PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Audio-visual stimuli for the evaluation of speech-enhancing algorithms

Giso GRIMM[(1)]; Gerard LLORACH[(1)]; Maartje M. E. HENDRIKSE[(1)]; Volker HOHMANN[(1)]

[(1)]Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, D-26111 Oldenburg, Germany,,
g.grimm@uol.de

Abstract

The benefit from speech-enhancing algorithms in hearing devices may depend not only on the acoustic environment, but also on the audio-visual perception of speech, e.g., when lip reading, and on other visual cues. In particular, the functioning of speech-enhancing algorithms depends on the motion behavior of the user, which in turn depends on visual cues. In this presentation we introduce various audio-visual stimuli used for evaluation of speech-enhancing algorithms in hearing devices. The stimuli include video recordings of the Oldenburg sentence tests (OLSA), real-time animation of lip movement for virtual animated characters, and complex audio-visual environments. We will discuss the effects of the material on speech perception and motion behavior, and outline applications of these stimuli. Results show that visual cues provide a benefit in terms of speech reception thresholds in the audio-visual OLSA. This benefit can not be found with the animated lip movement using a simple vocal tract model. However, it was found that the animations are sufficient to achieve natural motion behavior. This presentation is related to contributions at this conference of Llorach et al. on details of audio-visual speech test material [19], and Hendrikse et al. on natural motion behavior and its influence on hearing aid algorithm performance in complex listening conditions [12].

Keywords: Audio-visual stimuli, speech enhancement, speech material, virtual reality

1 INTRODUCTION

Understanding speech does not only involve hearing, but also seeing, e.g., for lip reading [3, 20, 2] or switching of attention [1], or other non-verbal behavior, e.g., facial expressions [22]. The benefit of speech-enhancing algorithms in hearing devices is often evaluated using audio-only stimuli (e.g., [16, 28, 26]). However, the benefit of such algorithms may depend on visual factors as well: hearing impaired and normal hearing listeners benefit from lip reading. Furthermore, typical conversational behavior, like looking at the active speaker, may support the listener in identifying and attending the target source in adverse listening conditions [5, 23, 1]. In the case of hearing aids, the head movements of the user may significantly interact with algorithm benefit: the directional benefit of fixed beamformers can only be exploited if the user is facing the target. The fact that head motion behavior significantly depends on the presence of visual cues [13] shows that audio-visual stimulation is an important factor in studying movement-algorithm interaction.

Evaluation of speech-enhancing algorithms may serve various goals. One aspect is the assessment of specific interaction with acoustic parameters, e.g., noise field properties. Here, the participant serves only as a sensor, and multi-model stimulation is not required. Another aspect is the prediction of real-life benefit of hearing devices. In this context, multi-modal assessment is more important. In the first place, being able to see the face of the speaker and lip-reading is more representative of a real-life situation. Secondly, the audiovisual simulations should not be limited to lip movements and facial expressions, as the head motion affects the benefit of hearing devices. Thus, audiovisual simulations should include realistic environments, audiovisual distractors and interaction with the participant. This way, systematic evaluation of speech-enhancing algorithms with audio-visual stimuli fills the gap between acoustic-only assessment and field tests where control over the stimuli is not possible.

For the assessment of behavior and the evaluation of algorithms which depend on the motion behavior of the user, interactivity is important: hearing aid algorithms may change their processing interactively. Examples of such algorithms are a directional microphone steered by the gaze [4], and a directional filter controlled by gaze-

based auditory-attention-decoding [7]. Furthermore, to achieve natural motion behavior, the immersion into a virtual audio-visual environment needs to be sufficiently high to create a sense of being present in the environment. Interaction of the participants with the virtual audio-visual environment may contribute to an increase of the sense of presence [21]. For example, in the case of audio-visual virtual environments, incorporation of self-motion in the rendering allows to look behind objects, and thus achieves a sense of depth.

2 METHODS

2.1 Audio-visual matrix speech test

The matrix speech test is an established method of testing speech intelligibility in noise [10], available in many languages [16]. The matrix test consists typically of non-predictable sentences of five words, e.g., *"Peter sees eleven old gloves."*. For each word position, ten different words are available. From this 5×10 matrix, lists of sentences are chosen.

Creating such a speech corpus requires big effort. The material needs to be phonetically balanced, and calibrated to achieve equal intelligibility for each word. Furthermore, for the benefit of natural transitions between words in more recent versions of matrix tests each combination of words is recorded as a full sentence for a reduced set of sentences. For example, the female version of the "Oldenburger Satztest" [27, 28] consists of set of 120 different sentences.

Two options of creating an audio-visual matrix speech test are possible: to re-record the whole speech corpus, using audio and video recordings, or to record video-only material matching an existing speech corpus. In this study, we decided to record videos and align them with existing speech material [27]. This way, it is possible to use an established and calibrated speech corpus. For the recording of the material, the original female speaker was hired. The acoustic speech corpus was presented to the speaker via in-ear headphones. Each sentence was repeated several times, and the speaker was asked to speak along with the recording. Videos were recorded together with a linear time code (LTC) signal and a microphone recording of the new utterance of the speech material.

For the editing, the video recording of each sentence was matched with the original speech corpus based on the time code. The best-matching recording was selected based on highest speech envelope similarity between the original corpus and the new recording, and by manual selection.

2.2 Automated animation of lip movement

For higher flexibility it is often advisable to use virtual simulated visual environments, with animated virtual characters [18, 13, 14]. Here, we use a very simple vocal tract model, which estimates the normalized signal power in three frequency bands [17]. A linear mapping was manually adjusted to control three different blend shapes of the animated characters: kiss blend shape, lips pressed blend shape, mouth open blend shape. This method has the advantage of being real-time applicable , e.g., it permits to speak through a microphone and generate the corresponding lip-syncing instantaneously. However, the visual quality is lower than alternative approaches based on phoneme classification.

2.3 Automated animation of conversational behavior

Head movement of animated virtual characters should be related to the conversation content. In our simulation, to simulate conversational head motion behavior in a multi-talker situation, the gaze of all animated characters belonging to the same conversation is controlled by the speech signals. Whenever a speech onset is detected in any of the character's audio content, a message is sent to the other characters to look at the mouth position of the active speaker. The head- and eye direction is controlled using inverse kinematics, to follow the gaze target position. This method is real-time capable, because it does rely only level estimators of the clean speech signals.

In addition to automated conversational gaze behavior, the gaze target position can also be controlled manually,

(a) cafeteria (b) lecture hall (c) living room (d) street (e) train station

Figure 1. Overview of the virtual audio-visual environments used in [11].

e.g., to visually guide the participants attention towards target direction and to indicate stimulus presentation.

2.4 Influence of visual cues on motion behavior

In the study of Hendrikse et al. [13] the influence of visual cues such as using animations instead of video recordings on motion behavior and general audio-visual perception was assessed. As measures, subjective ratings with the items 'quality of virtual experience', 'perceived intelligibility' and 'perceived listening effort' were collected. Furthermore, the gaze behavior was assessed by measuring the distance between the gaze and target direction.

2.5 Virtual audio-visual environments

A set of complex virtual audio-visual environments, containing not only speech targets, but also dynamic moving distractors, was developed [11, 6] (see Figure 1). This set contains conversations in a crowded cafeteria, on a street, in a train station, a presentation in a lecture hall, and a living room. All environments contain animated virtual characters. For reproduction of the acoustic part, the toolbox for acoustic scene creation and rendering (TASCAR) [9, 8] can be used. An overview of methods applicable for capturing and reproduction of audio-visual stimuli can be found in [18].

3 RESULTS

3.1 Audio-visual matrix sentence test - Benefit of visual cues on SRT

The speech reception thresholds (SRT) adapted to 80% correct sentence scores were measured in a matrix speech test (OLSA, [28]) with audio-only stimuli, audio+video stimuli, and audio+animation stimuli. The study was performed with 15 elderly normal-hearing participants (mean age 70.7 ± 5.4a, 11 female, 4 male). The visual stimulus, if any, was presented via a computer screen in front of the participants. The size of the image and the distance was corresponding to typical face-to-face communication. The acoustic stimulus was presented via one speaker in front of the user. The audio and video signals were synchronized. We used a video with frame numbers and the corresponding LTC recorded by an external camera for calibrating the offset. For a subset of the participants, also SRT at 50% intelligibility was measured.

The individual benefit of the visual stimuli on SRT at 80% intelligibility (shown in Figure 2, left panel) was 4.2 dB on average (2.5 dB standard deviation) for the video recordings, and -0.1 dB (0.7 dB standard deviation) for the animations. At 50% intelligibility, the benefit was 6.2 ± 1.5 dB for the video recordings, and -0.3 ± 0.6 dB for the animations (Figure 2, right panel). The audio-visual video condition is significantly easier to understand than the audio-only condition (t-test, 5% significance level). The individual benefit of the recorded video is up to 10 dB, which is comparable to the findings of [24] and [25]. No significant difference from zero can be found for the animation condition, which means that the animations do not provide any benefit in speech intelligibility, but also don't have a detrimental effect.

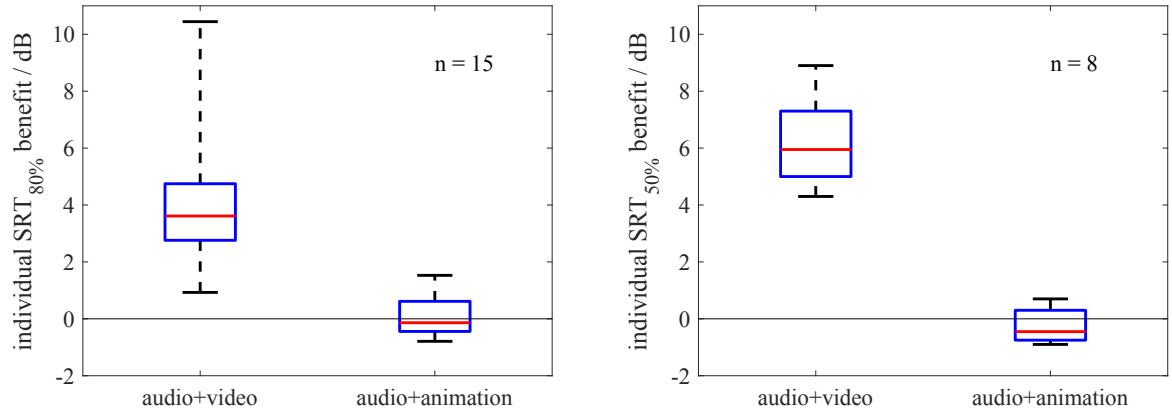

Figure 2. Benefit of the visual stimuli on speech intelligibility, for SRT at 80% intelligibility (left panel) and SRT at 50% intelligibility (right panel). Positive values correspond to lower SRTs and thus a larger benefit. Presenting videos along with the acoustic stimuli provides a significant benefit, the lip animations do not affect SRTs.

3.2 Visual target cues and motion behavior

In [13] the head movement and gaze behavior was measured in conversations between four people in different visual conditions. The conversations were recorded as videos. For the playback either the videos were shown (condition 'video'), the audio signals were presented alone (condition 'audio only'), or a simulation with animated characters in three different animation modes along with the original audio signals was presented. The animations combined lip-syncing and gaze animations. The lips and mouth were animated in the conditions 'no lip-syncing' (NLS), speech-driven lip-syncing (LS, [17]), or 'fish-mouth lip-syncing', i.e., periodic opening and closing of the mouth (FLS). The gaze of the virtual animated characters was controlled to look either to the participant (GS), or to the currently active target speaker, simulating a conversational gaze behavior (GT). In all animation conditions eye-blinks were simulated randomly.

The outcome of subjective ratings resulted in a similar ranking of the conditions for all tested items: For the items 'quality of virtual experience' and 'perceived intelligibility' the ranking was: 'video', 'LS-GT', 'LS-GS', 'NLS-GS', 'audio only', and 'FLS-GS'. The 'video' condition had the highest quality of virtual experience and the highest perceived intelligibility. For the item 'perceived listening effort', the ranking was: 'video', 'LS-GT', 'LS-GS', 'audio only', 'FLS-GS', and 'NLS-GS' (rank correlation with 'quality of virtual experience' of 0.83). The 'video' condition showed the smallest listening effort. The ranking of the objective measure 'gaze direction error' (e.g., RMS of angular distance from target speaker) is: 'LS-GT', 'video', 'LS-GS', 'FLS-GS', 'NLS-GS', and 'audio only'. The condition 'LS-GT' resulted in the smallest gaze direction error (rank correlation with 'quality of virtual experience' is 0.77).

The systematic differences between visual conditions indicate an influence of visual cues on the motion behavior. Even visual differences like type of lip animation, which was shown to have no effect on speech intelligibility, affects behavior, and may in combination with speech-enhancing algorithms thus also affect performance.

Another finding of Hendrikse et al. [14] and in many other studies (see [22] for an overview) is that listeners follow the active speaker during a large amount of time. However, a part of the gaze movements is achieved by turning the head, and the rest is achieved by the movement of the eyes. This means that directional microphones attached to the head will not provide the theoretically achievable maximum benefit due to misalignment. This effect can result in an SNR difference of several dB [14].

3.3 Visual distractors and motion behavior

In [2] it was shown that visual distractors may have a detrimental effect on speech perception. The effect is largest for moving distractors, which consisted of a video clip in [2]. In the motion behavior observed by [14] it can be seen that visual distractors, such as passing cars on a street, or a paper plane in a lecture hall, trigger gaze and head movement towards the distractor. This gaze behavior can be seen as a shift of attention from the primary auditory or audio-visual task to the distractor. Furthermore, in combination with speech-enhancing algorithms, this behavioral response to visual distractors may affect signal enhancement performance.

4 DISCUSSION AND FUTURE DIRECTIONS

As shown in a number of studies, it can be important to use audio-visual stimuli for the assessment of speech-enhancing algorithms. The choice of the stimuli depends on the task: as animated lip movement does not contribute to speech intelligibility, it can not be used for tasks which rely on the visual benefit. Despite not having an effect on speech intelligibility, animated lip movement is relevant for achieving natural motion behavior and experienced quality of audio-visual stimuli.

It can also be seen that animated head movement affects motion behavior and experienced quality of the virtual environments. This is extremely relevant when assessing algorithms which are controlled by the gaze, e.g., in gaze-based attention models [7]. For audio-visual stimuli it is also important to consist of controlled types and amount of visual distractors, in a similar way as auditory stimuli of assessment of speech-enhancing algorithms consist of acoustic target and noise components.

Future directions in this research field can be the improvement of stimuli. For example, the animated lip movement should be improved to provide a similar SRT benefit as recorded lip movement. To allow for interactive dynamic environments, ideally this benefit should be also reached in real-time applications.

To come closer to real-life situations, the realism of virtual audio-visual environments should be improved, as shown in [13]. On the acoustic side the generation of Lombard-speech [15] for conversation content in noisy environments would improve the stimuli. In the visual part, the movement of objects needs to be more natural (e.g., deceleration of cars in curves, smoother movement).

For simulating social interaction and human-to-human interaction, systems with interaction on a conversational level would be desirable, e.g., by the incorporation of systems with embodied conversational agents. As an intermediate step, animated virtual characters could be controlled by the voice and gestures of the experimenter, recorded and transmitted in real time from an acoustically and visually separated space.

A standardized exchange format would be desirable to allow for an exchange of stimuli and comparison of test methods between labs.

5 CONCLUSIONS

This contribution shows that interactive virtual audio-visual environments may provide a basis for the development of more realistic and ecologically valid tests of acoustic communication and speech-enhancing algorithms with subject-in-the loop paradigms. Visual features such as the simulation of lip movement and of typical conversational behavior were shown to be relevant for this task. Real-time animated lip-movement, however, was found to not be of sufficient quality yet to provide the intelligibility gains provided by video recordings.

ACKNOWLEDGEMENTS

Work funded by the German Research Foundation DFG project number 352015383 – SFB 1330 B1, and by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 675324 (ENRICH).

We thank Anja Gieseler for the collection of SRT data of the audio-visual OLSA.

REFERENCES

[1] V. Best, E. J. Ozmeral, and B. G. Shinn-Cunningham. Visually-guided attention enhances target identification in a complex auditory scene. *Journal for the Association for Research in Otolaryngology*, 8(2):294–304, 2007.

[2] J. I. Cohen and S. Gordon-Salant. The effect of visual distraction on auditory-visual speech perception by younger and older listeners. *The Journal of the Acoustical Society of America*, 141(5):EL470–EL476, 2017.

[3] N. P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of speech and hearing research*, 12(2):423–425, 1969.

[4] A. Favre-Felix, C. Graversen, R. K. Hietkamp, T. Dau, and T. Lunner. Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment. *Trends in hearing*, 22:2331216518814388, 2018.

[5] K. W. Grant. The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America*, 109(5):2272–2275, 2001.

[6] G. Grimm, M. M. E. Hendrikse, G. Llorach, and V. Hohmann. Gesture-lab youtube channel. https://www.youtube.com/channel/UCAXZPzxbOJM9CM0IBfgvoNg, 2018.

[7] G. Grimm, H. Kayser, M. M. E. Hendrikse, and V. Hohmann. A gaze-based attention model for spatially-aware hearing aids. In *Speech Communication; 13. ITG Symposium*, pages 231–235. VDE Verlag GmbH Berlin, Offenbach, Oct 2018. ISBN 978-3-8007-4767-2.

[8] G. Grimm, J. Luberadzka, T. Herzke, and V. Hohmann. Toolbox for acoustic scene creation and rendering (TASCAR): Render methods and research applications. In F. Neumann, editor, *Proceedings of the Linux Audio Conference*, Mainz, Germany, 2015. Johannes-Gutenberg Universität Mainz.

[9] G. Grimm, J. Luberadzka, and V. Hohmann. A toolbox for rendering virtual acoustic environments in the context of audiology. *Acta Acustica united with Acustica*, 105(3):566–578, 2019.

[10] B. Hagerman. Sentences for testing speech intelligibility in noise. *Scandinavian audiology*, 11(2):79–87, 1982.

[11] M. M. E. Hendrikse, G. Llorach, G. Grimm, and V. Hohmann. Realistic virtual audiovisual environments for evaluating hearing aids with measures related to movement behavior. *The Journal of the Acoustical Society of America*, 143(3):1745–1745, 2018.

[12] M. M. E. Hendrikse, G. Llorach, G. Grimm, and V. Hohmann. Realistic audiovisual listening environments in the lab: analysis of movement behavior and consequences for hearing aids. In *Proceedings of the 23rd International Congress on Acoustics*, Aachen, 2019.

[13] M. M. E. Hendrikse, G. Llorach, G. Grimm, and V. Hohmann. Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. *Speech Communication*, 101:70–84, Special Issue on Realism in Robust Speech and Language Processing, 2018.

[14] M. M. E. Hendrikse, G. Llorach, V. Hohmann, and G. Grimm. Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life. *Trends in Hearing*, 2019. in preparation.

[15] J.-C. Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524, 1993.

[16] B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener. The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, 54(sup2):3–16, 2015.

[17] G. Llorach, A. Evans, J. Blat, G. Grimm, and V. Hohmann. Web-based live speech-driven lip-sync. In *2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, Barcelona, Spain, 2016. IEEE.

[18] G. Llorach, G. Grimm, M. M. E. Hendrikse, and V. Hohmann. Towards realistic immersive audiovisual simulations for hearing research: Capture, virtual scenes and reproduction. In *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, pages 33–40. ACM, 2018.

[19] G. Llorach and V. Hohmann. Word error and confusion patterns in an audiovisual german matrix sentence test (OLSA). In *Proceedings of the 23rd International Congress on Acoustics*, Aachen, 2019.

[20] A. MacLeod and Q. Summerfield. Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, 21(2):131–141, 1987.

[21] K. L. Nowak and F. Biocca. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 12(5):481–494, 2003.

[22] K. Ruhland, S. Andrist, J. Badler, C. Peters, N. Badler, M. Gleicher, B. Mutlu, and R. Mcdonnell. Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In *Eurographics state-of-the-art report*, pages 69–91, 2014.

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

[24] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[25] N. Tye-Murray, M. S. Sommers, and B. Spehar. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and hearing*, 28(5):656–668, 2007.

[26] K. Wagener and T. Brand. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. *International Journal of Audiology*, 44(3):144–156, 2005.

[27] K. Wagener, S. Hochmuth, M. Ahrlich, M. Zokoll, and B. Kollmeier. Der weibliche oldenburger satztest. *The female version of the Oldenburg sentence test"), in Proceedings of the 17th Jahrestagung der Deutschen Gesellschaft für Audiologie, Oldenburg, Germany*, 2014.

[28] K. C. Wagener, V. Kühnel, T. Brand, and B. Kollmeier. Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil I–III (development and evaluation of a german sentence test). *Z Audiol*, 38(1-3), 1999.