

# McGurk stimuli for the investigation of multisensory integration in cochlear implant users: The Oldenburg Audio Visual Speech Stimuli (OLAVS)

Maren Stropahl<sup>1</sup> · Sebastian Schellhardt<sup>2</sup> · Stefan Debener<sup>1,3</sup>

Published online: 25 August 2016  
© Psychonomic Society, Inc. 2016

**Abstract** The concurrent presentation of different auditory and visual syllables may result in the perception of a third syllable, reflecting an illusory fusion of visual and auditory information. This well-known McGurk effect is frequently used for the study of audio-visual integration. Recently, it was shown that the McGurk effect is strongly stimulus-dependent, which complicates comparisons across perceivers and inferences across studies. To overcome this limitation, we developed the freely available Oldenburg audio-visual speech stimuli (OLAVS), consisting of 8 different talkers and 12 different syllable combinations. The quality of the OLAVS set was evaluated with 24 normal-hearing subjects. All 96 stimuli were characterized based on their stimulus disparity, which was obtained from a probabilistic model (cf. Magnotti & Beauchamp, 2015). Moreover, the McGurk effect was studied in eight adult cochlear implant (CI) users. By applying the individual, stimulus-independent parameters of the probabilistic model, the predicted effect of stronger audio-visual integration in CI users could be confirmed, demonstrating the validity of the new stimulus material.

**Keywords** Cochlear implants · McGurk illusion · Audio-visual integration

✉ Maren Stropahl  
maren.stropahl@uni-oldenburg.de

<sup>1</sup> Department of Psychology, Neuropsychology Lab, European Medical School, Carl von Ossietzky University of Oldenburg, Ammerländer Herrstraße 114-118, 26129 Oldenburg, Germany

<sup>2</sup> Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany

<sup>3</sup> Cluster of Excellence Hearing4all Oldenburg, Oldenburg, Germany

## Introduction

If individuals are presented with incongruent audio-visual syllables such as an auditory “Ba” and a visual “Ga”, the integration of both may result in the percept of a different syllable (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). However, it is known that the McGurk illusion depends strongly on the particular stimuli used (Basu Mallick, Magnotti, & Beauchamp, 2015; Jiang & Bernstein, 2011; MacDonald & McGurk, 1978). The free availability of well-characterized, high-quality stimulus material would be desirable to facilitate inferences across studies and individuals.

Talker characteristics (such as differences in articulation) and the particular syllable combination(s) used seem to influence the McGurk illusion. Whereas the effect of different syllable combinations on the fusion frequency has been known since MacDonald & McGurk (1978), the issue of talker characteristics has received less attention. It is suggested that talker characteristics also influence the perception of either auditory or visual speech (Demorest & Bernstein, 1992; Jiang & Bernstein, 2011). Influencing factors, such as added noise, for example, caused by reduced audio and video quality (Fixmer & Hawkins, 1998), might influence the perception of the McGurk illusion. Other factors such as eye blinks, make-up or differences in face size are not yet well investigated, but might also alter the perception of the McGurk illusion (Basu Mallick et al., 2015). Most studies recorded their own stimulus material, frequently using a female talker (e.g., Basu Mallick et al., 2015; MacDonald & McGurk, 1978; McGurk & MacDonald, 1976; Rouger, Fraysse, Deguine, & Barone, 2008; van Wassenhove, Grant, & Poeppel, 2005). For auditory speech perception, intelligibility is more biased towards female compared with male talkers (Bradlow, Torretta, & Pisoni, 1996), which may reflect the prevalence of specific phonological reduction phenomena such as vowel

centralization or reduced frequency of stop release. Freely available stimuli (cf. Table 1 in Basu Mallick et al., 2015) seem to vary strongly in the above-mentioned recording aspects, rendering post-processing for quality enhancement ineffective.

For a particular stimulus, most individuals seem to perceive the McGurk illusion almost always or almost never (Basu Mallick et al., 2015; Nath & Beauchamp, 2012; Tremblay & Champoux, 2007). Accordingly, the fusion distribution deviates from normality which renders comparisons based on mean and standard deviation measures invalid (Basu Mallick et al., 2015; Nath & Beauchamp, 2012; Stevenson, 2012; Strand, Cooperman, Rowe, & Simenstad, 2014). This might be less of a problem if within-subject comparisons are investigated (e.g., van Wassenhove et al., 2005) but becomes critical when the McGurk effect is compared between perceivers, as it is not possible to differentiate between the influence of the given stimulus and the perception of the participant viewing the stimulus. An elegant solution to overcome this problem is the noisy encoding of disparity (NED) model (Magnotti & Beauchamp, 2015), which characterizes each stimulus on its stimulus disparity, the estimated likelihood that the particular combination of auditory and visual syllables elicits the McGurk effect. The NED model also estimates two parameters to characterize a perceiver's audio-visual integration strength, the sensory noise of encoding the auditory and the visual component and the individual disparity threshold. The latter is the prior probability of a perceiver to encode the audio-visual incongruent stimulus as a fused percept. Both individual parameters are assumed to be consistent across

stimuli (Magnotti & Beauchamp, 2015). Importantly, the NED model enables the comparison of audio-visual integration abilities between perceivers, independent of stimulus effects. Additionally, the model provides the ability to predict for a perceiver the frequency of the McGurk effect by a certain stimulus that has not been presented to that person before.

Audio-visual integration plays a major role in hearing impaired individuals, such as cochlear implant (CI) users. CI users outperform normal hearing individuals in lip reading and audio-visual integration (Rouger et al., 2007; Schorr, Fox, van Wassenhove, & Knudsen, 2005; Strelnikov et al., 2015; Strelnikov, Rouger, Barone, & Deguine, 2009; Stropahl et al., 2015). Despite the better multisensory integration abilities, for incongruent multisensory information CI users focus more on the visual modality but did not show a higher fusion (Desai, Stickney, & Zeng, 2008; Rouger et al., 2008; Schorr et al., 2005; Tremblay, Champoux, Lepore, & Théoret, 2009). Yet most of these studies used stimulus material from a single talker and one syllable combination. Whether stimulus-specific effects biased these results cannot be ruled out as long as a stimulus-independent analysis strategy is not applied.

We have developed a new set of 96 different audio-visual stimuli, the Oldenburg audio-visual speech (OLAVS) stimuli. Evaluation included stimulus characterization with the NED model in normal-hearing individuals. A re-test was conducted approximately 1 year later to validate previous results. Furthermore, experienced CI users were studied to test the hypothesis of stronger audio-visual integration in CI users.

## Methods

### Stimuli

A set of eight syllables, which were selected from the second study of MacDonald & McGurk (1978), were recorded (Ba, Da, Ga, Ka, Ma, Na, Pa, Ta). The syllables were spoken from eight trained talkers (four females) ensuring high professionalism in narrating the material. Talkers were native German students at a mean age of 25 years. Individual salient features were avoided. Recordings were made in a sound-shielded booth with a Canon HF100 HD (CAM) high-definition camera with a resolution of 1920 × 1080 (MPEG4 H.264, 25fps) as well as the 26TK microphone (G.R.A.S.). Every syllable was repeated at least three times to assure high quality of the material (e.g., to later, offline, select takes without eye blinks and with most natural articulation). The audio stimuli were recorded as WAV-files (PCM), mono with 24Bit-rate, and a sampling rate of 48 kHz. Audio and video streams were synchronized and processed offline to optimize stimulus quality. The separately recorded audio stream was synchronized to the

**Table 1** Twelve syllable combinations were used as audio-visual stimuli, selected based on the original MacDonald & McGurk (1978) data. Syllables were tested in the unimodal conditions auditory only and visual only and in bimodal audio-visual combinations. The right column shows the response options presented to the participant as a four-alternative forced choice design (4-AFC)

A – V Stimulus	4-AFC options (A, V, Fusion1, Fusion2)
Ba-Da	Ba, Da, Ga, Pa
Ba-Ga	Ba, Ga, Da, Ma
Ba-Ka	Ba, Ka, Ga, Da
Ba-Na	Ba, Na, Ga, Da
Ba-Ta	Ba, Ta, Pa, Da
Ma-Ga	Ma, Ga, Na, Ba
Ma-Ta	Ma, Ta, Na, La
Pa-Da	Pa, Da, Ka, Ta
Pa-Ga	Pa, Ga, Ka, Ta
Pa-Ka	Pa, Ka, Da, Ta
Pa-Na	Pa, Na, Ka, Ta
Pa-Ta	Pa, Ta, Da, Ka

simultaneously recorded audio stream of the HD camera. Based on existing literature, the synchronization of an incongruent audio syllable and the corresponding visual speech (e.g., audio “Ba” and video “Ga”) was done manually by temporally aligning the incongruent audio file (“Ba”) to the visual speech of the video-based file on the consonant burst of the original, congruent (“Ga”) audio file (Grant & Greenberg, 2004; van Wassenhove et al., 2005). Software used for the synchronization was Praat (Boersma, 2002) and Adobe Premiere Pro (ADOBE). The final audio-visual videos all begin with a still image of the talker (last frame before movement onset) followed by the spoken syllable, giving a total duration of approx. 2 s for each clip. In total, 12 combinations of audio-visual syllables were created for each talker (see Table 1), giving a set of 96 tokens in total.

### Data acquisition

The stimulus set was evaluated with 24 normal-hearing (NH) students (15 females; mean age  $26 \pm 5.9$  years) without neurological or psychiatric conditions, normal hearing thresholds and normal or corrected-to-normal vision. For a re-test, 16 out of 24 normal-hearing participants accepted an invitation for a re-test, which was conducted approximately 1 year later. Additionally, audio-visual integration of a second group of eight CI users (four females; age range 19–75 years with a mean age of  $47 \pm 24.5$  years) was compared with the NH group. All CI users were post-lingually deafened and unilaterally implanted, five with the implant on the right ear. All participants received the implant at least 12 months before testing and seven CI users used a hearing aid on the non-implanted ear, which was activated during testing. The CI users showed a variety of hearing loss etiologies. Hereditary causes of hearing loss might have induced deafness in five CI users, which was further accompanied by loudness damage in three CI users. Two cases might have suffered from oxygen loss at birth and one CI user was diagnosed with a Gusher syndrome. The study was approved by the local ethical committee of the University of Oldenburg and conducted 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 in front of a computer monitor. Videos were sized to  $1920 \times 1080$  pixels presented on a 24-inch (c.60-cm) screen with a horizontal angle of  $30^\circ$  and a vertical angle of  $17^\circ$ . Audio signals were presented binaurally in a free-field setting. Loudness was set to a pre-defined comfortable loudness level of approximately 60 dB (A) and if necessary individually adjusted, to ensure good understanding of speech. The three experimental conditions, unimodal auditory ( $A_{\text{only}}$ ) and visual ( $V_{\text{only}}$ ) and incongruent audio-visual (McGurk) were randomly presented in a single-trial four-alternative forced-choice (4-AFC) procedure. The 4-AFC was designed such that the four response options always

represented the auditory and the visual component of the AV stimulus as well as two possible fusion options (see Table 1). Response options differed for each presented syllable combination. The two possible fusion options were selected based on the outcome of the original study by MacDonald & McGurk (1978), using those responses that showed the highest fusion amount. To validate the response options, plausibility of options was piloted before the experiment. In the audio-visual condition, the participants were instructed to indicate what they perceived aurally. Each of the 96 tokens was presented five times. For the re-test session 1 year later, the 16 participants were tested with the same experimental set-up as in the original experiment, but only a subset of the audio-visual incongruent condition was tested and the amount of trials for each AV token was changed from 5 to 15 repetitions.

### Data analysis

The NH group data were used to evaluate the OLAVS stimuli. Individual fusion frequencies were calculated, which were defined as the percept of a combination of the auditory and the visual component (originally defined by McGurk & MacDonald, 1976). For all incongruent McGurk trials, the two possible fusion response options of the 4-AFC were equally counted as fusion for a specific AV token. Fusion responses were averaged over trials to obtain the mean fusion for each specific AV token. In a first step, the hypothesis of differences in fusion percepts for different stimuli were investigated to further motivate the stimulus-independent analysis of the noisy-encoding of disparity model for the group comparison (NED, Magnotti & Beauchamp, 2015). To depict differences in fusion percepts due to different AV tokens, the effects of talkers and different syllable combinations were investigated. Therefore, the average over each of the eight talkers was computed for each AV syllable combination. Additionally, to reveal differences based on different talkers, the average over all syllable combinations for each talker was computed. The non-parametric Friedman ANOVA, accounting for a violation of normal distribution, was applied to investigate differences of the stimulus material. Furthermore, to accurately account for differences between *all* 96 individual AV tokens, the NED model by Magnotti & Beauchamp (2015) was applied. This model allows for separating individual and stimulus differences and is moreover able to reveal stimulus differences between large amounts of AV tokens, which was suitable for our stimulus set. Three parameters are estimated based on the behavioral fusion data. (1) The *audio-visual disparity* estimates for each AV token the likelihood of the two components to be fused to the McGurk illusion. The disparity is independent of the individual fusion frequency and inversely associated to its property to elicit the illusion. The design of the model restricts parameter values for fusion amounts of 0 % or 100 % to avoid parameter values

going against infinity (e.g., 0 % fusion is mapped to a maximum of 2.0 disparity). This restriction does not change the overall model fit but makes calculations of means and standard deviations plausible (Magnotti & Beauchamp, 2015). In a second step, the CI users were compared with the NH controls by the two individual parameters of the model. (2) The *individual sensory noise* describes the noise while processing the visual and auditory component of the audio-visual token. (3) The *disparity threshold* is defined as the prior probability of each perceiver to integrate auditory and visual features (resulting in a fusion). The individual disparity threshold as well as the sensory noise is independent of the stimulus disparity and assumed constant for a person across stimuli. The values for all parameters have arbitrary scaling and the same disparity unit. The model uses the approach of minimizing the squared error between measured data (fusion for each token averaged across trials) and the model prediction. In a first step, the stimulus disparity is calculated by transforming the mean fusion over subjects for each AV token into an initial value of the stimulus disparity ( $e^{1 - \text{mean}(\text{fusion of stimulus})} - 1$ ). Individual participant parameters are then estimated and the stimulus disparity fitted. A fitting procedure of participants and stimulus values is furthermore applied to obtain the best-fitting parameters. For more details cf. Magnotti & Beauchamp (2015). The model fitting was done in R based on source code provided by Beauchamp and colleagues (<http://openwetware.org/wiki/Beauchamp:NED>).

The validity of the model was tested in several ways. First, the root mean square error (RMSE) was computed to test the overall model fit carried out as presented in Magnotti & Beauchamp (2015). The RMSE was calculated for the fit of the fusion of each stimulus (across subjects) predicted by the model to the measured data. The same procedure was applied for testing the fit for the predicted fusion for each subject (across stimuli) compared to the measured fusion for each subject. Results of the NH model were compared with the data provided by Magnotti & Beauchamp (2015). Additionally, two assumptions about the NED were tested, following the procedure described in Magnotti & Beauchamp (2015). First, the model assumes that participant parameters are consistent across stimuli, meaning that, if a participant has a higher fusion compared with another participant for a specific AV token, the fusion should also be higher for that participant for another AV token. For testing this assumption, the rank for all participants for each stimulus was computed to compare it for each participant to his/her overall rank (averaged across stimuli). The second assumption is that stimulus parameters are consistent across participants. This means that, if one stimulus is weaker in eliciting the McGurk effect compared with another stimulus for one participant, then this should also be the case for another participant. For this analysis, the rank for each stimulus was calculated

for all participants to be compared with each stimulus's overall rank. Due to large amounts of correlation coefficients, results are reported as average correlation coefficients derived following Fisher's  $z$  transformation.

To further test the validity of the model data, a nested modeling approach was performed, testing the influence of single participants as well as single AV tokens. First, in a leave-one-subject-out validation approach, the model was fitted 24 times, each time excluding one of the participants ( $n = 23$  participants and  $n = 96$  AV token). Due to non-normal parameter distributions, the parameter medians (across 23 subjects) of the 24 models were tested against the original model performing the paired-samples sign rank test (noise, disparity threshold and stimulus disparity). In a second approach, a leave-one-stimulus-out validation was applied. The original model was compared with the estimated parameters of 25 different sub-models, each time leaving out one randomly chosen AV token. This procedure was applied consecutively for 25 randomly chosen AV tokens instead of for all 96 AV tokens. Accordingly, the parameters for 25 models were estimated, each containing all 24 participants and 95 (one left-out) AV token. Also, the sign rank test was used here to test the new model outcomes against the median of the original model for all three parameters.

The analysis of the re-test session included a comparison of the overall mean fusion for each of the 24 stimuli from the original session and the re-test session. After verifying for a normal distribution with the Kolmogorov–Smirnov test, a paired  $t$  test was performed, comparing the average fusion between the test and the retest trials for each subject.

To test for group differences, the correct phoneme identification was calculated for each condition and compared between groups. The Mann–Whitney  $U$  test (MWU-test) was applied to account for violation of the normality assumption and unequal group sizes. The same procedure was applied to account for differences in the response chosen within the incongruent McGurk trials. Furthermore, to compare audio-visual integration of CI users and NH controls in a stimulus-independent way, a combined NED model of the data from the CI users and the NH controls (in a single model) was computed. The model fit was also evaluated based on the RMSE for the fit of individual stimuli across subjects and on the RMSE for each subject across stimuli. Group differences in NED parameter outcome (sensory noise and individual disparity threshold) were compared in a first analysis with the Kruskal–Wallis test, the non-parametric one-factorial analysis of variance (without repeated measures), which is suitable if the data violates a normal distribution and has unequal group sizes. Significant effects were followed up post-hoc with the MWU-test.

For all statistical analyses, degrees of freedom are reported and, where necessary, a correction for multiple comparisons using the Bonferroni method was applied.

## Results

### Evaluation of stimulus material

The first part of the “Results” sections describes the evaluation of the stimulus material, which is based on the NH group only ( $n = 24$ ). The fusion frequency as well as the stimulus disparity obtained by the NED model will be reported. To first highlight the expected variability of the fusion frequency depending on either the 12 syllable combinations or the different talkers, the fusion frequency of the 24 NH controls was averaged for each of the stimuli over all eight talkers. To correct for multiple comparisons, an alpha level of 0.025 was chosen. The results are shown in Fig. 1a.

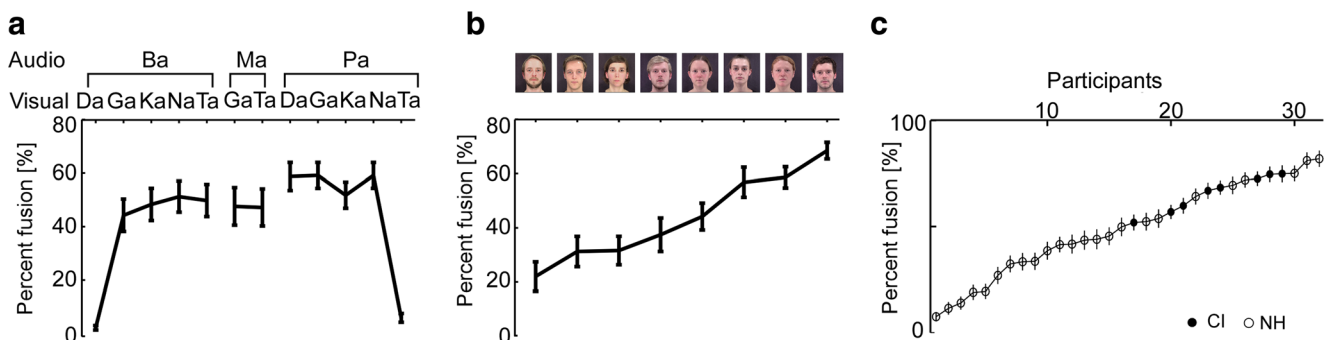
The mean fusion frequency was significantly different between different stimuli ( $\chi^2(11) = 127.66, p < .001$ ). Although most of the stimuli achieved a mean fusion frequency between 40 % and 60 %, individual fusion frequencies showed a different pattern (see Fig. 2). Over 80 % of the individual percepts were either rarely (<20 %) or very frequently a fusion (>80 %).

Secondly, the mean fusion frequency of all participants was averaged over all stimuli for each talker (Fig. 1b), which were significantly different for the talkers, ranging from 22 % to 68.5 % ( $\chi^2(7) = 103.90, p < .001$ ). To describe these stimulus differences more accurately, the NED disparity was estimated for all 96 stimuli separately. The stimulus disparity for the NH control group ranged from the minimum [0] to the maximum [2] possible value, with a median of  $M = 0.74$  (SEM 0.05). The results for all 96 stimuli are shown in Fig. 3a, with color indicating the talker. Figure 3b shows that the stimulus disparities for all eight talkers were distributed over the whole disparity range. The overall model fit was tested by calculating the root mean square error. The average RMSE for the subjects (across stimuli) for the NH group was comparable with the data provided by Magnotti & Beauchamp (2015),  $RMSE_{NH} = 0.019$  ( $RMSE_{M\&B} = 0.026$ ). The average RMSE for all stimuli (across subjects) was  $RMSE_{NH} = 0.042$  ( $RMSE_{M\&B} = 0.032$ ).

The NED model assumes that participant parameters are consistent across stimuli and that stimulus parameters are consistent across participants. These assumptions were tested with the same procedure as in Magnotti & Beauchamp, (2015), which is summarized in the “Methods” section. All average correlation coefficients reported were derived following Fisher’s  $z$ -transformation.

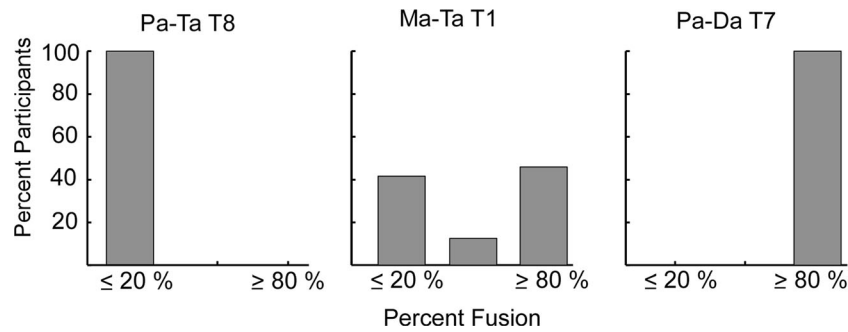
Testing if participant effects are consistent across stimuli showed that the assumption was met for the data of the NH group, revealing correlation coefficients ranging from  $-0.28$  to  $0.89$ . The corrected average correlation coefficient was  $r_{nh}(94) = 0.64 (\pm 0.04 \text{ SEM})$ , which is very similar to the value reported by Magnotti & Beauchamp ( $r = 0.65 \pm 0.04 \text{ SEM}$ ). The consistency of the stimulus parameter across participants was also relatively high, with a corrected average  $r_{nh}(22) = 0.74 (\pm 0.05 \text{ SEM})$  with coefficients ranging from  $0.28$  to  $0.86$ . Again, this result is similar to the one obtained by Magnotti & Beauchamp ( $r = 0.64 \pm 0.02 \text{ SEM}$ ).

To further test the validity of the model data, a nested modeling approach was performed, testing the influence of single participants as well as single AV tokens. First, in a leave-one-subject-out validation approach, the model was fitted 24 times, each time excluding one of the participants ( $n = 23$  participants and  $n = 96$  AV token). The three parameter medians of the 24 models were compared with the original model with all 24 participants by applying the sign rank test. Bonferroni correction for multiple comparisons was applied, giving an adjusted alpha level of  $0.017$  for the 3 comparisons. The sign rank test did not reveal a difference for any of the parameters ( $p_{noise} > .627$ ;  $p_{thresh} > .549$ ;  $p_{stimdisp} > .032$ ). Second, in another leave-one-stimulus-out validation approach, the influence of single AV tokens was tested. The original model was compared to the estimated parameters of 25 different sub-models, each time leaving out one randomly chosen AV token. Accordingly, the parameters for 25 models were estimated, each containing all 24 participants and 95 AV tokens. The corrected  $p$ -value for the 3 comparisons for a significant difference was  $p = .017$ . The medians of the model



**Fig. 1** Fusion proportion for the NH control group ( $n = 24$ ). Fusion is defined as neither the auditory nor the visual component but a new “fused” syllable. **a** The variability of the fusion proportion for the different syllable combinations, and **b** the talker dependency of the

fusion proportion. **c** The sorted fusion for NH controls (open circles) and CI users (filled circles) averaged over all stimuli. Error bars indicate the standard error of the mean



**Fig. 2** Examples of three AV tokens showing a skewed (*left and right*) or bimodal (*middle*) distribution of fusion perceptions. Plotted are the percentages of participants for a particular fusion proportion

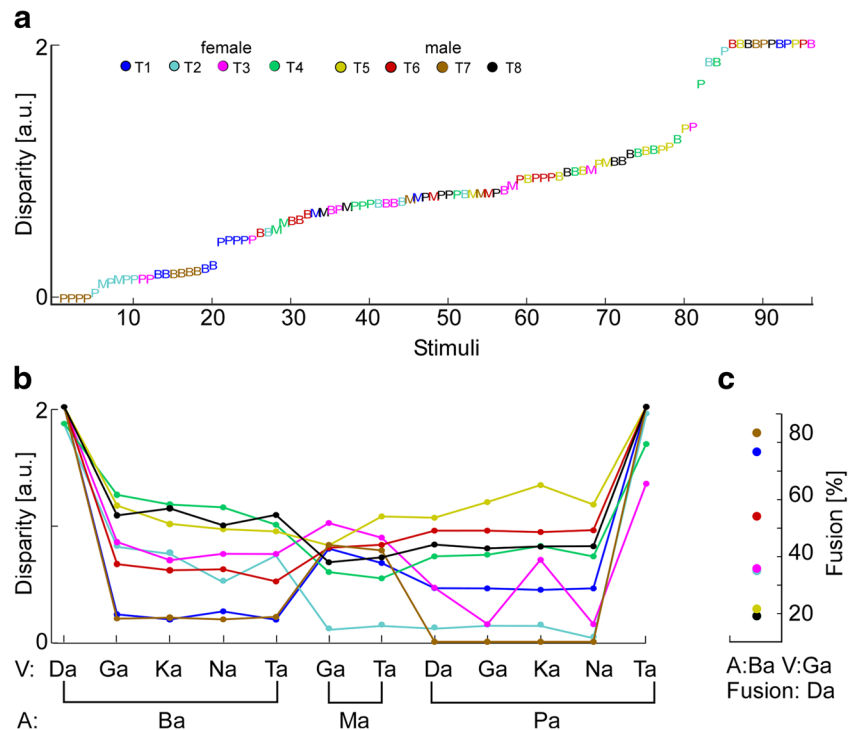
parameter did not differ significantly for the individual parameters ( $p_{noise} > .382$ ;  $p_{thresh} > .065$ ), but showed a significant difference for the overall stimulus disparity median ( $p_{stimdisp} < .017$ ).

The analysis of the amount of fusion in the re-test that was conducted 1 year later revealed no significant difference between the mean fusion for each participant for the test versus the re-test session ( $t = -1.23$ ,  $p = .237$ ).

**Group comparison**

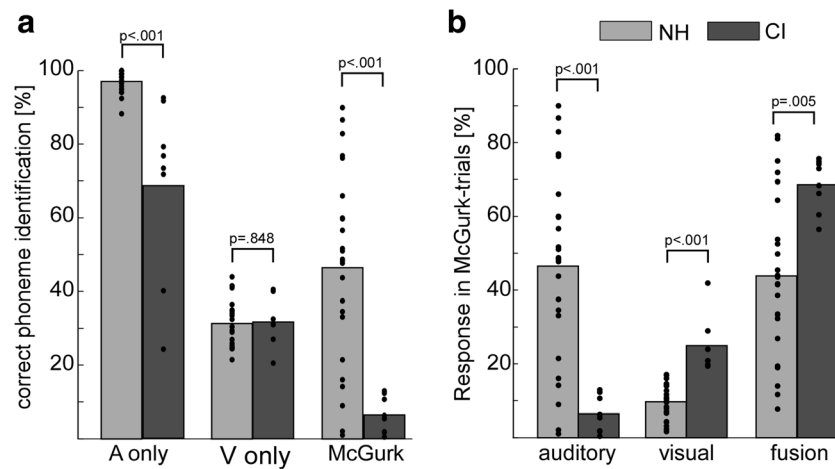
The second part of the results describes the comparison between CI users and NH controls. The group average results for the correct phoneme identification are shown in Fig. 4a. Group differences were investigated by applying the MWU

test. The Bonferroni corrected alpha level for the three comparison between CI users and NH controls for the correct phoneme identification was  $p = .017$ . 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 controls easily identified the auditory stimuli in the auditory condition. 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 (Fig. 4a) and no group difference in the ability to discriminate the  $V_{only}$  phonemes was observed ( $M_{NH} = 31.3\%$ ,  $M_{CI} = 31.69\%$ ,  $U = -.22$ ,  $p = .848$ ). When evaluating the results of the AV incongruent (McGurk) condition, the correct answer would be the audio



**Fig. 3** Stimulus disparity obtained by the NED model. **a** Ranked stimulus disparities for each of the 96 AV tokens. Plotted are the consonants of the auditory component of the McGurk token (e.g., *B* for *Ba*). Different talkers are color-coded. **b** Stimulus disparities for each AV

token shown for each talker. **c** The mean fusion averaged over all participants for each talker for the most common McGurk combination *A:Ba* with *V:Ga*. Fusion in this case is just the response *Da*, to allow comparison to other studies using this AV syllable combination



**Fig. 4** **a** The correct phoneme identification of the NH controls (light gray) and the CI users (dark gray) are shown for the three experimental conditions. In the incongruent audio-visual condition (McGurk), the

correct answer was by instruction the auditory stimulus. **b** Response types for the McGurk condition separated in the percept of the auditory, the visual, or the fusion component. Each dot indicates a single subject

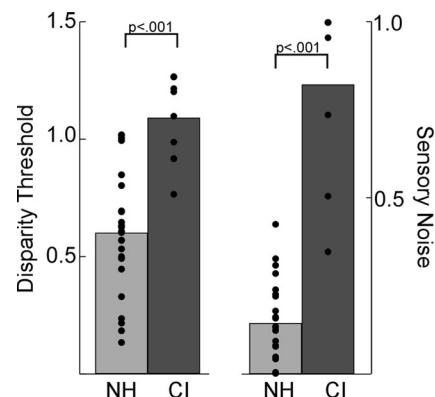
part of the AV token. A significant group difference could be observed for the McGurk condition. The NH controls correctly identified the audio stimuli despite the incongruent visual input with  $M_{NH} = 46.48\%$ . In contrast, the CI group showed a significantly lower number of correctly identified phonemes ( $M_{CI} = 6.43\%$ ,  $U = -3.53$ ,  $p < .001$ ). To further explore the McGurk condition, the response types of both groups were split up to assess if the participants either perceived the correct audio part of the AV token, or in the incorrect trials the visual part of the AV token, or a fused percept (see Fig. 4b). To test for group differences, the MWU was applied and the Bonferroni correction adjusted the alpha value to 0.017 for the three comparisons. The difference in the amount of the chosen audio component has already been shown previously. The NH controls reported for the incorrect trials significantly less the visual component ( $M_{NH} = 9.7\%$ ,  $M_{CI} = 24.94\%$ ,  $U = -4.18$ ,  $p < .001$ ). A fused percept was reported by the NH controls in  $M_{NH} = 43.83\%$  of the trials, which was significantly less compared with the CI users ( $M_{CI} = 68.62\%$ ,  $U = -2.74$ ,  $p = .005$ ). Nevertheless, comparing the fusion frequency between groups is difficult and should importantly be done independent of the stimulus material. Fused percepts were therefore further analyzed with the NED-model.

A combined NED model for CI users and NH controls was computed to compare individual audio-visual integration. The estimated NED parameters, the sensory noise and the disparity threshold are based on the fusion frequency of each individual for each AV token. To reveal group differences, the Kruskal–Wallis test was applied, followed up by a post hoc analysis, performing the MWU test on the group data. The Bonferroni adjusted alpha level was 0.025. The Kruskal–Wallis test revealed a difference between the groups for both NED parameters, disparity threshold:  $\chi^2(1) = 12.74$ ,  $p < .001$ , sensory noise:  $\chi^2(1) = 17.16$ ,  $p < .001$ . The MWU test revealed that the group of CI users showed a higher sensory noise ( $CI_{noise}$

median = 0.98; SEM = 0.09) compared with the NH controls ( $NH_{noise}$  median = 0.13; SEM = 0.02),  $U = -4.14$ ,  $p < .001$ . The individual prior probability to perceive the McGurk illusion was higher in the CI group as well ( $CI_{threshold}$  median = 1.15; SEM = 0.07) compared to the NH controls ( $NH_{threshold}$  median = 0.61; SEM = 0.05),  $U = -3.57$ ,  $p < .001$ . The group difference is shown in Fig. 5. The evaluation of the model fit for the combined data from the NH and CI group revealed an average RMSE for all stimuli (across subjects) of RMSE = 0.019 (compared with  $RMSE_{M\&B} = 0.026$ ) and an average RMSE for the subjects (across stimuli) of RMSE = 0.061 (compared to  $RMSE_{M\&B} = 0.032$ ).

**Discussion**

The present study evaluated newly developed audio-visual stimuli with a group of 24 NH individuals. Although we took special care to standardize recording and post-processing



**Fig. 5** Group mean individual NED-Model parameters disparity threshold (left) and sensory noise (right) for NH controls (light gray) and CI users (dark gray). Each dot indicates the value of a single subject

conditions and avoided as much physical variance as possible, clear stimulus-dependent fusion frequencies were evident, replicating previous reports (Basu Mallick et al., 2015; Jiang & Bernstein, 2011). The different syllables revealed mean fusion frequencies of approximately 40–60 %, which, however, is highly misleading because the mean values arose by a bimodal distribution of the perceivers. Most of the participants either never or always perceived a fused percept for a particular token (cf. Fig. 2). This is further evident from the elicited fusion frequency for each talker/each syllable combination showing a significant variability. The observed variability highlights the need for talker- and stimulus-independent measures and confirms recent studies indicating influences of both on the results of McGurk studies (e.g., Basu Mallick et al., 2015; Jiang & Bernstein, 2011). It may be inferred that stimulus-dependent factors average out if a large group of participants is presented with a huge amount of stimuli, which is, however, not a very efficient strategy in many experimental setups. These basic results emphasize the importance of comparing the fusion frequency independent of the presented stimuli. To address the concern, that five repetitions of AV tokens may not be sufficient to reveal a reliable behavioral pattern, a re-test was performed approximately 1 year later. We re-invited 16 out of 24 participants and presented them with a subset of 24 AV tokens, each repeated 15 instead of the initial 5 times. The analyses of the re-test session confirm a very high test–retest reliability and furthermore indicate that the initial five repetitions were sufficient to reveal a reliable behavioral pattern, even in a relatively small sample. We tested our stimuli in a closed response set, giving the participants a 4-AFC with the auditory and the visual component and two fusion options as response choices. There are studies showing that differences in the amount of fusion percepts occur between an open and a closed response set (Basu Mallick et al., 2015; Colin, Radeau, & Deltenre, 2005), and especially that fusion percepts are more likely in a closed AFC setting. Nevertheless, it has been found that the variability across stimuli and the extreme values for particular stimuli/participants can be observed in both settings (Magnotti et al., 2015). Jiang & Bernstein (2011) suggest evaluating new stimuli in an open-set before using an AFC. We tested our two proposed fusion options in a pilot study before the main experiment and participants did report positively on the response options presented. We encourage users to use the OLAVS stimuli to test the McGurk illusion in an open-response set.

The NED-model derived stimulus disparity delivers a metric that allows for selecting stimuli based on their ability to elicit the McGurk illusion independent of talker, talker gender and syllable combinations (Magnotti & Beauchamp, 2015). We therefore characterized our 96 stimuli on their disparity and found that the OLAVS stimuli covered the whole range of possible disparities, which makes this set suitable for addressing different research questions with the McGurk illusion. Testing the model fit by calculating the error of the estimated

and the measured fusion percepts for the subjects and for the stimuli revealed reliable results. Further positive outcomes of testing the NED model also support the validity of our results. Alternatively, the estimated disparity for our OLAVS stimulus set provides the possibility to select audio-visual stimuli according to their likelihood of eliciting the McGurk illusion. The OLAVS stimuli, including information about individual stimulus disparities, will be made freely available.<sup>1</sup>

Our main motivation for developing OLAVS was to assess differences between hearing-impaired and normal-hearing individuals in their audio-visual integration. To this end, we compared in a second step a small subgroup of CI users to the NH controls by the correct phoneme identification and by the individual parameters of the NED model, aiming to test if a more stimulus-independent confirmation of the predicted stronger audio-visual integration in CI users could be achieved. Our group of CI users showed a significantly higher frequency of fusion compared with the NH controls, whereas recent studies did not reveal an overall significantly higher fusion frequency for the CI users. An indication of a difference in the fusion frequency could nevertheless be revealed in those studies, depending on the CI benefit (Tremblay et al., 2009), the age at implantation (Schorr et al., 2005) and the duration of the CI experience (Desai et al., 2008). However, considering that the fusion frequency depends greatly on the stimulus material used, group comparisons within one study and across individuals and studies are difficult to generalize. Our result of a higher fusion frequency in CI users is based on a very small sample size but a large set of stimuli. Nevertheless, the combined model fit (data of CI users and NH controls in a single model) revealed a reasonable RMSE, which is in line with the evaluation of the original data (Magnotti & Beauchamp, 2015). Moreover, the interpretation of stronger audio-visual integration is supported by the NED model confirming group differences in the sensory noise of encoding the auditory and the visual component and, importantly, in the individual threshold of perceiving the McGurk illusion. It is known that the perception of the McGurk illusion becomes stronger with increasing age in normal-hearing adults (Sekiyama, Soshi, & Sakamoto, 2014). Age differences for CI users regarding the McGurk percept have not been intensively investigated so far. Even though CI individuals covered a wide age range (19–75 years), a clear association between age and model outcome was not found, and variance in model outcome parameters was similar between groups. Furthermore, the individual disparity threshold for the youngest CI user (1.27) descriptively did not differ much from the two oldest CI participants (1.21 and 1.10). It therefore appears unlikely that the group differences we observed are strongly biased by age. Nevertheless, future work investigating the role of age on audio-visual integration in hearing-impaired populations may reveal

<sup>1</sup> <https://www.uni-oldenburg.de/psychologie/neuropsychologie/links/>



interesting insights. In any case, our study is among the first to support differences in audio-visual integration abilities between hearing impaired and hearing individuals on a stimulus-independent level. We conclude that the OLAVS stimulus set can identify differences in audio-visual integration between hearing-impaired and normal-hearing individuals, in particular if the NED model is applied.

Furthermore, the response behavior in the incongruent cases in which the participants did not perceive a fusion percept was analyzed, meaning the analysis of the percentage of responses to the auditory or the visual part of the audio-visual McGurk stimulus. As expected from existing literature, another difference between CI users and NH controls occurred. In line with other studies, the group of CI users more often chose the visual component in the case of ambiguous audio-visual input (McGurk trials) compared to the NH controls (Desai et al., 2008; Rouger et al., 2008; Schorr et al., 2005; Tremblay et al., 2009). This indicates that CI users rely more on the visual information in ambiguous situations and not as much on the auditory input, which might be due to the coarse auditory information by the CI. NH individuals in contrast, if they did not perceive the McGurk illusion, more frequently chose the auditory component of the incongruent stimulus.

Neurocognitive studies have shown that the auditory cortex takes over visual functions during the time of auditory deprivation, a pattern of reorganization which has been considered maladaptive for speech recovery (Sandmann et al., 2012). On the other hand, CI users also show stronger activity in the auditory cortex while processing faces, and this effect has been related to better face recognition and lip-reading abilities (Stropahl et al., 2015). Furthermore, a recent study showed cross-modal reorganization of the visual cortex in CI users (Chen, Sandmann, Thorne, Bleichner, & Debener, 2015). The relationship between the reorganization of visual and auditory cortices in response to auditory sensory deprivation may be linked to speech recognition performance (see Stropahl, Chen, & Debener (2016) for a review). It is therefore important to investigate the integration of information from both modalities to further our understanding of neural changes due to auditory deprivation (and subsequent auditory restoration). A recent study comparing normal hearing and early-stage hearing-impaired individuals found further support for cross-modal reorganization (Campbell & Sharma, 2014), raising the interesting question how much sensory deprivation is needed to initialize cortical reorganization. In our view, the individual parameters of the NED model can help to associate altered patterns of audio-visual integration with neural markers. The sensory noise measure in particular might be related to trial-by-trial fluctuations in brain responses (partly reflecting “neural” noise), whereas the disparity threshold might be related to the amount of activity in cortical regions involved in audio-visual integration, such as the superior temporal sulcus (Magnotti & Beauchamp, 2015; Nath & Beauchamp, 2012; Nath, Fava, & Beauchamp, 2011).

## Conclusion

The OLAVS audio-visual stimuli presented here can be used to study the McGurk effect and track differences in audio-visual integration between individuals, such as clinically relevant groups. Stimulus selection based on the audio-visual properties can be done by applying the NED model, which provides a measure of disparity (Magnotti & Beauchamp, 2015). The comparison of the stimulus-independent, individual parameters of the NED model revealed a higher sensory noise and a higher threshold of perceiving the McGurk illusion for the CI users compared with the NH controls. This result provides evidence for stronger audio-visual integration in CI users. The OLAVS stimuli and their characterization by the NED model are freely available.

**Acknowledgments** This research has received funding by the grant I-175-105.3-2012 of the German-Israel Foundation and by the cluster of excellence Hearing4all Oldenburg, Germany.

## References

- Basu Mallick, D., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-015-0817-4
- Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glott International*, 5(9/10), 341–345. Retrieved from <http://dare.uva.nl/record/1/200596>
- Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3), 255–272. doi:10.1016/S0167-6393(96)00063-5
- Campbell, J., & Sharma, A. (2014). Cross-modal re-organization in adults with early stage hearing loss. *PLoS ONE*, 9(2). doi:10.1371/journal.pone.0090594
- Chen, L., Sandmann, P., Thorne, J. D., Bleichner, M. G., & Debener, S. (2015). Cross-modal functional reorganization of visual and auditory cortex in adult cochlear implant users identified with fNIRS. *Neural Plasticity*, 2016. doi:10.1155/2016/4382656
- Colin, C., Radeau, M., & Deltenre, P. (2005). Top-down and bottom-up modulation of audiovisual integration in speech. *European Journal of Cognitive ...* Retrieved from doi:10.1080/09541440440000168
- Demorest, M., & Bernstein, L. (1992). Sources of Variability in Speechreading Sentences: A Generalizability Analysis. *Journal of Speech, Language, and Hearing ...* Retrieved from <http://jslhr.pubs.asha.org/article.aspx?articleid=1779202>
- Desai, S., Stickney, G., & Zeng, F.-G. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *The Journal of the Acoustical Society of America*, 123(1), 428–440. doi:10.1121/1.2816573
- Fixmer, E., & Hawkins, S. (1998). The influence of quality of information on the McGurk effect. Retrieved from [http://www.isca-speech.org/archive\\_open/avsp98/av98\\_027.html](http://www.isca-speech.org/archive_open/avsp98/av98_027.html)
- Grant, K., & Greenberg, S. (2004). Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. *Seminars in Hearing*, 25(3), 241–255. doi:10.1055/s-2004-832858
- Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of*

- Experimental Psychology: Human Perception and Performance*, 37(4), 1193. Retrieved from <http://psycnet.apa.org/journals/xhp/37/4/1193/>
- MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3), 253–257. doi:10.3758/BF03206096
- Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*, 22(3), 701–709. doi:10.3758/s13423-014-0722-2
- Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, 1–6. doi:10.1007/s00221-015-4324-7
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748. doi:10.1038/264746a0
- Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage*, 59(1), 781–787. doi:10.1016/j.neuroimage.2011.07.024
- Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 31(39), 13963–13971. doi:10.1523/JNEUROSCI.2605-11.2011
- Rouger, J., Fraysse, B., Deguine, O., & Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Research*, 1188, 87–99. doi:10.1016/j.brainres.2007.10.049
- Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17), 7295–7300. doi:10.1073/pnas.0609419104
- Sandmann, P., Dillier, N., Eichele, T., Meyer, M., Kegel, A., Pascual-Marqui, R. D., ... Debener, S. (2012). Visual activation of auditory cortex reflects maladaptive plasticity in cochlear implant users. *Brain*, 135(2), 555–568. doi:10.1093/brain/awr329
- Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51), 18748–18750. Retrieved from <http://www.pnas.org/content/102/51/18748.short>
- Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: A heightened McGurk effect in older adults. *Frontiers in Psychology*, 5, 1–12. doi:10.3389/fpsyg.2014.00323
- Stevenson, R. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental ...* Retrieved from <http://psycnet.apa.org/journals/xhp/38/6/1517/>
- Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech ...* Retrieved from <http://jslhr.pubs.asha.org/article.aspx?articleid=1915675&resultclick=3>
- Strelnikov, K., Rouger, J., Barone, P., & Deguine, O. (2009). Role of speechreading in audiovisual interactions during the recovery of speech comprehension in deaf adults with cochlear implants. *Scandinavian Journal of Psychology*, 50(5), 437–444. doi:10.1111/j.1467-9450.2009.00741.x
- Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Démonet, J.-F. J.-F., Déguine, O., & Barone, P. (2015). Increased audiovisual integration in cochlear-implanted deaf patients: Independent components analysis of longitudinal positron emission tomography data. *European Journal of Neuroscience*, 41(5), 677–685. doi:10.1111/ejn.12827
- Stropahl, M., Plotz, K., Schönfeld, R., Lenarz, T., Sandmann, P., Yovel, G., ... Debener, S. (2015). Cross-modal reorganization in cochlear implant users: Auditory cortex contributes to visual face processing. *NeuroImage*, 121, 159–170. doi:10.1016/j.neuroimage.2015.07.062
- Stropahl, M., Chen, L.-C., & Debener, S. (2016). Cortical reorganization in postlingually deaf cochlear implant users: Intra-modal and cross-modal considerations. *Hearing Research*. doi:10.1016/j.heares.2016.07.005
- Tremblay, C., & Champoux, F. (2007). Evidence for a generic process underlying multisensory integration. *The Open Behavioral ...* Retrieved from <http://benthamopen.com/ABSTRACT/TOBSJ-1-1>
- Tremblay, C., Champoux, F., Lepore, F., & Théoret, H. (2009). Audiovisual fusion and cochlear implant proficiency. *Restorative Neurology and Neuroscience*, 28(2), 283–291. Retrieved from <http://europepmc.org/abstract/med/20404415>
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4), 1181–1186. doi:10.1073/pnas.0408949102