A virtual speaker in noisy classroom conditions: supporting or disrupting children’s listening comprehension?

Jens Nirme\textsuperscript{a}, Magnus Haake\textsuperscript{a}, Viveka Lyberg Åhlander\textsuperscript{b}, Jonas Brännström\textsuperscript{b} and Birgitta Sahlén\textsuperscript{b}

\textsuperscript{a}Division of Cognitive Science, Lund University, Lund, Sweden; \textsuperscript{b}Division of Logopedics, Phoniatrics and Vocology, Department of Clinical Sciences, Lund University, Lund, Sweden

\section*{ABSTRACT}

\textbf{Aim:} Seeing a speaker’s face facilitates speech recognition, particularly under noisy conditions. Evidence for how it might affect comprehension of the content of the speech is more sparse. We investigated how children’s listening comprehension is affected by multi-talker babble noise, with or without presentation of a digitally animated virtual speaker, and whether successful comprehension is related to performance on a test of executive functioning.

\textbf{Materials and Methods:} We performed a mixed-design experiment with 55 (34 female) participants (8- to 9-year-olds), recruited from Swedish elementary schools. The children were presented with four different narratives, each in one of four conditions: audio-only presentation in a quiet setting, audio-only presentation in noisy setting, audio-visual presentation in a quiet setting, and audio-visual presentation in a noisy setting. After each narrative, the children answered questions on the content and rated their perceived listening effort. Finally, they performed a test of executive functioning.

\textbf{Results:} We found significantly fewer correct answers to explicit content questions after listening in noise. This negative effect was only mitigated to a marginally significant degree by audio-visual presentation. Strong executive function only predicted more correct answers in quiet settings.

\textbf{Conclusions:} Altogether, our results are inconclusive regarding how seeing a virtual speaker affects listening comprehension. We discuss how methodological adjustments, including modifications to our virtual speaker, can be used to discriminate between possible explanations to our results and contribute to understanding the listening conditions children face in a typical classroom.

\section*{Introduction}

\section*{Background}

Oral instruction is a major part of how most children receive their basic education at school. For oral instruction to be effective, children have to be able to comprehend what the teacher is saying. Classrooms are however often challenging listening environments. Children have to listen and comprehend surrounded by competing speech that might mask or distract from their teacher’s speech \cite{1,2}.

Noise in classrooms can have serious consequences, putting large groups of children at disadvantage. Academic performance of children aged 7–11 in the London area was negatively correlated with the noise levels in their schools \cite{3}. In our study, the focus is on listening (passage) comprehension in noise, and how it is affected by seeing the face of a speaker.

It is well established that \textit{speech recognition}, a prerequisite for comprehension, is impaired by noise. Speech recognition is typically assessed as the ability to repeat heard words \cite{4} or sentences \cite{5}. Important for our study, children’s speech recognition has been found to be more sensitive to noise compared to adults \cite{6}. Speech recognition does not require any semantic processing, but it can be mediated by semantic context \cite{4}.

Speech recognition can also be aided by visual context. School children can typically see their teacher speaking, which enables audiovisual integration that can affect speech processing in several ways. Adults’ speech recognition in noise has been found to improve when a speaker’s face is visible \cite{7}. Seeing lip movements have a direct effect on speech recognition, that is not always beneficial for disambiguating speech. The so-called McGurk effect refers to how seeing synchronous lip movements affect perception of spoken syllables \cite{8}. Besides lip reading, head movements coordinated with speech prosody (“visual prosody”) have been shown to facilitate recognition of syllables in spoken Japanese \cite{9}.

In the current study, we compare effects on comprehension of speech matched with the movements of a digitally rendered face. Technological advancements have enabled computers to both produce speech and present a speaker visually. The terms Virtual Humans, Virtual Agents, and Embodied Conversational Agents \cite{10} have been used to describe such implementations and their potential value in educational or instructional applications have been pointed out \cite{11}, as well as their importance as research tools \cite{12,13}. Adults’ speech recognition has been found to be facilitated with digitally rendered faces \cite{14–17}. However, the facilitation is generally weaker than from a video
recording of a speaker. Since the educational scenario that we aimed at creating in the current study involves no interaction but fixed roles of the teacher (speaker) and children (listeners), we use the term virtual speaker.

Recognition is, as mentioned, one important prerequisite to comprehend speech. Comprehension however also require listeners to extract and infer meaning from speech, as well as relating it to previous knowledge and encoding it for later retrieval. Even at noise levels where speech is accurately recognized, comprehension can be impeded. Studies have found that adults were less able to answer questions about content included in lectures they had heard in noise [18], and that the effect is stronger for children [19]. Similarly, another study assessed listening comprehension as the ability to follow oral instructions and found children (but not adults) to be affected in the presence of competing speech [20]. Yet another study tested children listening in noisy classroom conditions [21] on five different aspects of listening comprehension included in a standardized listening comprehension test [22] identifying the main idea, defining vocabulary, recalling details, inferring information, and identifying the most pertinent information. Most children performed below the standardized 95% confidence interval for their age on the three latter aspects [21].

In a review paper from 2012, Mattys et al. map how factors defining the speech signal, listening environment and listener capabilities affect comprehension differently, and discuss links to general cognitive capacities [23]. It has been demonstrated that strong working memory capacity can alleviate the negative effect of noise on reading comprehension [24] and that comprehension of both typical and hoarse voices was positively correlated with performance on the “Elithorn’s Mazes” test (probing executive functioning including attentional and inhibitory control) [25]. Increased cognitive load also requires listeners to expend more effort, possibly inducing fatigue and stress. Thus, presence of competing speech has not only been linked to decreased performance but also to both physiological indices of effort [26] and self-reported perceived listening effort and “frustration” (interpreted as indicator of stress) [27].

In this context, effective audiovisual integration might challenge a child’s cognitive capacities, and make the visual channel a distraction rather than a support. One study found executive functioning in adults to strongly predict “visual enhancement” of the ability to recognize spoken sentences in competing speech [28]. Moreover, at least one study has found that visual presentation of the speaker increased load on working memory [29]. Another study investigated listening effort, and found no general effect of visual presentation, but that participants with strong lip-reading ability and working memory capacity were more likely to benefit from visual cues, when listening in multi-talkers babble noise. The authors interpret their results as indicating that audio-visual integration depends on more general cognitive resources [30].

The study presented in this paper investigates how children’s listening comprehension is affected by multi-talker babble noise with or without visual presentation of a speaker. For our implementation, we used a virtual speaker, speaking with a hoarse voice. Attempting to compensate for a noisy classroom environment, teachers often strain their voices to a point where the voice becomes hoarse. Speakers generally tend to speak about 10 dB louder than the background and in a higher pitch [31]. A survey of 22 randomly selected schools throughout the south of Sweden found that 37% of teachers suffered at least occasionally from voice problems [32]. The rationale behind the use of a degraded voice was thus to preserve ecological validity as well as to make our results comparable to parallel studies using similar methods [33]. The aim of the study was not to evaluate effects of voice quality. It is, however, worth noting that degraded voice quality can itself (in silent background) increase listening load [34] and impede children’s comprehension [35].

**Study objective**

The main purpose of the current study was to investigate if a virtual speaker can serve as a visual support for listening comprehension in adverse conditions that approximate a classroom environment.

**Research question 1**

Will children’s comprehension and perceived listening effort suffer when listening to a narrative read in multi-talker babble noise compared to in a quiet setting?

**Research question 2**

Is children’s comprehension and perceived listening effort mitigated or accentuated when they can see a virtual speaker?

**Research question 3**

What role does executive functioning play for children’s comprehension and perceived listening effort?

**Material**

**Listening comprehension**

To investigate our research questions regarding effects of noise, visual presentation of the speaker and executive functioning on listening comprehension we used material from the listening comprehension part of the Clinical Evaluation of Language Fundamentals (CELF4), a commonly used assessment tool for language skill [36]. The listening comprehension part of CELF4 involves listening to short narrative texts. Each narrative is accompanied by five questions. The questions are divided into three “content” questions (about details or facts explicitly mentioned: “What did the students receive more than pizza and soft drinks?”), one “inference” question (requiring extrapolation based on the content, “Why do you think Pricken ran in the opposite direction?”) and one “summary” question (a simple question about the overall theme of the narrative, “How do you think it ended?”). Predefined guidelines for scoring (defining correct versus incorrect answers) are provided for each question. Due to the non-specific answer definitions, the
The children responded to the two questions using Visual-... as simply as possible in order to be administrable to children.

The background noise used in the A–N and AV–N conditions was constructed from recordings of four children reading separate stories. These recordings were normalized and combined, and the resulting multi-talker babble noise track added to the voice track with a signal-to-noise ratio (SNR) of +10 dB. The purpose of the multi-talker babble noise track was to approximate classroom conditions, with several competing simultaneous speakers.

Virtual speaker

In order to study the effect of visual presentation of the speaker, we created a virtual speaker based on facial and postural animation captured at the same time as the voice recordings using an ASUS Xtion Pro Live 3D-sensor. The 3D-sensor captures depth maps in $640 \times 480$ pixels resolution via an active infrared sensor as well as $1280 \times 1024$ pixels (RGB) video at 30 frames per second. Facial animation, orientation of head and torso, and gaze direction was extracted in Faceshift (software specialized for facial motion capture). The captured movements were then implemented on a digital character generated with Autodesk Character Generator, and video frames ($1024 \times 768$ pixels) of a frontal view of head and upper torso of the virtual speaker rendered in Autodesk Maya 2014 (Figure 1). Finally, the video and audio tracks were combined into video files (AVI multimedia container format) using Avidemux 2.5, with Xvid video compression and uncompressed audio. The fidelity of the lip movements in the final videos was deemed sufficient by an expert lip reader (post-experiment evaluation), though a minor issue with some pronunciations of “l/i” was noted.

Perceived listening effort

To probe perceived listening effort, we used a short self-report questionnaire. The first question (Q1) was “How did listening to this text make you feel?” (“Hur kände du dig när du lyssnade på den här texten?”). The second question (Q2) was “Did you think the task was easy or difficult?” (“Tyckte du att den här uppgiften var lätt eller svår?”). The questions were formulated as simply as possible in order to be administrable to children. The children responded to the two questions using Visual-

Figure 1. A video frame with the virtual speaker, as presented in conditions AV–Q and AV–N.

Analogue Scales (VAS) [38]. Endpoints were represented by negative and positive emoticons and responses were sampled by measuring (in millimeters) where on the lines the children had made a mark. Obtained values thus ranged from 0 that was the most negative (effortful) to 100 that was the most positive (least effortful).

Executive functioning

We also included the Elithorn’s Mazes (EM) test of executive functioning (part of the Wechsler Intelligence Scale for Children [39]). The test involves completing mazes of increasing difficulty, by connecting the correct number of dots without lifting the pen from the paper. The possible scores (including time bonus) range from 0 to 56 (best). The reason to schedule this test at the end of the experimental procedure was to avoid fatigue during the main listening comprehension part.

Methods

Participants

We conducted an experiment with children recruited from elementary schools in the Scania region in the south of Sweden. Out of the 61 participants, six were excluded (three failed to pass the hearing screening (described below), one did not complete the EM test, and two scored abnormally low on the listening comprehension test and also failed on more than half of the “summary questions”. The median age of the 55 included participants (34 female, 21 male) was 104 months, ranging between 100 and 111 months.

Procedure

The procedure included a short hearing screening using a Grason-Stadler GSI 66 audiometer and TDH 39 headphones. Hearing levels poorer than 20 dB Hearing Level on any of the frequencies 0.5 kHz, 1 kHz, 2 kHz, 3 kHz, 4 kHz, 6 kHz, or 8 kHz resulted in exclusion. The screening was followed by the actual listening comprehension test. The two variables “mode of presentation” and “auditory setting” were fully crossed; each participant received one narrative in each of the
We analyzed the participants’ scores on the different categories (content and inference) of the listening comprehension questions and the two self-assessments of perceived listening effort as mixed-design full-factorial models with three factors. Specifically, our models were built with auditory setting (Q as reference level) and mode of presentation (A as reference level) as within-subject factors and EM score (centered on global mean) as a between-subject factor (see Equation 1). All analyzes were performed in R [40]. Regression analyses were performed using the lmerTest package [41], and the coefficient of determination of mixed models (conditional $R^2$) [42] were calculated using the MuMin package [43]. Since some children had to be excluded, the combinations of narratives and conditions were not entirely balanced (with regard to our use of a Latin square design), which prevented us from ruling out possible effects of specific narrative or order. However, in the official instructions, the listening comprehension scores of the different narratives are summed uniformly towards the total CELF score. We therefore proceeded with our planned mixed-design analyses.

\[
[\text{Outcome variable}] \sim \text{Auditory Setting} \times \text{Mode Of Presentation} \times \text{EM} + (1|\text{Participant}) + \epsilon \quad (1)
\]

**Results**

**Listening comprehension – the content questions**

The proportions of correct answers to the content questions were analyzed as a mixed linear regression model (see Equation 1, conditional $R^2 = 0.226$). The analysis revealed a strongly significant negative effect of noise ($\beta = -0.265, t = -4.96, p < .001$), but no significant effect of audio-visual presentation ($\beta = -0.036, t = -0.674, p = .50$). There was also a marginally significant interaction between mode of presentation and auditory setting ($\beta = 0.133, t = 1.76, p = .080$), indicating that presentation of the virtual speaker somewhat reduced the negative effect of noise (Figure 2). Furthermore, our analysis found a positive effect of EM score in quiet settings ($\beta = 0.013, t = 2.20, p = .029$), together with an opposite (negative) interaction effect of EM in noise ($\beta = -0.017, t = -2.11, p = .037$). This indicates that the observed benefit of strong executive functioning with regard to listening comprehension (content) in quiet settings was not present in noise (Figure 3). No significant effect was found for the three-way interaction between EM, auditory setting, and mode of presentation ($\beta = 0.004, t = 0.351, p = .73$).

**Listening comprehension – the inference question**

Since the four experimental conditions each yielded only one correct/incorrect data point, proportions of correct answers to the inference question were analyzed as a logistic regression model (see Equation (1), conditional $R^2 = 0.059$). We did not find any significant effects neither of auditory setting ($\beta = 0.299, Z = 0.627, p = .53$), mode of presentation ($\beta = 0.121, Z = 0.258, p = .80$) nor their interaction ($\beta = 0.035, Z = -0.520, p = .60$). We found however, a positive effect of EM score ($\beta = 0.112, Z = 2.08, p = .038$) and an opposite (negative) interaction effect of EM with mode of presentation ($\beta = 0.169, Z = -2.36, p = .018$), indicating that there is a benefit of strong executive functioning with regard to listening comprehension (inference) only when listening without visual presentation of the virtual speaker (Figure 4).
No significant effect was found for the three-way interaction between EM, auditory setting, and mode of presentation (\(b = 0.141, Z = 1.42, p = .16\)).

Perceived listening effort

The VAS ratings related to perceived listening effort were analyzed as mixed linear regression models. The ratings of Q1 (“How did listening to this text make you feel?”, conditional \(R^2 = 0.572\)) did not yield any significant effects. All conditions produced ratings around 80\% towards the positive end. For the ratings of Q2 (“Did you think the task was easy or difficult?”, conditional \(R^2 = 0.385\)) we found a strongly significant negative effect of auditory setting (\(b = -17.77, t = -3.91, p < .001\)). Ratings following narratives in the A–N (\(M: 52.0, SD: 30.7\)) and AV–N conditions (\(M: 55.0, SD: 32.9\)) were on average more negative (difficult) than after the A–Q (\(M: 69.7, SD: 25.8\)) and AV–Q conditions (\(M: 64.5, SD: 27.1\)). No significant effects of mode of presentation (\(b = -5.16, t = -1.14, p = .26\)), EM (\(b = -0.539, t = -0.927, p = .36\)) or interactions were detected.

Summary of results

The presence of multi-talker babble noise when listening to the narratives impaired the participating children’s performance on the related listening comprehension content questions, and made them perceive the task as more difficult. The impaired performance on content questions after listening in noise was somewhat mitigated by audio-visual presentation. Furthermore, children who scored high on EM also performed better on the listening comprehension content questions but only in the absence of noise, as well as on the listening comprehension inference question with audio-only presentation.

Discussion

We investigated how seeing an animated virtual speaker affected listening comprehension and perceived listening effort (research question 1) with or without background babble noise (research question 2), and how comprehension in these conditions might be related to executive functioning (research question 3) in children. Our results were partly inconclusive.

Research question 1

Listening comprehension was impaired by background noise, as was to be expected from studies finding noise to adversely influence speech recognition [6] and comprehension [18–20]. Note that the SNR used (+10 dB) was generous compared to many previous studies [18,19,21], but we still observed a clear effect on comprehension, compared to no noise.

The degraded quality of the hoarse voice used in all conditions probably added to the difficulties of listening in noise, if we extrapolate from the results from some of our earlier studies [1].
**Research question 2**

The effect of noise tended to be reduced by the visual presentation of the virtual speaker, but our results only showed a marginally significant effect. Visual support only compensated for about half of the reduction in performance on the content questions, when listening in noise. Previous findings indicating that seeing a speaker’s face (virtual or real) facilitates speech recognition [14–17] support the prediction of a general positive effect of audio-visual presentation on comprehension, which obviously in part depends on recognizing the verbal content of the speech. Visual presentation has also previously been found to facilitate comprehension of semantically challenging sentences [44]. The absence in our results of any effect of mode of presentation in quiet settings could indicate a ceiling effect; without distracting noise, recognition of spoken content was no bottleneck for successful comprehension. The inconclusive effect of audio-visual presentation in background noise could have different explanations. Apart from recognition, comprehension also depends on several parallel processes and, although visual sensory information is integrated already at an early stage of speech perception, studies of audio-visual integration indicate that it is sensitive to perceptual and cognitive context. The McGurk effect has for example been found to be more dominant in a noisy environment [45], but less prevailing as cognitive load increases [46] or in situations where visual attention is divided [47]. Previous studies have indicated a complex role of visual information in speech processing [30], and that audio-visual integration can result in an increased load on resources manifested by hindered performance on a parallel memory task [29].

Our participants’ age (8- to 9-year-olds) might also have affected the result. Previous research shows age dependent effects of audio-visual speech. One study found 5- to 9-year-olds to be less sensitive to audio-visual verbal distractors compared to both 4-year-olds and 10- to 14-year-olds while performing a picture naming task [48]. In another study, 10- to 11-year-olds reported it to be more difficult to hear their teacher’s speech when they could not see her/his face [49]; this in contrast to 6- to 7-year-olds who did not report any such difficulties. These findings indicate that we possibly would have found a stronger effect of audio-visual presentation if our participants had been a few years older.

Qualitative aspects of the presentation of the virtual speaker possibly prevented benefits that could have been observed by visual presentation of a real speaker As mentioned, our lip-reading expert did note minor issues with some pronunciations of /f/. The appearance of the virtual speaker might also have distracted the participating children. The term “uncanny valley” is used to discuss how virtual characters that approach (but not quite replicate) human visual appearance may be perceived as unpleasant [50]. The virtual speaker in the current study was therefore designed not to appear far too photorealistic to decrease the risk of “falling into” the uncanny valley. It is also worth noting that the recorded speaker was reading the texts from a piece of paper placed next to the camera rather than addressing a listener. This could also add to an unnatural appearance, as much of non-verbal behavior typically present in face-to-face communication (such as hand gestures) was not produced. The responses to the first question addressing perceived listening effort (Q1) did, however, not indicate any adverse reactions to seeing the virtual speaker. We are currently working on a follow-up study comparing the virtual speaker with the real video of the actual speaker recorded in parallel to the motion capture recordings. Preliminary results indicate that the benefit of seeing the virtual speaker when listening in noise is at least at the same level as seeing the actual speaker [51].

On the other hand, if the audio-visual presentation of the speaker induces an increased “perceptual load” analogous to discrimination based on conjunction of features, perceptual load theory would predict that the visual information should facilitate semantic processing [52]. With less available resources to process the noise, it would thus be “filtered” out at an earlier perceptual stage. It is however unclear if perceptual load has such an effect across modalities [53].

**Research question 3**

The results of the content questions indicate that a strong capacity for executive functioning (as measured by EM) did facilitate comprehension in quiet auditory settings. The material was presented with a (provoked) hoarse voice in all conditions, and our finding that processing of hoarse speech can be facilitated by strong executive functioning is in line with previous results [54]. In the presence of background noise, however, no such benefit was observed. We have no explanation for this observation, and refrain from speculation.

**Limitations**

The inference question results showed no significant effects of noise or visual presentation of the virtual speaker, but a benefit from strong executive functioning in the absence of the virtual speaker. The absence of any effects of noise for the interference question may be due to the fact that the inference questions were too generally stated (in the Swedish translation of CELF4), making it possible for the children to infer the answers without extracting the information directly from the narratives. It is also worth noting that the analysis was based on only one question per child and condition. A previous study had indeed observed impaired performance (compared to standardized confidence intervals) on questions requiring their participants to “inferring information” [21].

The perceptual and cognitive load associated with other aspects of listening comprehension such as identifying the most pertinent information [21] or following verbal instructions [20] may differ from the questions related to narrative content tested in the current study. A more differentiated operationalization, accounting for different aspects and involved sub processes is needed to specify how visual information can support listening comprehension under different conditions.

Another limitation of the study was the low number of participants in combination with rather few data points per
participant and condition and especially so for the inference questions. Furthermore, the combination of narratives, order and conditions was as mentioned not perfectly balanced. All conditions were presented with a hoarse voice (to establish a higher ecological validity), and we cannot assume that our results also hold for listening to typical (clear) voices. Extending the experimental design with control conditions with a clear voice could have made our results more revealing. However, the number of available narratives for the age group would then have forced us to abandon the within-subject design, risking loss of statistical power.

Future work

Our future plans include, besides the aforementioned follow-up study comparing effects of real and virtual speakers, to further investigate how audiovisual information can support listening in situations with surrounding noise. Factors defining the appearance and movement of virtual speakers can be systematically varied, while still be presented in rich, multimodal contexts. This can allow us to investigate the contribution of different aspects such as lip-movements, "visual prosody", or distracting visual features. These are all factors that would be difficult to control systematically in natural face-to-face settings or using video stimuli. We also want to develop an applicable objective measure of cognitive load by introducing a secondary task compatible with the virtual speaker paradigm. Performance on parallel tasks performed while listening is indicative of remaining available cognitive resources. One example of such a parallel task is to connect items by drawing lines on a paper [34], something that requires vision. The challenge would be to find a task that is possible to perform while also looking at a speaker.

Conclusions

Our results indicate that children listening in noise to some extent benefit from seeing a speaker's face, but the results are inconclusive. We consider virtual speakers as promising research instruments to help us disambiguate between possible explanations of our results and to contribute to the understanding of how audio-visual interaction works in adverse and ecologically valid listening environments.

Acknowledgements

The authors gratefully acknowledge the master students, Sonny Aldenklink and Stephanie Meier, who collected the data; the Linneaus' environment Cognition, Communication and Learning at Lund University for financial support; Professor Agneta Gulz at Lund University Cognitive Science division for valuable comments on the proof, Marianne Gullberg and Joost van de Weijer at Lund University Humanities Lab for advice on stimulus production and statistical analyses.

Disclosure statement

The authors report no declarations of interest.

Funding

This work was supported by Vetenskapsrådet (Swedish Research Council) [grant no. 349-2007-8695].

Notes on contributors

Jens Nirme, PhD student with the Educational Technology Group (ETG), div. of Cognitive Science (LUCS) at Lund University since 2014, funded by the multidisciplinary research environment: Thinking in Time: Cognition, Communication and Learning (CCL). Main research interest in multimodal verbal communication. Also working with motion capture at the Lund University Humanities Lab.

Magnus Haake, Associate Professor in Cognitive Science. Part of the ETG (The Educational Technology Group) at the div. of Cognitive Science, Lund University, and the Noise & Voice-group at the div. of Logopedics, Phoniatrics, and Audiology, Lund University. Research areas: Cognitive Science, Learning Sciences, Educational Technology, and Social and Behavioral Psychology.

Viveka Lyberg Åhlander, Speech Pathologist since 1999, PhD (Lund University) in 2011. Researching teachers’ voice health in classroom sound environments, with current focus on children's perception, comprehension and cognitive capacities. Board member of the Sound Environment Centre at Lund University and the advanced study group AVACO at the Pufendorf Institute.

Jonas Brännström, DMSc., Associate professor, Dept. of Logopedics, Phoniatrics, and Audiology, Clinical Sciences Lund. Major research interests: Acceptance of background noise, Psychoacoustics and dichotic listening, Audiological rehabilitation.

Birgitta Sahlén, Professor and research group leader at Dept. of Speech Language Pathology, Lund University. Heading several language intervention projects aimed at improving language learning environments for children with weak or vulnerable language (language disorder, with hearing loss) and multilingual children.

References

1. Lyberg-Åhlander V, Brännström KJ, Sahlén BS. On the interaction of speakers' voice quality, ambient noise and task complexity with children's listening comprehension and cognition. Front Psychol. 2015;6:871.
2. Bradley JS, Sato H. The intelligibility of speech in elementary school classrooms. J Acoust Soc Am. 2008;123:2078–2086.
3. Shield BM, Dockrell JE. The effects of environmental and classroom noise on the academic attainments of primary school children. J Acoust Soc Am. 2008;123:133–144.
4. Kalikow DN, Stevens KN, Elliott LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J Acoust Soc Am. 1977;61:1337–1351.
5. Hagerman B. Sentences for testing speech intelligibility in noise. Scand Audiol. 1982;11:79–87.
6. Neuman AC, Wrobleswski M, Hajicek J, et al. Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults. Ear Hear. 2010;31:336–344.
7. Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. J Acoust Soc Am. 1954;26:212–215.
8. McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976;264:746–748.
9. Munhall KG, Jones JA, Callan DE, et al. Visual prosody and speech intelligibility head movement improves auditory speech perception. Psychol Sci. 2004;15:133–137.
10. Cassell J, Pelachaud C, Badler NI, et al. Embodied conversational agents. Cambridge (MA): MIT press; 2000.
11. Johnson WL, Rickel JW, Lester JC. Animated pedagogical agents: face-to-face interaction in interactive learning environments. Int J Artificial Intelligence Educ. 2000;11:47–78.
12. Blascovich J, Loonis J, Beall AC, et al. Immersive virtual environment technology as a methodological tool for social psychology. Psychol Inq. 2002;13:103–124.
13. Lingonblad M, Londos L, Nilsson A, et al. Virtual Blindness-A Choice Blindness Experiment with a Virtual Experimenter. In International Conference on Intelligent Virtual Agents. Cham: Springer International Publishing; 2015. p. 442–451.
14. Agdoros F, Beskow J, Dahlquist M, et al. Synthetic faces as a lip-reading support. In: ICSP; 1998.
15. Mötönen R, Olivié JI, Kuluji J, et al. Parameterized visual speech synthesis and its evaluation. In: 10th European IEEE Signal Processing Conference; 2000 Sep 4; Tampere, Finland. p. 1–4.
16. Ross LA, Saint-Amour D, Leavitt VM, et al. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex. 2007;17:1147–1153.
17. Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am. 2000;108:1197–1208.
18. Ljung R, Söörqvist P, Kjelberg A, et al. Poor listening conditions impair memory for intelligible lectures: implications for acoustic classroom standards. Building Acoust. 2009;16:257–265.
19. Valente DL, Plevinský HM, Franco JM, et al. Experimental investigation of the effects of the acoustical conditions in a simulated classroom on speech recognition and learning in children. J Acoust Soc Am. 2012;131:232–246.
20. Klatte M, Lachmann T, Meis M. Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting. Noise Health. 2010;12:270.
21. Schafer EC, Bryant D, Sanders K, et al. Listening comprehension in background noise in children with normal hearing. J Educ Audiol. 2013;19:58–64.
22. Bowers L, Huisingh R, LoGiudice C. The listening comprehension test 2. East Moline: LinguiSystems Inc; 2006.
23. Mattys SL, Davis MH, Bradlow AR, et al. Speech recognition in adverse conditions: a review. Lang Cogn Process. 2012;27:953–978.
24. Söörqvist P. The role of working memory capacity in auditory distraction: a review. Noise Health. 2010;12:217.
25. Lyberg-Åhlander V, Haake M, Brännström J, et al. Does the speaker’s voice quality influence children’s performance on a language comprehension test? Int J Speech Lang Pathol. 2015;17:63–73.
26. Koelewijn T, Zekveld AA, Festen JM, et al. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. Ear Hear. 2012;33:291–300.
27. Mackersie CL, Cones H. Subjective and psychophysiological indexes of listening effort in a competing-talker task. J Am Acad Audiol. 2011;22:113–122.
28. Jansen S, Chaparro A, Downs D, et al. Visual and cognitive predictors of visual enhancement in noisy listening conditions. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; Los Angeles (CA): SAGE Publications; 2013 September. Vol. 57. p. 1199–1203.
29. Mishra S, Lunner T, Stenfält S, et al. Visual information can hinder working memory processing of speech. J Speech Lang Hear Res. 2013;56:1120–1132.
30. Picou EM, Ricketts TA, Hornsby BW. Visual cues and listening effort: Individual variability. J Speech Lang Hear Res. 2011;54:1416–1430.
31. Lane H, Tranel B. The Lombard sign and the role of hearing in speech. J Speech Hear Res. 1971;14:677–709.
32. Åhlander VL, Rydell R, Löfqvist A. Speaker’s comfort in teaching environments: voice problems in Swedish teaching staff. J Voice. 2011;25:430–440.
33. von Lochow H, Lyberg-Åhlander V, Sahlen B, et al. The effect of voice quality and competing speakers in a passage comprehension task: perceived effort in relation to cognitive functioning and performance in children with normal hearing. Logop Phoniatr Vocol. 2018;43:32–41.
34. Imhof M, Välikoski TR, Laukkonen AM, et al. Cognition and interpersonal communication: The effect of voice quality on information processing and person perception. Studies Commun Sci. 2014;14:37–44.
35. Morton V, Watson DR. The impact of impaired vocal quality on children’s ability to process spoken language. Logop Phoniatr Vocol. 2001;26:17–25.
36. Semel EM, Wiig EH, Secord W. CELF 4: clinical evaluation of language Fundamentals. Pearson: Psychological Corporation; 2006. Available from: http://www.pearsonassessment.se/celf-4/.
37. Whitting S, Rydell R, Åhlander VL. Design of a clinical vocal loading test with long-time measurement of voice. J Voice. 2015;29:261–e13.
38. Gift AG. Visual analogue scales: measurement of subjective phenomena. Nurs Res. 1989;38:286–287.
39. Wechsler D. WISCIV. Integrated. Wechsler intelligence scale for children – Fourth Edition – integrated [Manual]. London (UK): Pearson Assessment; 2004.
40. R Core Team. A language and environment for statistical computing. 2016. Available from: https://www.R-project.org/.
41. Kuznetsova A, Brockhoff PB, Christensen RH. Package ‘lmerTest’. R package version. 2015:2.0. [cited 2017 May 15]. Available from: https://cran.r-project.org/web/packages/lmerTest/index.html.
42. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol Evol. 2013;4:133–142.
43. Barton K. MuMin: multi-model inference. R package version. 2013. [cited 2017 May 15]. Available from: https://CRAN.R-project.org/package=MuMin.
44. Lieberg D, McLean J, Goldfield A. Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: Dodd B, Campbell R, editors. Hearing by eye: the psychology of lip-reading. Hillsdale (NJ): Lawrence Erlbaum Associates Inc. p. 97–113.
45. Sekiyauma K, Tohikura YI. McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. J Acoust Soc Am. 1991;90:1797–1805.
46. Buchan JN, Munhall KG. The effect of concurrent working memory task and temporal offsets on the integration of auditory and visual speech information. Seeing Perceiving. 2012;25:87–106.
47. Andersen TS, Tippanna K, Laarni J, et al. The role of visual spatial attention in audiovisual speech perception. Speech Commun. 2009;51:184–193.
48. Jerger S, Damian MF, Spence MJ, et al. Developmental shifts in children’s sensitivity to visual speech: a new multimodal picture–word task. J Exp Child Psychol. 2009;102:40–59.
49. Dockrell JE, Shield B. Children’s perceptions of their acoustic environment at school and at home. J Acoust Soc Am. 2004;115:2964–2973.
50. Seyama S, Nagayama RS. The uncanny valley: effect of realism on the impression of artificial human faces. Presence. 2007;16:337–351.
51. Borring O. Har barn på lågstadiet svårare att f
52. Lavie N. Distracted and confused?: selective attention under load. Trends Cogn Sci. 2005;9:75–82.
53. Parks NA, Hillire MR, Corballis PM. Steady-state signatures of visual perceptual load, multimodal distractor filtering, and neural competition. J Cogn Neurosci. 2011;23:1113–1124.
54. Lyberg-Åhlander V, Holm L, Kastberg T, et al. Are children with stronger cognitive capacity more or less disturbed by classroom noise and dysphonic teachers? Int J Speech Lang Pathol. 2015;17:577–588.