Acta Logopaedica (2024), Vol. 1 | 23–31
DOI: 10.58986/al.2024.13972

Acta Logopaedica

ORIGINAL ARTICLE

Developing a teacher agent to explore the role of gaze direction on young school children's listening comprehension

K. Jonas Brännström[1], Sebastian Waechter[1], Jens Nirme[2], Susanna Whitling[1]*, Viveka Lyberg-Åhlander[1,3], Birgitta Sahlén[1]

[1]Logopedics, Phoniatrics and Audiology, Department of Clinical Sciences in Lund, Lund University, Lund, Sweden
[2]Division of Cognitive Science, Lund University, Lund, Sweden
[3]Department of Speech Language Pathology, Faculty of Arts, Psychology and Theology, Åbo Akademi, Turku, Finland

*Address for correspondence: Susanna Whitling, Logopedics, Phoniatrics, and Audiology, Department of Clinical Sciences in Lund, Lund University, SE-221 85 Lund. Email: susanna.whitling@med.lu.se

Publication date: 25 April 2024

Abstract
Teachers' gaze and school children's language proficiency play a role for listening comprehension in the classroom. The influence of multimodal conditions on listening comprehension may be hard to study in a systematic manner with real speakers. The aim of this study was to develop a method to study the influence of a speaker on passage comprehension performance using a digitally animated character. Twenty-four primary-school children were tested. A narrative passage was presented to the participants by a teacher agent. Half of the participants were randomised to a direct gaze condition and the other half to an averted gaze condition. The method could be used with the intended population, but we detected no differences in listening comprehension due to gaze behaviour in our sample. The method needs further development, such as dynamic gaze behaviour for the teacher agent before it could be used to demonstrate the influence of gaze behaviour on children's listening comprehension.

Keywords: listening comprehension; classroom; gaze behaviour; teacher agent

Introduction

Multimodal interactions influence listening comprehension (Holler *et al.*, 2014). In an instructional/educational situation, a child is listening to a visible teacher, where lip movements may give clues to word recognition (Sumby and Pollack, 1954) and head and eyebrow movements may give clues to prosody and emphasis (Swerts and Krahmer, 2008). Previous research shows that children improve their listening comprehension in adverse listening conditions when they are able to see the speaker's face (Rudner *et al.*, 2018). Other visible movements may have more indirect effects on comprehension and memory via affective (Helminen *et al.*, 2016) or attentional processes (Holler *et al.*, 2014). The current work is focused on the speaker's gaze. It has been shown that recall of narratives (Otteson and Otteson, 1980) as well as factual presentations (Sherwood, 1987) can improve by listener-directed gaze.

However, the influence of multimodal interactions on listening comprehension may be hard to study in a systematic manner in everyday conditions. In a previous study, Nirme *et al.* (2020) showed that it was possible to examine the effect of audio-visual integration on children's listening comprehension in noise by using a digitally animated virtual speaker. The findings indicated that seeing the virtual speaker facilitated listening comprehension in young school children under noisy conditions (see also Nirme *et al.*, 2019). The advantage of using a virtual speaker is that specific aspects of the speaker's behaviour can be systematically altered in any given way. For example, it is possible for the researcher to lower the presentation level of a virtual speaker's voice while all other aspects of the virtual speaker's behaviour, such as gaze, gestures, or intonation, remain unchanged. This is opposed to human speakers who may unconsciously also make additional subtle behavioural changes if instructed to speak softer. Hence, adopting virtual, instead of human, speakers in a research setting can allow for the possibility to measure the effects of small isolated changes in appearances or behaviour, for example gaze behaviour, to study their impact on comprehension. Previous work has demonstrated the feasibility of this method, showing that listener-directed gaze influences subjective impressions of the speaker (Bente *et al.*, 2007) and the quality of communication (Garau *et al.*, 2003).

The aim of this pilot study was to develop a method to test the influence of a speaker's gaze – directed (DG) at or averted (AG) from a listener – on passage comprehension performance in primary school children using a digitally animated character in the role of a teacher retelling a narrative, henceforth a *teacher agent*. To test the feasibility of this new method and to indentify potential areas of future improvement, language test results and ratings were collected from the participants . To the best of our knowledge, this has not been developed for young school children previously.

Method

Participants

Twenty-four children (15 boys and 9 girls) with a mean age of 9.4 years (SD = 0.7, minimum age = 8.2, maximum age = 10.1) were recruited from a single school in southern Sweden. The school district's proportion of L2 learners was only approximately 10%, and the proportion of parents with tertiary education was almost 90%. Hearing screening showed that all children had normal hearing for pure tones (i.e. 20 dB HL) at 0.25, 0.5, 1, 2, 3, 4 and 6 kHz, except for two children, who had a hearing threshold of 25 dB HL at one of the frequencies in one ear (both children reported that they had recently had a cold). The project was approved in advance by the Swedish Ethical Review Authority (Registration No. 2019-02665, Amendment 1).

Development of the animated teacher agent

A narrative comprehension task was used to develop the animated teacher agent. The task assessed auditory passage comprehension. A detailed description of the development of this task is reported in Carlie *et al.* (2021). In its original version,

https://doi.org/10.58986/al.2024.13972

the task consists of two equally complex passages of about 2 minutes and 45 seconds long, followed by 10 multiple-choice content questions for each passage. Each question has four response options and possible correct scores range from zero to 10. Only one of these passages was used in the present study.

To develop the teacher agent, a combination of digital animation techniques was used. These were driven by both original audio signal (one passage taken from Carlie *et al.*, 2021) and new video recording of the same speaker reading the same as in the original audio signal. This was done in order to use the same original audio recording as in a previous study (Carlie *et al.*, 2021). The COVID precautions prevented the speaker to be physically present, and she therefore recorded herself using a Samsung S20 smartphone camera, in 3840 × 2160 pixels resolution at 60 frames per second. The camera was oriented vertically, framing a frontal view of the speaker's head and torso. The speaker listened through the entire audio recordings of each narrative before recording the video to ensure consistency in terms of speech rate and prosody between audio and video recordings. Texts were displayed slightly to the right of the camera position as a memory aid. The speaker looked at the camera, or at the displayed text, throughout the recordings except during a few pauses in the speech. Two video recordings were initially made. The video recording that most closely aligned with the speech was selected. For the narrative used, the duration of the audio recording was 165 seconds and the corresponding selected video recording was of 169 seconds.

To synchronize audio and selected video recordings, the onset in milliseconds of each syllable was manually annotated in the ELAN software (version 5.9, https://archive.mpi.nl/tla/elan). These timestamps were then exported to comma-separated text files. From the video, head movement and facial expressions (e.g. eye-blinks and squints, eyebrow movements, smiles and frowns) were detected using the OpenFace software (version 2.2.0, https://github.com/TadasBaltrusaitis/OpenFace), and exported as FBX-files (for 3D animation). Mouth and lip movements detected by OpenFace were deemed not to sufficiently match the audio, and were therefore complemented with mouth and lip movements generated from the audio file using the FaceFx software (version 2018, https://facefx.com/). Lip movements corresponding to [b], [f], [v] and [«:] were adjusted manually guided by the audiofile.

The 3D model used for animations was generated using Autodesk Character Generator (https://charactergenerator.autodesk.com/) to match the voice of the speaker in terms of age and gender. Several versions were generated using the parameters available in the software (hair style and colour, skin complexion, facial features and clothes). The selected model was deemed sufficiently realistic and free of distracting features. Separate animation files (in FBX format) implementing 6 degrees of freedom (DoF) head movement and facial animation blendshapes (Lewis *et al.*, 2014) were imported, combined and applied to the 3D model in Autodesk Maya (version 2020, https://www.autodesk.se/products/maya). The animation sourced from the video recordings was transformed using Maya's built-in scripting engine, to match speech in the audio recordings, according to the annotated timestamps.

A virtual camera was positioned 2 m in front of the teacher agent (in proportion to the dimensions of its movements), slightly below the eye line and with a view angle of 18°, to frame the teacher agent's face and torso in the rendered videos, centred on the teacher agent's eyes. The gaze behaviour of the teacher agent, corresponding to

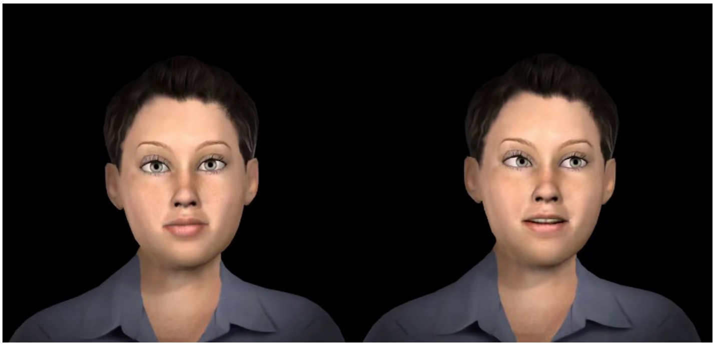

Figure 1. A frame from the rendered video showing the teacher agent in the direct gaze (left) and averted gaze (right) conditions.

the two conditions, direct gaze (DG) and averted gaze (AG), was realized by directing the eyes at the centre of the viewport of the virtual camera (adjusted to avoid cross-eyedness), or 60 cm to the right (see Figure 1). In both conditions, the speaker kept the gaze at these points in space throughout the narration, apart from a few instances (four times over the course of the presentation) of the gaze being lowered to the left side, coinciding with pauses in the narration and aligning naturally with eye blinks and head movements of the speaker. Note that due to the natural head movements of the speaker, the gaze of the speaker did not appear static but resembled a TV presenter reading from a teleprompter.

Subjective ratings of the teacher agent

After listening to the passage and after answering 10 content questions, subjective ratings of the teacher agent were made by the children. This was done by the test leader reading three predefined statements about the teacher agent ('I thought that the teacher was good at telling the story', 'I thought that it was easy to listen to the teacher', and 'I liked the teacher'), and asking the child to indicate to what degree they agreed or disagreed to the statements read. Participants indicated their attitude to the statements read by pointing to one of the four filled circles after each statement, where the largest bubble represented 'I fully agree' (scored as 3), the second largest bubble represented 'I partially agree' (scored as 2), the second smallest bubble represented 'I partially disagree' (scored as 1), and the smallest bubble represented 'I fully disagree' (scored as 0). For the purpose of the present study, the scores for these three questions were also summarised prior to statistical analysis, which means that possible scores ranged between zero and 9.

Nonword repetition

The Crosslinguistic Nonword Repetition Test (Chiat, 2015, Boerma *et al.*, 2015; Chiat, 2015) was used to test implicit phonological processing, which was associated with language skills as vocabulary and predicted the development of grammar essential for listening comprehension. The test consisted of 16 nonwords of increasing syllable length (four consisting of two syllables, four consisting of three syllables, four consisting of four syllables, and four consisting of five syllables). In the present, study a recorded version of the nonwords spoken with language-specific prosody was used (details are given in Brannstrom *et al.*, 2021). The children were instructed to listen to the made-up nonwords in earphones and to try to repeat aloud each word they had heard.

The nonwords were presented at 65 dB SPL from a laptop connected with Sennheiser HDA200 earphones. Each correctly reported nonword received a score of 1 while an incorrect nonword received a score of zero. The possible scores ranged between zero and 16.

Procedures

All participants underwent the following procedure individually and one at a time: (1) otoscopy, (2) hearing screening, (3) narrative passage read by a teacher agent with following multiple-choice questions, (4) ratings of teacher agent impression and (5) nonword repetition task. The entire procedure was carried out by a licensed audiologist in a quiet room at the children's school. All participants underwent otoscopy prior to hearing screening in order to determine that their outer ear canals were not occluded by cerumen, and that there was no indication of ongoing infection. Thereafter, the audiologist controlled that the participant had normal pure tone hearing thresholds in both ears.

Participants were then seated in front of a laptop and instructed that a teacher appearing on the screen would tell them a story, and that they would be asked questions about the story after listening to it. The passage (a narrative from Carlie *et al.*, 2021) was read to the participants by the teacher agent. The passage was presented diotically at 65 dB SPL using a laptop connected to Sennheiser HDA200 earphones. Half of the participants were randomised to the direct gaze condition and the other half to the averted gaze condition. After listening to the passage, children answered multiple-choice questions followed by their subjective ratings of the teacher agent. Finally, the nonword repetition task was performed.

Results

The DG group consisted of six girls and six boys aged 8.2–10.1 years (M = 9.4; SD = 0.7). The AG group consisted of three girls and nine boys aged 8.2–10.1 years (M = 9.4; SD = 0.7). The nonword repetition scores ranged between 14 and 16 (M = 15.25; SD = 0.62) for the DG group and between 13 and 16 (M = 14.92; SD = 0.67) for the AG group. Non-parametric statistics were used in the following analyses due to the small sample size.

The difference in performance on the passage comprehension task for the DG (M = 8.1; SD = 1.3) and AG (M = 8.3; SD = 1.9) groups was not significant as assessed using the Mann–Whithey U test (U = 58.5, $P = 0.425$, Cohen's D = 0.152).

The difference in the summarized subjective ratings of the perception of the teacher agent for the DG (M = 7.0; SD = 1.1) and AG (M = 6.3; SD = 1.2) groups was not significant as assessed using the Mann–Whithey U test (U = 50, $P = 0.190$, Cohen's D = 0.584), although a medium effect size was observed, indicating that teacher DG was favoured. The frequency distributions for each of the statements are shown in Table 1. Similar distributions are seen for the statement 'Easy to listen

Table 1. Frequency distributions for the subjective ratings of the teacher agent for DG (n = 12) and AG (n = 12) groups. Actual frequencies are shown for the three statements (see text).

	DG response distribution				AG response distribution			
Statement	Fully disagree	Partially disagree	Partially agree	Fully agree	Fully disagree	Partially disagree	Partially agree	Fully agree
"Good at telling the story"	0	0	5	7	0	0	8	4
"Easy to listen to"	0	0	6	6	0	0	5	7
"Liked the teacher"	1	1	8	2	1	6	4	1

DG = direct gaze; AG = averted gaze AG

to', while less favourable ratings are seen in the AG group for the other two statements. However, it was unclear whether there was a difference between these distributions, since no statistical analysis was conducted due to low cell counts (several below n = 5). However, visual inspection of the ratings for the third statement 'liked the teacher' indicated a more positive attitude towards the teacher for the DG group if we consider participants who rated 'fully agree' and 'partially agree' as positive towards the teacher. Ten children had a positive attitude in the DG group and five children children in the AG group.

To assess the potential influence of age, nonword repetition ability, and subjective rating of the teacher agent on passage comprehension performance, Spearman's rank correlation coefficients were calculated between the variables such as age, nonword repetition, subjective ratings of the teacher agent, and passage comprehension performance. No significant correlations were found (rho[22] < + 0.281, $P > 0.193$).

Discussion

The present study reported the development of a method to test the influence of a speaker's gaze – directed at or averted from a listener – on passage comprehension performance in primary school children. It also reported the use of this method in a sample of young primary school children when identifying potential areas of future improvement. As indicated by the findings, the present method showed no effect of differences in gaze using a virtual speaker. One possible reason for the lack of effect between the two conditions is the low number of participating children. Another reason is that it is possible that our choice to use an optimal listening condition (no adverse listening condition, such as background noise present) has resulted in a listening task that is too easy resulting in performance ceiling effects. This may be particularly true for school children in a school district with a high socioeconomic status (SES) as measured by parents' education. The benefits of a gaze directed towards the listener may only appear when the speech signal is degraded (Mattys *et al.*, 2012; Nirme *et al.*, 2019) or under more cognitively demanding multimodal task conditions. It is therefore possible that the effect of gaze behaviour may appear if we had used a more complex or demanding task, including degraded listening conditions or with less language proficient listeners (cf. listening comprehension performance in Carlie *et al.*, 2021).

Yet another reason for the absence of any observed effect is the lack of interactivity and social context in which the narrative was presented. Previous studies (with adults) have demonstrated stronger effects of speakers looking at the listener in real life (Sherwood, 1987) or video conferencing (Lanthier *et al.*, 2022), compared to video conditions. Further development of the teacher agent could enable it to simulate some interactive behaviour, such as reacting to the gaze of the listener. Using agents to study listening comprehension also has the advantage that they can precisely control factors such as gaze behaviour that would be difficult for a real speaker to naturally reproduce or vary across individually tested participants, since most gaze behaviour is beyond conscious control. It would be possible that increasing interactivity and social context to the teacher agent along with manipulations of body posture would influence performance. However, adding many types of manipulations at the same time would make it hard to discern the influence of specific behaviours on performance.

However, the perhaps most compelling reason that we could not detect the influence of gaze behaviour on passage comprehension is the use of static gaze behaviour. In real life conversation where gaze behaviour indicates turn taking between speakers, the dynamic shifting of gaze direction from the recipient provides cues to the recipient that the speaker wants to hold the turn (Jokinen *et al.*, 2010). When the speaker signals ending of turn, the gaze shifts to the recipient (Novick *et al.*, 1996). In the present study, static gaze behaviour was used. That may differ from what the participants experience in everyday communication affecting the ecological validity of the present study design.

Subjective ratings and non-systematic behavioural observations

We found no differences in summarized subjective ratings between the groups. However, there is a descriptive tendency to rate the teacher agent less favourably for two of the three rated statements about the teacher ('good at telling a story' and 'easy to listen to'). It could indicate that some individuals experience averted gaze as more detrimental than others do. As for the third statement ('liked the teacher'), twice as many individuals in the DG group as in the AG group fully or partially agreed to the statement. Attitudes towards the teacher may be important for school children's motivation to listen and merit further exploration in future studies.

While there were no differences between the AG and DG groups in terms of how much they remembered from the passage, certain behavioural differences across the groups were noted by the test leader. All children in the DG group kept looking at the teacher agent for the entire reading of the passage (approximately 2 minutes and 45 seconds), while several children (5 out of 12) in the AG group quickly changed the direction of their gaze towards the computer keyboard, the chair they were sitting on or something outside the window. As all the information needed to succeed on the passage comprehension task was auditory, this behaviour did not manifest as lower scores in the present settings. Future studies could use eye tracking to systematically study the listener's gaze behaviour in relation to an agent's gaze behaviour. Another approach is to film the listener during the task to explore more holistic behaviours as well.

Future relevance

It is plausible that children with more limited language proficiency could be supported by a multimodal DG condition. It has been shown that if the speaker uses a

https://doi.org/10.58986/al.2024.13972

combination of directed gaze, and even better a combination of gaze and gestures, listeners' listening comprehension increases (Holler *et al*., 2014). It is well-known that teachers' body communication plays an important role in student's listening comprehension and learning (Dockrell and Marshall, 2015). This fact should also be better acknowledged in formal test administration. During the school years, a range of language assessments are administered to students by professionals, that is teachers and speech pathologists. Most of the tests are still administered auditorily, the majority by live voice. It is noreworthy that most test manuals simply recommend examiners to use 'a neutral voice'. Differences between examiners as for their gaze direction, gesture, speech rate etc. can strongly jeopardize reliability of test results.

Conclusions

The purpose of the present study was to test the concept of using virtual speaker (teacher agent) among primary school children. The method needs further development before it could be used to demonstrate the influence of gaze behaviour on children's listening comprehension.

Acknowledgement. This research received financial support from The Swedish Research Council under Grant Number 2018-04160.

References

Bente, G., Eschenburg, F. and Kramer, N. 2007. Virtual gaze. A pilot study on the effects of computer simulated gaze in avatar-based conversations. In: International Conference on Virtual Reality 2007. Berlin, Germany: Springer, pp. 185–194.

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F. and Blom, E. 2015. A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *J Speech Lang Hear Res,* 58, 1747–1460.

Brannstrom, K.J., Rudner, M., Carlie, J., Sahlen, B., Gulz, A., Andersson, K. and Johansson, R. 2021. Listening effort and fatigue in native and nonnative primary school children. *J Exp Child Psychol,* 210, 105203.

Carlie, J., Sahlen, B., Nirme, J., Andersson, K., Rudner, M., Johansson, R., Gulz, A. and Brannstrom, K. J. 2021. Development of an auditory passage comprehension task for Swedish primary school children of cultural and linguistic diversity. *J Speech Lang Hear Res,* 64, 3883–3893.

Chiat, S. 2015. Non-word repetition. In: Armon Lotem, S. and De Jong, J. (eds.) *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment.* Bristol, UK: Multilingual Matters, pp. 125–150.

Dockrell, J.E. and Marshall, C.R. 2015. Measurement issues: assessing language skills in young children. *Child Adolesc Ment Health,* 20, 116–125.

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

Helminen, T.M., Pasanen, T.P. and Hietanen, J.K. 2016. Learning under your gaze: the mediating role of affective arousal between perceived direct gaze and memory performance. *Psychol Res,* 80, 159–171.

Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M. and Ozyurek, A. 2014. Social eye gaze modulates processing of speech and co-speech gesture. *Cognition,* 133, 692–697.

Jokinen, K., Nishida, M. and Yamamoto, S. 2010. On eye-gaze and turn-taking. In: Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction. pp. 118–123.

https://doi.org/10.58986/al.2024.13972

Lanthier, S.N., Zhu, M.J., Byun, C.S., Jarick, M. and Kingstone, A. 2022. The costs and benefits to memory when observing and experiencing live eye contact. *Visual Cognition,* 30, 70–84.

Lewis, J.P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F.H. and Deng, Z. 2014. Practice and theory of blend-shape facial models. *Eurographics (State of Art Rep),* 1, 2.

Mattys, S.L., Davis, M.H., Bradlow, A.R. and Scott, S.K. 2012. Speech recognition in adverse conditions: a review. *Lang Cognitive Proc,* 27, 953–978.

Nirme, J., Haake, M., Lyberg Ahlander, V., Brannstrom, J. and Sahlen, B. 2019. A virtual speaker in noisy classroom conditions: supporting or disrupting children's listening comprehension? *Logopedics Phoniatrics Vocology,* 44, 79–86.

Nirme, J., Sahlen, B., Ahlander, V.L., Brannstrom, J. and Haake, M. 2020. Audio-visual speech comprehension in noise with real and virtual speakers. *Speech Commun,* 116, 44–55.

Novick, D.G., Hansen, B. and Ward, K. 1996. Coordinating turn-taking with gaze. In: Proceedings of the Fourth International Conference on Spoken Language Processing, pp. 1888–1891.

Otteson, J.P. and Otteson, C.R. 1980. Effect of teacher's gaze on children's story recall. *Perceptual Motor Skills*, 50(1), 35–42, 50.

Rudner, M., Lyberg-Ahlander, V., Brannstrom, J., Nirme, J., Pichora-Fuller, M.K. and Sahlen, B. 2018. Listening comprehension and listening effort in the primary school classroom. *Front Psychol,* 9, 1193.

Sherwood, J.V. 1987. Facilitative effects of gaze upon learning. *Perceptual Motor Skills,* 64, 1275–1278.

Sumby, W.H. and Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am,* 26, 212–215.

Swerts, M. and Krahmer, E. 2008. Facial expression and prosodic prominence: effects of modality and facial area. *J Phonet,* 36, 219–238.