrative?*" (VAS.2) and "*[...] to answer the questions?*" (VAS.3).

For our study, we were mainly interested in the VAS.1 and VAS.2 ratings. The third was a dummy item, included to help the children distinguish the demands of listening from their general performance on the test. For the analysis, ratings were encoded with one decimal precision on a scale from -1.0 (most difficult) to (1.0 easiest).

3.2.5 *Word cloud: Perceived social traits*

To probe how the children perceived the speaker, we constructed a 'word cloud' in the form of 18 valence words (items) spread out in an elliptic pattern. The 18 valence words were based on 9 pairs describing opposite positive vs. negative social traits (translated from Swedish): agreeable – annoying; bright – dull; kind – mean; amusing – boring; certain – uncertain; good self-confidence – poor self-confidence: nice – nasty; normal – strange; cocky – wimpish. The valence of the words were evaluated and pre-tested by 13 adults.

For the questionnaire, the items were randomly distributed across a single A4 page as a 'word cloud' with the instruction to circle any items that described the speaker. The children were further told to circle as many items as they wanted describing the speaker. The distribution of the items (words) in the word cloud were identical for all children. The number of encircled (negative and positive) words was calculated for each child.

The rationale for the design of a 'word cloud' consisting of opposite items (valence words) describing a set of social traits, stems from previous experiences using Likert scales. Repeatedly, the use of Likert scales to probe different types of social traits resulted in confined and strongly positive or negative weighted data sets (Brännström et al., 2015). The approach to ask the children to encircle appropriate words in a word cloud was thus an attempt to circumvent this.

3.3 Procedure

Preceding the experiment, the teachers involved in the study distributed and collected informed consent forms from the children's guardians. Before starting the actual data collection, the experimenter practiced the experimental procedure with a few adults and children volunteers (not related to the participating schools). Video recordings of the practice sessions were reviewed by a senior researcher together with the experimenter. Identified inconsistencies or ambiguities in the verbal instructions were addressed and the script used for data collection was then reviewed and finalized.

For the actual experiment, the children participated one-by-one. First, the experimenter greeted the child and led her/him from their regular classroom to a quiet non-adjacent room were the experiment was conducted. Next, the child was asked to take a seat in front of a table after which the experimenter gave a brief description of the sequence of tests s/he was going to experience. The participant was also told that s/he would receive a small reward for her/his participation, and that s/he were free to interrupt the experiment and leave at any time.

The first (and main) part of the experiment was the passage comprehension test. Each child was assigned to one of the three experimental conditions (VIRTUAL, VIDEO, or AUDIO) and one of the three predefined orders of the three narratives (ABC, CAB, and BCA) of the passage comprehension test.

Both the assignment to the experimental condition and the order of narratives was determined by stepwise following a balanced list, rotating the experimental condition for every participant as well as rotating the order of the three narratives by cycling through three distinct orders (ABC, CAB, and BCA) for every set of three participants. This was done to balance the occurrence of each narrative as the first, second, or third in the sequence of narratives, thus minimizing the risk of observed adaptation effects over trials to be confounded by potential variations in the level of difficulty across the three narratives and their associated questions. The order the children performing the tests was determined by their class teacher, who was blind to the order determined by the balanced experimental list as well as the experimental design as such. Furthermore, the experimenter was careful to interact with the children during the experiment according to a standardized and minimized scheme in order to minimize his influence on their performance.

Starting main the data collection (the speech comprehension test) a laptop (*HP Elitebook, 14''*) was placed on the table with the screen approximately 60 cm in front of the child. A pair of circumaural sound-attenuating earphones (*Seinheisser HDA 200*) were placed covering the children's ears and attenuating environmental sounds by at least 30 dB. The speech signal was presented at a sound pressure level (SPL) of 70 dB. The equipment had been calibrated according to IEC 60318-2 and ISO 389-8 with a *Brüel and Kjaer 2209 Sound Level Meter* and a *4134 microphone* in a *4153 coupler* (IEC: 1998, ISO: 2004). A 1 kHz tone with the same average RMS as the speech signal (see above) was used to verify the sound pressure levels.

While responding to the associated comprehension questions (CELF) following each narrative, and before the next narrative starts, the headphones were removed. Questions were given verbally by the experimenter in the order they

appear in the CELF test sheets (three content questions followed by one inference question and one summary question). Answers were noted on a test sheet concealed from the child.

To become familiarized with the task, each child listened to a practice narrative (audio-only) and answered the associated CELF-questions prior to the experimental trials. Next, following the practice narrative, the child was presented with the sequence of the three experimental narratives (A, B, and C) in the order determined by the balanced experimental list).

After completing the passage comprehension test, the headphones and laptop were removed. The children were then given a pencil and presented with the self-assessment questionnaire (VAS ratings). After completing the three self-assessment items in the questionnaire, they were presented with the word cloud and asked to review all the words and encircle those describing the speaker in the passage comprehension test.

At the very end of the procedure, the children underwent a pure tone hearing screening using an *Entomed SA210 Audiometer* and *PD-81 Headphones*. Children requiring levels higher than 25 dB to hear tones on one or more of the frequencies 250 Hz, 500 Hz, 1 kHz, 2 kHz, 3 kHz, 4 kHz, or 6 kHz were excluded from any analysis.

3.4 Ethical considerations

Informed consent was obtained from the legal guardians of all children who participated in the study. The duration procedure was kept short, to assure that it would not be overly taxing or distracting from their regular curriculum. Children were informed they were participating on a voluntary basis and allowed to terminate the experiment and leave at any point. The loudness in the headphones was calibrated below any potentially hazardous level. Children not fulfilling inclusion criteria but whose parents consented to their participating were still allowed to perform the test procedure in order not to experience any feeling of exclusion. (Their data was afterwards excluded from the data analyses.)

The study adheres to the Helsinki ethical guidelines (General Assembly of the World Medical Association, 2014). Identities of the participating children were anonymized directly following the data collection. Summary data files, identification keys, and original data collection forms were stored separately; the former on password protected folders with restricted access and the latter in separate locked cabinets.

3.5 Analysis

All analyzes were performed in *R version 3.5.2* (R Core Team, 2018). Between group effects were analyzed by linear regression comparing the aggregate scores per child. Mixed (within-subject by group) effects were analyzed by logistic binomial regression modelling using the *lme4* R package (Bates, Mächler, Bolker, & Walker, 2014). Coefficients of determination of mixed models (conditional and marginal R^2: Nakagawa, Johnson, & Schielzeth, 2016) were calculated using the *MuMIn* R package (Barton, 2013). Coefficients of determination for Poisson models (variance function-based, Zhang, 2017) were calculated using the *rsq* R package.

4 Results

In total, 102 students performed the experiment with seven being excluded from the data analysis due to not passing the hearing screening. Of the remaining 95 participants included in the analysis, 32 were assigned to the VIRTUAL (virtual speaker) condition, 32 to the AUDIO (audio-only) condition and 31 to the VIDEO (real video) condition. Their ages ranged from range 96 months to 124 months with an equal age distribution between groups (Kruskal-Wallis $\chi^2 = 3.11$, $p = .21$). The gender distribution (in total 48 females and 47 males) was also matched between groups (Chi-squared test: $\chi^2 = 0.147$, $p = .929$). The order of the narratives was balanced between the groups (Chi-squared test: $\chi^2 = 0.364$, $p = .985$).

To address research questions RQ1a and RQ1b (whether visual presentation of a virtual speaker facilitate speech comprehension in babble noise compared to an audio-only presentation, and to a comparable degree to a video of a real speaker) and RQ2 (whether there are any adaption effects related to exploiting visual speech clues) we analyzed the scores on content questions with logistic (binomial) regression, modelling the logarithm of the odds of correct answers as a sum of group (with VIRTUAL as reference level) and interactions between group and a narrative's order in the sequence centered on the middle (second) narrative and a random intercept factor per participant. The test revealed a significant negative effect of AUDIO compared to VIRTUAL ($\beta = -0.586$, $z = -2.379$, $p = .017$) and no significant effect of VIDEO compared to VIRTUAL ($\beta = -0.266$, $z = -1.064$, $p = .287$). The mean proportional score aggregated over all narratives was .691 ($SD = .184$) for the VIRTUAL group, .569 ($SD = .218$) for the AUDIO group and .638 ($SD = .229$) for the VIDEO group. This result indicates that seeing a virtual speaker facilitated speech comprehension, and no worse than seeing a video of a real speaker. There was a significant positive effect of order within the VIRTUAL group ($\beta = .374$, $z = 2.267$, $p = .0234$), while no significant effect of order was found within the AUDIO condition ($\beta = .094$, $z = .612$, $p = .54$) or VIDEO condition ($\beta = .128$, $z = .803$, $p = .422$). The conditional R^2 for the model was .137. This result indicates an adaptation effect to the audiovisual stimuli, for the VIRTUAL group. Fig. 2 shows the distributions (M and SE) of correct answers for the three groups over the sequence of three narratives.

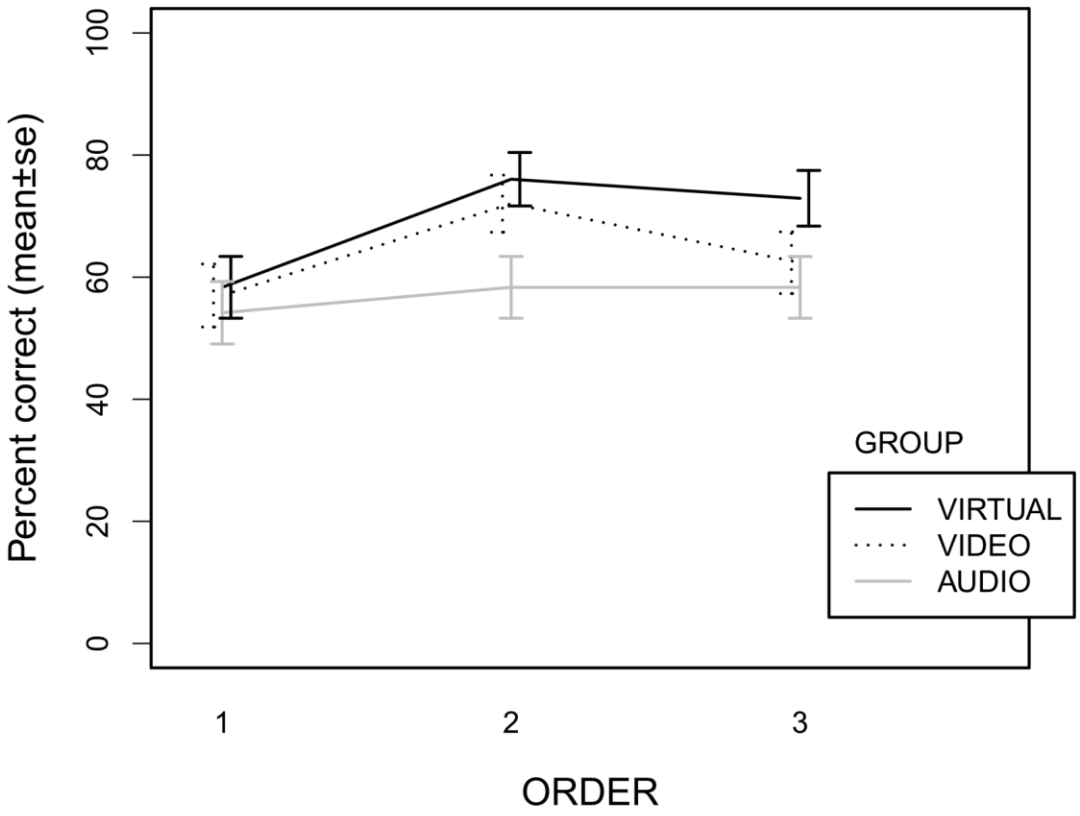

Fig. 2. The mean proportion of correct answers to the content questions for each of the three experimental groups (corresponding to the percentage of correct answers to content questions) over the order of the three subsequent narratives.

To investigate RQ3 – if seeing the virtual speaker increases the perceived effort involved in listening and sustaining to the narrative – we analyzed the VAS ratings on the first and second question in the short questionnaire ("*How easy or difficult was it to hear what the speaker was saying* (VAS.1) */ concentrate on listening to the narrative?* (VAS.2)"). The ratings on both questions tended to the 'easy' side of the (unmarked) center point of the VAS for all three groups, i.e. the mean rating was greater than 0.0 (the midpoint between difficult [-1.0] and easy [1.0]). There was, however, great variance between individual ratings within all groups both on VAS.1 (VIRTUAL: M = .226, SD = .478, AUDIO: M = .105, SD = .467, VIDEO: M = .256, SD = .416*)* and VAS.2 (VIRTUAL: M = .220, SD = .543, AUDIO: M = .107, SD = .626, VIDEO: M = .194, SD = .469*)* ratings. A linear regression modelling the ratings on VAS.1 as a function of group (R_{adj}^2 = -0.001, $F(2,92)$ = .982, p = .379) revealed no significant differences, neither between the VIRTUAL and AUDIO conditions (β = -0.121, t = -1.066, p = .289) nor between the VIRTUAL and VIDEO conditions (β = .030, t = .261, p = .794). Likewise, a linear regression modelling the ratings on VAS.2 as a function of group (R_{adj}^2 = -0.014, $F(2,92)$ = .370, p = .691) revealed no significant differences, neither between

the VIRTUAL and AUDIO conditions (β = -0.113, t = -0.822, p = .413) nor between the VIRTUAL and VIDEO conditions (β = -0.026, t = -0.186, p = .413). These results give no indication that visual presentation of a virtual speaker increases (self-assessed) listening- or attentional effort, compared to an audio-only presentation or a video of a real speaker..

To address RQ4a – whether the virtual speaker elicit more negative or positively perceived social traits compared to only hearing the audio or watching a video recording of a real speaker – we analyzed the number of encircled positive (min = 0, max = 6, median = 2) and negative words (min = 0, max = 4, median = 0) in the word cloud. Kruskal-Wallis tests revealed that significantly more positive than negative words encircled by participants in the AUDIO group (χ^2 = 6.57, p = .010), but not in the VIRTUAL (χ^2 = 5.95, p = .114) or VIDEO (χ^2 = 3.62, p = .164). A Poisson regression modelling number of positive words as a function of group (pseudo R^2 = .018) revealed no significant differences between the VIRTUAL and AUDIO conditions (β = .186, z = 1.290, p = .197) or between the VIRTUAL and VIDEO conditions (β = .098, z = .659, p = .510). A Poisson regression modelling number of negative words as a function of group (pseudo R^2 = .050) did reveal a significant difference between the VIRTUAL and AUDIO conditions (β = -0.794, z = -2.469, p = .014) but not between the VIRTUAL and VIDEO conditions (β = -0.407, z = 1.417, p = .165). This result gives some indication that the visual presentation of the virtual speaker elicited a more negative perception of the speaker. Fig. 3 gives an overview of the encircled words for the three groups.

To address RQ4b - regarding the how perceived social traits affect speech comprehension outcome – we looked at the correlations between number of encircled words and aggregated comprehension score per participant (pooled for all groups) and found the number of positive words to be positively correlated with score (Spearman's ρ = .210, p = .041) but no significant correlation between the number of negative words and score (Spearman's ρ = -0.050, p = .631).

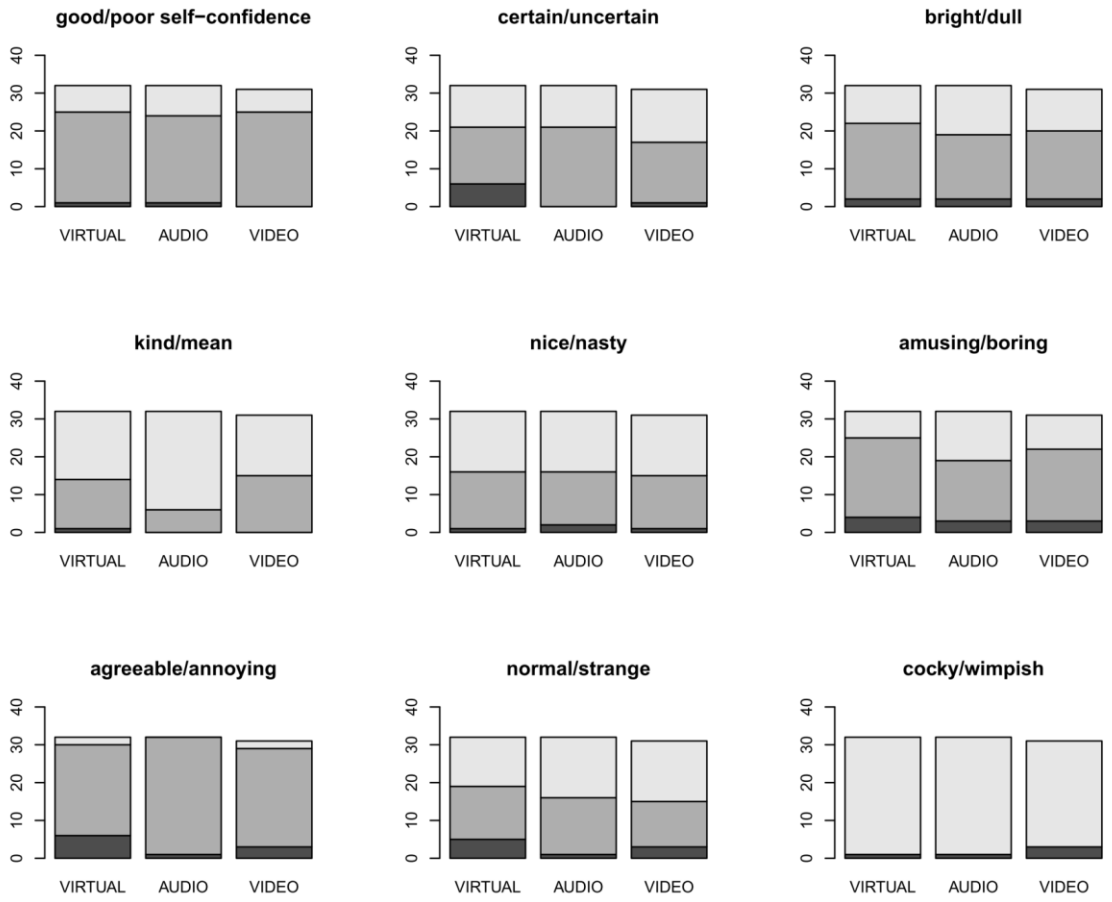

Fig. 3. Distribution of the encircled words (describing social traits) in the word cloud for the experimental conditions VIRTUAL (virtual speaker), AUDIO (audio-only), and VIDEO (video of real speaker). Black areas represent number of participants encircling the negative word, white areas number of participants encircling the positive word, and grey number of participants encircling none of the words in the pair.

5 Discussion

In order to investigate the effects of a virtual (digitally rendered) speaker or presenter, we performed an experimental study comparing a virtual speaker with a real video and an audio-only representation of the same speaker. The virtual speaker was generated from motion capture data and lip movement animations based on depth map, video, and audio recordings of a real speaker. The video recording used for the generation of the virtual speaker was the same video that was used in the 'real video' condition and all three conditions (virtual video, real video, and audio-only) shared the same audio recording. The primary outcome variable was scores on a listening comprehension test, comparing the virtual speaker video presentation with the corresponding real video and audio-only presentations. The purpose for this was to corroborate the inconclusive results with regards to the effect of the

virtual speaker obtained in a previous study (Nirme et al., 2018), and to investigate if the exploitation (for speech comprehension) of visual speech cues from a virtual speaker requires adaptation. For the two secondary (and more explorative) outcome variables, we measured self-assessed listening and attentional effort of watching/listening to the three presentational formats using VASs (Visual Analog Scales) and perceived social traits elicited by the three presentational formats by means of a 'word cloud' selection test (see section *Materials and Methods*).

5.1 Research questions

Results confirmed a positive effect on listening comprehension in background multitalker babble noise of watching the virtual speaker compared to listening to the speech in the audio-only condition (see RQ1a). The difference in mean scores between the groups (virtual speaker and audio-only) was 12.2%, comparable to the 9.7% found in the previous study (Nirme et al., 2018). This result, thus, supported our prediction that an increase in statistical power (i.e. greater number of narratives and content questions per participant) would produce a more robust result, even though the effect on speech comprehension score (in babble noise) was moderate and there still a lot of unexplained variance in our model.

Next, relating to research question 1b (RQ1b), the results did not indicate any difference in listening comprehension between watching a digitally rendered video with the virtual speaker and the corresponding real video recording (identical to the video recording used to generate the virtual speakers). Thus, our prediction for RQ2b, that the video with a virtual speaker would differ negatively compared to a video with a real speaker, was not supported – at least not in ways that are relevant or sufficient to weaken support of speech comprehension. One should, however, be careful before generalizing from this result. The outcome might have been different using high-quality video recordings or a virtual speaker with scripted or synthesized animation schemes having less fidelity to the speakers naturally occurring movements.

Another possible explanation for this lack of difference between the virtual and real video presentation (RQ1b) in terms of speech comprehension is related to a potential advantage of using virtual speakers as presenters. The negative impact on audiovisual speech processing by incorrect or 'awkward' visual cues in the artificially generated virtual speaker might be counteracted by other speech relevant visual cues presented in a more distinct or idealized way. A virtual speaker will by definition never surpass a video recording in terms of realism. It could, however, provide stronger or more salient visual cues while filtering out potentially distracting natural occurring visual noise in the animation process. Previous studies have found some support for the potential to emphasize and exaggerate certain visual features to improve or investigate audiovisual speech processing (Alghamdi et al., 2017) and within the production of animated film, the strategy of emphasizing basic emotional expressions and key gestures by exaggeration is manifested as one of the 12 basic principles of animation (Lasseter, 1987; Thomas and Johnston, 1995). Using a virtual speaker like the one in the current study also has the potential advantage – as a research tool – to be able to modulate specific movement or features without necessarily drawing explicit attention to them (as might be a risk when presenting stylized abstractions of facial movement or edited video recordings).

Speech recognition is typically studied in controlled experimental settings with a simple repetition task, allowing strict control of external factors (e.g. signal-noise ratio, phonemes, and word familiarity) as well as internal factors (e.g. load on a specific cognitive or perceptual capacities and perceptual priming). A similarly reductionist approach to studying speech comprehension risks lowering ecological validity. Results obtained when studying different effects on aspects of comprehension in separate settings (speech recognition, attention, encoding, or recall)may not be generalizable to coordinated multimodal processing in a real-world environment such as a classroom. Some internal factors crucial to comprehension, such as previous word and genre knowledge or associations, are particularly hard to control for. As for external factors, control is still feasible – but stimuli consisting of, for example, longer passages (compared to isolated words or sentences) are more difficult to produce without inconsistencies for different audiovisual experimental conditions. Video editing enables precise control of stimuli, but the range of possible experimental manipulations that can be said to approximate a realistic listening situation is limited. Also, the cognitive and perceptual load associated with listening in unpredictable and rich perceptual environments (such as classrooms) might deviate from the corresponding cognitive and perceptual load in controlled settings, making findings less generalizable. How to balance experimental control and ecological validity is an important and lasting debate within behavioral science (Brunswik, 1956; Berkowitz & Donnerstein, 1982; Conway, 1991; Araújo, Davids, & Passos, 2007). In contrast to using manipulated video recordings, a virtual speaker can easily be presented from any angle or even in an immersive 3-dimensional virtual environment (Bailenson, et al., 2008) which might accentuate certain visual features. It should here be pointed out that the viewing angle at which the children watched the virtual speaker in this study was different than the frontal view used in our previous study as we strived to match the virtual and real video as close as possibly except for the experimental difference in appearance and movements. The gaze behavior of the virtual speaker, was animated (based on recordings) to be as similar as possible to the real speaker. Eye contact seems to have a crucial pragmatic role in face-to-face verbal communication (Richardson & Dale, 2005; Hanna & Brennan, 2007).

Turning to the second research question (RQ2), the analysis of the interaction between the experimental conditions and the order of the narratives revealed a significant improvement on comprehension questions over the sequence of narratives for the virtual group alone. This result supported our hypothesis that there is an adaptation effect involved in exploiting visual speech cues from a virtual speaker, resulting in improved speech comprehension in babble noise.

A possible explanation for the absence of adaptation effect for the videos of the real speaker is because it represents a more familiar stimulus. This would be in line with findings in speech recognition research showing stronger adaptation effects for atypical stimuli such as exaggerated lip movements (Alghamdi et al., 2017) and point light representation (Rosenblum, Johnson & Saldaña, 1996). Trude and Brown-Schmidt (2012) also found indications that adaption to linguistic accents was speaker-specific and cued by the visual representation of the speaker. The adaptation effect for the group presented to the virtual speaker observed in the current study was mainly driven by an increased benefit associated with the transition from the first to the second narrative. Moreover, it is worth pointing out that at the first narrative in the sequence, presumably before any adaptation has taken effect, the mean listening comprehension score was almost identical for all of the three experimental conditions.

As for our third research question (RQ3), we found no indication of visual presentation (neither real nor virtual) distracting or disrupting attention to the spoken content or making listening more effortful – as measured by self-assessment on visual analog scales (VASs). Direct measures of attention or effort, such as performance on a parallel task (Mishra et al. 2013a), pupil dilation (Koelewijn, Zekveld, Festen, & Kramer, 2012), or gaze tracking (Gullberg & Holmqvist, 2006) could provide stronger evidence. Both virtual speaker and video presentation adds perceptual information that is irrelevant to speech processing and could be described as increasing "incidental processing" (Mayer & Moreno, 2003). However, if audiovisual speech integration is a primarily independent and automatic process, one would not expect any increased load by adding irrelevant visual input. It is also worth mentioning that the 'principles of multimedia learning' (Mayer & Moreno, 2003) or the underlying cognitive load theory (Sweller, 1998) are not directly concerned with audiovisual speech processing or an embodied presentation of a speaker, but of the 'multimodal' presentation of educational content.

Regarding research question 4a (RQ4a), the word-cloud responses and the analysis of the summed valence scores gave no clear indications that the social traits of the virtual speaker was perceived any differently than of the real speaker presented in the video, but that the virtual speaker elicited more negative social traits than listening to the voice alone. However, despite perceiving the virtual speaker more negatively, participants' speech comprehension in noise (see RQ1a) benefited from seeing it. This indicates that, at least in this particular listening task, the benefit of seeing the virtual speaker was not driven by any 'social agency effect' (Moreno et al., 2001) but as audiovisual cues to speech processing. This is in line with the dominant theories that propose separate pathways for different aspects of face perception, as well as the absence of any effect of visual presentation in quiet settings in our previous study (Nirme et al., 2018). Regarding research question 4b (RQ4b), the weak correlation between the number of positive words selected to describe the speaker and speech comprehension scores on the other hand hints that the perception of the speaker as a social agent might be related to engagement with speech content. While Moreno et al. (2001) proposes that social agency arises through reciprocal interaction with an animated agent, as opposed to passively listening to a speaker as in the current study and the animated agents in the study by Domagk (2010) that were not interactive beyond simply introducing topics. Nevertheless, this study showed that positive social traits can affect learning outcomes.

The results of our analysis of the perceived social traits should however be interpreted in light of the nature of our measure. The contexts of the subjective judgements were arguably different between the three groups; the audio-only condition leaves much to the listener's own imagination, while an actual virtual speaker is a specific and definitive statement of the speaker's visual look (Gulz & Haake, 2006; Haake, 2009). The virtual speaker in this experiment constitutes a single, specific instance out of a broad range of possible renditions, why it is fully possible that other visual instances of a virtual speaker could different results.

5.2 Limitations

The study had some limitations that are worth pointing out. The experimental procedure was purposefully short. The rationale for this was to limit the duration of the experimental procedure per child, partly to be able to include

enough students to get statistical power – especially with regard to RQ1 – with the main purpose to further investigate the inconclusive findings of the previous study (Nirme et al., 2018). Furthermore, given the study's second main focus on adaptation effects (RQ2), we did not want to interfere or interrupt the listening comprehension part of the procedure and avoid confounding any adaptation effects with fatigue.

The study incorporated two ad-hoc subjective measures of perceived effort and perceived social traits. We did not include any corresponding direct measures, nor other baseline measures related to listening comprehension such as working memory capacity or other executive functioning (Lyberg-Åhlander et al, 2015; Sörqvist, 2010). Again, this was partly due to prioritizing the length of the procedure in order to avoid fatigue in the participating children. Furthermore, baseline measure probing working memory would most certainly be correlated with comprehension outcomes (Daneman & Merikle, 1996; Kim, 2016). The randomization in the current study should however allow us to assume equal effects of working memory capacity between groups and cancel out any working memory induced effects in the interpretation of our results. It is also unclear how working memory capacity might interact with listening comprehension under audiovisual conditions (Mishra et al., 2013a,b) and a viable topic for future study, however beyond the scope of the current work.

We did not include a control condition without babble noise, which might have allowed us to directly address whether seeing the virtual speaker benefits comprehension by allowing listeners to overcome the effect of babble noise or by some more general effect such as increasing engagement with the listening task. However, our previous study (Nirme et al., 2018) using a 2×2 within-subject design (each participant listening with and without noise and with and without virtual speaker) revealed no effect of seeing the virtual speaker in the absence of background multitalker babble noise as well as a clear main effect of background multitalker babble noise. The latter has also been demonstrated in a number of previous studies (Klatte, Lachmann, & Meis, 2010; Ljung et al., 2009; Valente et al., 2012).

A higher number of participants might have revealed more subtle differences in the comprehension and perception of the real and virtual speakers. We nevertheless deemed the number of participants (102) within the sampled age group to be sufficient to address our current research questions and on par with previous studies investigating listening comprehension (Klatte, Lachmann, & Meis, 2010; Valente et al., 2012). In the current study we focused on the specific age group (8- to 10-year-olds) in order to have our results comparable to that of our previous study (Nirme et al., 2018) and for the selected narratives in the comprehension test to remain applicable. Including other age groups in our study would possibly have modified the outcome. It is, for example, well established that children are more sensitive to noise than adults (*Fallon,Trehub, & Schneider, 2000; Valente et al., 2012*). Other findings have also indicated varying effects of audio-visual speech processing in relation to age (Dockrell & Shield, 2004; Jerger, Damian, Spence, Tye-Murray, & Abdi, 2009).

Both the age and working memory capacity have been linked to listening effort and fatigue (Pichora-Fuller et al., 2016) as well as mind wandering and metacognitive self-regulation (McVay & Kane, 2009; Mrazek, Phillips, Franklin, Broadway, & Schooler, 2013), and might thereby have influenced the observed adaptation effect in our sample by confounding the progression of comprehension score over the order of the narratives. However, as

mentioned above, the duration of the experimental procedure was kept relatively short in order to minimize such effects.

It is also probable that the type of noise in the current study (background multitalker babble noise) might have had some effect on the outcome. We argue that this type of babble noise is representative for the emulation of the listening conditions children might encounter in real school settings and that it is more challenging to overcome compared to static noise (e.g. white noise). While static noise causes 'energetic masking', babble noise also causes 'informational masking' as the source of the noise cannot be as readily distinguished from the speaker's voice (Brännström et al., 2018; Brungart, 2001). Informational masking has been shown to be particularly challenging for children (Wightman & Kistler, 2005). Effect of seeing the speaker while listening in babble noise are also less studied where previous studies targeting audiovisual speech processing have mostly involved static noise (e.g. Agelfors et al., 1998; Möttönen et al., 2000).

6 Conclusions

Even if there are limitations to the study, the results support that watching a virtual speaker can benefit speech comprehension in realistic classroom noise. The improvement in speech comprehension from seeing the virtual speaker compared to only hearing the voice, despite being associated with more negative social traits, suggests that audiovisual integration can support speech comprehension independently of children's social perception of the speaker.

Regarding our result indicating an adaption effect – that the benefit of the virtual speaker for speech comprehension come into effect after the first trial – it is worth considering the need to be introduced and familiarized with virtual speakers, both in research and real-world applications.

Unlike a real speaker or a video recording of a real speaker, the use of a virtual (digitally rendered) speaker makes it possible to manipulate a wide range of behavioral factors in a precise and controlled way. In this, virtual speaker technology can provide a promising research tool, that can be used to deconstruct and investigate audiovisual support of speech processing or other aspects of communication and learning. The natural compatibility of a virtual speaker with a virtual environment further expands the level of experimental control and range of possible research questions in paradigms that connect expertise from different disciples (Veletsianos, Heller, Overmyer, & Procter, 2010). Beyond the implementation as research tools, virtual agents may have direct pedagogical benefits (Gulz, 2004). However, what is the optimal mode of presentation depends on their intended pedagogical role (Gulz & Haake, 2006).

7 References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K. E., & Öhman, T. (1998). Synthetic faces as a lipreading support. *Proceedings of the 5th International Con-ference on Spoken Language Processing (ICSLP'98)*. Sydney, Australia, 3047–3050.

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.

Al Moubayed, S., Beskow, J., & Granström, B. (2009). Auditory visual prominence. *Journal on Multimodal User Interfaces, 3*(4), 299-309.

Araújo, D., Davids, K., & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comment on Rogers, Kadar, and Costall (2005). *Ecological Psychology: A Publication of the International Society for Ecological Psychology, 19*(1), 69-78.

Bailenson, J. N., Yee, N., Blascovich, J., Beall, A. C., Lundblad, N., & Jin, M. (2008). The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *The Journal of the Learning Sciences, 17*(1), 102141.

Barton, K. (2013) MuMIn: multi-model inference. R package version 1.43.6. Available from: https://CRAN.R-project.org/package=MuMIn

Bates, DM., Mäechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(7), 2414-2417.

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. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science, 14*(6), 592-597.

Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning, 5(2)*, 123-152.

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.

Bradley, J. S., & Sato, H. (2008). The intelligibility of speech in elementary school classrooms. *The Journal of the Acoustical Society of America, 123*(4), 2078-2086.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Bruce, V., & Young, A. (1986). Understanding face recognition. British Journal of Psychology, 77(3), 305-327.

Brännström, K. J., Holm, L., Lyberg-Åhlander, V., Haake, M., Kastberg, T., & Sahlén, B. (2015). Children's subjective ratings and opinions of typical and dysphonic voice after performing a language comprehension task in background noise. *Journal of Voice*, *29*(5), 624-630.

Brännström, K. J., von Lochow, H., Åhlander, V. L., & Sahlén, B. (2018). Immediate passage comprehension and encoding of information into long-term memory in children with normal hearing: The effect of voice quality and multitalker babble noise. *American journal of Audiology*, *27*(2), 231-237.

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

Burleigh, T. J., & Schoenherr, J. R. (2014). A reappraisal of the uncanny valley: categorical perception or frequency-based sensitization? *Frontiers in Psychology, 5*(1488), 1–19

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. Nature Reviews. *Neuroscience, 6*(8), 641-651.

Cassell, J., Sullivan, J., Prevost, S. & Churchill, E. (eds) (2000) *Embodied Conversational Agents.* Cambridge, MA: MIT Press.

Cassell, J., & Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, *13*(1-2), 89-132.

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

Choi, S., & Clark, R. E. (2006). Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. *Journal of Educational Computing Research, 34*(4), 441-466.

Clark, R. C., & Mayer, R. E. (2016). e-*Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning.* (3rd ed., pp. 179-200). London: John Wiley and 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.

Conway, M. A. (1991). In defense of everyday memory. *The American Psychologist, 46*(1), 19-26.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422-433.

Dockrell, J.E. & Shield, B. (2004). Children's perceptions of their acoustic environment at school and at home. *The Journal of the Acoustical Society of America. 115*(6), 2964-2973.

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

FaceShift Studio 2015 [Computer software] (2015) [Discontinued]

Fallon, M., Trehub, S. E., & Schneider, B. A. (2000). Children's perception of speech in multitalker babble. *The Journal of the Acoustical Society of America, 108*(6), 3023-3029.

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, 53*, 18-33.

Garau, M., Slater, M., Pertaub, D.-P., and Razzaque, S. (2005). The Responses of People to Virtual Humans in an Immersive Virtual Environment. *Presence: Teleoperators and Virtual Environments, 14*(1), 104–116.

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.

General Assembly of the World Medical Association. (2014). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *The Journal of the American College of Dentists, 81*(3), 14.

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*(1), 124-132.

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.

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 and 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. (2005). Social enrichment by virtual characters – differential benefits. *Journal of Computer Assisted Learning, 21*(6), 405-418.

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.

Haake, M. (2009). Embodied pedagogical agents – from visual impact to pedagogical implications. PhD thesis in design sciences, Lund University. Lund, Sweden: E-husets tryckeri.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology, 11*(2), 79-87.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*(4), 596-615.

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.

Heidig, S., & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review, 6*(1), 27-54.

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, 57*(1), 1199-1203. Los Angeles, CA: SAGE Publications.

Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: A new multimodal picture–word task. *Journal of Experimental Child Psychology, 102*(1), 40-59.

Johnson, W. L. & Lester, J. C. (2015). Face-to-face interaction with pedagogical agents. Twenty years later. *International Journal of Artificial Intelligence in Education, 26*(1), 25-36.

Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America, 93*(1), 510-524.

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.

Kim, Y. S. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology*, *141*, 101-120.

Klatte, M., Lachmann, T., & Meis, M. (2010). Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting. *Noise and Health, 12*(49), 270-282.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291-300.

Kristiansen, J., Lund, S. P., Persson, R., Shibuya, H., Nielsen, P. M., & Scholz, M. (2014). A study of classroom acoustics and schoolteachers' noise exposure, voice load and speaking time during teaching, and the effects on vocal and mental fatigue development. *International Archives of Occupational and Environmental Health, 87*(8), 851-860.

Lasseter, J. (1987). Principles of traditional animation applied to 3D computer animation. *ACM SIGGRAPH Computer Graphics, 21*(4), 35-44.

Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., & Bhogal, R.S. (1997). The Persona Effect: Affective impact of animated pedagogical agents. *Proceedings of CHI 97,* 359-366. New York: ACM.

Lewis, J. P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F. H., & Deng, Z. (2014). Practice and theory of blendshape facial models. *Eurographics State of the Art Reports, 1*(8), 199–218.

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 *International Conference on Intelligent Virtual Agents (IVA 2015)*, 442-451. Berlin: Springer.

Ljung, R., Sörqvist, P., Kjellberg, A., & Green, A. M. (2009). Poor listening conditions impair memory for intelligible lectures: Implications for acoustic classroom standards. *Building Acoustics*, *16*(3), 257-265.

Lusk, M. M. & Atkinson, R. K. (2007), Animated pedagogical agents: does their degree of embodiment impact learning from static or animated worked examples? *Applied Cognitive Psychology, 21*, 747-764.

Lyberg-Åhlander, V., Brännström, K. J., & Sahlén, B. S. (2015). On the interaction of speakers' voice quality, ambient noise and task complexity with children's listening comprehension and cognition. *Frontiers in Psychology, 6*, 871.

Lyberg-Åhlander, V., Haake, M., Brännström, J., Schötz, S., & Sahlén, B. (2015) Does the speaker's voice quality influence children's performance on a language comprehension test? *International journal of speech-language pathology. 17*(1), 63-73.

Lyberg-Åhlander, V., Rydell, R., & Löfqvist, A. (2011). Speaker's comfort in teaching environments: Voice problems in Swedish teaching staff. *Journal of Voice: Official Journal of the Voice Foundation*, *25*(4), 430-440.

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.

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695-710.

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. (2014). Principles based on social cues in multimedia learning: Personalization voice image and embodiment principles. In *The Cambridge Handbook of Multimedia Learning* (2nd ed.), 345-368. New York, NY: 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, 18*(3), 239-252.

Mayer, R. E., & Moreno, R. (2003). *Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist, 38*(1), 43-52.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 196-204.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013a). Seeing the talker's face supports executive processing of speech in steady state noise. *Frontiers in Systems Neuroscience, 7*, 96.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013b). 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., Spires, H. A., & Lester, J. C. (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*(2), 177-213.

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

Möttönen, R., Olivés, J. L., Kulju, J., & Sams, M. (2000, September). In *Proceedings of the 10th European Signal Processing Conference (Eusipco2000)*, 1-4. Tampere, Finland.

Mrazek, M. D., Phillips, D. T., Franklin, M. S., Broadway, J. M., & Schooler, J. W. (2013). Young and restless: validation of the Mind-Wandering Questionnaire (MWQ) reveals disruptive impact of mind-wandering for youth. *Frontiers in Psychology, 4*, 560.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science, 15*(2), 133-137.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 20170213.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies,43,* 223–239.

Neuman, A. C., Wroblewski, M., Hajicek, J., & Rubinstein, A. (2010). Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults. *Ear and Hearing, 31*(3), 336-344.

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.

Norman, D. (1988). The design of everyday things. NewYork: Basic Books

Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. *Speech Communication*, *51*(7), 630-639.

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. (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. Y. (2011). Visual cues and listening effort: Individual variability. *Journal of Speech, Language, and Hearing Research, 54*(5), 1416-1430.

R Core Team (2013). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. Available from: https://www.R-project.org/

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

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045-1060.

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.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147-1153.

Semel, E. M., Wiig, E. H., & Secord, W. (2004). *CELF-4: Clinical Evaluation of Language Fundamentals Screening Test.* San Antonio, TX: Pearson, PsyhCorp.

Shield, B. M., & Dockrell, J. E. (2008). The effects of environmental and classroom noise on the academic attainments of primary school children. *The Journal of the Acoustical Society of America, 123*(1), 133-144.

Sörqvist, P. (2010). The role of working memory capacity in auditory distraction: A review. *Noise and Health, 12*(49), 217-224.

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.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285.

Thomas, F., & Johnston, O. (1995). *The illusion of life: Disney animation*. New York: Hyperion.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*(1), B13-B22.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes, 27*(7-8), 979-1001.

Valente, D. L., Plevinsky, H. M., Franco, J. M., Heinrichs-Graham, E. C., & Lewis, D. E. (2012). Experimental investigation of the effects of the acoustical conditions in a simulated classroom on speech recognition and learning in children. *The Journal of the Acoustical Society of America, 131*(1), 232-246.

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.

von Lochow, H., Lyberg-Åhlander, V., Sahlén, B., Kastberg, T., & Brännström, K. J. (2018a). The effect of voice quality and competing speakers in a passage comprehension task: Perceived effort in relation to cognitive functioning and performance in children with normal hearing. *Logopedics Phoniatrics Vocology*, *43*(1), 32-41.

von Lochow, H., Lyberg-Åhlander, V., Sahlén, B., Kastberg, T., & Brännström, K. J. (2018b). The effect of voice quality and competing speakers in a passage comprehension task: Performance in relation to cognitive functioning in children with normal hearing. *Logopedics Phoniatrics Vocology*, *43*(1), 11-19.

Whitling, S., Rydell, R., & Åhlander, V. L. (2015). Design of a clinical vocal loading test with long-time measurement of voice. *Journal of Voice*, *29*(2), 261.e13 - 261.e27.

Yung, H. I., & Paas, F. (2015). Effects of cueing by a pedagogical agent in an instructional animation: A cognitive load approach. *Journal of Educational Technology & Society, 18*(3), 153.

Zhang, D. (2017). A coefficient of determination for generalized linear models. *The American Statistician, 71*(4), 310-316.