Looking for a more effective way of learning: reading versus listening.
How different modes of presentation of factual texts using synthesized speech influence text learning for Swedish 8th-graders

Olga Fredriksson

handledare/supervisors
Agneta Gulz
Magnus Haake

KOGM20

2015-09-15
Looking for a more effective way of learning: reading versus listening.
How different modes of presentation of factual texts using synthesized speech influence text learning for Swedish 8th-graders

Olga Fredriksson
deograciass@me.com

Traditionally reading new material in a school textbook has been a standard way of learning and sometimes listening to a teacher has been a way to get introduction to a topic. Lately, due to growing knowledge about cognitive aspects of learning, increasing interest has been paid to different ways of listening to new material. New technologies for listening have been developed, one of them being that of animated pedagogical agents.

The main objective of this study was to compare the learning outcome for 8th grade students when they i) read a social science text on their own, ii) listened to and read the text simultaneously, iii) just listened to the text and iv) watched a video with an animated virtual human reading the text. Synthesized speech was used for all speech instances.

Overall results of the experiment revealed that reading on one’s own, closely followed by reading and listening to a text simultaneously, were more effective ways of learning than the situations where no text support was available. Nevertheless, the students with the lower scores on the reading tests were favored by auditory presentation modality on the whole, especially by listening to a text while reading it simultaneously. Listening to a virtual human did not favor students’ learning as a whole. It was demonstrated, though, that listening to a virtual human helped many students with lower scores on auditory tests as well as on reading tests to better understand and remember the contents of the texts.

The main conclusion of this paper is that it is not advisable to use one single teaching method for a whole class. Teachers should apply various pedagogical approaches in a class of students, taking into consideration their individual learning styles, thus, differentiating the mode of presentation of new material for students.

1 Introduction

There are two common ways of introducing new material in theoretical subjects at school: one of them is when students read themselves about a phenomenon in a book, another is that a teacher introduces new material in lecture form. In both cases the teacher expects that the students will understand and remember the material if they are thorough, attentive and interested in learning.

The reality is, however, that some students have difficulties reading the material on their own, and others have difficulties listening to their teachers. There are also students who actually do learn best by listening, but have difficulties in concentrating on listening while being in a large classroom with many other students. These students need to be shielded from external stimuli to be able to learn.

Computers today offer extensive possibilities for individualizing learning. It has already been demonstrated that watching a video lecture on a computer helps many students, because they can concentrate better on the task or they can watch the video at home and repeat the material (Schacter & Szpunar, 2015). However, video recording has several disadvantages: it is time consuming for teachers to produce their own videos and it is difficult to make changes of the recorded material in order to adapt it to the needs of individual students. There are also teachers who have specific ways to read or speak that makes it difficult for students to apprehend what they hear and that makes the video-recorded lectures less attractive. Some other teachers simply do not feel comfortable with being video-recorded and abstain from using new techniques.

One solution to this problem can be the application of synthetic speech and/or animated agents, e.g. virtual humans. Synthetic speech has been used for some time already and has through the years become better and better. Using this technique you can easily scan a text, make some minor corrections and listen to it at a preferred speech rate, pausing or rewinding. You can also follow a written text either on screen or on paper. A scanned text can also be read as automatically generated synthetic speech by a virtual human (Mattheyses & Verhelst, 2015), making it possible to adapt and adjust the agent, by choosing a male or female speaker, a visual appearance, speech rate, the tone of voice etc. Facial and/or hand gestures and head movements add a communicative aspect to the listening process. The students may be able to study on a one-to-one basis and several learning styles may be covered by this mode of presenting new material.
Synthetic speech is widely recommended and used in schools today (e.g. ViTal, CD-ORD etc.), Swedish services of read-in teaching media and fiction, such as Inläsningstjänst and Legimus, use synthetic speech for their daily on-line newspapers and, to some extent, for talking books and audio books for children.

Another major issue concerning the presentation of new material for students is concurrent reading and listening. Swedish Agency for Accessible Media, MTM, recommend simultaneous reading and listening in order to reinforce understanding and remembering of the contents (http://mtm.se). Recent research has demonstrated that reading and listening to some extent have similar underlying processes (Gazzaniga, M. S., Ivry, R.B., & Mangun, G. R., 2009). As a consequence of the increased knowledge in this field of research, Swedish national comprehension tests in Swedish and English allow certain groups of students to listen to the texts when they are tested for reading comprehension.

But there are studies that argue that reading and listening to a text does not give the positive results one is hoping for. Visual information may interfere with auditory information, and what happens in such cases is that a student unconsciously chooses either reading or listening, thus eliminating the synergy effect (Hilbert, Nakagawa, Puci, Zech, & Buhner, 2014; Moussa-Inaty, Ayres & Sweller, 2012). These claims are based on an assumption that two channels, visual and auditory, cannot be used for the same task; that it creates a redundancy effect.

The aspects of presenting new material and their impact on learning described above have prompted this research. The present study investigated the learning outcome after four different modes of presenting new material: reading, reading and listening simultaneously, listening without any text support and listening to a virtual human. Synthetic speech was used for all the listening modes.

The first question raised in this study was whether reading and listening to a text simultaneously on a computer screen would give a better learning outcome for students compared to reading on their own, due to the auditory reinforcement (hypothesis 1). The second question examined was the effect of synthetic speech on the learning outcome. In particular, it was proposed that reading a text on one’s own would give a better learning outcome than listening to synthetic speech without any visual text support (hypothesis 2). If this was true, would listening to an animated agent with synchronized facial gestures and head movements, a virtual human, improve the learning outcome (hypothesis 3)?

Additionally, some specific questions, concerning several sub-groups of students, were investigated in the current paper. Reading is a complicated process and in every class there is a certain number of students who have some kind of reading impairments. Will these students be favored by listening instead of reading and which listening mode will be the most appropriate, if this was true (hypothesis 4)? According to Hilbert et al. (2014), individual learning styles are tightly connected to individual differences of perceiving visual and auditory signals (presentation modality), and working memory (processing modality). Thus it was proposed that the students with the highest scores on visual no verbal working memory tests would be benefitted by listening to virtual humans, as the virtual humans’ visual and communicative aspect would enhance understanding and remembering (hypothesis 5). The final question that seemed of interest for this paper was whether auditory working memory influenced the learning outcome, especially, when the synthetic speech was concerned. A group of students with lower results on the auditory working memory tests were selected and it was proposed that these students would show better results on the virtual humans’ tests than on listening without support (hypothesis 6).

To sum up the questions stated above, there are two primary aims of this study: i) to investigate differences in comprehension between reading and listening to synthetic speech, and ii) to examine whether the application of virtual humans using synthetic speech can provide an effective way of learning for secondary school students.

The next chapter will provide a theoretical overview of some aspects of reading, listening and comprehension. It will also lay out the role of working memory and attention for comprehension. Some important features of pedagogical agents (virtual humans) and synthetic speech, will be highlighted.

2 Theoretical background

Reading versus listening

Research into differences and similarities between written and spoken language has a long history. The question has been studied from many distinct angles depending on the scientific discipline, psychology, cognitive sciences, neuroscience, linguistics etc. The objective of this chapter is not to try to cover it all, but to present a short review of some aspects of the topic, that are relevant to the current study.

Reading is a process that, according to Banich and Scalf (2003), is scattered over several regions of the brain, that involves acoustic processing of phonological information as well as syntactic processing, semantic processing of meaning and speech production, to name a few. Reading differs from listening so that when we read we have to process both the visual information (symbols), including orthographic decoding, and acoustic information, including phonological decoding. We also have to activate auditory and visual word forms simultaneously, before we can activate lexical and conceptual information and understand what we read (Gazzaniga et al., 2009).

When we listen to spoken language, acoustic stimuli that we perceive, are transformed into auditory signals and transferred along neural pathways to the primary auditory cortex for further processing. Auditory-temporal processing, i.e. the rate at which we can process auditory information, is considered to be a low-level processing and the phonological awareness – a higher-level processing (Gazzaniga et al., 2009). The subjects, who have difficulties with auditory-temporal processing, i.e. on the low levels of processing, at the later stages, when acquiring the ability to read, have also demonstrated difficulties connecting phonemes and graphemes, though not always (Banich & Scalf, 2003).

Carroll (2008) points out that when we listen, the environment around us is crucial. It is important not to be distracted
by the external stimuli, auditory or visual, to be able to understand and encode the information. The way a person speaks - her stressing, intonation, rate of speech, articulation, dialect - is vital. We get help of the lips movements, facial gestures, and head movements.

Furthermore, Carroll (2008) suggests that speaking is temporal and sometimes very fast, while reading is spatial. This means that we cannot adjust the rate of a speaker as we can adjust the rate of our own reading, according to text difficulty. We can also re-read the text if we need to. Co-articulation leads to overlapping of phonemes and, when we listen, we guess a word, after we have heard the whole word, using top-down strategies. This can be compared to reading, when we guess words from the beginning, as in the "garden path", using bottom-up strategies. But when we intend to comprehend a text as a whole, we apply both strategies, top-down and bottom-up, and we do it simultaneously, thus using strategies for both written and spoken languages.

Comprehension

The ultimate goal of reading and/or listening is to understand what is written or said, i.e. comprehension. As noted by Elwér (2014), reading comprehension is a product of decoding and linguistic comprehension. The former is the aspect of reading accuracy and the latter is seen as a combination of listening comprehension, vocabulary, grammar and morphology. The aspect of decoding is not of interest in the current study and it will be left at that. What is of interest, though, is the aspect of comprehension. In the following, all reference made to reading and/or listening will have comprehension in view.

Catts, Adlof, and Weismer (2006) propose that the primary way to measure linguistic comprehension is by listening comprehension tasks, because reading comprehension is more and more influenced by oral language the older the children become. Catts, Hogan, and Adlof (2005) argue that, in general, reading and listening comprehension tasks do not differ much between them and that listening and reading are strongly correlated, because both depend on the same underlying processes. Catts et al. (2005) quote a broad board of research suggesting that phonological ability is not the decisive factor in comprehension, neither for reading nor for listening. Similarly, Nation, Clarke, Marshall, and Durand (2004) found that comprehension depends on receptive vocabulary and semantic processing, the two areas where poor comprehenders show deficits.

According to Zwaan and Radvansky (1998), a reader or a listener can create a coherent mental model of the situation described in a text, which helps them to remember the text better and make better inferences. The researchers claim that understanding of a situation helps to integrate the information in the text using modality-independent cognitive procedures. The previous knowledge of the situation positively influences the learning outcome. Those who are good at constructing situation models reveal good comprehension ability for written, spoken and visual information, an ability that exceeds any certain specific modality, for example, deficiency in visual word recognition. Hilbert et al. (2014) explain this phenomenon so that the linguistic information is presented in a sequential manner and many things can go wrong. However, in the case of the situation model the situation described in a text is visualized as coherent event and it helps the reader/listener to fill in the gaps.

Working memory, dual channels, dual tasks

Gazzaniga et al. (2009) define learning as a process of acquiring new information either by a simple uptake or repeatedly, and the outcome of learning is memory. That is a memory is generated when something is learned. Memories are processed in different ways, suggest Gazzaniga et al. When you read or listen you receive visual and acoustic signals into the sensory memory, which lasts from milliseconds to seconds. When the information has to be retained for somewhat longer time, seconds or minutes, it is held in short-term memory. Then encoding of new information takes place. This process of acquisition is followed by the process of consolidation of the memories. Finally, the information passes into the long-term memory for storage and is retrieved from there, when necessary.

According to Gazzaniga et al. (2009), working memory, i.e. maintaining and manipulation of the information, is considered as a concept of short-term memory. The information in working memory is stored or processed either on the way from sensory memory into long-term memory or when it is retrieved from there. One of the most widely accepted models of the working memory is the model of Alan Baddeley and Graham Hitch, which proposes the existence of three basic components: the central executive that controls two subordinate systems: i) the phonological loop for processing of acoustic information, and ii) the visuospatial sketch pad for the visual and visuospatial information. The subordinate systems of the working memory, according to this model, are modality specific, dedicated either to the processing of acoustic or visual information.

Hilbert et al. (2014) distinguish between presentation modality and processing modality. Presentation modality is defined as either an acoustic or optical input. Processing modality depends on individual’s cognitive strategies. Hilbert et al. claim that acoustic or optical information does not automatically go to corresponding verbal or nonverbal working memory and thus the presentation mode makes very little difference on the memory results. Instead the incoming information is re-coded in accordance with the individual’s cognitive strategies thus dividing all people into verbalizers and visualizers. It was also found that visualizers used the optical input channel more often, when performing dual tasks. The verbalizers, on the other hand, did not show any significant preference for the acoustic channel in such tasks, compared to visualizers. Hilbert et al. explain this finding that the acoustic information can be maintained in the articulatory loop, a part of the phonological loop, without any conscious attention. These findings can have an important impact on the process of concurrent reading and listening.

A considerable amount of literature has been published on the question of dual channels. According to Paivio (2006) and the dual coding theory, we create internal representations, based on the sensory input, verbal (linguistic) and nonverbal (imagery). What we perceive we have to comprehend and remember. Paivio claims that we remember best objects and pictures, followed by concrete words and, finally, by abstract words. This means that the nonverbal code is stronger than the verbal code. But the two sensory systems, though functionally
Both independent, are involved in dual encoding of the information, verbally and visually, producing synergy effects on recall. This happens when we, for example, can imagine (visualize) concrete words. Abstracts words that are difficult to imagine, are usually not dually coded, unless they are linked and cued by concrete words in phrases or texts.

Preferences with respect to verbal or visual encoding of sensory information are very individual and can depend on many factors, like imagery skills, age or linguistic abilities (Paivio, 2006). But they also depend on the task itself. Westerberg (2004) holds that to perform two tasks simultaneously demands more working memory than two single tasks, performed one after another, due to increased work load and effort. According to Westerberg, it was demonstrated that brain activation during a dual task with two different input modalities (verbal and visual) was the same as during two single tasks, thus creating no mental energy saving. Therefore, Westerberg suggests that working memory is partly domain-free, i.e. a dual task of visual or verbal type, which needs working memory, will be equally difficult/easy to perform.

What we hear or read, we have to keep in working memory, sentence after sentence, so that we get a feeling of coherence of the information we receive. Gathercole and Baddeley (1993) suggest that we comprehend a written or a spoken text first after we have read/heard it, either the sentence or the whole text. The information in this case is first stored in the working memory, so to say on-line, for future use, until the listener can retain the meaning of the message. The capacity of the working memory can in this case play an important role and it can influence the amount of the text remembered. Similarly, Leamy and Sweller (2011) point out that the length of a text is of big importance. In an experiment they have tested texts of different length. The results showed that shorter texts gave better learning outcome.

The question if simultaneous reading and listening (a dual task) produces a redundancy effect or synergy of two verbal aspects (visual and auditory) for better comprehension and recall has been treated by the cognitive load theorists. Leamy and Sweller (2011) hold that if the material is presented in dual modality, both audio and visual, and one modality only recapitulates the material in another modality, no modality effect will be obtained. This leads to a redundancy effect that is brought by simultaneous audio and visual presentations of identical text (Moussa-Inaty et al., 2012). What happens in these cases, according to Moussa-Inaty et al. (2012), is that when redundant information is presented, working memory resources unnecessarily attend to both sources of information in attempt to integrate these two sources. It can also be so that visual information received while reading a text makes a negative impact on the phonological loop (auditory aspect). Only if the two modalities complement each other can cognitive load diminish. This is due to the fact that the written information is permanent while the spoken information is transient. You can re-read the written information to be able to understand and remember the meaning. You have to memorize the spoken information to be able to return to it and analyze. But when you try to transform the permanent information into transient, it leads to cognitive overload.

**The role of attention**

Both processes, auditory as well as visual, need a good deal of attention which requires working memory. Smith and Kosslyn (2009) distinguish failures of attending to information as failures of selection in space and in time. The former refers to activities around us, the latter refers to incoming information, usually the auditory or visual.

Attention failure, known as divided attention, occurs when two tasks are similar and/or the level of difficulty of each task is high. When students read a scientific text and listen to it simultaneously, the auditory and visual modalities seem to be different and do not interfere with one another. On the other hand, listening and reading are both verbal activities which makes them both of the same modality, task likeness.

The task of reading a factual text is difficult in itself, more difficult than listening to a story, where you can imagine the situation and yourself in it. It means that there is a greater risk for attention failure, when you work with a scientific text, because of cognitive load. All this results in processing difficulties and memory overload and, therefore, shallow encoding. (Smith & Kosslyn, 2009).

**Virtual Humans - the audio-visual synthesis**

During the learning process new information can be delivered in different ways, through text, speech, diagrams, movies or pictures. In some situations, a text is expected to be worked on by the student herself and a speech can be delivered by a teacher in a classroom. But these tasks can be left over to the computers and task-oriented agents, who will deliver spoken information on a one-to-one basis, thus individualizing the instruction (Veletsianos & Russell, 2013). Therefore, one of the important questions raised in this paper was whether the use of animated pedagogical agents could provide a better learning environment for students and, consequently, a better learning outcome.

Literature on pedagogical agents and their significance for education in general and learning in particular is extensive. According to Veletsianos and Russell (2013), various studies have shown the impact of pedagogical agents, seen from both social and cognitive perspectives. It has been shown that virtual agents can lower learners’ anxiety, have the role of tutoring or conversational partners, increase learners’ interest and motivation. Furthermore, some researchers claim that pedagogical agents improve comprehansion due to nonverbal gestures and positive facial expressions. Others claim that they make it easier to focus on the cognitive aspects of learning, i.e. perception, attention, and memory (see Veletsianos & Russell, 2013, for a detailed review).

There are, nevertheless, studies that have pointed out limitations of the pedagogical agents used in education (Veletsianos & Russell, 2013). Virtual agents can only be part of the learning environment, a supplement, used for certain tasks. In one study it was shown that more than half of the participating students (59 %) preferred relation-oriented agents instead of task-oriented ones (Gulz, 2005). It was also reported that very human-like agents can distract listeners as the listeners start to believe that the agents are human, and, consequently, get irritated of their strangeness (Matthesyes & Verhelst, 2015). Furthermore, it was demonstrated that the looks of a virtual agent was of the utmost importance (Gulz & Haake, 2006) as well as the voice (Dunsworth & Atkinson, 2007). According to Dunsworth and Atkinson, some studies
have suggested that virtual agents increase extraneous cognitive load due to peripheral information (facial expressions), thus interfering with the task (listening to the text), caused by split-attention.

However, it is not the objective of this study to account for overall benefits and limitations of pedagogical agents. The main question in this thesis is if virtual agents can enhance text comprehension, compared to reading on one's own, listening or both, and what is it that should be included in the design, so that virtual humans will become a more effective alternative.

**Gestures and prosody as a form of integration of verbal and non-verbal aspects**

Listening is considered to be closely connected with the gestures we make when we speak. The motor theory speech perception (Liberman & Mattingly, 1985) claims that speech perception and speech production are closely linked, representing two distinct systems at the same time. This theory is based on an evolutionary perspective, according to which language originated from manual gestures, shifting later to the mouth and, subsequently, developing into the articulatory system (Coello & Bartolo, 2013).

Earlier theories of speech perception suggest that the acoustic signals are perceived and processed independently by the auditory perception area. According to the motor theory of speech perception, however, the acoustic signals generate motor representations of the corresponding articulatory gestures. These motor representations are located in the motor and premotor cortex, well away from the auditory cortex. Liberman and Mattingly (1985) therefore claim that both speech perception and speech production are motoric due to the mapping of perception and action. Or, in other words, speech production is a part of the gestural system as well as of the auditory system (Gentilucci & Campione, 2013).

When we listen to someone delivering speech we integrate both auditory and visual modality, the latter being in the form of gestures, hand, face and head gestures. While we get a linguistic message through auditory modality, the visual modality provides us with different types of cues, articulatory, prosodic and other extralinguistic signals (Granström & House, 2005). Some parts of human speech production system are visible to the receiver, clearly or sometimes, such as the lips, lower jaw, tongue and teeth. So when we listen to somebody, we receive information along two ways, acoustic and visual, that is called audio-visual speech signal (Mattheyses & Verhelst, 2015). Expressed in another way, speech is a multimodal way of communication and both the auditory and the visual information are decoded by the receiver (Mattheyses & Verhelst, 2015). This multimodality increases the complexity of the process as the decoding of each modality is influenced by another modality, thus producing perception effects that can prevent or obstruct understanding of the spoken message.

Takehashima and Gyoba (2014) studied audio-visual integration and found that a visual task performance (visual target identification) could be enhanced by sound. This occurred when visual stimuli were presented in the right visual field and processed in the left hemisphere, involved in linguistic processing. The researchers, therefore, suggest that auditory stimuli reinforce visual processing at the early stages. They hold that poor visual quality can be compensated by synchronous auditory information, a phenomenon known as auditory facilitation effect. This way visual object representation becomes stronger. To achieve this effect attention is required. However, research on visual speech synthesis (Ouni, Cohen, Ishak, & Massaro, 2007) demonstrated that communication becomes better when the synthetic voice is accompanied by the face of a speaker. It is this face-to-face communication that improves intelligibility of the synthetic speech. Earlier research in this area revealed that auditory perception, when accompanied by an animated agent, increased even in the presence of background noise (Coen & Massaro, 1994).

Wagner, Malisz, and Kopp (2014) point out that a speaker or a reader usually has an audience, and since she/he communicates with this audience, interaction occurs. Communication consists of gestures, not only hand gestures, but also head movements and facial gestures. Gestures support communicative efficiency, because they give complementary information.

There is some research on facial gestures used in conversations and dialogues (Bavelas, Gerwing, & Healing, 2014). The present study investigated reading and listening to scientific texts. Facial expressions and head movements of a speaker in these situations can be quite different, than those used in a conversation or a dialogue. As this experiment focused on learning factual texts, no special attention was given to emphatic content or expressions of inner state of gestures. On the other hand, gestures can be used to clarify or highlight something or point out what is important (Wagner et al., 2014), and this can be applied to reading and listening to scientific texts. Wagner et al. hold that people use head and facial gestures when they want to enhance information in one modality, for example, auditory, by visual modality of gestures, thus emphasizing a word or several words. Even reading a scientific text a teacher would bend her head, shake it, raise her voice, make mouth movements or move the eyebrows. (Bavelas et al., 2014). We have to distinguish between facial actions and facial expressions, the former referring to formal musculature of the face and the functions these changes bring into interaction; the latter revealing the inner state of the talker/reader.

With respect to this study it was important to pay attention to how head and face gestures are linked to prosody, grammar, or semantics. Even in this area there is not much research done. Bavelas et al. (2014) claim that some research proposed that facial gestures serve linguistic functions. According to Birdwhistell (1983), gestures are "restricted to those acts or actions which can be immediately interpreted by actor or viewer" (p.358). They can be expressed as vocalized behavior in form of intonation patterns, phrase superfixes, and voice qualifiers. As the relationship between the auditory and the visual modalities is very close, experienced linguists can imagine movements just by listening to a telephone call. Similarly, Cassell (2012) holds that gestures serve different purposes, like stressing of words or stating questions, when eyebrows are raised at the end of a question. Some of them are hand gestures. Beasts are, for example, small movements to emphasize or stress words, i.e. they serve linguistic purposes. Beasts are also produced with hands, but can presumably be made by the head. Cassell (2012) claims that it is important not to have too many nonverbal behaviors, because they can produce a wrong impression on the listener.

Another important factor that can influence audio-visual perception is prosody. These are speech elements that involve
syllables, words, phrases and whole sentences (Rao, 2012). Their duration, intonation and melody add naturalness to human speech. Prosody helps us to understand the message and sort out perceptual ambiguities. Mattheyses and Verhelst (2015) divide prosody into visual and auditory prosody. Granström and House (2005) also suggest that prosodic information the listener receives is not limited only to acoustic features such as prominence and phrasing. Visual signals add up to this information making it complete and easier to take in. These visual signals can be produced by the lips, jaw, tongue, eyes, eyebrows, eye closure and head movements. In other words, the speaker uses audio-visual verbal as well as non-verbal modalities thus facilitating the reception of information.

*Synthetic speech and virtual humans*

Using synthetic speech is not the same as listening to a natural speaker. According to the scoping review by Drager, Reichele & Pinkovski (2010), twelve percent of children within special education use some kind of speech-generating devices for listening to enhance their learning and increase their independence, e.g. synthesized speech devices. At the same time, Drager et al. found that children in general demonstrate lower levels of intelligibility when listening to synthesized speech instead of natural speech, when compared to adults. Intelligibility is the quality of the acoustic signal to transmit sounds and it is closely connected with comprehension of speech and memory. The working memory of children has smaller capacity than that of adults. Natural speech is easier to understand whilst synthesized speech often lacks natural acoustic cues, among other qualities. This makes heavy demands on cognitive resources thus decreasing comprehension and memory. The researchers also found that there were age and practice effects that influenced comprehension of synthetic speech. It was demonstrated that with increasing age the intelligibility improved and that only after two sessions of listening to sentences in synthesized speech comprehension progressed significantly.

On the other hand, many studies (see Mattheyses & Verhelst, 2015, for more detailed overview of the subject) have demonstrated that when you listen to synthetic speech, accompanied by a talking head, the understanding increases and the listener becomes more comfortable in his or her interaction with the machine. But of course this “naturalness and realism” (Mattheyses & Verhelst, 2015, p.185) depends a lot on i) the quality of synchronization of the acoustic and visual modes (timing), ii) on the quality of synthetic speech (prosody), and iii) on the quality of animation (gestures).

Animation can be carried out in 2D (cartoon-like) or 3D dimensions and the animated figures can be designed as very human-like. Mattheyses and Verhelst (2015) argue that, on the one hand, perceivers feel comfortable communicating with such virtual agents. Static photorealism (static pose) is put against dynamic realism. On the other hand, there is a risk of the "uncanny valley effect", when the perceiver associates the agent with a human, but notices some strange behavior, if it is not perfectly synchronized, thus feeling irritation with this strange "real" person (Mattheyses & Verhelst, 2015). This effect makes it more important to attain high quality or produce less human-like agents. In this case 2D figures would be better, because, according to Mattheyses and Verhelst, they will appear more familiar, resembling television broadcast and movies. In other words, there is no perfect solution and animation should be chosen after needs.

Another side of the problem is techniques for designing co-speech gestures. There are several techniques, all of them having both advantages and disadvantages: terminal-analogue systems, anatomy-based, performance-driven and so on. Speech gestures are also programmed in accordance with co-articulation strategies for speech synthesis, like rule-based or data-driven. Mattheyses and Verhelst (2015) hold that audio-visual synthesis has to sound realistic and that can be achieved when the virtual speaker's “acoustic waveform plays together with the video signal” (p.210). Auditory and visual prosody makes the message sent by the virtual speaker livelier, because "facial gestures add extra metacognitive information to the speech" (p.210).

*Summary*

Summarizing what is said in this chapter, two main aspects were considered in the current study: reading versus listening, on the one hand, and the use of virtual humans with synthetic speech in education, on the other hand. Reading and listening seem to have similar underlying processes with the same objective: to receive new information and remember it. When you understand what you read or hear you remember it better. The memory of the received information is the learning outcome. The presentation mode, acoustic or visual, is of importance when intelligibility is concerned. But when the information has safely reached its destination, it is the cognitive strategies (visualizing, verbalizing, creating the situation model, previous knowledge) and linguistic abilities (phonology, semantics, syntax and pragmatics) of the receiver that influence the learning outcome. The third factor is task demands on the memory capacity (dual vs. single task, length of the text, text difficulty etc.) that can change the learning outcome.

Comprehension of the new text material delivered in oral form does not always rely upon language, but there can be visual non-verbal cues that aid to understand the contents better. But the auditory reinforcement can also improve the visual perception. In education the audio-visual combination can be achieved by the use of pedagogical agents, in particular, virtual humans. This technology has both advantages and disadvantages. Audio-visual technology can help individualize instructions, but it can also distract the audience and impede learning. Gestures and prosody are two important factors when virtual humans are concerned. The use of synthetic speech alone may be difficult for children to comprehend, but the combination of synthetic speech and animation can facilitate perception and, consequently, comprehension and learning.

*Summary of the hypotheses*

The following hypotheses were tested in this study:

Hypothesis 1 (H1) proposed that reading and listening to a text simultaneously on the computer screen (LR) will give a better learning outcome compared to reading on one's own (R): \( LR > R \).
Hypothesis 2 (H2) proposed that reading a text on one’s own (R) will give a better learning outcome than listening without any visual text support (L): R > L.

Hypothesis 3 (H3) proposed that listening to a virtual human with synchronized head and face gestures (LV) will not only give a better learning outcome than just listening (L), but also a better outcome compared to the other two conditions (R, LR): LV > L and LV > R, LR.

Additionally, three hypotheses on the sub-group level were suggested:

Hypothesis 4 (H4) proposed that students in the sub-group 1 (SG1), with poorer learning outcome on the reading condition (R), will perform better on at least one of the three listening conditions: R vs. LR, L, LV.

Hypothesis 5 (H5) proposed that students in the sub-group 2 (SG2), with the highest scores on the visual non-verbal working memory tests, will be favored by listening to virtual humans (LV) compared to reading (R), listening and reading (LR) and just listening (L): LV vs. R, LR, L.

Hypothesis 6 (H6) proposed that students in the sub-group 3 (SG3), with the lowest scores on auditory verbal working memory tests, will perform better on listening to and watching virtual humans (LV) than on just listening to the texts (L): LV vs. L.

3 Method

Participants

Participants were 8th grade students (n = 76), recruited from a secondary school in the south of Sweden (25 females, 51 males, age 14-15). 63 students were native Swedish speakers and 13 students had another mother-tongue than Swedish. The students in the latter group were either born in Sweden or came to Sweden before the age of seven, i.e. before the start of compulsory school, and have thus followed the main curriculum of the Swedish educational system from the beginning. Of the total amount of participants two students were diagnosed with dyslexia, three with SLI (specific language impairment), and two students with ADHD (attention deficit hyperactivity disorder). The participants came from seven different classes in two different schools in the area with a total number of 156 students (males=89, females=67). All the students in the 8th grade were invited to participate in the experiment and 92 students entered the experiment voluntarily and well aware of the tasks and procedures. No check of their learning abilities was made before the experiment. As the experiment stretched over a period of two weeks, the number of the students who completed all seven tests (three pre-tests and four treatment conditions; see below for more information about the tests) was reduced to 76 due to involuntary absence. Two students did not accomplish one of the pre-tests and were excluded from sub-groups’ analyses.

Procedure

In order to test the equipment, the procedure, and the texts a pilot study with six students was conducted two weeks before the experiment. For the main experiment all the participants were grouped randomly, to avoid unnecessary bias of group effects and to enhance statistical validity. For that purpose a blocking design for males/females was used, thus minimizing the confounding factor of gender. This could also partially solve the bias of the prevailing number of male participants. Furthermore, all the participants were divided into two blocks (Block1 and Block2), each block corresponding to the week of participation. Each student obtained a number drawn from a bowl, thus allocating the students into one of the two blocks. Within each block the students were divided into four groups, consequently producing four groups per week: A, B, C, and D.

To eliminate external factors, the exterior conditions of the experiment were kept as equal as possible. The tests were carried out by the students on the same type of computers (Dell Latitude E5440, screen size 15 inches, operative system Windows 7) using ear-phones to eliminate disturbance. The same classroom was used for all the experiments. The students, between 12 and 15 in each group, were seated so that they could not see each other’s computer screens. This was close to the natural settings of a classroom and at the same time this design with relatively few students in each group formed a controllable environment for the experiment. All tests were installed on a USB memory stick before the experiment, and could be used locally to avoid internet disturbances.

All three pre-tests were carried out during the first day of each experimental week according to a specially arranged schedule. During the following two days the four conditions were implemented, two each day, to minimize tiredness and experience of monotony. The students listened to the same instructions before each test and were given possibility to ask questions. The time limits were set for reading/listening tasks and answering tasks (5 minutes). In the former case (reading), the time limits were determined by the audio recording time for the respective text. The same procedure was applied the following week for the Block 2 students. After each text the students were given a paper with questions to be answered in writing. To make meaningful comparisons the scores for each student were calculated as a percentage of the whole.

The structure of the whole study

The total experiment consisted of three pre-tests and four experimental conditions (Table 1). The working memory pre-tests were carried out in order to investigate the questions concerning working memory impact on the reading and/or listening conditions for a number of sub-groups.

Pre-tests

- Visual nonverbal memory (Patterns): the participant is presented with different patterns on the computer screen for several seconds and the participant has to allocate the shown patterns in correct order from memory.
Table 1. The outline for the three pre-tests and all the experimental conditions

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Participant</th>
<th>Pre-tests</th>
<th>Pre-tests</th>
<th>Pre-tests</th>
<th>Conditions</th>
<th>Conditions</th>
<th>Conditions</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pre-test 1: Patterns</td>
<td>Pre-test 2: Letters Numbers synthetic speech</td>
<td>Pre-test 3: Numbers synthetic speech</td>
<td>Reading condition: R</td>
<td>Treatment condition 1: LR</td>
<td>Treatment condition 2: L</td>
<td>Treatment condition 3: LV</td>
</tr>
<tr>
<td>A</td>
<td>A01</td>
<td>visual non-verbal memory</td>
<td>verbal working memory</td>
<td>verbal working memory</td>
<td>silent reading on one’s own</td>
<td>listening and reading simultaneously</td>
<td>listening without text support</td>
<td>listen to and watch a virtual human</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>B01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>C01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>D01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Verbal working memory for storage and processing (Letters; human voice): the participant is instructed to listen to and remember an alphabetical letter, followed by listening to and answering short questions. Afterwards the participant is asked to write down the letters in correct order.

- Verbal working memory for storage (Numbers; synthetic speech): the participant is presented with a digit span backwards task, i.e. listen to a chain of numbers and write them down in reverse order.

**Experimental design of treatment conditions**

For the four treatment conditions, this study used within-subjects design (repeated measures design). The question of the design choice, i.e. between-subjects or within-subjects design, was central for the experiment. Using independent measures design (between-subjects), with separate experimental groups for each condition, where each participant was tested only once, had advantages such as the factors of fatigue and monotony could be eliminated. On the other hand, the objective was to have a greater number of participants in each condition of the experiment, which was difficult to achieve in such a design.

Within-subjects design required all the chosen students to participate in all the treatment conditions and the effects of tiredness and boredom were controlled by counterbalancing, using a Latin squares design (Table 2) and carrying out the experiment over two days. This way all the conditions could be carried out in any order. Another important advantage of such a design was the possibility to analyze results on the individual level, which could not be done in a between-subjects design.

In this experiment one independent variable was used, the mode of presentation of a new text, on four different levels: reading (R), listening and reading simultaneously (LR), just listening without text support (L) and watching/listening to an
animated character on the computer (LV). The dependent variable was the learning outcome.

As stated in the Theoretical background, the learning outcome is memory, generated when something is learned. It was not the goal of the present study to divide the process of learning into a number of steps and study them separately; perception of visual and auditory signals, intelligibility, comprehension, or recall, among others. Certain steps have been mentioned in this paper, commented upon and some of them were analyzed, but the dependent variable of learning outcome was chosen as a generic term in this study, referring to different components of the learning process. It was measured by the scores on the multiple-choice tests.

Table 2. The Latin squares design for the four conditions. Each group (A, B, C, and D) performed all the conditions (R, LR, L, and LV) in different order during the two-days’ testing. The texts that were treated in each condition (see below for the description of the texts) are indicated in brackets. The same procedure was carried out for both Block 1 and Block 2.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>R(T1)</td>
<td>LR(T2)</td>
</tr>
<tr>
<td>B</td>
<td>LR(T1)</td>
<td>LV(T2)</td>
</tr>
<tr>
<td>C</td>
<td>LV(T1)</td>
<td>L(T2)</td>
</tr>
<tr>
<td>D</td>
<td>L(T1)</td>
<td>R(T2)</td>
</tr>
</tbody>
</table>

Text choice and design

The four texts that were used in this study were chosen from three textbooks in Social Sciences: Religion (Olofsson, Uppström, 2014), History (Nilsson, Olofsson & Uppström, 2013) and Social Studies (Linder, Homlin Nilsson, Friborg & Isacsson, 2014). These textbooks were recently bought by the school and had not been previously used by the students. Lix-value was used to assess the comparability of the texts. Lix-value is a readability index, frequently used in Sweden (http://www.lix.se). The index is based on the mean of words per sentence plus the ratio of longer words in percent (words longer than 6 letters). The index varies between very easy (< 30) and very difficult (> 60) (http://me.se). The lix-value of the texts, chosen for the present experiment, is just under the moderate difficulty level (lix 40 – 50) used in standard newspaper articles. The difficulty of the content is not measured by lix, but as all four texts originated from factual study books, adequate for the students of the 8th grade, it was assumed that they were of the similar difficulty from the point of view of the contents. The lix-value and the topics of the texts are presented below:

Text 1 (348 words): lix 40; about a proposal to name EU for Europe’s United States; 3:04 min.
Text 2 (266 words): lix 38; about a French successor to the Swedish throne Bernadotte; 2:21 min.
Text 3 (361 words): lix 39; about two explorers of the 19th century, Giuseppe Garibaldi and Henry Morton Stanley; 3:22 min.
Text 4 (409 words): lix 37; about the invention of the first vaccine against smallpox; 3:40 min.

Multiple choice questions were used for all the texts to eliminate the confounding factor of writing difficulties or tiredness. Multiple choice tests are broadly used in schools, including national tests in social sciences.

The questions to the texts, composed by the experimenter and revised by a certified teacher of Social Sciences, were of three types: one caption question, seven fact questions, and two questions on the contents as a whole. Thus, the questions to all the texts had the following structure:

1. The caption question (multiple choice between 3 alternatives; only one could be chosen); measurement of an overall understanding of the text.
2. The fact questions (multiple choice of 3 alternatives; one or more could be chosen); measurement of facts recall using bottom-up strategies.
3. The questions about the contents in general (multiple choice of 4 alternatives; one or more could be chosen); measurement of understanding and recall of the essentials of the text using top-down strategies.

The texts were read by the synthetic/artificial voice Elin (Vital 5.1.2) at 126 words per minute.

Design of Virtual Humans

The theoretical paradigm outlined so far in Chapter 2 in this paper lead to some methodological implications when making a comparison between traditional ways of studying a text and modern approaches of using new technology within artificial intelligence, or more specifically, virtual humans.

The technical aspects of designing and programming of virtual humans was not part of this study and was carried out by Jens Nirme at Cognitive Sciences, Lund University. Paralleling the technical side, the author of this paper has contributed to this work by investigating the theoretical background of speech communication in general and the susceptibility of information presented by virtual humans in particular, thus providing insights into these issues.

Motion capture technique was considered as the first alternative. This technique is based on video-recording of a real human being. His/her body movements are then reprogrammed and used to animate virtual humans. This procedure, however, takes considerable time and even though it has important scientific value, it is hardly of any practical value for schools to date. What is needed instead is a fast way of converting texts into animated speech, a procedure and technology that can help to reduce teachers’ workload (Mattheyses & Verhelst, 2015). This was a reason why it was decided to use
virtual humans with speech synthesis. As speech synthesis is used more and more frequently in schools, it was assumed that applying this technology in the experiment would procure higher generality of the experiment and increase its external validity. Furthermore, speech synthesis was already used in conditions LR and L and comparable parameters for virtual humans were necessary.

To facilitate the programming of virtual humans only facial gestures and head movements were used (so called talking heads), thus skipping hand gestures. To create facial animation from audio files a commercial application FaceFX (http://facefx.com) was used.

As for the audio files, the texts were transformed into synthetic speech (MP3 standard) by ViTal (version 5.1.2) at a speech rate of 126 words per minute. The research on synthesized speech presentation rate by von Berg, Panorska, Uken, and Qeadan (2009) revealed that the lowest rate for young adults aged 14-24 was 113 wpm (words per minute) and the fastest was 156 wpm. The students in the current experiment were 14-15 years of age so they could be considered as young adults. As the speech rate of 126 wpm was a default setting in ViTal, it was left at that.

Model analysis

In the analysis of the data it was predicted that there could be or should be some within-subject variance due to either the presence of confounding variables (some random factors, unrelated to the experimental conditions, that might have influenced one condition but not the other) or due to individual differences of the participants (age differences, learning style differences, performance differences etc.). The statistical methods for the analyses were chosen considering the above circumstances.

The analyses were performed in several steps: first analysis and evaluation of the model as a whole was conducted followed by post hoc tests to target hypotheses H1 to H3. Additionally, analyses of several aspects on the sub-group level were made to target hypotheses H4 to H6.

The analysis of the experimental model as a whole included only experimental conditions (R, LR, L, and LV) and no pre-tests (see Table 1), as the pre-tests were used for establishing sub-groups for hypotheses H4 to H6. First, validity of the four conditions and the pre-tests were verified, using Mauchly's test of assumption of sphericity, to provide information about the equality of variances of the differences between the tests. Mauchly's test was based on the assumption that the variation between pairs of several experimental conditions was similar at $p > .05$. If this was achieved it would mean that the variation between group results could be considered non-significant. Then, a one-way repeated measures ANOVA (analysis of variance) was conducted in order to test the significance of the model. The modes of presentation, i.e. reading, reading and listening, listening, and watching the virtual human (R, LR, L, and LV), were regarded as an independent variable on different levels and scores on the tests as dependent variables. $F$-ratio assessed the systematic variance of the dependent variable compared to the error within this model.

Hypothesis testing

To test hypotheses H1 to H3 a one-way ANOVA with pairwise $t$-test post hoc comparisons between all conditions were conducted. When more than one comparison is made, the higher is the risk to obtain a significant result due to chance and Bonferroni correction for multiple comparisons was used to control such possible false-positive results.

Hypotheses H4 to H6 were also tested by a one-way ANOVA with pairwise $t$-test post hoc comparisons, for selected sub-groups in accordance with the claims in the hypotheses. As group comparisons can conceal significant individual differences, it was decided to double-check some of the results on the individual level, comparing results for different tests per student.

In order to reveal relationships between the two verbal pre-tests Letters (natural voice) and Numbers (synthetic speech) and conditions L and LV for the whole group, Pearson's product-moment coefficients were calculated. Even though correlation analyses do not evaluate causality, but only reveal possible relations between two variables, it was valuable to locate the variables with significant correlations to be able to cross-check the results and strengthen the accuracy of the conclusions.

All the analyses were carried out in the statistical analysis program SPSS.

Ethical considerations

The value of the present experiment and the results it generates were considered of great future value for secondary education in general and were expected to bring forward important general knowledge. The experiments required participation of minors, students of the 8th grade, being 14-15 years of age. Therefore, a number of ethical considerations was taken into account.

The experiment was approved by the principal and the school education management, as they considered the results of the study to be of vital importance for the future school development. All the students and their parents were informed by a letter about the experiment and its objectives. All the students were aware and repeatedly reminded that their participation in the experiment was strictly voluntary. They were also informed that they could quit the study at any time and that all the individual results from the study were confidential. The results and conclusions of the experiment were duly presented to all the students and the teaching staff.

4 Results

Analysis of the model as a whole

Figure 1 (see next page) presents histograms for all the four conditions, where conditions R, LR, L, and LV were respectively compared to the expected normal distribution. From the histograms, it could be assumed that the data in all the four conditions was normally distributed and, therefore, could be analyzed parametrically.
Mauchly’s test indicated that the assumption of sphericity for the four experimental conditions (R, LR, L, and LV) had not been violated ($p = .440$) and that the relations between pairs of the four experimental conditions were similar. Furthermore, Mauchly’s test for the three pre-tests together with the four experimental conditions also showed that the assumption of sphericity was valid for all seven tests ($p = .143$).

**Results for hypotheses H1 to H3**

The analysis of the main effect of the model showed that there was a significant difference between reading on one’s own (R: $M = 0.68, SD = 0.16$) and reading and listening (LR: $M = 0.68, SD = 0.15$), on the one hand, versus listening (L: $M = 0.60, SD = 0.18$) and listening to a virtual human (LV: $M = 0.58, SD = 0.18$), on the other hand. Confidence intervals of the means are illustrated in the error bar chart (Figure 2). The significance of the results of the whole model was tested by a one-way repeated measures ANOVA, applied for within-subjects designs. It showed the $F$-ratio to be significant on the 0.001 level ($F(3,225) = 12.45; p < .001$), i.e. the mode of presentation of the material created a significant difference in the learning outcome, favoring reading (R) and simultaneous reading and listening (LR). Hypotheses H1 to H3 were evaluated by post hoc tests in the form of paired $t$-tests, adjusted for multiple comparisons by Bonferroni method (Table 3 on the next page).

**Hypothesis 1 (H1):** There was no significant difference between the R and LR conditions ($p > .05$; cf. Table 3), which gave no support for hypothesis 1 (H1: LR > R), claiming that reading and listening to a text simultaneously (LR) would give a better learning outcome than reading on one’s own.

**Hypothesis 2 (H2):** There was a significant difference between the R and L conditions ($p < .001$; cf. Table 3), confirming hypothesis 2 (H2: R > L), thus suggesting that reading a text on one’s own (R) will give a better learning outcome than listening without any text support (L).

Figure 1. The four conditions (R, LR, L, and LV) compared to a normal distribution curve

Figure 2. Error bar chart of means for the four conditions at 95% confidence interval
Hypothesis 3 (H3): There was no significant difference between the LV and L conditions (p > .05; cf. Table 3), and hypothesis 3 could thus not be supported (H3: LV > L), claiming that listening to a virtual human (LV) would give a better learning outcome compared to the L condition. As for the comparison between LV versus R and LR respectively, both R and LR were significantly larger than LV thus further contradicting hypothesis 3.

Table 3. Paired t-tests of the four conditions, Bonferroni corrected for multiple comparisons. The first value shows the mean difference and the second value the significance level

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>L</th>
<th>LV</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.006 / 1.000</td>
<td>0.079 / .001</td>
<td>0.101 / .000</td>
</tr>
<tr>
<td>LR</td>
<td>0.073 / .006</td>
<td>0.095 / .000</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>0.022 / 1.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results for hypotheses H4 to H6

Hypothesis 4 (H4): This hypothesis claimed that the students who had poorer results on the reading condition (R), would have better results on all or at least one of the listening tests (H4: R vs. LR + L + LV; cf. Figure 3). Twenty-one students in the lower quartile (subgroup 1, n = 21), with scores for the reading condition (R) ranging from 0.33 to 0.56, were qualified (total score range for R: 0.33 - 0.94). Mauchly’s test of the assumption of sphericity for all the four conditions for the 21 students was not violated (p = .991), indicating similar variations within the groups. To test hypothesis 4 the results for the reading condition in this sub-group (R: M = 0.47, SD = 0.08), were compared to the means of the respective listening condition (LR: M = 0.60, SD = 0.17; L: M = 0.49, SD = 0.18; LV: M = 0.50, SD = 0.17), and to their aggregated means (LR + L + LV: M = 0.53, SD = 0.14). Paired samples t-test for the four conditions respectively, Bonferroni corrected for multiple comparisons, revealed significant differences between the reading condition (R) and the listening and reading simultaneously condition (LR) (t(20) = -3.279, p = .004). The difference between R and the aggregated results for the three listening conditions (LR + L + LV) was, however, non-significant (t (20) = -1.844, p = .080).

In order to unmask possible individual differences, the results for this sub-group (SG1) were also analyzed on the individual level. For this purpose, the scores were checked per person, comparing the results on reading with the results on other tests. This analysis revealed that 14 of 21 students in this sub-group (67 %) had better results on the combined listening conditions than that of reading.

The results on the individual level for sub-group 1 (SG1; the lowest results on the R condition) were juxtaposed to the results on the listening conditions for the students in the highest quartile (SG1a: score range 0.80 - 0.94; n = 24). Concerning the results per student for this sub-group (the highest scores on R), it showed that 25 % of the students had better results on at least one listening test (6 out of 24). The results on the individual level for both reading sub-groups thus demonstrated that, in general, there was a relatively large number of students who seemed to have a different style of learning: they preferred to listen to new material. Interestingly though, listening preferences within each sub-group differed from each other, as demonstrated in Figure 4.

Figure 3. Error bar chart R vs. Combined listening conditions (LR+L+LV; 95% confidence interval)

Figure 4. Comparisons of the results for the three listening tests for the two sub-groups, sub-group 1 and sub-group 1a, demonstrating that the students in both sub-groups were favored by the LR condition. The distribution of the scores was nevertheless different: those who had the lowest scores on the reading condition (SG1) also had slightly better results on LV than on L, compared to the students with the best results on R (SG1a). Worth mentioning is also that SG1a had better overall results on all the listening conditions.
**Hypothesis 5 (H5):** This hypothesis proposed that the students who scored best on the visual nonverbal pre-test Patterns would be benefited by listening to virtual humans (LV) compared to reading (R), listening and reading (LR), and just listening (L) (H5: LV vs. R, LR, and L). The students in this sub-group (SG2) were 16 students with the highest scores on the pre-test Patterns (10 scores out of 10). Additionally, 14 students (SG2a) with the lowest scores on the test (scores 1-6 out of max 10) were chosen in order to be able to juxtapose the results. Mauchly’s test of the assumption of sphericity for the above comparisons was valid (SG2; p = .102; SG2a: p = .201). The means of scores for the learning outcome in the LV condition for the respective sub-group were compared to the means of the other three conditions combined: (SG2 (LV): $M = 0.57$, $SD = 0.19$) vs. combined means for the same students (SG2 (R+LR+L): $M = 0.68$, $SD = 0.11$) suggesting a significant effect ($t(15) = -2.448$, $p = .027$), though contrary to H5. The results for this sub-group (SG2; high scores on the visual memory pre-test Patterns) demonstrated that the students with the best scores on the visual non-verbal ability pre-test Patterns achieved significantly poorer results, when listening to a virtual human (LV), compared to the other ways of learning. In other words, these results, though statistically significant, did not support hypothesis 5.

The sub-group with the lowest results on the Patterns pre-test (SG2a) showed no significant difference between LV and the other three conditions: (SG2a (LV): $M = 0.53$, $SD = 0.20$) vs. combined means for the same students (SG2a (R+LR+L): $M = 0.59$, $SD = 0.12$; $t(13) = -1.125$, $p = .281$).

**Hypothesis 6:** This hypothesis proposed that the students with the lowest scores on the auditory verbal memory tests (pre-test Letters and pre-test Numbers, respectively) would perform worse on just listening to the texts (L), than on listening to and watching virtual humans (LV) (H6: LV vs. L). The score means are presented in Figure 5.

![Figure 5. Students who performed poorly on the pre-tests Letters (sub-group 3; SG3) were not favored by listening to the virtual humans (LV). The students who did poorly on the pre-test Numbers (sub-group 3a; SG3a) showed similar preference for L and LV conditions (both conditions with no text support)](image)

**Pre-test Letters sub-group (SG3).** For the purpose of testing this hypothesis the results for the sub-group 3 (SG3) of 15 students with the lower scores on the pre-test Letters (SG3: scores 1 - 30 out of max 36) were analyzed (SG3 (L): $M = 0.55$, $SD = 0.24$; SG3 (LV): $M = 0.49$, $SD = 0.22$) indicating no significant effect ($t(14) = 1.124$, $p = .280$). When comparing the students’ results for each student, however, it was found that 53% of the students in the SG3 subgroup had the same or better scores on LV compared to L (8 out of 15) and 40% of these students had better results on LV.

**Pre-test Numbers sub-group (SG3a).** The results for the sub-group 3a of 23 students with the lower results on the pre-test Numbers (SG3a: scores 0 - 4 out of max 10) were evaluated (SG3a (L): $M = 0.57$, $SD = 0.22$; SG3a (LV): $M = 0.57$, $SD = 0.19$) showing no significant effects ($t(22) = 0.000$, $p = 1.000$), why hypothesis 6 could not be confirmed. However, when the results were checked on the individual level for the SG3a group, it was demonstrated that 65 %, i.e. 15 students of 23, had the same or better scores on virtual humans’ tests (LV) than on just listening tests (L) and 57% of the students had better results on LV condition.

The results on the individual level, compared to group results, unmask individual preferences of the students and their individual learning styles. They also demonstrate the significance of animated agents for some students.

**Correlation analysis**

Bivariate Pearson’s product-moment correlation analyses were carried out to evaluate possible relationships between the working memory pre-tests and the four experimental conditions.

A significant medium correlation was noted between the pre-test Numbers and the reading condition (R) ($r = .359$, $p = .001$) and between the pre-test Patterns and the listening and reading condition (LR) ($r = .327$, $p = .004$). There was also a weak but significant positive relationship between the pre-test Numbers and the listening condition (L) ($r = .261$, $p < .023$). Finally, the pre-test Letters showed a weak but significant correlation with the virtual human listening condition (LV) ($r = .233$, $p < .042$).

**5 Discussion**

The main objective of this study was to investigate how distinct modes of presentation of a text (reading, listening or watching a virtual human) influenced the learning outcome. Two input modalities were used, auditory and visual, which allowed manipulations of the modes of presentation as follows: visual and oral verbal presentation (reading (R), concurrent reading and listening (LR)), oral verbal presentation - listening without text support (L) and visual nonverbal and oral verbal presentation listening to and watching a virtual human (LV). The receptiveness to the modes of presentation of a new text, using different input modalities, was the learning outcome. This dependent variable was measured by the number of correct answers during a written test of free recall after each condition.

In this part of the paper some questions concerning the design of the experiment will be discussed first, to explain the choices made in this study and its limitations. That will hopefully clarify some issues discussed later.
Design considerations

In order to acquire pure scores for the tests and to maximize reliability of the experiment, an attempt was made to eliminate both the non-random (systematic) bias, e.g. placing all the students in the similar situation, and to cancel the random (non-systematic) error by a relatively large number of participants. The valid value of the F-ratio for the whole model, being greater than 1, indicated that the experimental manipulation (systematic variation) had a considerably stronger effect on the results compared to other extraneous factors, the ones outside of our control (non-systematic variation).

Nevertheless, the author of this paper is aware that the experiment could be conducted in a more controlled form. At the same time, it seemed important to create an environment similar to everyday school surroundings and, in that way, increase ecological validity. This was obtained by using more representative and natural environment, a classroom, instead of one-to-one settings, which was the original alternative for this experiment. The use of randomized within-subjects design ruled out several possible confounds as the same students participated in all the experimental conditions, and individual differences were held constant across the tests.

The risk that the scores for the learning outcome could include scores for something else than what it was meant to (e.g. previous knowledge of topics or practice effects of listening to synthetic speech), was considered. It was assumed that the pre-tests of visual nonverbal and auditory verbal working memory could minimize this bias, thus increasing the validity of the experiment. The pre-tests, being different from the four conditions, did not interact with the experimental conditions, thus increasing the external validity and the generality of the results.

The aspect of previous knowledge could not be controlled to full extent, but this might be a problem with all text-based tests, including national comprehension tests in Swedish schools. The selection of the texts from new school books and similar lex-value were intended to eliminate this factor. Another measure to decrease the factor of previous knowledge could have been the use of several texts for each condition, but this measure would have increased students’ tiredness and test monotony. One more possible confounding factor was left as such, the students’ motivation factor.

One of the considerations, concerning the internal validity, was the question if the results of the experiment were influenced by some other extraneous factors than already mentioned above, the factors that lead to the change of behavior of the students during the tests. Among these factors were:

- different strategies used by different students to remember the texts, e.g. re-reading of part or the whole
- getting tired after several tests
- the fact of being tested influencing the participants and leading to a changed behavior: either doing their best/worst or showing anxiety
- the experimenter’s influence: the experimenter behaving differently by the end of a session or at the end of the day
- allocation of the students in random and new for them groups, leading to attempts to assert themselves

All these factors could decrease the internal validity of the results. A comprehensive summary of the precautions that were taken to minimize these risks is presented below:

- randomization against group threats
- Latin squares design to procure equal text distribution and counterbalancing of time for tests
- different days for pre-tests and tests, to avoid the students’ tiredness
- the text length carefully chosen to be able to give the students fair opportunity to learn, instead of introducing breaks between tests
- the experimenter’s effort to read the before-hand written instructions in the same manner and the same voice
- not too much information given to the students, to avoid too much test anxiety as well as test indifference

The gender perspective was not considered in this experiment, but the fact that the majority of the participants were male students is worth mentioning. The number of male students in the whole 8th grade was slightly larger compared to the number of the female students (57 % male students). It was obvious from the beginning that there were more male students who volunteered for the experiment. At the same time the students who quit during the experiment were mainly female students, resulting in that 67% of all the participants were male students. The reason for this might be that the male students were more curious to discover new learning possibilities. But it could also mean that they were more eager to skip classes.

The discussion of the results that follows will be divided into two sections. First the results for the reading condition will be contrasted with the results on the listening conditions. Simultaneous reading and listening will be discussed in particular. Then the results concerning the virtual humans will be juxtaposed to the results of other conditions.

Reading versus listening

The overall results revealed that the students were better off when they read the texts themselves, without any listening support. Auditory reinforcement in form of listening to a text simultaneously, while reading it, did not produce any significant difference, compared to the reading condition (H1). Listening without text support and watching virtual humans were found to be the least effective modes of text presentations.

One of the explanations to the better learning outcome after reading may be found in different aspects of written and spoken language. The students could read the text several times during the applied period of time, which could be an advantage for fast readers. Likewise, they could read the text at their own pace and understand and thus remember better (text permanency), compared to the listening without text support. In the L condition the students could listen to the text only once, without any possibility to adjust the speech rate after their individual needs and likes (text transiency).

Another explanation to the better scores on the reading condition may be that students in general are more accustomed
to reading on their own, which is a traditional and a well-established way of learning. Simultaneous reading and listening is usually provided for students with reading difficulties but not for a student population as a whole. It is likely that the majority of the students who participated in the study were not accustomed to this reinforcement.

There seems to be no consensus among researchers, considering reading versus reading and listening. Hale, Skinner, Winn, Oliver, and Allin, (2005) have claimed that the strategy of reading and listening simultaneously enhances reading fluency and comprehension. Interestingly, their results are contradictory to the findings of Balass, Nelson and Perfetti (2010), according to which less-skilled readers learn better when the information is presented in written form (orthographically) rather than in spoken form (phonologically). The cognitive load theorists hold that the best way of learning is reading (Moussalat et al., 2012), instead of simultaneous listening and reading: if not because of cognitive load (redundancy effect ensuing from two presentation modalities), then because of visual interference into auditory processes.

Although the results for the whole group favored the reading condition, it was somewhat surprising that the sub-group of students with the lower scores on the reading tests was particularly favored by the listen-and-read presentation modality (H4), even though the difference between the combined means of the three listening conditions (LR+L+LV) and reading (R) was not statistically significant. However, the three listening conditions were found to be a more effective way of learning for 67% of the students in this sub-group, when the results were checked on the individual level.

This discrepancy could be attributed to individual differences within this sub-group, as indicated by research on reading (Catts et al., 2006; Hilbert et al., 2014; Leahy & Sweller, 2011). For example, among the students in the sub-group there were five students with SLI (specific language impairment; \( n = 3 \)), and ADHD (attention deficit hyperactivity disorder; \( n = 2 \)). Two participants with dyslexia had a much better outcome after reading than listening and they fell outside the boundaries of this group. Dyslexia is considered as impairment on a phonological level, caused by (or causing) auditory difficulties (Banich & Scalf, 2003) and as synthesized speech lacks many important acoustic cues, these students could have had difficulties to properly apprehend the synthesized speech. Thus, reading was probably an easier and a more effective way for them to learn. Though students with SLI and ADHD often lag behind their peers in reading, their problems seldom have phonological basis and this may explain their better comprehension of synthetic speech. Furthermore, there could be students in this group who scored poorly on reading because of attentional or motivational reasons.

The question of circularity was considered here, i.e. that the students who had the lowest scores on the R condition would predictably have higher scores on the listening conditions. The same could be valid for SG1 group: those with the highest scores on R would have lower scores on the listening conditions. And so was the fact. On the other hand, it could also be a question of individual differences and, therefore, individual preferences, as mentioned above. Put simply: some people learn better when they listen and others when they read. That is why it was of interest in this study to investigate which of the listening conditions was the most suitable for these two sub-groups. It was found that the poor readers were helped more by listening to virtual humans than the good readers. But listening and reading simultaneously was the best alternative for both sub-groups of students.

For 33% of students in the sub-group 1 (the lowest scores on reading) listening reinforcement did not help and these students obtained low results on R as well as on all of listening conditions, LR, L, and LV. This may depend on the cognitive overload due to task likeness and attendance to several tasks simultaneously (Smith & Kosslyn, 2009). The explanation can also be found in comprehension difficulties, either due to insufficient vocabulary and semantic processing or to the lack of previous knowledge of the situation, among other things. In this connection it is worth mentioning that the students with the highest results on the reading condition (sub-group 1a) had better results on the listening tests, compared to the sub-group 1, demonstrating overall better comprehension of the texts, which is in line with the research by Elwér et al. (2013).

One of the questions in this study, concerning reading versus listening, was to compare reading on one’s own (R) with listening to a text, delivered in synthesized form, without any written text support (L, H2). The results were to elucidate if this alternative was advisable for students in secondary schools. On the one hand, the overall results demonstrated that listening to synthesized speech without text support was not suitable for students at this age. On the other hand, the sub-group with low scores on reading (SG1) demonstrated slightly better results on the listening condition than on reading, even though the difference was not statistically significant (H4). These findings bring forward the question of intelligibility of the synthesized speech by children (Drager et al., 2010). Listening to synthesized speech alone was expected to substantially worsen understanding and, consequently, remembering, due to lack of both natural acoustic cues and visual cues. Somehow, students with low scores on reading were not bothered much by this factor. This can be explained so that some of these students might be using synthesized speech in their learning environment already and they demonstrated practice effects, compared to the other listeners. If this is the case, practicing can be a solution. It can be arranged by teachers for all the students, thus offering a greater number of students a more effective way of learning. Furthermore, listening to a written text using synthesized speech can be more time-effective than video-recording, using natural voice, because teachers can transfer any text into synthesized speech in no time.

**Watching virtual humans**

The current study found that the students, when listening to a virtual human (H3), obtained the lowest scores, compared to all other conditions. The difference between LV and L was statistically non-significant, but, nevertheless, the scores on the L condition were slightly better. On the sub-group level (H5), neither the students with high nor low scores on the visual non-verbal working memory pre-test Patterns were favored by listening to the virtual humans. No significant correlation was found either between the pre-test Patterns and the LV condition. These results may be explained by the combination of shortcomings in synthetic speech and poor quality of
virtual humans. The low scores may also have to do with the novelty of a virtual human as a "teacher" for the students. The students seemed to be more curious to watch the animated agents' head and facial gestures than listen to the texts, thus increasing extraneous cognitive load and decreasing comprehension and learning.

Considering the quality of virtual humans in this experiment, several aspects should be mentioned. It was a female speaker and the students could not change the gender; the synchronization of eye movements with the speech was not complete; clothing and colors were neutral. Students' liking of the agent, that, according to social agency theory (Meyer & DaPra, 2012), leads to students' better engagement in learning and better learning outcome, so called embodiment effect, could not be achieved. On the other hand, too many non-verbal behaviors could also have produced a wrong impression and looked maybe unnatural, according to the findings by Cassell (2012). And it was what we had in mind, designing the virtual agents: avoiding the "uncanny valley effect", i.e. an intent to prevent that the virtual human could be associated with a human and distract the students. But a lot more could be done (and will surely be done in the future) to improve both acoustic and visual reception of a text: wrinkles on the forehead, more lively eyebrows movements, better synchronization techniques etc., as investigated by Granström and House (2005).

Elwér, Keenan, Olson, Byrne, and Samuelsson (2013) stated that oral language deficits can influence both listening comprehension and reading comprehension. Poor comprehension is considered by some researchers to be caused by deficits in verbal memory (Elwér, Gustafson, Byrne, Olson, & Keenan, 2015). Others (Nation et al., cited by Elwér et al., 2015) argue that poor results on verbal memory tasks are caused by oral language deficits. Whatever the causal direction, since listening comprehension is considered as a part of linguistic comprehension (Catts et al., 2005; Elwér, 2014), auditory ability in general and auditory working memory in particular (i.e. the presentation modality and/or processing modality, according to Hilbert et al., 2014) should be deemed an important determinant of students' learning.

In this light, the most striking result that emerged from the data on the sub-group level was that the students with the lowest scores on the reading condition (H4) had second best scores on LV of the three listening tests and these scores were better than those for reading (R: M = 0.47; LV: M = 0.50). Additionally, it was demonstrated that on the individual level 57 % of the students with low scores on auditory working memory pre-test Numbers were benefitted by listening to the virtual humans (H6). Together these results provide important insights into the significance of multimodal (audio-visual) method that can be based on two assumptions: that it is easier to remember a non-verbal code (Paivio, 2006) and that for certain students it is easier to remember acoustically presented information (Hilbert et al., 2014), even though it is delivered as synthetic speech.

Interestingly, though, that for many students listening to virtual humans produced no audio-visual synergy effects (H1, H3, and H5). A possible explanation to this can be that the capacity of the working memory was overloaded because of the dual task phenomenon, watching an animated figure and listening to a text. Another reason could be difficulties in adapting to synthetic speech as the intelligibility of the speech is dependent on the quality of the audio-visual synthesis. Some subjects could, for example, be stirred by the quality of the speech due to inexperience with speech synthesizers or it could also be the question of individual perception, as the experiment was about a perception-related task. This is why the pre-tests Numbers and Letters seemed to be important determiners of auditory verbal perception. It was expected that these pre-tests would be a good control measure for the outcome of the L and LV tests. Contrary to expectations, only a weak relationship was found between the pre-test Numbers and L, and Letters and LV.

Summarizing the above, listening alternatives, where virtual humans are used for text presentation, should be considered as a new pedagogical approach in education for some categories of students in the future. The idea of virtual humans is based on the fact that each student has a computer and listens to a text through ear-phones on a one-to-one basis, thus eliminating some of the external stimuli, concentrating on the material to learn. Attention will be then more focused on the material. The student can adapt speaker's performance to her own speech rate and make pauses at her own liking. The imperfectness of the synthetic speech can be thus partly eliminated. Practicing can increase the positive outcome. Proper facial gestures, head positioning and vocalization can improve virtual humans' performance even more. One can come as close as it is possible to a "perfect teacher", and it will, in a certain way, resemble social communication due to increased interaction.

**Future research**

This study provided a broad overview of issues concerning modes of presentation of new material for students of secondary schools. Future research with a more profound focus on particular issues might produce interesting findings within a number of areas, especially concerning virtual humans.

In the present study the aspect of integration of virtual humans in education was one of the central ones, with focus on presentation modality. Nevertheless, the assumption that pedagogical agents can be interactive partners in a learning process is important. Even if it has been claimed that listening to a text, performed by a virtual human on a one-to-one basis, increases students’ motivation and enhances learning (Veletsianos & Russell, 2013), we cannot claim that in the current experiment it happened due to improvement of the communication process between the learner and the machine, or growing social interaction due to embodiment of cognition, like in classical Hutchins’ pilot cabin (Hutchins, 1995). The present experiment was more concerned with the question of integration of pedagogical agents into the learning process, rather than the question of interaction. It is assumed though that the aspect of interaction should be the ultimate goal of development of pedagogical agents for education. To develop further the aspect of interaction, virtual humans can be programmed so that they would be able to answer students’ questions and carry out digital tests, among others. This is a desirable scenario for the future.

Another suggestion for future research would be a similar study, but with younger children, or children of different ages,
to be able to establish age limits for which visual speech synthesis is most propitious. The question of speech presentation rate should be further expounded. It would be also interesting to study practice effects of listening to synthetic speech.

Future research could also be conducted to investigate the alternatives listen-and-read versus listen, both using synthetic speech, but in a more controlled environment, studying, for example, the redundancy effect, using two presentation modalities, auditory and visual. Other studies could explore the effects of presentation modality versus processing modality, or comparing synthetic speech and natural voice in L and LV conditions, and, why not, for different levels of text difficulty. Further investigation of relationship between auditory/visual working memory and the virtual humans' alternative might be interesting in this connection.

Finally, as a larger number of the participants were male students it would be interesting to investigate the receptiveness to the modes of presentation of new material from the gender perspective, focusing in particular on the virtual humans aspects.

Concluding remarks

The results of the study as a whole yielded a series of conclusions that can have significant practical applications. Reading seems to be the most successful mode of presentation of novel material from the point of view of comprehension and learning, when considering an entire learner population, such as an entire class. Concurrent reading and listening is nearly as effective as reading.

It also became apparent that when the outcomes had been classified and grouped according to individual scores, the results for different groups of students appeared in a different light. From the present study alone we could draw the conclusions that the students who score lowest on reading and auditory tests benefit from either listening to the texts, using synthetic speech, and/or to the texts delivered by virtual humans. In this paper an attempt was made to explain some of these results. But there are many possibilities, related to complex cognitive underlying processes that lead to individual differences. If we continued looking for more comparisons, we would have undoubtedly made more discoveries. And this could have led to even more topics for future research.

As for the current study, the point has been made. The results have demonstrated that teaching an entire class of students in the same way may not be advisable. In a classroom there are students with various learning styles who need different pedagogical approaches. In practice teachers can propose the best way of learning for each student, after carrying out some tests and by checking each student's results, in other words, implementing the methods provided in this study. Hopefully, the development of the virtual humans and synthetic speech will continue and virtual humans will soon become one of the effective pedagogical approaches which will broaden the individualization concept.

6 Acknowledgments

I enjoyed working on this project very much and I would like to thank those who helped me to see it through. First of all I wish to thank all the students and staff who participated in this experiment. I could not have made it without them. I would like to thank Mattias Jakobsson, a social science teacher, who helped me to choose appropriate texts for the experiment, reviewed the test questions and assisted with the initial arrangements. I would like to give my warmest thanks to my supervisors, Agneta Gulz and Magnus Haake, who patiently followed my zig-zag marathon through the research field, returning me back on track, when I went astray. Special thanks are, of course, to Magnus Haake and Jens Nirme for their help and assistance with the technical and statistical aspects of the project as well as their valuable comments and discussions during the whole project.

References


