Audio-visual speech stimuli for the study of lip-reading and multi-sensory integration abilities in hearing-impaired individuals

MAREN STROPAHL[1,*] AND STEFAN DEBENER[1,2]

[1] *Department of Psychology, Carl von Ossietzky University, Oldenburg, Germany*

[2] *Cluster of Excellence "Hearing4all", Oldenburg, Germany*

Audio-visual integration of speech is frequently investigated with the McGurk effect. Incongruent presentation of auditory and visual syllables may result in the perception of a third syllable, reflecting fusion of visual and auditory information. However, perception of the McGurk effect depends strongly on the stimulus material used, making comparisons across groups and studies difficult. To overcome this limitation we developed a large set of audio-visual speech material, consisting of eight different speakers (4 females and 4 males) and 12 syllable combinations. The quality of the material was evaluated with 24 young and normal-hearing subjects. The McGurk effect was studied in eight adult cochlear implant (CI) users and compared to 24 normal-hearing individuals using a probabilistic model. The comparison confirmed previous reports of stronger audio-visual integration in CI users. The audio-visual material developed in this study will be made freely available.

INTRODUCTION

In daily life situations the integration of information from multiple senses is necessary to interact with the environment (Driver and Noesselt, 2008). In real-life communication most of the speech signal is encoded by the auditory input. Nevertheless, it has been shown that visual information such as lip movements can improve speech intelligibility especially in noisy situations (Sumby and Pollack, 1954). Audio-visual integration therefore plays a major role for communication and auditory restoration. Cochlear implants (CIs) are biomedical devices that allow individuals with a profound sensorineural hearing loss to regain parts of their hearing ability. Despite the electrical input, CI users are able to show improved speech recognition shortly after implantation (Sandmann *et al.*, 2014). Nevertheless, speech understanding in noisy situations remains difficult for the majority of CI users (Fu *et al.*, 1998). The deficit of CI users in their auditory processing may also be reflected in a different use of visual speech cues compared to normal hearing (NH) controls. There is evidence that CI users are better in lip reading and in integrating audio-visual stimuli (Rouger *et al.*, 2007; Stropahl *et al.*, 2015).

*Corresponding author: maren.stropahl@uni-oldenburg.de

Proceedings of ISAAR 2015: Individual Hearing Loss – Characterization, Modelling, Compensation Strategies. 5th symposium on Auditory and Audiological Research. August 2015, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C. Dalsgaard, L. Tranebjærg, and T. Andersen. ISBN: 978-87-990013-5-4. The Danavox Jubilee Foundation, 2015.

One way to investigate audio-visual integration is the McGurk effect which became a popular tool over the past decades (MacDonald and McGurk, 1978; McGurk and MacDonald, 1976). If individuals are presented with incongruent audio-visual syllables such as an auditory “Ba” and a visual “Ga” they may perceive neither the auditory nor the visual component but a third, different syllable (e.g., “Da”). This perception of a fused auditory and visual component is called the McGurk effect. Behavioral studies with CI users showed a bias towards the visual component of incongruent audio-visual McGurk stimuli and an altered audio-visual integration (Rouger *et al.*, 2008; Tremblay *et al.*, 2009). For by far most studies investigating the McGurk effect, research groups recorded their own stimulus material which comprises typically only one male or female speaker and very few syllables (MacDonald and McGurk, 1978; McGurk and MacDonald, 1976; Rouger *et al.*, 2008; van Wassenhove *et al.*, 2005). Basu Mallick *et al.* (2015) recently reported that the perception of the McGurk illusion strongly depends on the stimuli used. They recorded eight different McGurk stimuli from four female and four male speakers and compared the amount of fusion percepts for a large sample of 165 participants. A high variability of the amount of fusion of individuals was clearly evident across the different stimuli. Furthermore most of the participants (77%) almost always or almost never perceived the illusion, so the distribution deviates from normality (Basu Mallick *et al.*, 2015).

To account for stimulus differences and to correctly identify individual differences the noisy encoding of disparity (NED) model was proposed (Magnotti and Beauchamp, 2015). The NED model classifies each stimulus in its estimated likelihood that the auditory and the visual component evoke the McGurk effect (stimulus disparity). Furthermore the model estimates two individual parameters: the sensory noise of encoding the audio and the visual component and the individual disparity threshold which is the prior probability of an individual to encode the audio-visual incongruent stimulus as a fused percept. The individual disparity threshold is a fixed value along the stimulus disparity. Both individual parameters are assumed to be consistent across stimuli (Magnotti and Beauchamp, 2015). The two individual parameters of the model allow researchers to compare groups in their audio-visual integration independent of the presented stimulus. Using this approach, we developed a large battery of audio-visual stimuli and applied the NED model. This enabled us to investigate audio-visual integration in hearing and hearing-impaired individuals and describe group effects independent from stimulus effects. Specifically, a subgroup of eight adult, experienced CI users was compared to a control group (N = 24).

METHODS

Stimuli

To test the McGurk illusion, a set of audio-visual stimuli was recorded. Eight syllables were selected. The selection was based on the second study of MacDonald and McGurk (1978). The syllables were spoken from eight trained speakers (four females) with education in singing or theater playing, ensuring high professionalism in narrating the material. A Canon HF100 HD (CAM) high definition camera with a resolution of 1920 × 1080 (MPEG4 H.264, 25fps) was used, as well as the 26TK microphone (G.R.A.S.). Audio and video materials were synchronized offline and processed to optimized stimulus quality. The audio-visual videos obtained all start with a still image of the speaker (last frame before movement onset), followed by the spoken syllable, giving a total duration of approx. 2s for each clip. In total twelve combinations of audio-visual incongruent stimuli were used to test the McGurk illusion.

Data acquisition

To evaluate the recorded stimulus set, a control group of 24 NH students (15 females; mean age 26 ± 5.9 years) was tested. The participants did not report any neuropsychological abnormalities, had normal hearing thresholds and normal or corrected-to-normal vision. A second group consisted of eight CI users (four females) that were all post-lingually deafened and had received their implant at least one year before testing. All CI users were unilaterally implanted and seven used an additional hearing aid on the non-implanted ear which was activated during testing. The mean age of the CI group was 47 ± 24.5 years. The CI users showed a variety of hearing loss etiologies. Five CI users had presumable hereditary causes of hearing loss which was in three cases further impelled by loudness damage, two cases might have undergone a probable oxygen loss at birth, and one CI user suffered from a Gusher syndrome. The study was conducted in accordance with the local ethical committee guidelines of the University of Oldenburg and in agreement with the declaration of Helsinki. Participants gave written informed consent before the experiment. Participants were seated in a sound-shielded booth 1.5 m away in front of a screen. Audio signals were presented binaurally in a free-field setting. Three different conditions were tested in randomized order; the participants had to respond in a four-alternative forced-choice to either auditory only or visual only syllables or the percept for incongruent audio-visual syllables. Participants were instructed to select one of the four syllables presented on the screen after each trial. In the audio-visual condition the participants were instructed to indicate what they perceived aurally. Each stimulus was presented five times for each of the eight speakers, giving a total number of 800 trials (120 audio only (A_{only}), 200 visual only (V_{only}), 480 A-V incongruent (McGurk)).

Data analysis

The correct phoneme identification frequency was calculated for each condition and compared between groups. To test group differences, the Mann-Whitney-U-Test (MWU-Test) was applied. This non-parametric test is suitable for not normally distributed data and unequal group sizes. To further analyze the results and to account for group differences, the NED model by Magnotti and Beauchamp (2015) was applied. The probabilistic model allows separating individual and stimulus differences. The NED uses the individual fusion proportion for each stimulus which was defined as neither the auditory component nor the visual component but a percept of a new combination of the auditory and the visual component (originally defined as an illusion by McGurk and MacDonald (1976)). Three parameters are estimated based on the behavioral fusion data: (1) The audio-visual disparity for each stimulus estimating the differences between the auditory and the visual component and therefore the likelihood of the two components to be fused to the McGurk illusion; (2) The individual sensory noise describing the noise while processing the visual and auditory component of the audio-visual stimulus. The sensory noise is assumed to be constant for a person across stimuli; (3) The disparity threshold as the prior probability of each individual to integrate auditory and visual features (resulting in a fusion percept). The individual disparity threshold is independent of the stimulus disparity. The NED model considers stimulus differences and therefore allows comparing multi-sensory integration across individuals and across groups. The model fitting was done in R based on source code provided by Magnotti and Beauchamp (2015).

RESULTS

Correct phoneme identification

The group average result for correct phoneme identification is shown in Fig. 1. For the NH controls, the correct identification in the A_{only} condition was overall very high, with a mean of M_{NH} = 97.1%. NH individuals easily identified the audio stimuli which confirms the good quality of the audio material. The CI users on the other hand showed a significant reduction in correctly identified phonemes (M_{CI} = 68.7%, U = -4.09, p < .001). The V_{only} condition revealed a clearly diminished correct identification rate for both groups. As can be seen in Fig. 1, the groups did not differ in their ability to discriminate the V_{only} phonemes (M_{NH} = 31.3%, M_{CI} = 31.69%, U = -.22, p = .848). When evaluating the results of the AV incongruent (McGurk) condition, the correct answer would be the audio stimulus. A significant group difference could be observed for the McGurk condition. The NH controls correctly identified the audio stimuli despite the incongruent visual input with M_{NH} = 46.48%. In contrast the CI group showed a lower number of correctly identified phonemes (M_{CI} = 6.43 %), U = -3.53, p < .001. To further explore the difference in the McGurk condition, the response types of both groups were split up to determine if the participants either perceived the correct audio stimulus, the visual stimulus or a fused percept (see Fig. 2). The NH controls reported for the incorrect trials in M_{NH} = 9.7% the visual component and reported in M_{NH} = 43.83% of the trials a fused percept. The CI users reported the visual component in M_{CI} = 24.94% of the trials and a fused percept in M_{CI} = 68.62%. The CI users hence showed an overall stronger reliance on the visual component and a higher proportion of fusing the auditory and the visual component. Comparing the amount of fusion between groups and independent from the stimulus material is important. Fused percepts were therefore further analyzed with the NED model.

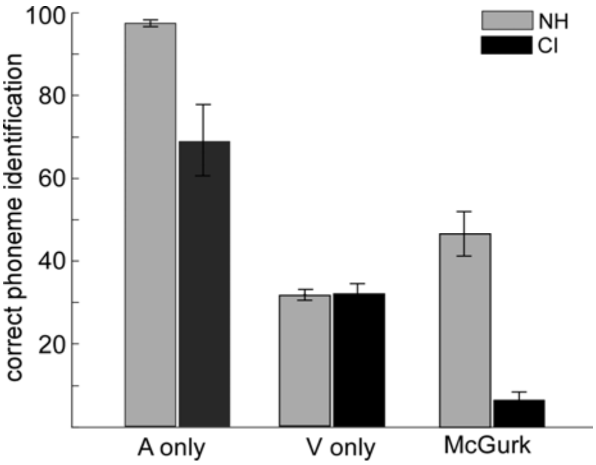

Fig. 1: Correct phoneme identification (with standard error of mean) of the NH controls (grey) and the CI users (black) for the three conditions of audio only (A only), visual only (V only) and the incongruent audio-visual combination (McGurk). CI users showed a significant deficit in understanding the correct phoneme in the A only and the McGurk condition. The visual only condition did not reveal a group difference.

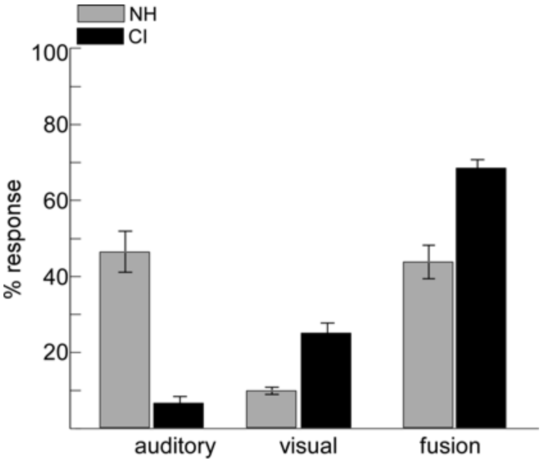

Fig. 2: Response types (with standard error of mean) for the incongruent audio-visual condition (McGurk) separated for the two groups (NH grey, CI black). The correct answer was the auditory component. For wrong answers, NH controls reported more often a fused percept and barely the visual component, whereas CI users were more focused on the visual component and showed a higher amount of fusion percept compared to NH controls.

Group comparison based on the NED-model

The comparison between the NH control group and the CI group was based on the NED model, which accounts for stimulus differences. The estimated parameters are based on the amount of fusion for each individual for each stimulus. The individual parameters, which are stimulus independent, are the sensory noise and the disparity threshold. Both parameters were estimated for each individual and the mean of the groups was compared. The MWU-Test revealed a significant group difference in the sensory noise of encoding the auditory and the visual component (*U* = -4.18, *p* < .001) as well as in the individual prior probability to perceive the McGurk illusion (*U* = -3.57, *p* < .001). The group difference is shown in Fig 3.

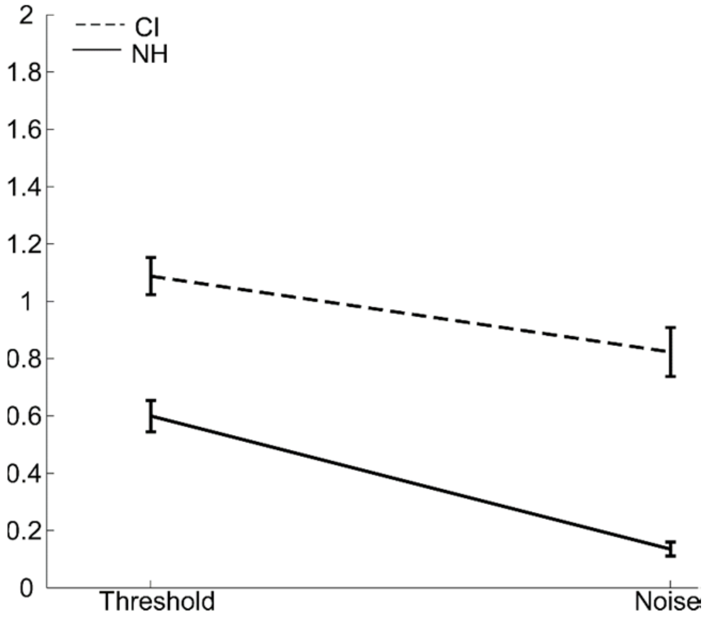

Fig. 3: Group comparison of the NED-Model parameters sensory noise (Noise) and individual disparity threshold (Threshold) plotted with standard error of mean. CI users (dashed lines) showed a significantly higher noise as well as a higher disparity threshold compared to the NH controls (solid line), which reflects differences in audio-visual speech integration.

DISCUSSION

The present study evaluated the McGurk effect tested with newly developed audio-visual stimuli. A group of NH controls and a small subgroup of CI users were compared. CI users showed a deficit in identifying the correct syllable in the A_{only} condition and showed an altered response behavior in the incongruent conditions compared to NH controls. The NED model revealed further group differences in the sensory noise of encoding the auditory and the visual component as well as in the individual probability of perceiving the McGurk illusion which further indicates differences in audio-visual integration between hearing-impaired and hearing individuals. Importantly, these measures aim to reveal a more stimulus-independent characterization of audio-visual integration.

The CI users showed a deficit in the auditory only condition which might be due to the degraded input of the CI (Fishman *et al.*, 1997). Moreover, syllables provide sparse linguistic information compared to meaningful words, hence they may be more problematic to identify for the CI users (Rouger *et al.*, 2008). Interestingly, the visual only condition revealed no group differences between the groups although previous studies suggested superior lip reading abilities even after many years of CI use (Rouger *et al.*, 2007; Stropahl *et al.*, 2015). However, also for the visual only condition the stimuli were meaningless syllables providing only little linguistic information. The better lip reading abilities might therefore result from a strong integration of lexical, semantic, and syntactic information usually provided by the audio-visual stimulus for example in daily-life communication (Rouger *et al.*, 2008). The ability of CI users to identify the correct phoneme (based on the auditory percept) in the audio-visual incongruent conditions was significantly reduced compared to NH controls. By splitting up the responses of the incongruent condition it could be shown that the CI users relied more often on the visual component in the case of ambiguous auditory input, which is in line with other studies (Rouger *et al.*, 2008; Tremblay *et al.*, 2009). In contrast the NH controls relied more often on the auditory component of the incongruent stimulus. The fact that CI users reported more fusion percepts indicates an altered, possibly stronger pattern of audio-visual integration. This interpretation is supported by the NED analysis, which also showed a significant difference in audio-visual integration of the CI users. In a study by Tremblay *et al.* (2009) the CI users did not show an overall higher fusion in the incongruent conditions, whereas descriptively the better CI users showed higher fusion proportions. Nevertheless, the amount of fusion highly depends on the stimulus material used (Basu Mallick *et al.*, 2015) which makes group comparisons within one study and across individuals and studies rather difficult if the amount of fusion is considered without taking into account stimulus effects.

We plan to make the stimulus material freely available in the near future. This will allow others to select McGurk stimuli most appropriate for specific research questions. Furthermore, an extended study investigating audio-visual integration of CI users is under way. Identifying the neural correlates of the stronger McGurk illusion in CI users may help to guide hearing restoration rehabilitation efforts.

REFERENCES

Basu Mallick, D., Magnotti, J.F., and Beauchamp, M.S. (**2015**). "Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type," Psychon. B. Rev., doi:10.3758/s13423-015-0817-4.

Driver, J. and Noesselt, T. (**2008**). "Multisensory interplay reveals crossmodal influences on "sensory-specific"brain regions, neural responses, and judgments," Neuron, **57**, 11-23.

Fishman, K.E., Shannon, R.V., and Slattery, W.H. (**1997**). “Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor,” J. Speech Lang. Hear. Res., **40**, 1201-1215.

Fu, Q.-J., Shannon, R.V., and Wang, X. (**1998**). “Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing,” J. Acoust. Soc. Am., **104**, 3586-3596.

MacDonald, J. and McGurk, H. (**1978**). “Visual influences on speech perception processes,” Percept. Psychophys., **24**, 253-257.

Magnotti, J.F. and Beauchamp, M.S. (**2015**). “The noisy encoding of disparity model of the McGurk effect,” Psychon. B. Rev., 22, 701-709. [Source code: http://openwetware.org/wiki/Beauchamp:NED]

McGurk, H. and MacDonald, J. (**1976**). “Hearing lips and seeing voices,” Nature, **264**, 746-748.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., and Barone, P. (**2007**). “Evidence that cochlear-implanted deaf patients are better multisensory integrators,” Proc. Nat. Acad. Sci. USA, **104**, 7295-7300.

Rouger, J., Fraysse, B., Deguine, O., and Barone, P. (**2008**). “McGurk effects in cochlear-implanted deaf subjects,” Brain Res., **1188**, 87-99.

Sandmann, P., Plotz, K., Hauthal, N., de Vos, M., Schönfeld, R., and Debener, S. (**2014**). “Rapid bilateral improvement in auditory cortex activity in postlingually deafened adults following cochlear implantation,” Clin. Neurophysiol.. **126**, 594-607.

Stropahl, M., Plotz, K., Schönfeld, R., Lenarz, T., Sandmann, P., Yovel, G., De Vos, M., and Debener, S. (**2015**). “Cross-modal reorganization in cochlear implant users: Auditory cortex contributes to visual face processing,” NeuroImage, **121**, 159-170.

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

Tremblay, C., Champoux, F., Lepore, F., and Théoret, H. (**2009**). “Audiovisual fusion and cochlear implant proficiency,” Restor. Neurol. Neuros., **28**, 283-291.

van Wassenhove, V., Grant, K.W., and Poeppel, D. (**2005**). “Visual speech speeds up the neural processing of auditory speech,” Proc. Nat. Acad. Sci. USA, **102**, 1181-1186.