Speech Communication 101 (2018) 70–84

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters

Maartje M.E. Hendrikse[a,*], Gerard Llorach[a,b], Giso Grimm[a], Volker Hohmann[a,b]

[a] Medizinische Physik and Cluster of Excellence 'Hearing4all', Universität Oldenburg, Germany
[b] Hörzentrum Oldenburg GmbH, Oldenburg, Germany

ARTICLE INFO

Keywords:
Head- and eye movement
Hearing aids
Evaluation
Audiovisual environments
Animations

ABSTRACT

Recent studies of hearing aid benefits indicate that head movement behavior influences performance. To systematically assess these effects, movement behavior must be measured in realistic communication conditions. For this, the use of virtual audiovisual environments with animated characters as visual stimuli has been proposed. It is unclear, however, how these animations influence the head- and eye-movement behavior of subjects. Here, two listening tasks were carried out with a group of 14 young normal hearing subjects to investigate the influence of visual cues on head- and eye-movement behavior; on combined localization and speech intelligibility task performance; as well as on perceived speech intelligibility, perceived listening effort and the general impression of the audiovisual environments. Animated characters with different lip-syncing and gaze patterns were compared to an audio-only condition and to a video of real persons. Results show that movement behavior, task performance, and perception were all influenced by visual cues. The movement behavior of young normal hearing listeners in animation conditions with lip-syncing was similar to that in the video condition. These results in young normal hearing listeners are a first step towards using the animated characters to assess the influence of head movement behavior on hearing aid performance.

1. Introduction

Established procedures for hearing-aid evaluation in the laboratory use stationary acoustic-only conditions with fixed source and receiver positions (e.g., Luts et al., 2010). These procedures do not reflect natural dynamic communication conditions, which might explain why directional microphones perform well in the stationary laboratory conditions, but not in real life (Bentler, 2005; Cord et al., 2004). Furthermore, recent achievements in hearing-aid development result in algorithms that interact with or rely on head- and eye-movement behavior (e.g., Tessendorf et al., 2011; Abdipour et al., 2015). To properly evaluate these novel algorithms, realistic head- and eye movement is required, and it is assumed that this can only be achieved by including visual cues in the test environment. The long-term goal of our research is therefore to create a laboratory-based, audio-visual test environment in which ecologically valid movement behavior and ecologically valid task performance can be measured under comfortable conditions for both normal-hearing listeners (to provide reference data) and hearing-aid users. The ecological validity of a measure indicates that it is meaningful under real life conditions. As a first step towards this goal, this study measured the influence of visual cues on head- and eye-movement behavior and looked at the influence on task performance and perception. Young, normal-hearing listeners were included in this study to develop and evaluate the methods, and to provide reference movement data for later studies involving older subjects and subjects with hearing aids.

One approach to investigate movement behavior experimentally is to perform recordings in the field (Tessendorf et al., 2012). However, in the field it is difficult to achieve reproducible test conditions and to systematically assess the factors underlying movement behavior. An alternative approach is to use virtual environments that offer high-quality, plausible sound-field simulations in the laboratory (e.g., Grimm et al., 2016; Cubick and Dau, 2016). Using virtual environments in the laboratory makes experiments more realistic and closer to field tests, but sufficiently controlled to be reproducible. However, when compared to clinical tests, the higher complexity and variability of the virtual environments make certain experiments less sensitive to small effects on, e.g., speech-reception thresholds. Whereas the sound reproduction in laboratory virtual environments has been evaluated extensively, also for hearing aids (e.g., Grimm et al., 2016; Cubick and Dau, 2016), it is not yet clear to what extent movement behavior in such environments is ecologically valid.

* Corresponding author.
E-mail address: maartje.hendrikse@uni-oldenburg.de (M.M.E. Hendrikse).

https://doi.org/10.1016/j.specom.2018.05.008
Received 30 June 2017; Received in revised form 25 May 2018; Accepted 31 May 2018
Available online 15 June 2018

Presenting visual stimuli makes the test environment more realistic. It may also influence the listener's head- and eye-movement behavior. Furthermore, it is well known that seeing the speaker[1] can improve speech intelligibility (Sumby and Pollack, 1954). Speech intelligibility improves because of lip reading (e.g., Erber, 1969; Macleod and Summerfield, 1987) and seeing the movements of the speaker's mouth and face, which focuses listeners' attention on specific spectro-temporal locations in the speech waveform, enhancing sensitivity to the acoustic information. This results in an additional increase in speech intelligibility that is different from lip reading (Grant, 2001; Schwartz et al., 2004). Gestures made by the speaker also help to improve speech intelligibility (Munhall et al., 2004; Drijvers and Özyürek, 2017). Visual information is also important for social aspects, such as gaze direction for conversational turn-taking (Vertegaal et al., 2001), and for using facial expressions to judge emotions (Ekman and Oster, 1979).

The use of animations as visual stimuli permits more freedom when creating different real-life scenarios and avoids making time-consuming video recordings (involving related privacy issues). Animations are finding their way into audiological research (Meister et al., 2016), and a number of aspects need to be considered. It is known that both the appearance and behavior of the animated characters are important for the perceived realism and that those two need to be in balance, as higher realism of the appearance may lead to higher expectations of the behavioral realism (Slater and Steed, 2002). Furthermore, it is important to stay away from the so called "Uncanny Valley" (Mori, 1970), where a too human-like appearance can cause feelings of unease when there is also an abnormal feature (Seyama and Nagayama, 2007). As mentioned above, behavioral features such as mouth movements, gestures, gaze direction and facial expression can increase speech intelligibility for human speakers and should therefore also be included in animated characters to achieve speech perception similar to that for real speakers. Studies of these behavioral features in virtual characters usually focus on their effect on speech intelligibility (e.g., Fagel and Clemens, 2004; Meister et al., 2016), or on the perceived realism by the user (e.g., Lee et al., 2002; Le et al., 2012). However, it remains unclear how these behavioral features affect the user's movement.

This study focuses on the effect of different visual cues from animated characters on head- and eye movement, on combined speech intelligibility and localization task performance, and on perceived speech intelligibility, perceived listening effort and the general impression of the audiovisual environments. It seeks a level of realism of the animated characters that results in ecologically valid movement behavior. In the laboratory, using video recordings of real people, either 2D recordings, stereoscopic recordings or 3D recordings with light-field cameras (Akeley, 2012), is the most realistic condition that can reproducibly be achieved. In this study, 2D video recordings of real people projected onto a cylindrical screen were used to compare to the animations. Animation conditions that induce movement behavior similar to the video condition were considered more ecologically valid. Furthermore, a condition in which the speakers cannot be seen (audio-only) was included as an additional reference. Two important features of the animated characters, lip-syncing and gaze direction, were analyzed in detail to determine which information is necessary to measure ecologically valid movement behavior and the task performance of the listener. Not only were concordant lip-syncing (speech-driven) and gaze direction (towards active speaker) tested, but also discordant lip-syncing ('fish mouth') and gaze direction (randomized), to determine how these features influence movement behavior, task performance, and perception.

To formulate a hypothesis on the effect of visual cues on movement behavior in challenging listening environments, information is needed about strategies that could drive movement behavior. In multi-talker conversations, most of the time (62%) people look at the individual they listen to (Vertegaal et al., 2001). This is probably due to social factors and because of the information that can be obtained by looking at the speaker's mouth movements, gestures, gaze direction and facial expression. Thus looking at the active speaker could be one strategy to increase speech intelligibility. Another strategy would be to move the head to maximize the signal-to-noise ratio (SNR). Studies have shown that although hearing-impaired listeners successfully increase SNR or speech level by head movements, young normal-hearing listeners have difficulty spontaneously finding a beneficial head orientation, and half of them do not move at all (Brimijoin et al., 2012; Grange and Culling, 2016). The results of the aforementioned study were obtained without visual information, and it is not clear which effect visual information has on orienting behavior. A third possible strategy could be one driven by localization. It is known that target localization is important for understanding the target speech (Yost et al., 1996) and that head movements are important for localization (Wallach, 1939). Kim et al. (2013) studied how people move during a localization task and found that when localizing, people turn their head towards the sources. Of course this strategy is mainly important if the active speaker cannot be seen, otherwise localization is trivial. The first two strategies are often conflicting, because the optimal head direction can be too far away from the target direction to see the target speaker's face. Both strategies usually require some movement, which is important for localization. Based on these movement strategies, it is expected that people will look at the active speaker if the speaker is visible. However, if there is discordant or no lip-syncing, this strategy will provide no speech intelligibility benefit and people are more likely to follow the SNR-optimization strategy. If the speakers are not visible, people might turn their head toward the active speaker when following the localization strategy, or turn their head to optimize SNR.

Therefore, the first hypothesis of the current study is that the animation conditions with speech-driven lip-syncing induce movement behavior similar to that in the video condition. As a second hypothesis, the task performance is expected to increase when adding lip-syncing and gaze direction towards the active speaker. Perception of the different visual conditions will be influenced by how realistic the visual conditions are. Therefore, as a final hypothesis, the most realistic animation condition, with speech-driven lip-syncing and gaze direction towards the active speaker, is expected to receive the best subjective ratings of the animation conditions, and it is expected that subjects feel comfortable in this condition. Note that, since there is no comparison to the video condition for the speech intelligibility and localization task (see Section 2), the ecological validity of the task performance will not be assessed. It can however be determined if, as stated in the second hypothesis, the task performance in the animation conditions increases compared to the task performance in the audio-only condition.

2. Method

Movement behavior, task performance and perception were tested and evaluated using three tasks. First, a listening task was carried out (Task 1). The aim of this task was to measure natural movement behavior. Subjects were asked to listen to a conversation between four persons in cafeteria background noise and subsequently answer multiple-choice questions about the content. Second, a German adaptation of the coordinate response measure (CRM) task (Moore, 1981; Bolia et al., 2000) (Task 2) was used to measure combined speech intelligibility and localization performance, as well as movement behavior. Finally, to measure how subjects perceived the different visual conditions, they were asked to compare and rate the visual conditions from the first listening task using a questionnaire (Task 3). The conditions used in the tests included an audio-only condition in which the speakers could not be seen, a video condition with recordings of real people, and conditions using animated characters. The animated characters could have different lip-syncing: concordant lip-syncing (speech-driven),

[1] In this paper the term 'speaker' is used to refer to persons or animated characters in the tasks, otherwise the term 'loudspeaker' is used.