Using fNIRS to study audio-visual speech integration in post-lingually deafened cochlear implant users

XIN ZHOU[1,2,*] HAMISH INNES-BROWN[1,2], AND COLETTE MCKAY[1,2]

[1] *Bionics Institute, Melbourne, Australia*

[2] *Medical Bionics Department, University of Melbourne, Melbourne, Australia*

The aim of this experiment was to investigate differences in audio-visual (AV) speech integration between cochlear implant (CI) users and normal hearing (NH) listeners using behavioural and functional near-infrared spectroscopy (fNIRS) measures. Participants were 16 post-lingually deafened adult CI users and 13 age-matched NH listeners. Participants' response accuracy in audio-alone (A), visual-alone (V), and AV modalities were measured with closed-set /aCa/ non-words and with open-set CNC words. AV integration was quantified by using a probability model and a cue integration model that predicted participants' AV performance given minimal or optimal integration. Using fNIRS, brain activation was measured when listening to or watching A, V, or AV speech with or without multi-talker babble. For fNIRS, evidence of AV integration was measured using the *principle of inverse effectiveness (PoIE)* model (comparing the difference in activation in two brain regions between A and AV modalities in quiet and noise conditions). Behavioural AV integration was similar in the two groups for CNC words but poorer in the CI group compared to NH group for consonant perception. Our fNIRS data did not demonstrate any AV integration in either NH listeners or CI users, by testing the PoIE.

INTRODUCTION

Neuroplasticity and changes in speech processing strategies have been reported in cochlear implant (CI) users (see review by Anderson *et al.*, 2016). These changes are thought to be due to hearing loss and increased reliance on lip-reading before implantation, and the introduction of distorted hearing input after cochlear implantation. In this study, the audio-visual (AV) integration ability of CI users was of special interest. Rouger *et al.* (2007) used a cue integration model to quantify AV integration ability and claimed that CI users had better AV integration ability than normal listening (NH) listeners when the latter were listening to vocoded speech. Using electroencephalography (EEG) measures, Schierholz *et al.* (2015) investigated changes in response in auditory cortex of CI users and NH listeners when visual-alone (V) cues were added to audio-alone (A) object stimuli compared to the response in A condition. Changes of response in auditory cortex in that study were interpreted as the amount of AV integration. Compared to the older NH listeners, Schierholz *et al.*

*Corresponding author: xzhou@bionicsinstitute.org

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

(2015) found that older CI users had larger AV integration responses in auditory cortex. Results from the above two studies suggested that CI users may have better AV integration ability and more neural AV integration response than NH listeners.

We investigated whether functional near-infrared spectroscopy (fNIRS) could reveal AV integration in experienced CI users and we hypothesized that CI users have increased AV speech integration compared to age-matched NH listeners, using both behavioural and fNIRS measures. To reveal AV integration in fNIRS measures, we used the principle of inverse effectiveness (PoIE) first found in a study of Meredith and Stein (1983). This rule assumes that when V cues are added to A stimuli, enhancement of neural responses should be greater when the effectiveness of stimuli in each modality is low compared to high. The PoIE was derived using the dynamic response of multisensory neurons to stimuli of different effectiveness levels (Perrault *et al.*, 2005) and has also been applied to in functional magnetic resonance imaging (fMRI) and EEG studies (Holmes, 2007; James *et al.*, 2012). In this study, we investigated two regions of interest (ROIs), i.e., left superior temporal sulcus (LSTS) and left occipital cortex (LOC) where the PoIE has been previously demonstrated in NH listeners using fMRI (Laurienti *et al.*, 2005; Stevenson *et al.*, 2009). Using A, V, and AV speech stimuli, we tested the PoIE of fNIRS responses in the 2 ROIs of NH listeners and CI users, separately. We hypothesised that compared to NH listeners, CI users would show larger fNIRS measures of AV integration activation in at least one of the two ROIs.

METHOD

Participants

Sixteen post-lingually deafened adult CI users and 13 aged-matched NH listeners were recruited for this study. All the participants were native English speakers, with no history of diagnosed neurological disorder, and with normal or corrected-to-normal vision. All CI users had a right-ear implant and experience of using the CI for more than 12 months. The ages of participants in the CI and old NH group ranged from 45 to 82 (mean ± SD: 69.0 ± 9.1) and 52 to 76 years (mean ± SD: 64.9 ± 7.1), respectively, with no significant mean difference in age (t = 1.38, p = 0.179). To develop the cue integration model for AV speech integration, an additional 16 young NH listeners were also recruited, with ages ranging from 21 to 39 years (mean ± SD: 28.7 ± 5.3). All participants provided their written informed consent.

Speech stimuli

Two types of speech stimuli were used to measure AV integration ability. The first type were 12 consonant tokens in the form of /aCa/, with the 12 consonants being 'B', 'D', 'F', 'G', 'K', 'M', 'N', 'P', 'S', 'T', 'V', 'Z'. The second type were Consonant-Nucleus-Consonant (CNC) words (Peterson and Lehiste, 1962). For all the consonant and CNC word stimuli, the A and V components of video recordings were separated. The levels of all the auditory consonant/CNC stimuli were normalized to the same root mean square (RMS) level.

Speech tests and AV integration ability

Speech tests were conducted in A, V, and AV modalities, using software Max/Msp (https://cycling74.com). Visual stimuli were presented on an LCD monitor at a 1.5-m distance and in front of the participant. Auditory stimuli were delivered to the right-ear processor of CI users via direct audio input accessory or the right-side insert earphone of NH listeners. The level of sound directly input to the CI processor or earphone was set equivalent to 65 dBA (F_{max}). Speech sounds in the A and AV modalities were presented with babble noise at a participant-dependent signal-to-noise ratio (SNR), at which each participant could achieve 50% of the consonants or 50% of the phonemes in the CNC words correct in the A condition (denoted SNR50%). For each individual participant, SNR50% was first determined using an adaptive procedure. During the consonant discrimination task, 12 consonants in the same modality were presented sequentially in a pseudo-random order with four repeats. In total, 48 consonants in A, V, and AV modality were presented. Participants responded using a touch-screen with 12 buttons corresponding to the 12 consonants. No feedback about response accuracy was provided. For the CNC word identification task, 60 different CNC words in each modality were presented in a pseudo-random order in blocks of 20 stimuli. Participants were required to verbally repeat back the word they recognised each time. For both types of speech stimuli, the order of A, V, and AV modalities was randomly chosen.

AV integration for each participant was quantified using a probability model (Blamey *et al.*, 1989) and a cue integration model (Rouger *et al.*, 2007). The probability model estimates participants' AV performance P_{AV}^{est} when auditory and visual speech processing are independent, i.e., *minimum integration* happens (Eq. 1), where, P_A and P_V are response accuracies in A and V, respectively.

$$P_{AV}^{est} = P_A + P_V - P_A * P_V \quad \text{(Eq. 1)}$$

The cue integration model predicts AV performance when *optimal cue integration* happens between the two modalities. The cue integration model assumes that to be able to understand speech information, we need to recognize at least a certain number (T) of cues correctly. Further, our perception of the cues has a Poisson distribution (Eq. 1), where λ is the average number of cues that we recognise.

$$P(n > T) = \Sigma_{k=n} \left(\lambda^k e^{-\lambda}\right)/k! \quad \text{(Eq. 2)}$$

Threshold T depends on the type of speech stimuli, regardless of modality. When optimal integration happens, the number of cues recognised in the AV modality (λ_{AV}) equals the sum of those recognised in A (λ_A) and V (λ_V) modalities, i.e. $\lambda_{AV} = \lambda_{\mathrm{A}} + \lambda_{\mathrm{V}}$. Based on participants' performance in A (P_A) and V (P_V) modalities, λ_{A} and λ_{V} can be estimated using Eq. 2.

To apply the cue integration model, we tested a group of young NH listeners to obtain the stimulus-dependent T values which best fit the data for young NH listeners' AV performance, i.e. $P_{AV}^{est} = P_{\mathrm{AV}}$, with $\lambda_{AV}^{est} = \lambda_{\mathrm{A}} + \lambda_{\mathrm{V}}$. We then applied these T values to

old NH listeners and CI users to assess whether the older NH listeners or CI users had better or worse AV integration than the young NH listeners.

fNIRS imaging

Data collection

In this study, a continuous-wave fNIRS device (NIRScout, NIRX medical technologies, LLC) with 16 LED illumination sources and 16 photodiode detectors was used. fNIRS measures the concentration changes of oxygenated (HbO) and deoxygenated (HbR) haemoglobin in the blood. To (partly) remove the signals recorded from extracerebral tissue, two 1.3-cm 'short' channels that were located in the anterior temporal cortex of each side were used. For fNIRS imaging, data were recorded from the two ROIs, i.e. LSTS and LOC, as shown in Fig. 1.

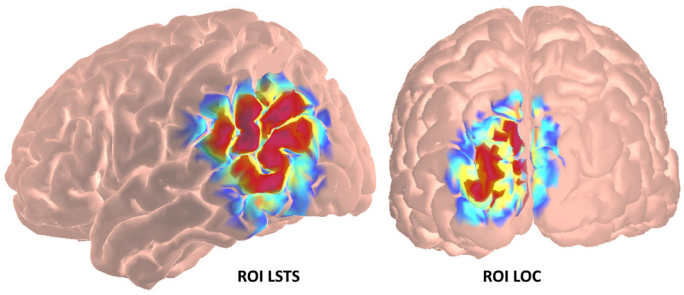

Fig. 1: ROIs where fNIRS responses were measured, i.e., LSTS and LOC.

A block-design was used for fNIRS data collection, with the length of a stimulus block being 14.5 s. Each stimulus block was preceded and followed by a 25-s white fixation cross on the black screen of the CRT monitor. To ensure participants remained focused on the experiment, they were asked to perform a recognition task at the end of each block. Seven blocks of stimuli in each modality were presented.

Six testing periods of fNIRS data were collected, with the first three testing periods using consonant stimuli, and the second three using CNC word stimuli. For each type of speech stimuli, the first testing period used blocks of A and AV stimuli in quiet, and the second testing period used blocks of A and AV stimuli with babble noise. When A stimuli were presented, there was a static picture of the female speaker on the monitor. For these two testing periods, 7 blocks of A and AV stimuli were played in pseudo-random order. The SNR of the babble noise was presented at participant-dependent levels (SNR50%) previously determined for behavioural speech tests. In the third testing period, 7 blocks of stimuli in V modality were presented, with no auditory input through the earphone or CI processor. The recording of response in V modality is supplementary, to check that responses in the ROIs in two A and V modalities correlate with responses in AV modality.

Data analysis

fNIRS data analysis consisted of signal pre-processing and signal processing. Signal pre-processing included 1) identifying and removing step-like artefacts that were caused by sudden loss of contact between optodes and skin, 2) excluding channels that had poor data quality, 3) estimation of haemodynamic response and band-pass filtering to remove environmental noises. Short-channel-separation was further conducted to remove the extracerebral response from the long channels within the ROIs. This was done by first extracting the first principal component (PC_1) of HbO or HbR from the 2 short channels by using principal component analysis (PCA). Channels with short distance were assumed to only measure responses from the extracerebral tissues, and PC_1 was assumed to be the systemic response that would exist globally. A general linear model (GLM) was then used, as shown in Eq. 3, to remove the PC_1 signal from the response in each long channel (Y_L). Within Eq. 3, HRF was the experimental specific haemodynamic response function model for different types of stimuli. β was the coefficient of HRF and α was the coefficient of PC_1 estimated from GLM; ε was the residual noise.

$$Y_L = [HRF_1, HRF_2] * [\beta_1, \beta_2]' + \alpha * PC_1 + \varepsilon \quad \text{(Eq. 3)}$$

After short-channel-separation in the long channels, the averaged hemodynamic response across the 7 blocks was then estimated for stimuli of each modality. Outlier blocks of response were excluded. Only the HbO response were used for further statistical analysis. To test our hypothesis that the fNIRS data in old NH listeners would show the PoIE, the inequality in Eq. 4 was used. The left and right sides of Eq. 4, represent the differences between HbO responses in the AV and A modalities when the auditory background was quiet (Q) and with noise (N), respectively.

$$(A_Q V - A_Q) < (A_N V - A_N) \quad \text{(Eq. 4)}$$

RESULTS

AV integration: Behavioural performance

Figure 2 plots the speech test results in three modalities for young, old NH listeners, and CI users when responding to consonants (first row) and CNC words (second row). Black dashed lines and magenta dash-dot lines plot the probability model and the cue integration model predicted AV performance, respectively, in each group. For the cue integration model, the stimulus-dependent T thresholds of 1 and 3 for consonant and CNC word stimuli, respectively, which were obtained based on the best fit for young NH listeners' performance, were applied to old NH listeners and CI users. Figure 2 shows that when responding to consonant stimuli (first row), the cue integration model (magenta dash-dot line), fits old NH listeners' AV performance (red dots) well but CI users' performance was lower than predicted by the young NH based model. In contrast, the probability model (black dash line) fits CI users' performance (red dots) well, i.e., CI users showed essentially independent use of A and V cues in AV mode. These results showed that when responding to consonant stimuli, old NH listeners had comparable AV integration with young NH listeners (optimal cue integration), while

our experienced CI users had less AV integration ability than NH listeners. When responding to CNC word stimuli, as shown in Fig. 2 (second row), the cue integration model (magenta dash-dot line) fits the AV performance (red dots) of both CI users and old NH listeners well, i.e., both old NH listeners and CI users had optimal integration compared to young NH listeners.

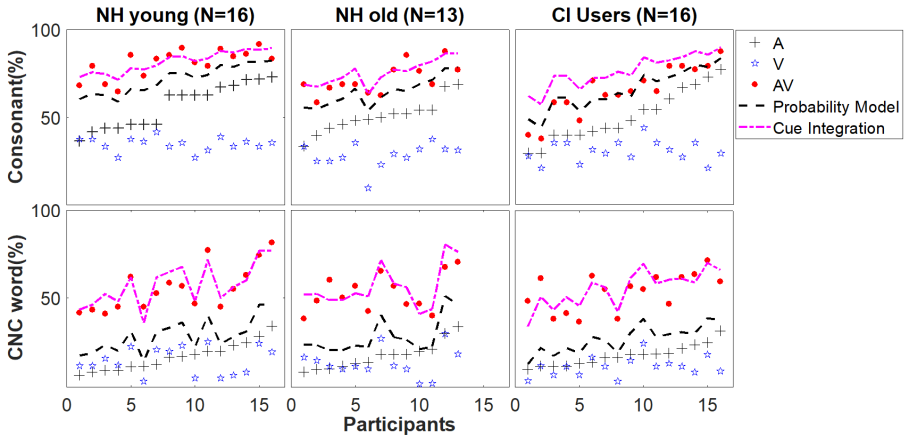

Fig. 2: Audio-visual (AV) speech perception of consonants and CNC words in NH listeners and CI users.

AV integration: fNIRS imaging

Figure 3 shows the fNIRS response in ROI LOC of age-matched NH listeners (first row) and CI users (second row) when responding to consonant stimuli. Red lines and shaded areas plot the mean and standard error of mean (SEM) of HbO; blue plots HbR response. Vertical dashed lines indicate the stimulus onset and offset. From left to right, each column plots $(\mathrm{A_Q V} - \mathrm{A_Q})$, $(\mathrm{A_N V} - \mathrm{A_N})$, and $(\mathrm{A_N V} - \mathrm{A_N}) - (\mathrm{A_Q V} - \mathrm{A_Q})$ measures, respectively. A pairwise running one-tailed t-test was performed on the HbO response between quiet and noisy conditions, using Eq. 4. Permutation *t*-tests (Groppe *et al.*, 2011) were done to control familywise error rate for multiple comparisons. No significantly larger response was found in the noisy condition than in quiet, i.e., no occurrence of the PoIE in the fNIRS responses in ROI LOC, in either CI users or NH listeners. The same statistical analysis was done for responses in two ROIs and to two types of speech stimuli. The PoIE of AV integration was not significantly demonstrated for NH listeners or CI users for either speech stimulus type, in either of the ROIs.

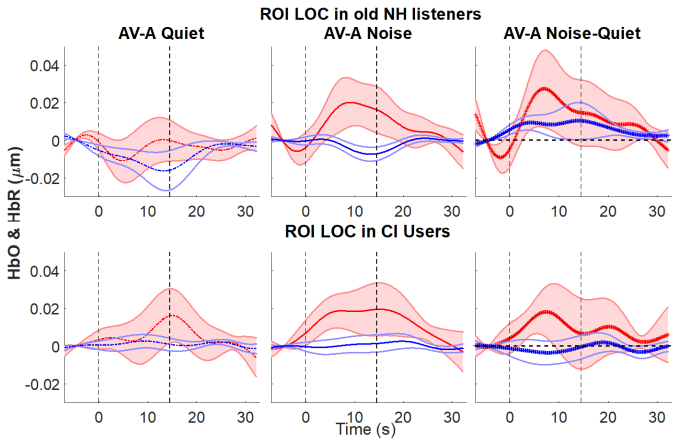

Fig. 3 fNIRS response of the old NH listeners and CI users in the ROI LOC when responding to consonant stimuli.

DISCUSSION AND CONCLUSION

This study examined AV speech integration in CI users and old NH listeners using behavioural and fNIRS measures. Using behavioural measures, CI users had poorer AV integration compared to old NH listeners when responding to consonant stimuli, but had comparable AV integration ability when responding to CNC word stimuli. For fNIRS imaging, no PoIE of AV integration was observed in either of the two ROIs for either CI users or age-matched NH listeners.

Our behavioural results that CI users had comparable or poorer AV speech integration ability than NH listeners could be because, first, they were CI users who have years of experience of using their implant and no longer relied on lip-reading for speech perception. Thus, these CI users showed no super-normal lip-reading ability or AV integration ability than NH listeners. Further, when responding to consonant stimuli, CI users' performance in AV modality was mainly dependent on their performance in A modality. As shown by the cue integration model that, participants only needed to recognise more than one cue from the consonant stimuli to make a correct response. When responding to consonant in AV modality, CI users selectively attended to A cues and ignored V cues. This selective attention maladaptively affected their AV integration.

Our fNIRS results that no PoIE being observed in either group could be because, first, large variance of response existed in each group, which derived from both experimental measures and individual's difference in fNIRS response. As to reveal this inverse effectiveness of AV integration, fNIRS measures were estimated from responses recorded in four different conditions, resulting in too much noise in the data. Also, largely variant AV integration responses have been reported in old NH listeners, due to their wider AV integration window (Diederich *et al.*, 2008). Further, because

of the limited spatial resolution of fNIRS compared to that of fMRI, the ROIs in this study were larger and less focussed than those in the fMRI studies that showed the PoIE. All these reasons make it challenging to reveal the PoIE of AV integration in our old NH listeners and CI users.

REFERENCES

Anderson, C.A., Lazard, D.S., and Hartley, D.E.H. (**2016**). "Plasticity in bilateral superior temporal cortex: effects of deafness and cochlear implantation on auditory and visual speech processing," Hear. Res., **343**, 138-149.

Blamey, P.J., Cowan, R.S., *et al.* (**1989**). "Speech perception using combinations of auditory, visual, and tactile information." J. Rehabil. Res. Dev., **26**, 15-24.

Diederich, A., Colonius, H., *et al.* (2008). "Assessing age-related multisensory enhancement with the time-window-of-integration model," Neuropsychologia, **46**, 2556-2562.

Groppe, D.M., Urbach, T.P., *et al.* (**2011**). "Mass univariate analysis of eventrelated brain potentials/fields I: A critical tutorial review," Psychophysiology, **48**, 1711-1725.

Holmes, N.P. (**2007**). "The law of inverse effectiveness in neurons and behaviour: multisensory integration versus normal variability," Neuropsychologia, **45**, 3340-3345.

James, T.W., Stevenson, R.A., *et al.* (**2012**). "Inverse effectiveness and BOLD fMRI," *The New Handbook of Multisensory Processes* (Stein BE, ed.), pp. 207-222.

Laurienti, P.J., Perrault, T.J., *et al.* (**2005**). "On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies," Exp. Brain Res., **166**, 289-297. doi: 10.1007/s00221-005-2370-2

Meredith, M.A., and Stein, B.E. (**1983**). "Interactions among converging sensory inputs in the superior colliculus," Science, **221**, 389-391.

Perrault, T.J., Vaughan, J.W., *et al.* (**2005**). "Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli," J. Neurophysiol., **93**, 2575-2586.

Peterson, G.E., and Lehiste, I. (**1962**). "Revised CNC lists for auditory tests," J Speech Hear. Disord., **27**, 62-70.

Rouger, J., Lagleyre, S., *et al.* (**2007**). "Evidence that cochlear-implanted deaf patients are better multisensory integrators," Proc. Natl. Acad. Sci. USA, **104**, 7295-7300. doi: 10.1073/pnas.0609419104

Schierholz, I., Finke, M., *et al.* (**2015**). "Enhanced audio–visual interactions in the auditory cortex of elderly cochlear-implant users," Hear. Res., **328**, 133-147.

Stevenson, R.A., and James, T.W. (**2009**). "Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition," NeuroImage, 44, 1210-1223. doi: 10.1016/j.neuroimage.2008.09.034