Abstract

Objective: The speech intelligibility benefit of visual speech cues during oral communication is well-established (Sumby and Pollack, 1954). Therefore, an ecologically valid approach of auditory assessment should include the processing of both auditory and visual speech cues. This study describes the development and evaluation of a virtual human speaker designed to present speech auditory-visually. *Design:* A male and female virtual human speaker were created and evaluated in two experiments: a visual-only speech reading test of words and sentences and an auditory-visual speech intelligibility sentence test. *Study sample:* A group of five hearing, skilled speech reading adults participated in the speech reading test whereas a group of young normal hearing participants (N=35) was recruited for the intelligibility test. *Results:* Skilled speech readers correctly identified 57 to 67% of the words and sentences uttered by the virtual speakers. The presence of the virtual speaker improved the speech intelligibility of sentences in noise by 1.5 to 2 dB. *Conclusions:* These results demonstrate the potential applicability of virtual humans in future auditory-visual speech assessment paradigms.

Introduction

People communicate continuously, in either a spoken or a written way. During oral communication, listeners do not solely rely on auditory input. In addition, they process visual speech cues including mouth movements, which improves both the detection and discrimination of speech (Bernstein, Auer and Takayanagi, 2004; Grant and Seitz, 2000a; Ma et al., 2009; Schwartz, Berthommier and Savariaux, 2004; Sumby and Pollack, 1954). This visual speech benefit even sustains when listeners do not directly face the speaker, as shown by research of Grange and Culling (2016). They presented sentences together with spatially separated noise and demonstrated that head orientations up to 30° sideways of the speaker did not influence the speech intelligibility benefit obtained by speech reading. In addition, speech reading is known to be automatic and therefore an intrinsic part of speech processing (Rosenblum, 2008; Woodhouse, Hickson and Dodd, 2009). The multimodal nature of speech processing is established by brain imaging studies showing activation of the superior temporal sulcus, a brain area related to multimodal integration (Calvert, Campbell and Brammer, 2000; Wright et al., 2003; Okada et al., 2013). Also, research shows that facial movements can alter activity in the sensory auditory cortex, in both humans and primates (Ghazanfar, 2009; Chandrasekaran, Lemus and Ghazanfar, 2013). More evidence for the high impact of visual cues during speech perception is provided by the fact that even incongruent visual facial cues influence the perception of speech segments. This is illustrated behaviourally by the McGurk effect where an auditory /bɑ/ syllable is often perceived as /dɑ/ when presented together with an incongruent visual /gɑ/ (McGurk and MacDonald, 1976). A neuroimaging study by Arnal, Wyart and Giraud (2011) revealed deviant neural processes when people were faced with incongruent cues, compared to congruent auditory-visual speech processing. Mismatches between visual speech signals

and auditory information activated fast and local brain regions around the primary sensory areas, whereas matching cues activated slower, higher order language regions in the brain to process speech at a semantical and syntactical level. Visual cues do not only influence speech perception per se, they also affect the cognitive load during listening by altering the listening effort during speech perception (Gosselin and Gagné, 2011; Stevens et al., 2013).

When the goal is to obtain an ecologically valid measure of daily speech processing abilities, adding accurate visual speech cues to the speech signal is of major concern. A straightforward way of implementing these cues into an auditory speech test is by presenting videomaterials. Unfortunately, recording videos of human speakers does not allow application of existing speech materials, even though many optimised and high quality speech intelligibility-measurement materials and procedures have been developed during the last two decades. New recordings would require a time-consuming optimisation, evaluation and validation of the speech material (Akeroyd et al., 2015), including characterisation of speech perception characteristics like the speech reception threshold (SRT), i.e. the signal-to-noise ratio (SNR) at which listeners understand 50% of the presented speech signal. In addition, visual cues that could steer the gaze of the listener, like lighting conditions, clothing or head movements of the speaker, are hard to control in videos. Hence, the interest in implementing speaking virtual human-like 3D characters in auditory assessment emerged. In this paper, we will refer to these 3D characters as “virtual humans”, even though most of the time the character consists of a dynamic head only.

In the last decades, many research projects related to human-computer interactions, language or auditory-visual speech processing and automated dialogue systems implemented auditory-visual speech synthesis. Human-computer interactions

have for example been integrated in a tutoring system aimed at enhancing science learning in elementary school (Ward et al., 2013). In this project, children who received face-to-face tutoring from a virtual human speaker developed by Ma et al. (2006), achieved larger learning gains compared to students in control classrooms. Virtual human speakers were also implemented in speech and language tutoring applications like training systems on pronunciation (Dey, Maddock and Nicolson, 2010; Peng et al., 2018) or on speech perception and production for hearing impaired individuals (Massaro and Light, 2004). More recently, Schreitmüller et al. (2017) demonstrated the applicability of "MASSY", a virtual human developed by Fagel and Clemens (2004), to investigate lipreading, visual benefit and auditory-visual integration of speech in both normal hearing and hearing impaired individuals. Even telephone conversations have been extended with virtual humans to augment speech intelligibility (Siciliano et al., 2003; Salvi et al., 2009).

Visual modelling of virtual humans can be divided in three categories: terminal-analog systems, anatomy-based systems or performance-driven animations (Mattheyses and Verhelst, 2015). The first approach refers to virtual humans in which the visual speech gestures are simulated by directly controlling the facial parameters that are linked to visual articulators (Massaro and Cohen, 1990; Salvi et al., 2009), whereas in anatomy-based systems the deformations are the result of an indirect modelling of the facial anatomy like skin, muscles and bones (Sifakis, Neverov and Fedkiw, 2005). In performance driven animations, facial motions are learned by mapping gestures from a real human speaker onto a 3D facial map by means of e.g. motion capture techniques (Williams, 1990; Fagel, Bailly and Elisei, 2007). Two approaches can also be combined, as is the case for "MASSY", where motion capture techniques are applied to determine the facial parameters of a terminal-analog system (Fagel and Clemens, 2004).

All visual synthesis methods have their advantages and selection of one of them is based on an intrinsic trade-off between the desired realistic picture and intelligibility of the virtual human on the one hand and the ease of creating it on the other hand (Mattheyses and Verhelst, 2015). According to Beskow (2004), terminal-analog approaches can outperform performance-driven methods in terms of speech intelligibility, although the latter tend to provide more natural mouth movements. However, anatomy-based systems or performance-driven animations typically require larger computational work.

The aim of the present report is to describe the development of a speaking virtual human and validate the quality of the provided visual facial information to define its potential applicability in future auditory assessment paradigms. Taking into account that we have a large amount of speech materials at our disposal, we opted for a modified terminal-analog method that allowed off-line manual mapping of speech phonemes to predefined mouth movements. In our case, this method was more efficient compared to e.g. online terminal-analog mapping or performance-driven auditory-visual mapping methods such as “MASSY”. Furthermore, we aimed for a realistic looking virtual human that can be easily integrated in 3D virtual environments to assess, for example, auditory-visual speech perception in complex daily listening situations. The need for multimodal, ecologically valid listening situations that can be applied in clinical practice is high among both researchers and clinicians (Pichora-Fuller et al., 2016). Therefore, realistic and dynamic 3D listening situations were created at our research department and the virtual human described in this study was especially designed to be implemented in these environments. We opted for modern, well-supported 3D rendering software systems to ensure the virtual humans and listening scenarios can be modified easily, according to the needs of researchers and clinicians. Furthermore, our virtual

human is innovative in that it represents a full virtual body instead of a talking head only, which is more realistic to implement in a 3D listening scenario.

We predicted that the mouth movements of our virtual human would be comparable to those of a real human speaker. The extent to which the virtual human features are realistic was evaluated in a speech reading experiment. The amount of visual benefit in speech understanding was assessed by comparing an auditory-visual (AV) speech intelligibility (SI) test to an auditory-only (AO) SI test.

Methods

Participants

Two groups of participants were recruited. First, a total of five hearing adults (one male, four female, aged 24, 25, 26, 29 and 56) participated in the speech reading study. All were native Flemish/Dutch speakers and had minimum one year of experience in speech reading by either following a speech reading course or teaching it to others. Second, a group of 35 young adults (13 male, 22 female) aged 18 to 30 (mean age = 22 years, SD = 2 years) took part in the AV SI experiment. All had bilateral normal audiometric pure-tone thresholds less than or equal to 20 dB HL at 250, 500, 1000, 2000, 4000 and 8000 Hz. Individuals of both groups reported normal or corrected vision. The Medical Ethical Committee of the University Hospitals and University of Leuven approved the experiment and materials (approval number B322201526677) and written consent was obtained from each participant.

Stimuli

Auditory stimuli

In order to capture a wide range of speech materials in the speech reading experiment,

both meaningful Flemish/Dutch words and sentences were presented to the participants. Monosyllabic (MS) consonant-vocal-consonant (CVC) words were selected from the NVA- and Lilliput-speech material (Wouters, Damman and Bosman, 1994; van Wieringen and Wouters, 2015) in order to cover all Flemish/Dutch phonemes. They were grouped together in three lists of 14 CVC-words. Four original lists of 10 disyllabic (DS) words were selected from the BLU-material (Wouters, Damman and Bosman, 1994) and six lists of 10 sentences from the male and female LIST (Jansen et al., 2014; van Wieringen and Wouters, 2008). The LIST was opted for because of its high standardisation and validation and widespread use in clinical practice in Flanders since 2008. The LIST-sentences are comparable to the American English HINT-sentences (Nilsson, Soli and Sullivan, 1994) and are semantically and syntactically predictable sentences, representative of daily communication, such as “Milk is healthy”. The sentences do not include questions, exclamations or proverbs and exist of two to three keywords (i.e. nouns, adjectives, numerals, adverbs and main verbs) and six to seven other words (i.e. articles, pronouns, prepositions, conjunctions, interjections and linking and auxiliary verbs). The total amount of syllables (89 to 90) and keywords (32 to 33) is balanced between lists. For the AV SI experiment, sentence lists from the female LIST material were randomly selected and presented in both speech weighted noise (SWN) and multitalker babble noise (MTN) (Francart, van Wieringen and Wouters, 2011).

Visual models

Three visual models were used to present the stimuli in the speech reading experiment (Figure 1, A). One 25-year-old male speech therapist was video-recorded uttering the different speech materials. He was instructed to pronounce the stimuli with the same

speaking rate as the original material. Recordings were made with a Canon XA10 professional camcorder and a shotgun microphone (Audio-Technica AT875R). The speaker was seated in a silent room in front of a neutrally colored background and was captured from shoulders to head.

In addition to the video model, two virtual humans were created by means of open source 3D graphics and animation software packages Blender (Blender, 2015, Version 2.76b) and Make Human (Make Human, 2015, Version 1.0.2). Blender and Make Human provide a predefined virtual human model consisting of a mesh, which simulates the skin of the virtual human, and bones attached to this mesh, which can be transformed to change the appearance of the mesh. To ensure a natural looking human appearance when bones are moved, multiple bones are grouped into a single configuration holding constraints on the transformations of the individual bones. As such, modification of a single configuration transforms groups of related bones. Starting from the predefined model, these configurations were modified to create a resting mouth shape position as well as distinct mouth shapes needed for the articulation of different speech sounds. A number of 17 mouth shapes was opted for based on the coding scheme of a semi-automated text to speech synthesis software, Annosoft Lipsync Tool (Annosoft Lipsync Tool, 2014, Version 4.0260), which was afterwards used to map the auditory speech signal to the mouth shapes. The mouth shapes were designed by manually adjusting the position and rotation of 12 facial configurations for the lower jaw, the lips and the tongue, based on Flemish/Dutch articulatory rules and expert knowledge. Three additional configurations were adjusted to obtain a realistic looking face, i.e. the chin appearance, retraction of the nose wings and the thickness of the upper cheeks. The resulting mouth shapes were static and stored individually in a facial bone-structure of a virtual human in Blender, containing transformation values for

each individual bone. The mouth shapes were first created for a neutral virtual human template (Figure 1, B) and later on copied to a male and female virtual human corresponding more closely to the voice timbre of the auditory stimuli used in our experiments (Figure 1, A). Mouth shapes were afterwards mapped to the audio files of the speech material by means of Annosoft Lipsync Tool. Both the original audio material and the audio signal of the video recordings of the male speech therapist were used for mapping. Plain text was automatically converted into a mouth shape sequence according to a many-to-one phoneme-to-mouth shape mapping scheme. The proposed mouth shapes were revised manually for selection, timing and intensity (Figure 2). The resulting sequences were saved as data strings in an output file, together with information on the interpolation between specific mouth shapes based on Annosoft's internal model of co-articulation. Finally, the output files were loaded in the 3D engine software Unity (Unity, 2016, Version 0.43.4) and parsed into objects holding temporal information on the start and intensity of mouth shapes as well as previously determined transformation values of individual bones for each mouth shape. Co-articulation between two neighbouring mouth shapes was accomplished by an interpolation calculation of the transformation of the individual bones, according to Annosoft's co-articulation model.

Procedure and analysis

Speech reading experiment

Participants were seated in a darkened room in front of an acoustically transparent screen (173 x 300 cm) on which both the male and female virtual human and the video recordings were presented alternately in complete silence. The visual models were shown at eye level and the participant sat at a distance of 1.0 m to the screen. The

mouth of the visual models was 3.8 cm wide, comparable to the real mouth size of a speaker during a typical two-side table conversation. Stimuli were presented through software programs Processing (Processing, 2016, Version 3.0.2), Pure Data (Pure Data, 2015, Version 0.43.4) and Unity (Unity, 2016, Version 0.43.4) in three blocks of approximately half an hour each. The order of the lists, words and sentences within a list and the presentation order of the visual models were randomised between participants.

Before the start of each list, the experimenter announced the type of speech material, i.e. MS words, DS words or sentences. The participants were asked to attentively watch the visual presentation of the speech stimuli and to verbally report all the phonemes they could identify. The video or virtual animation was paused by the experimenter after every stimulus and repetitions of the stimuli were provided when participants asked for it. Multiple answers were allowed, but no feedback was given. The experimenter encouraged the participants to always give a response and entered all answers manually into a computer. When multiple answers were given, the most correct answer was taken into account. For the sentences, only the keywords were analysed.

Intelligibility scores of the three visual models were calculated based on the number of mouth movements correctly identified by the participants. Therefore, more general viseme categories, i.e. categories of phonemes that cannot be distinguished on visual features alone (Fisher, 1968), were defined that contained all 17 subtle varying mouth movements (Tables 1 and 2). The viseme categories used for our study were based on previous work of van Son et al. (1994), up to date the most extensive research on categorisation of Dutch phonemes into visemes. Consonants were grouped according to place and manner of articulation (Table 1). Different from the categories of van Son, we created a separate category for the /w/. The pronunciation of this consonant differs

between Flanders, i.e. the Flemish/Dutch-speaking part in Belgium used in our research, and the Netherlands, with the Flemish one being uttered bilabial in contrast to the labiodental Dutch /w/. Furthermore, we gathered all nonlabial front and back consonants into one single category, because not all individuals are skilled enough to observe the differences between these consonants (van Son et al., 1994). In contrast to van Son et al. (1994), vowels were not categorised according to vowel duration, but to lip opening and lip rounding only (Table 2). One extra vowel was added as a separate category, being the mid-central vowel schwa.

MS and DS words were analysed according to a CVC or CVCCVC structure respectively. This means that both presented and answered stimuli were fit into a structure with three or six possible phoneme chunks. One chunk could consist of a single consonant or vowel or multiple phonemes such as a consonant cluster. For each chunk, a viseme category code of the corresponding presented stimuli was assigned, e.g. V1 for the vowel /e/ (Tables 1 and 2). When a consonant cluster consisted of phonemes of a different viseme category, a combined code was used for the corresponding chunk, e.g. C14 for the cluster /pr/. When the code of the presented and answered chunk aligned, a score of one was assigned to that chunk. Omissions and insertions received a score of zero if they resulted in a different viseme code. All scores of one list were summed up and divided by the total amount of chunks in that list, resulting in a % correct score. The same procedure was applied to keyword scoring of the LIST sentences. The only difference was that keywords did not always follow a CVC or CVCCVC structure. Therefore, keywords were first split into syllables and each syllable was fit into a CVC structure of which the first and initial chunk could be empty. An example of keyword scoring can be found in Table 3.

Auditory-visual speech intelligibility

In a second experiment, female LIST sentences were uttered by the female virtual human only in a virtual restaurant scenario projected on the screen in front of the participant. The set-up was identical to the speech reading experiment, except for the addition of two loudspeakers. Sentences were presented through a loudspeaker at 0°, standing at eye level 0.11 m behind the screen. A speaker in the zenith, right above the participant, was used for presentation of both SWN and MTN.

SRT's, i.e. the SNR at which 50% of the keywords is correctly repeated, were obtained by presenting lists of 10 sentences in noise, starting at a 60 dB A noise level and 48 dB A speech level, adaptively adjusting the speech level by 2 dB according to the answer of the participant. Only keywords had to be repeated correctly. The SRT was calculated by averaging the SNR's of the last five sentences plus the 11^{th} fictive sentence of a single list. Sentences were presented both AO and AV in SWN and MTN. In the AO condition, participants watched a scenario similar to the restaurant but without the virtual human. Each condition was conducted twice and the order was randomly assigned between participants.

Results

All results were analysed in the R software interface (R Core Team, 2017). Additional packages were used when needed, including the nlme package for general linear models (Pinheiro et al., 2017), the ez package for ANOVA-analysis (Lawrence, 2016) and the WRS2 package for robust paired samples t-tests (Mair, Schoenbrodt and Wilcox, 2017). Alpha-levels were set at 0.05 and p-values were corrected for multiple comparisons applying the Bonferroni method.

Speech reading experiment

Figure 3 compares the identification scores obtained in the speech reading experiment for the three visual models, i.e. the video recording of a real person and the two virtual humans, for the three different speech materials. The same participants were tested in all conditions. To account for within-individual variances, we analysed the repeated-measures data using a general linear model. The type of visual model had a significant effect on speech reading scores, $\chi^2(2) = 16.00$, $p < 0.01$, as did the type of speech material used in our experiment, $\chi^2(2) = 53.00$, $p < 0.01$. No significant interaction effect between visual model and speech material was observed, $\chi^2(4) = 5.00$, $p = 0.24$. Contrasts revealed that (1) scores obtained through the video recordings were significantly better compared to the ones from the two virtual humans, $b = -6.00$, $t(8) = -8.00$, $p < 0.01$ but (2) speech reading scores were comparable for the male and female virtual human, $b = -2.00$, $t(8) = -1.60$, $p = 0.15$; also (3) speech reading scores of the sentences were significantly lower compared to both MS and DS word scores, $b = 7.00$, $t(24) = 10.20$, $p < 0.01$ albeit (4) scores on MS and DS words were not significantly different from each other, $b = 2.00$, $t(24) = 1.90$, $p = 0.07$.

Since the scores of the male and female virtual humans did not differ significantly from each other, they were averaged per participant and subsequently divided by the speech reading score of the video model for the corresponding speech material. In this way, individual proportion correct scores per speech material were obtained (Figure 4) and individual differences in speech reading ability were ruled out. In what follows, these individual proportion correct scores are referred to as ‘normalised proportion correct scores’. Normalised proportion correct scores were analysed through a non-parametrical Friedman’s ANOVA because of non-normal data distributions. The analysis revealed significant differences between the scores $\chi^2(2) = 7.60$, $p = 0.02$,

although post-hoc paired samples robust t-tests could not confirm any significant difference: the normalised proportion correct scores of MS words (*Mdn* = 0.85) were equal to the scores of DS words (*Mdn* = 0.85), $p = 1.00$, $r = 0.02$, DS words (*Mdn* = 0.85) to sentences (*Mdn* = 0.62), $p = 0.37$, $r = 0.75$ and MS words (*Mdn* = 0.85) to sentences (*Mdn* = 0.62), $p = 0.52$, $r = 0.79$.

Based on the raw identification scores (Figure 3), scores on the sentences were significantly lower than the speech reading scores obtained on the MS and DS words. To investigate this difference in more detail, two extra sentence lists were presented to four out of five participants for the female virtual human only. In the previous experiment, sentences had to be speech read completely and no additional information was given on the possible words of the sentence. Having to remember the phonemes of an entire sentence could have impaired the speech reading scores, since the cognitive load on short-term memory during this task is higher compared to the memorisation of single words. Therefore, for the two extra sentence lists, sheets were handed over to the participant right before the start of the experiment with the non-keywords of the sentences printed on it. In this way, the participant had to speech read the keywords only, achieving a difficulty level more closely related to the MS and DS words task. Scores on the two extra sentence lists were averaged per participant and compared to scores on the MS words, DS words and regular sentences (Figure 5). Due to non-normal distributions, differences between all of the speech materials were investigated by means of a non-parametrical Friedman's ANOVA, which revealed significant differences $\chi^2(3) = 9.30$, $p = 0.03$. Post-hoc comparisons were conducted via paired samples robust t-tests. Speech reading scores on the sentences (*Mdn* = 37.5%) appeared to be significantly different from the scores on the extra sentences (*Mdn* = 57.2%), $p = 0.03$, $r = 0.76$. On the contrary, scores on the MS words (*Mdn* = 67.1%) and DS words

(*Mdn* = 60.8%) did not differ significantly from the scores on the extra sentences, p = 0.55, 0.42 and r = 0.49, 0.24 respectively.

Auditory-visual speech intelligibility

Figure 6 shows the SRT's of the AV SI experiment for the young adult group (N = 35) in the AO and AV condition for both noise materials. The SRT in the AO SWN condition (*Mean* = -12.1 dB SNR, *SE* = 0.3 dB) was better than expected based on research of van Wieringen and Wouters (2008), who found a norm SRT of -8.0 dB SNR ± 0.2 dB. This can be explained by the fact we spatially separated the noise source from the target sound. Listening in free field with two ears instead of one allowed our participants to make use of head diffraction effects and binaural listening phenomena like binaural squelch and binaural redundancy to improve speech understanding. The benefit of adding visual speech cues to the auditory signal was analysed through a factorial repeated-measures ANOVA and revealed a main effect of both mode $F(1, 34) = 33.58$, $p < 0.01$ and noise $F(1, 34) = 220.87$, $p < 0.01$. Addition of visual cues reduced the SRT in SWN on average by 1.4 dB (SE = 0.4 dB) and in MTN on average by 1.8 dB (SE = 0.4 dB). Overall, SRT's were significantly better in SWN (AV: *Mean* = -13.5 dB, *SE* = 0.3 dB; AO: *Mean* = -12.1 dB, *SE* = 0.3 dB) compared to SRT's obtained in MTN (AV: *Mean* = -10.3 dB, *SE* = 0.4 dB; AO: *Mean* = -8.5 dB, *SE* = 0.4 dB), which is expected based on research on the acoustical and informational masking effects of different noise types (Francart, van Wieringen and Wouters, 2011). There was no significant interaction $F(1, 34) = 0.58$, $p = 0.45$.

Discussion

The presented work describes the development and evaluation of two virtual human speakers, designed to be implemented in future auditory assessment methods where

speech will be presented auditory-visually in order to obtain an ecologically valid measure of speech processing. Both virtual humans, a male and female speaker, yielded the same intelligibility results, indicating they can be used interchangeably. Yet, present implementation of the virtual humans resulted in slightly lower speech reading scores than the videotaped speaker. This agrees with many reports showing that natural speakers almost always outperform their synthetic counterparts (Benoît and Le Goff, 1998; Geiger, Ezzat and Poggio, 2003; Siciliano et al., 2003; Fagel and Clemens, 2004; Fagel, Bailly and Elisei, 2007; Dey, Maddock and Nicolson, 2010).

Direct comparison of speech reading scores on our virtual humans with scores on other virtual human speakers is not straightforward, since modelling techniques differ greatly and evaluation methods are not uniform. In addition, the large variability of individuals' speech reading skills hampers direct comparison of raw intelligibility scores across participants. Therefore, we calculated normalised proportion correct scores, which are suitable to compare to other research. Fagel et al. (2007) set up a forced-choice speech reading experiment with German CVC-sequences presented through a performance-driven virtual human and natural video images. Proportion correct scores were about 0.60 (virtual human: 19.4%, video: 32.5%), which is equivalent to our results, i.e. a normalised proportion correct of 0.62 when speech reading sentences. Nevertheless, we found a normalised proportion correct score of 0.85 for speech reading both MS and DS words, outperforming the scores on CVC-sequences in the research of Fagel et al. (2007). Geiger et al. (2003) developed virtual human speakers by pasting synthetic mouth regions onto real faces, preserving the original head and eye movements. The proportion correct score of phonemes for English single words and sentences was about 0.71 (virtual human: 21.2%, video: 30.0%), which is similar to our results. Comparable scores were also found by Siciliano

et al. (2003), who reported phoneme scores correct of 0.58 (virtual human: 13.6%, video: 23.4%) for English CVC sequences presented by their virtual model. Although normalised proportion correct scores were similar between studies, it seems that our virtual humans resulted in higher raw speech reading scores compared to other virtual implementations (Geiger, Ezzat and Poggio, 2003; Siciliano et al., 2003; Fagel, Bailly and Elisei, 2007). In a study of Schreitmüller et al. (2017) for example, normal hearing and cochlear-implanted individuals obtained average scores of 12% and 38% words correct on speech reading nonsense matrix sentences. In our study, individuals scored on average 70%, 64% and 46% visemes correct on MS words, DS words and keywords of meaningful sentences respectively (Figure 3). However, it should be noted that scoring criteria and methods varied across studies.

Both MS and DS words as well as sentences were presented to our participants to investigate potential differences in difficulty. Phonemes and thus visemes were harder to identify when presented in sentences compared to single words. Memory storage of the presented phonemes could have impaired scores on the sentences, since the amount of phonemes per sentence was much higher than in MS and DS words. In other words, task-related difficulties rather than inconclusive mouth movements may have caused these lower results. The notion of increased memory load is in line with participants' spontaneous remarks during the experiment. Overall, participants asked for more repetitions during sentence presentations since they found it hard to remember all presented phonemes. Higher identification scores on two extra sentence lists, relative to scores on the regular sentences, further support this idea. In these lists, the amount of phonemes to be memorised was reduced by revealing all non-keywords of the sentence upfront. One could argue that revealing extra information on the words did not only reduce the needed memory capacity, but also provided more contextual information,

facilitating speech reading because of semantic and morphosyntactic redundancy (Martin et al., 1983; Grant and Seitz, 2000b). Interestingly, the overall benefit of context was smaller than expected based on research of Boothroyd and Nittrouer (1988), who designed a formula to predict word recognition in sentences based on recognition of isolated words, taking sentence context into account as a free parameter. According to their formula, assuming a sentence context factor of 2.7, and based on the median scores of our participants on the MS and DS words, i.e. 67% and 61% correct respectively (Figure 5), we expected our participants to achieve sentence speech reading scores of 92% to 95%. In our study, sentence scores were about 57%. This could be due to the fact we only provided non-keywords, which were mostly articles and were less informative than keywords, typically nouns and verbs. Work from Martin et al. (1983) supports this idea, showing that words are easier to speech read when they are highly predictable. Moreover, other relevant signal cues like word boundaries are hard to extract in absolute silence which was the case in our experiments. Nevertheless, in an auditory-visual environment, these cues help to gain access to relevant context information and thereby improve performance (Grant and Seitz, 2000b).

In terms of improvement in speech intelligibility, we found that the presentation of our female virtual human reduced the SRT's on average by 1.4 dB in SWN and 1.8 dB in MTN, in contrast to an auditory-only listening condition. This visual speech benefit is in agreement with research demonstrating that visual cues improve speech understanding in noisy conditions (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987, 1990; Grant, Walden and Seitz, 1998) and provides support for the applicability of our virtual humans in future research. Whereas many studies have focused on improvements in percentage correct speech intelligibility, only a few measured changes in SRT's, yielding benefits in the range of 2 to 6 dB when video

images are presented (Bernstein and Grant, 2009; Brungart, Sheffield and Kubli, 2014). These benefits are in line with our results, given the fact we used synthetic human speakers instead of real video images. Moreover, our benefit could have been larger when presenting sentences at lower SNR's approximating intelligibility levels below the SRT, since speech reading scores tend to be higher the more noise is presented (Benoît and Le Goff, 1998). This idea is supported by the principle of inverse effectiveness, which assumes that auditory-visual integration becomes more prominent when unimodal processing of either auditory or visual cues declines, as is the case in noisy conditions (Tye-Murray et al., 2010).

In general, the amount of correctly identified visemes and the significant visual benefit obtained by our virtual humans establishes the realism and quality of both the male and female virtual human. They differ from previously developed virtual humans in that the mouth movements are incorporated in full body virtual humans, as opposed to virtual talking heads, and that they are designed with easily accessible 3D rendering software systems, facilitating the implementation in ecologically valid 3D scenarios. Hence, they open avenues for the use of full-body virtual humans in a clinical tool to assess auditory-visual speech processing abilities in complex daily listening situations. Nevertheless, improvements can still be made in terms of intelligibility. A detailed analysis of the answers based on virtual human and video projection is shown in the confusion matrices in Figure 7, revealing which categories of visemes were harder to identify and thus need further investigation. Omitted answers and answers that could not be assigned to one of the stimulus categories (e.g. answering with a vowel to a consonant stimulus) were left out. When it comes to consonants, the most striking result was the difficulty in recognising the /w/ (C3), which tended to be identified as a nonlabial consonant (C4). In terms of vowels, V3 and V4 categories seemed to be

regularly interchanged compared to the video images. This is not surprising, given that both vowel categories differ subtly in terms of the amount of lip opening only. In addition, V1 seemed harder to recognise, possibly because these vowels encompass less pronounced mouth movements. Refinements of all above categories will be made in future research, possibly elevating the intelligibility of our virtual humans to an even higher level.

Acknowledgements

The authors would like to thank all the participants who took part in this study, as well as the students from the Speech Language Pathology and Audiology Sciences educational program for their help with data collection. Sam Denys is gratefully acknowledged for his help with the speech reading video recordings. This work was supported by the Oticon Foundation and a TBM-FWO grant from the Research Foundation-Flanders under Grant [T002216N].

Declaration of interest

The authors report no conflicts of interest.

References

Akeroyd, M. A., Arlinger, S., Bentler, R. A., Boothroyd, A., Dillier, N., Dreschler, W. A., Gagné, J. P., Lutman, M., Wouters, J., Wong, L. and Kollmeier, B. (2015). 'International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests: ICRA Working Group on Multilingual Speech Tests', *Int J Audiol*, 54(sup2), pp. 17–22.

Annosoft Lipsync Tool. (2014). Available at: http://www.annosoft.com/.

Arnal, L. H., Wyart, V. and Giraud, A. L. (2011). 'Transitions in neural oscillations reflect prediction errors generated in audiovisual speech', *Nat Neurosci*, 14(6), pp. 797–801.

Benoît, C. and Le Goff, B. (1998). 'Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP', *Speech Commun*, 26(1), pp. 117–129.

Bernstein, J. G. and Grant, K. W. (2009). 'Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners', *J Acoust Soc Am*, 125(5), pp. 3358–3372.

Bernstein, L. E., Auer, E. T. and Takayanagi, S. (2004). 'Auditory speech detection in noise enhanced by lipreading', *Speech Commun*, 44(1), pp. 5–18.

Beskow, J. (2004). 'Trainable articulatory control models for visual speech synthesis', *Int J Speech Technol*, 7(4), pp. 335–349.

Blender. (2015). Available at: https://www.blender.org/.

Boothroyd, A. and Nittrouer, S. (1988). 'Mathematical treatment of context effects in phoneme and word recognition', *J Acoust Soc Am*, 84(1), pp. 101–114.

Brungart, D. S., Sheffield, B. M. and Kubli, L. R. (2014). 'Development of a test battery

for evaluating speech perception in complex listening environments', *J Acoust Soc Am* 136(2), pp. 777–790.

Calvert, G. A., Campbell, R. and Brammer, M. J. (2000). 'Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex', *Curr Biol*, 10(11), pp. 649–657.

Chandrasekaran, C., Lemus, L. and Ghazanfar, A. A. (2013). 'Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection', *Proc Natl Acad Sci U S A*, 110(48), pp. E4668–E4677.

Dey, P., Maddock, S. C. and Nicolson, R. (2010). 'Evaluation of A Viseme-Driven Talking Head', in *Proc. Theory and Practice of Computer Graphic*, pp. 139-142.

Fagel, S., Bailly, G. and Elisei, F. (2007). 'Intelligibility of natural and 3D-cloned German speech', in *International Conference on Auditory-Visual Speech Processing, AVSP*. Hilvarenbeek, The Netherlands, pp. 56–61.

Fagel, S. and Clemens, C. (2004). 'An articulation model for audiovisual speech synthesis — Determination, adjustment, evaluation', *Speech Commun*, 44(1), pp. 141–154.

Fisher, C. G. (1968). 'Confusions among visually perceived consonants', *J Speech Lang Hear Res*, 11(4), pp. 796-804.

Francart, T., van Wieringen, A. and Wouters, J. (2011). 'Comparison of fluctuating maskers for speech recognition tests', *Int J Audiol*, 50(1), pp. 2–13.

Geiger, G., Ezzat, T. and Poggio, T. (2003). Perceptual Evaluation of Video-Realistic Speech. Tech Report: CBCL Paper 224/AI Memo 2003-003, MIT Artificial Intelligence Laboratory, Cambridge.

Ghazanfar, A. A. (2009). 'The multisensory roles for auditory cortex in primate vocal communication', *Hear Res*, 258(1), pp. 113–120.

Gosselin, P. A. and Gagné, J. P. (2011). 'Older adults expend more listening effort than young adults recognizing audiovisual speech in noise', *Int J Audiol*, 50(11), pp. 786–792.

Grange, J. A. and Culling, J. F. (2016). 'Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions', *J Acoust Soc Am* 140(6), pp. 4061–4072.

Grant, K. W. and Seitz, P. (2000a). 'The use of visible speech cues for improving auditory detection of spoken sentences', *J Acoust Soc Am*, 108(3), pp. 1197–1208.

Grant, K. W. and Seitz, P. F. (2000b). 'The recognition of isolated words and words in sentences: Individual variability in the use of sentence context', *J Acoust Soc Am*, 107(2), pp. 1000–1011.

Grant, K. W., Walden, B. E. and Seitz, P. F. (1998). 'Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration', *J Acoust Soc Am*, 103(5), pp. 2677–2690.

Jansen, S., Koning, R., Wouters, J. and van Wieringen, A. (2014). 'Development and validation of the Leuven intelligibility sentence test with male speaker (LIST-m)', *Int J Audiol*, 53(1), pp. 55–59.

Lawrence, M. A. (2016). 'ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0'. Available at: https://cran.r-project.org/package=ez.

Ma, J., Cole, R., Pellom, B., Ward, W. and Wise, B., (2006). 'Accurate visible speech synthesis based on concatenating variable length motion capture data'. *IEEE Trans Vis Comput Graph*, 12(2), pp.266-276.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J. and Parra, L. C. (2009). 'Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space', *PLoS One*, 4(3), p. e4638.

MacLeod, A. and Summerfield, Q. (1987). 'Quantifying the contribution of vision to

speech perception in noise', *Br J Audiol*, 21(2), pp. 131–141.

MacLeod, A. and Summerfield, Q. (1990). 'A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use', *Br J Audiol*, 24(1), pp. 29–43.

Mair, P., Schoenbrodt, F. and Wilcox, R. (2017). 'WRS2: Wilcox robust estimation and testing'.

Make Human. (2015). Available at: https://www.makehuman.org/.

Martin, L. F. A., Clark, G. M., Seligman, P. M. and Tong, Y. C. (1983). 'A lip-reading assessment for profoundly deaf patients', *J Laryngol Otol*, 97(4), pp. 343–350.

Massaro, D. W. and Cohen, M. M. (1990). 'Perception of synthesized audible and visible speech', *Psychol Sci*, 1(1), pp. 55–63.

Massaro, D. W. and Light, J. (2004). 'Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss', *J Speech Lang Hear Res*, 47(2), pp. 304–320.

Mattheyses, W. and Verhelst, W. (2015). 'Audiovisual speech synthesis: An overview of the state-of-the-art', *Speech Commun*, 66, pp. 182–217.

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

Nilsson, M., Soli, S.D. & Sullivan, J. A. (1994). 'Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise', *J Acoust Soc Am*, 95, pp. 1085-1099.

Okada, K., Venezia, J. H., Matchin, W., Saberi, K. and Hickok, G. (2013). 'An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex', *PLoS One*, 8(6), p. e68959.

Peng, X., Chen, H., Wang, L. and Wang, H. (2018). 'Evaluating a 3-D virtual talking head on pronunciation learning', *Int J Hum Comput Stud*, Elsevier Ltd, 109, pp. 26–40.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L. and Naylor, G., (2016). 'Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL)', *Ear Hear*, 37, pp.5S-27S.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and Team, R. C. (2017). 'nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131'.

Processing. (2016). Available at: https://processing.org/.

Pure data. (2015). Available at: https://puredata.info/.

R Core Team (2017). 'R: A language and environment for statistical computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org.

Rosenblum, L. D. (2008). 'Speech Perception as a Multimodal Phenomenon', *Curr Dir Psychol Sci*, 17(6), pp. 405–409.

Salvi, G., Beskow, J., Al Moubayed, S. and Granström, B. (2009). 'SynFace — Speech-Driven Facial Animation for Virtual Speech-Reading Support', *EURASIP J Audio Speech Music Process*, 2009(1), p. 191940.

Schreitmüller, S., Frenken, M., Bentz, L., Ortmann, M., Walger, M. and Meister, H. (2018). 'Validating a Method to Assess Lipreading, Audiovisual Gain, and Integration During Speech Reception With Cochlear-Implanted and Normal-Hearing Subjects Using a Talking Head', *Ear Hear*, 39(3), pp. 503-516.

Schwartz, J. L., Berthommier, F. and Savariaux, C. (2004). 'Seeing to hear better: evidence for early audio-visual interactions in speech identification', *Cognition*, 93(2), pp. B69–B78.

Siciliano, C., Williams, G., Beskow, J. and Faulkner, A. (2003). 'Evaluation of a Multilingual Synthetic Talking Face as a Communication Aid for the Hearing Impaired', in *International Congress of Phonetic Sciences, ICPhS*. Barcelona, Spain, pp 1-4.

Sifakis, E., Neverov, I. and Fedkiw, R. (2005). 'Automatic determination of facial muscle activations from sparse motion capture marker data', in *ACM Trans Graph*, 24(3), pp. 417–425.

Stevens, C. J., Gibert, G., Leung, Y. and Zhang, Z. (2013). 'Evaluating a synthetic talking head using a dual task: Modality effects on speech understanding and cognitive load', *Int J Hum Comput Stud*, 71(4), pp. 440–454.

Sumby, W. H. and Pollack, I. (1954). 'Visual Contribution to Speech Intelligibility in Noise', *J Acoust Soc Am*, 26(2), pp. 212–215.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J. and Hale, S. (2010). 'Aging, Audiovisual Integration, and the Principle of Inverse Effectiveness', *Ear Hear*, 31(5), pp. 636–644.

Unity. (2016). Available at: https://unity3d.com/.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E. and Weston, T. (2013). 'My Science Tutor : A Conversational Multimedia Virtual Tutor', *J Educ Psychol*, 105(4), pp. 1115–1125.

van Son, N., Huiskamp, T. M. I., Bosman, A. J. and Smoorenburg, G. F. (1994). 'Viseme classifications of Dutch consonants and vowels', *J Acoust Soc Am*, 96(3), pp. 1341–1355.

van Wieringen, A. and Wouters, J. (2008). 'LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands', *Int J Audiol*, 47(6), pp. 348–355.

van Wieringen, A. and Wouters, J. (2015). 'What can we expect of normally-developing

children implanted at a young age with respect to their auditory, linguistic and cognitive skills?', *Hear Res*, 322, pp. 171–179.

Williams, L. (1990). 'Performance-Driven Facial Animation', in *ACM SIGGRAPH Computer Graphics Conf.*, 24(4), pp. 235–242.

Woodhouse, L., Hickson, L. and Dodd, B. (2009). 'Review of visual speech perception by hearing and hearing-impaired people: clinical implications.', *Int J Lang Commun Disord*, 44(3), pp. 253–270.

Wouters, J., Damman, W. and Bosman, A. J. (1994). 'Vlaamse opname van woordenlijsten voor spraakaudiometrie', *Logopedie: informatiemedium van de Vlaamse vereniging voor logopedisten*, 7(6), pp. 28–34.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J. and McCarthy, G. (2003). 'Polysensory Interactions along Lateral Temporal Regions Evoked by Audiovisual Speech', *Cereb Cortex*, 13(10), pp. 1034–1043.

Tables

Table 1. Categorisation of Flemish/Dutch consonants.

Consonants	Category	Place and manner of articulation
/p, b, m/	C1	Bilabial plosive/nasal
/f, v/	C2	Labiodental fricative
/w/	C3	Bilabial fricative/semi-vocal
/t, d, n, j, l, s, z, ʃ, k, ɣ, r, x, ŋ, h/	C4	Nonlabial consonants

1

2 Table 2. Categorisation of Flemish/Dutch vowels.

Vowels	**Category**	**Lip opening and lip rounding**
/i, ɪ, e, ɛ/	V1	Closed and half-closed, unrounded
/ɛɪ, ɑ, a/	V2	Half-open and open, unrounded
/u, y, ʌ/	V3	Closed and half-closed, rounded
/ɔ, o, ø/	V4	Half-open, rounded
/œy, ɔu/	V5	Closing, rounding diphthongs
/ə/	V6	Middle, unrounded

3
4

Table 3. Scoring example of keywords of a LIST-sentence.

	Stimulus (S) and answer (A)						
Words and [keywords]	S:	Ze [viel] van de [trap].					
	A:	Ze [vreesde] voor de [trap].					
Phoneme chunk	S:	v	ie	l	tr	a	p
	A:	vr	ee	sde	tr	a	p
Viseme category code	S:	C2	V1	C4	C4	V2	C1
	A:	C24	V1	C4V6	C4	V2	C1
Scoring: 4/6 = 67%		0	1	0	1	1	1

Figure captions

Figure 1. Panel A: video-recorded male speaker (left) and male and female virtual human model (middle and right) used in the speech reading experiment (all models) and the AV SI test (female virtual human only). Panel B: neutral virtual human model uttering the vowel /o/.

Figure 2. Screenshot of the Annosoft Lipsync Tool for the mapping of the monosyllabic word ‘hang’ with a: auditory waveform, b: time bar, c: selected mouth shapes, d: timing and intensity of selected mouth shapes and e: plain text of speech stimulus.

Figure 3. Individual percentage correct identification scores on the speech reading experiment for different speech materials and visual models. The lines connect mean scores per visual model, error bars represent SE of the mean, N = 5.

Figure 4. Boxplots representing the normalised proportion of visemes correct when speech reading the virtual humans, N = 5.

Figure 5. Individual percentage correct scores and line connecting mean scores for all speech materials, for speech reading the female virtual human. Each symbol represents a different participant, N = 4.

Figure 6. Boxplots representing the SRT’s of the speech intelligibility experiment in different noise conditions (MTN = multitalker babble noise, SWN = speech weighted noise) and different listening modes (AO = Auditory-Only, AV = Auditory-Visual) for young normal hearing participants, N = 35.

Figure 7. Confusion matrices for the different consonant (top row) and vowel (bottom row) categories for all visual models: A = virtual humans, B = video. Numbers indicate percentage correct answers. Consonant and vowel categories are defined in Table 1 and Table 2.