AUDIOVISUAL SPEECH PERCEPTION IN NORMAL-HEARING INDIVIDUALS AND COCHLEAR IMPLANT USERS

Luuk Pieter Harrie van de Rijt

2021

Audiovisual speech perception in normal-hearing individuals and cochlear implant users

ISBN: 978-94-6421-274-7

Layout by: Bregje Jaspers | ProefschriftOntwerp.nl
Printed by: Ipskamp Printing, Enschede

This research was supported by EU FP7-PEOPLE-2013-ITN iCARE (grant 407139, A. Roye), EU Horizon 2020 ERC Advanced Grant ORIENT (grant 693400, A.J. van Opstal), Cochlear Benelux NV (L.P.H. van de Rijt, M.M. van Wanrooij), the Radboud University Medical Center (L.P.H. van de Rijt, E.A.M. Mylanus), and Radboud University (M.M. van Wanrooij).

Financial support for the publication of this thesis was provided by: ALK-Abelló B.V., Advanced Bionics, Beter Horen, Cochlear, Daleco Pharma B.V., EmiD B.V., GN Hearing Benelux, Goedegebuure Slaaptechniek B.V., MediTop Medical Products, Phonak.

Audiovisual speech perception in normal-hearing individuals and cochlear implant users

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op vrijdag 11 juni 2021
om 10.30 uur precies
door

Luuk Pieter Harrie van de Rijt

geboren op 17 juni 1989

te Nijmegen

Promotoren:
Prof. dr. A. John van Opstal
Prof. dr. Emmanuel A.M. Mylanus (Universiteit Gent, België)

Copromotor:
Dr. Marc M. van Wanrooij

Manuscriptcommissie:
Prof. dr. Raymond van Ee
Prof. dr. Pim van Dijk
Dr. Bas van Dijk

Audiovisual speech perception in normal-hearing individuals and cochlear implant users

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Friday, June 11, 2021
at 10.30 hours
by

Luuk Pieter Harrie van de Rijt

born on June 17, 1989

in Nijmegen (the Netherlands)

Supervisors:
Prof. dr. A. John van Opstal
Prof. dr. Emmanuel A.M. Mylanus (Ghent University, Belgium)

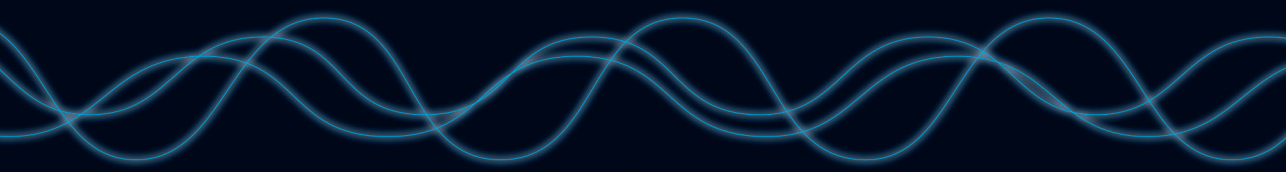
Co-supervisor:
Dr. Marc M. van Wanrooij

Doctoral Thesis Committee:
Prof. dr. Raymond van Ee
Prof. dr. Pim van Dijk (University of Groningen)
Dr. Bas van Dijk

TABLE OF CONTENTS

CHAPTER 1

Introduction



SCOPE

Over the past 30 years, clinical research in hearing rehabilitation yielded substantial hearing benefits to profoundly deaf individuals by focusing on the acoustic efficacy of cochlear implants (CIs). The most relevant benefit of cochlear implantation is recovery of auditory speech understanding. Nevertheless, speech understanding of CI users remains highly perturbed in noisy environments. Under these demanding situations, non-acoustic sources of information on speech, such as lipreading, might help. The aim of this thesis is to investigate how CI users incorporate both their listening and lipreading abilities when trying to understand speech in noisy environments.

This introduction explains first how the healthy hearing system works in understanding speech with a detailed description of the anatomy and physiology of the healthy hearing organ. Subsequently, we describe how, in case of severe deafness, the cochlear implant might partially restore hearing. This neuroprosthesis allows most deaf individuals to accurately perceive speech in quiet environments. However, listening in more challenging, everyday situations (e.g., in traffic, at a party) remains problematic for CI users. After this general problem statement, the benefit of lipreading in improving speech understanding both for hearing-impaired and for normal-hearing listeners in noisy environments is elucidated. Therefore, we will explain how lipreading could potentially provide an additional useful stream of information. Lipreading can be integrated with the - degraded - acoustic information to improve speech understanding. This is the main topic of this thesis. Through behavioral experiments, we will demonstrate how visual information is incorporated by normal-hearing listeners (Chapter 2) and CI users (Chapter 3) in speech perception. Finally, through the use of the non-invasive neuroimaging technique called functional near-infrared spectroscopy (fNIRS; further elaborated in Chapter 4), we studied neural correlates of audiovisual speech perception in CI users and in normal-hearing listeners (Chapter 5),

HEARING – AUDITORY SENSORY SYSTEM

Perceiving and being able to produce sounds provides the opportunity to communicate efficiently and unambiguously. Speech is a complex stream of sound; the acoustics vary in temporal, spectral and intensity aspects.

Sound is characterised as a pressure wave that is produced by a mechanically vibrating source. It propagates through the air at a speed of 343 m/s. The main characteristics of a sound pressure wave are described by its amplitude and its spectrum as a function of time. The amplitude of the pressure wave is measured in pascal (in Pascal; Pa, or N/m^2), and can be

expressed on a logarithmic scale in decibels (dB). The human ear can perceive sounds from extremely low sound pressure levels (SPLs) of approximately 0 dB SPL (reference: 20 µPa at 1000 Hz); up to high SPLs of approximately 130 dB SPL (±60 Pa), which span a huge dynamic range of nearly 13 orders of magnitude.

The ear transforms these pressure waves into neural signals (Fig. 1.1).

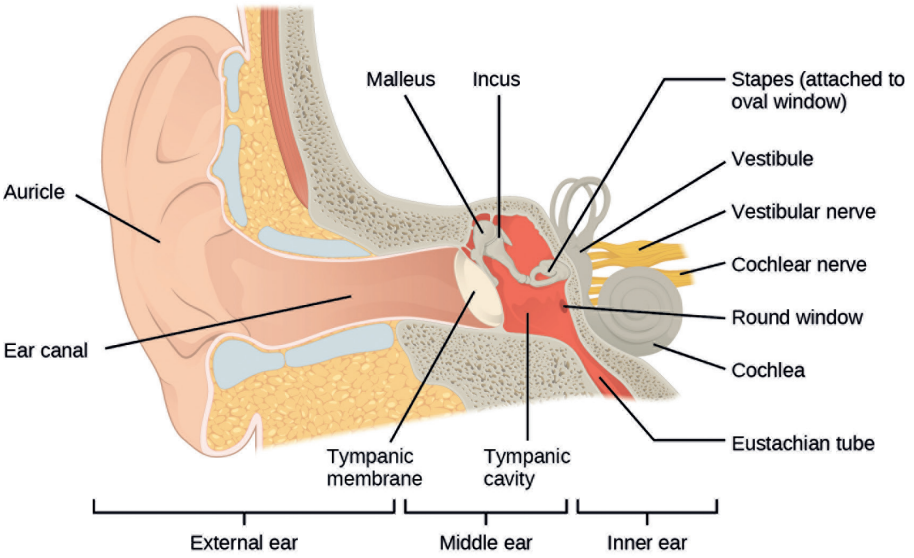


Figure 1.1. Anatomical parts of the external, middle and inner ear.*1 (Figure used with permission from source).

The external ear comprises the auricle (pinna) and the external auditory canal. Both structures conduct the sound to the tympanic membrane, which is the first structure of the middle ear. The sound pressure changes cause the tympanic membrane to vibrate. These vibrations are transmitted by the ossicular chain of the middle ear. The ossicular chain consists of three ossicles: the malleus, the incus and the stapes, which are serially located between the tympanic membrane and the cochlea. The function of the middle ear is to effectively transmit the energy of the airborne sound waves (low acoustic impedance) into fluid borne waves (high impedance) in the cochlea (impedance matching). The stapes is connected with the cochlear oval window by its footplate.

From the stapes footplate, the longitudinal acoustic waves are transmitted into transversal fluid motions in the cochlea (Fig. 1.2). In Greek, the word cochlea means 'snail', and refers to its snail-

*1 Mary Ann Clark, Matthew Douglas JC. *Biology 2e*. Houston, Texas: OpenStax; 2018. https://openstax.org/books/biology-2e/pages/36-4-hearing-and-vestibular-sensation licensed under Creative Commons Attribution 4.0 International License (CC BY 4.0)

shell like shape. This shell is subdivided into three fluid-filled compartments: scala vestibuli, scala media and scala tympani. The fluids of the scala vestibuli and scala tympani meet at the apex of the cochlea through the so-called helicotrema. The scala media is separated from the scala vestibuli by Reissner's membrane and from the scala tympani by the basilar membrane.

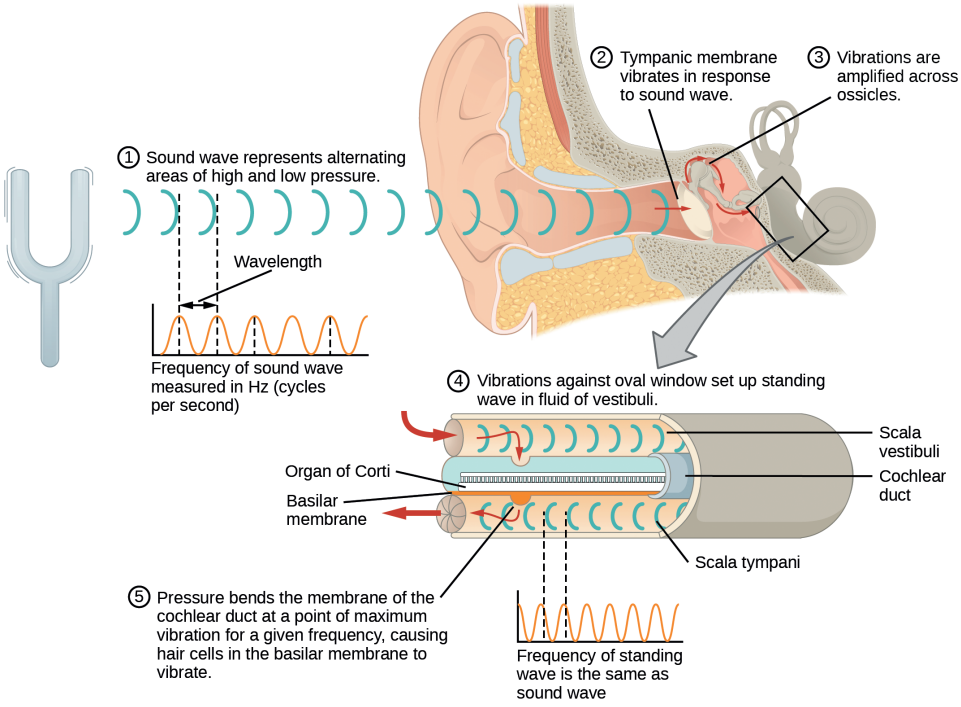


Figure 1.2. Schematic illustration of a sound wave travelling through the human ear.[*2] (Figure used with permission from source).

The acoustic pressure wave, induced by the vibrating stapes footplate at the oval window, is transformed into a dynamic pressure difference across the basilar membrane. This transversal pressure difference, which varies with the same frequency as the harmonic input at the stapes, induces 'resonances' of the elastic basilar membrane. Due to the variation of elastic (mechanical) characteristics of the basilar membrane along its entire length, Georg Von Békésy[*3] showed that different frequencies yield a maximal acoustic response of the basilar membrane (resonance) at different, frequency-related locations. This property of the cochlea to process sounds of specific frequencies at specific positions along the basilar membrane is called tonotopy.

[*2] Mary Ann Clark, Matthew Douglas JC. *Biology 2e*. Houston, Texas: OpenStax; 2018. https://openstax.org/books/biology-2e/pages/36-4-hearing-and-vestibular-sensation.

[*3] Géorg von Békésy (1899-1972) was a Hungarian biophysicist who was awarded the Nobel prize in Physiology or Medicine (1961) for his groundbreaking work on cochlear mechanics.

The scala media encloses the organ of Corti[*4], which rests on the basilar membrane. It contains two types of sensory hair cells (cells with stereocilia) that respond to the basilar membrane motions in essentially different ways. When stimulated, the stereocilia of the outer hair cells (OHC) deflect, providing locally a positive feedback to the basilar membrane motion (Fig. 1.3). That results in a much sharper and also (compressive) nonlinear tuning characteristic of the basilar membrane, and hence recruitment of the auditory nerve. The OHC function is under 'sound-intensity related feedback control' from the central nervous system: at higher sound intensities the OHCs become more and more unresponsive and the basilar membrane response more and more linear. This mechanism accounts for the wide dynamic range of the human auditory system, and its high (down to 0.1%) frequency selectivity.

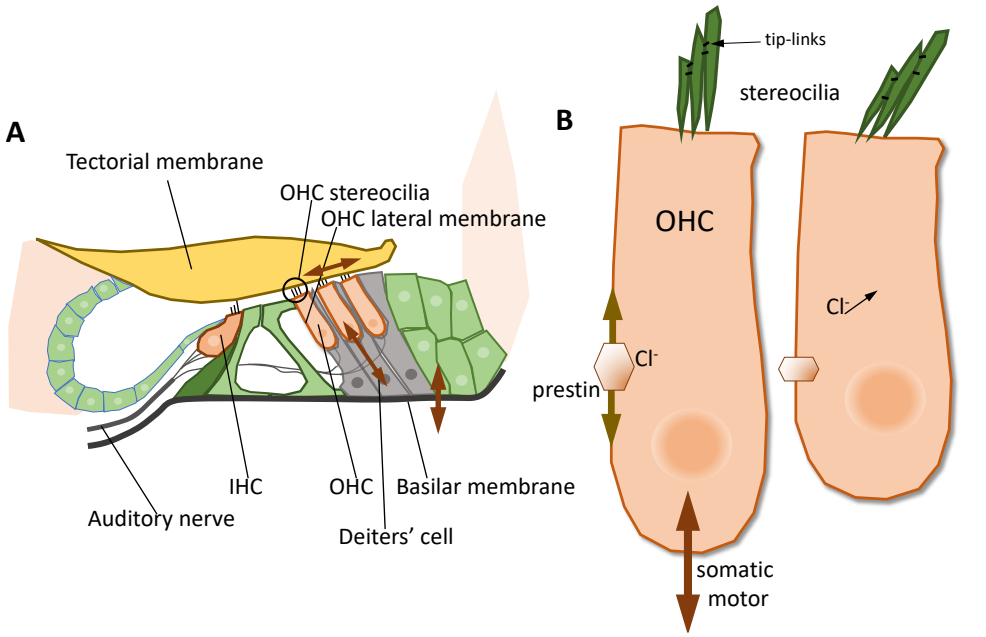


Figure 1.3. Schematic overview of the cochlea.

The outer hair cells act as a nonlinear intensity-dependent local mechanical amplifier, and respond with a rapid mechanical shortening-lengthening of its cell-body in phase with the sound's frequency when the stereocilia are deflected with respect to the tectorial membrane (figure used with permission from author)[*5].

In order to effectively convert the sound pressure waves into neural signals, a transduction process takes place in the cochlea. The mechanical response of the stereocilia of the inner hair cells convert their movements into an electric potential of the inner hair cell membrane, such that when there is movement of the basilar membrane relative to the tectorial membrane

[*4] The organ of Corti was named after its discoverer the Italian anatomist Alfonso Corti (1822-1876) (see: Leonhardt 1999, 217).
[*5] J. van Opstal (2016). *The Auditory System and Human Sound-Localisation Behavior. Elsevier, Academic Press*

they depolarize. This depolarization is synaptically transmitted as a depolarization of the spiral ganglion cells, situated in Rosenthal's canal in the modiolus of the cochlea, which contain the cell bodies of the primary auditory neurons (part of the auditory nerve). From here the action potentials travel through cochlear nerve to the cochlear nucleus in the brainstem, from where they are transmitted to other stages in auditory pathway, from auditory brainstem (the superior olivary complex, lateral lemniscus) and midbrain (inferior colliculus) to (sub) cortical auditory areas (medial geniculate body, and primary and higher auditory cortex).

HEARING - IMPAIRMENT AND REHABILITATION

A hearing impairment is a deficit in the auditory system. The most common type of hearing loss is a sensorineural hearing impairment, which involves a defect in either the cochlea (which are numerous, for example at the level of the hair cells, or gene mutations of cochlear proteins), the auditory nerve or the associated neural structures. Due to this defect, the neural correlates of speech is suboptimal and not properly processed. Hearing impairment[*6] can be classified in four categories; slight (26-40 dB HL), moderate (41-60 dB HL), severe (61-80 dB HL), and profound (>81 dB HL). Deafness refers to severe-to-profound sensorineural hearing loss. Deafness is further subdivided in either congenital, prelingual, or postlingual deafness, depending on the onset of deafness either present at birth, before language development or after language development, respectively.

If the hearing loss is severe-to-profound, a conventional hearing aid (behind-the-ear device) might provide insufficient benefit. Patients with severe-to-profound hearing loss may be rehabilitated with another type of hearing device, namely a CI (Fig. 1.4). A CI is an implanted device that directly stimulates auditory nerve fibers by an electric current. A CI comprises 2 parts; an externally worn sound processor (Fig. 1.4A), that is transcutaneous coupled to an implanted receiver with an electrode array (Fig. 1.4A). The array of electrodes is surgically positioned preferably in the scala tympani. The sound processor decodes the received sounds, which are processed and transmitted to the electrode array. The local currents produced by the individual electrodes of that array stimulate the (remaining) nerve endings, bypassing the damaged or missing sensory hair cells, in such a way that the tonotopical organization of the cochlea is mimicked (Fig. 1.4B).

[*6] According to the World Health Organization, based on the Pure Tone Averages (PTA) threshold over 0.5, 1, 2 and 4 kHz.

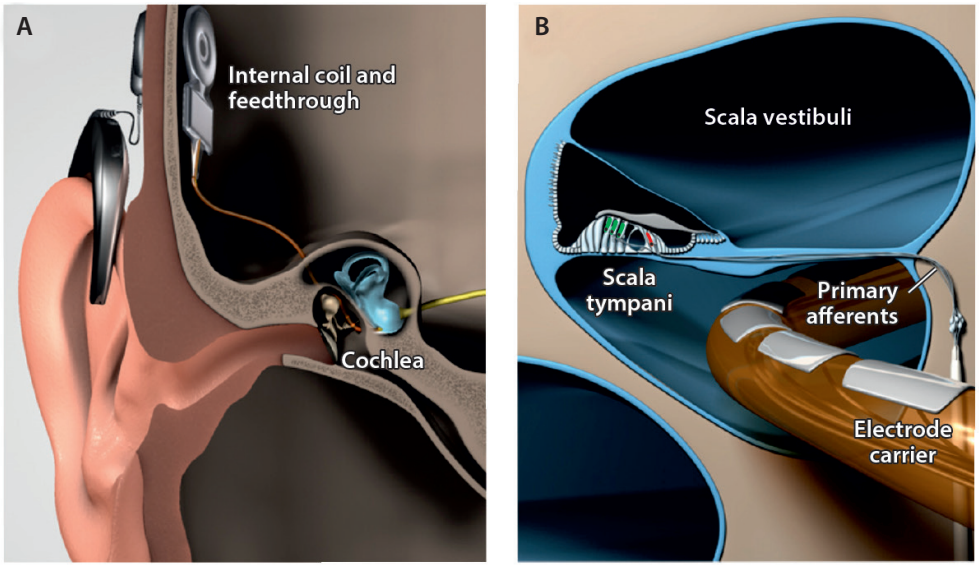


Figure 1.4. Cochlear implant.
A) internal and external parts of a CI. **B)** the inserted electrode array (illustration used with permission from the author)[1]

CIs allow proper speech perception for many recipients.[2,3] Over the last decades, the number of implanted subjects per year has increased steadily owing to relaxation of the criteria for implant candidacy, following improved outcomes as a result of technological advancements.[4] Multiple studies demonstrate improvements in speech perception and, related to that, better quality of life.[5,6] Nowadays, cochlear implantation is considered as the standard of care for severe-to-profound sensorineural hearing loss, in adults and children. There is hardly any dispute about the efficacy of cochlear implantation. However, variation in how well words are recognized across CI-users is obvious.[2,3,6,7] As an example, in a word identification task the entire range of test scores (0-100%) has been reported.[6]

Factors that might affect auditory outcome performance of CI wearers are numerous, but they can be broadly divided in three categories. One category consists of the clinical characteristics of these individuals, such as residual hearing, consistent use of hearing aids before implantation, and duration of the severe-to-profound hearing loss.[3,4,7–11] Another category of factors that are thought to influence the CI-outcome are device-related such as number of active electrodes, electrode design, signal processing and coding strategies. A third category consists of surgery-related factors; such as depth of electrode insertion and positioning of the electrode array in the cochlea with respect to the distance to the modiolar wall and neuronal structures.[12]

Despite the heterogeneity of all these factors, most CI recipients benefit substantially from cochlear implantation. However, the CI bypasses a sophisticated sound processing system that normally yields a tremendous benefit to the dynamic range of hearing and frequency tuning of the auditory system. Because of that, the CI provides limited information compared to our normal-hearing ear for various reasons. In essence, the spectral-temporal resolution of an acoustic signal processed by a CI is tremendously degraded compared to normal hearing, and that is primarily due to the following:

(i) a small number of electrodes (max. of about 22 channels, compared to approximately 3000 frequency channels in a normal-hearing ear) limits the spectral resolution
(ii) only the envelope of the acoustic signal is transmitted by the coding strategy of most conventional CIs, which means that the temporal fine-structure information is lost.

Therefore, unless the CI processor can somehow deal with the loss of information, CI users still have difficulty in understanding speech. So, while a CI enables acoustic processing, the quality (i.e., spectral-temporal resolution) is limited. It is therefore remarkable that a CI still provides proper speech intelligibility in quiet, since it entirely bypasses the intricate mechanisms of the sensory hair cells. In noisy environments, however, CI users might need additional sources of information, if available. Under such demanding acoustic conditions, they could incorporate lipreading information to improve speech understanding. It has been shown that visual speech stimuli can be (even) more informative than auditory speech stimuli under difficult listening conditions.[13] CI-users might thus benefit from the integration of both the auditory and visual streams simultaneously in order to comprehend speech under noisy conditions.

MULTISENSORY INTEGRATION – AUDIOVISUAL SPEECH PERCEPTION

The perception of someone talking typically offers congruent visual and acoustic inputs. In various everyday situations one can obtain reliable visual information from looking at the speaker (such as conversations in a noisy situation, or watching television). Vision thus carries substantial, linguistically relevant cues about the acoustic signal.

In quiet listening conditions, the acoustic speech signal provides sufficient information by itself for proper speech perception for normal-hearing individuals, with lipreading providing little to no additional benefit. However, in complex listening situations (e.g., in a busy restaurant), the acoustic information might be severely perturbed, and now a matching visual information stream provided by lipreading may significantly improve speech understanding. In fact, speech perception improves considerably when words in noise are presented synchronously with congruent lip movements.[14,15]

Research with deaf individuals has also demonstrated an added benefit of lipreading.[16] For deaf individuals, the auditory input is severely degraded. They often need to rely on other sources of information, such as visual information, to understand speech. Individuals with acquired deafness adapt to this "deaf situation" and might develop/improve lipreading abilities.[13,17,18] After cochlear implantation, they still have difficulty to understand auditory speech in noisy environments, so they need extra sources of information to understand speech properly. In this thesis we studied this audiovisual integration. In particular, how normal-hearing individuals and CI-users integrate auditory and visual information into a coherent percept.

STATISTICAL FACILITATION

To quantitatively compare how individuals integrate audiovisual cues, several models have been proposed. In these models, the audiovisual condition is quantitatively compared with a prediction in which the audiovisual response does not result from a neural integration of auditory and visual information streams, but merely results from a statistical summation effect. Finding improved performance (i.e. better speech recognition) in an audiovisual condition does not necessarily mean that the brain has integrated the auditory and visual inputs. Indeed, having both modalities available already increases the probability of stimulus recognition (Fig. 1.5).

In this latter, so-called statistical facilitation model, the recognition of a word in an audiovisual trial follows from either the auditory or the visual information source, which are considered to form independent, parallel processing channels. In that case, the probability of word recognition is given by the sum of the auditory and visual recognition probabilities, which is in general determined by:

$$P_{sum} = 1 - P_{fail} = P_A + P_V - P_A \times P_V \tag{1.1}$$

where P_{sum} is the probability to successfully recognize a word according to the summation model, P_A is the probability to recognize a word in the auditory-only condition, and P_V is the probability of recognizing a word in the visual-only condition. The subtraction term $-P_A \times P_V$ accounts for double counting of the probability overlap (see Fig. 1.5).

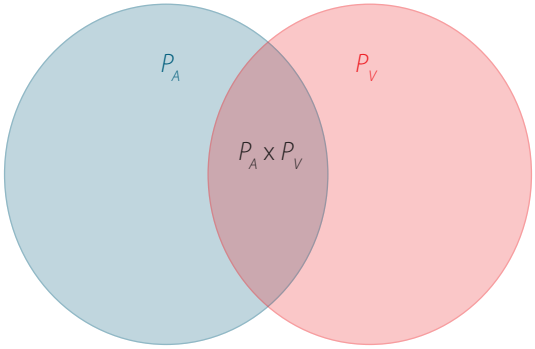


Figure 1.5. Probability summation model displayed in a Venn diagram.

The blue circle (P_A) illustrates the probability to recognize a word in the auditory-only condition, the red circle (P_V) illustrates the probability to recognize a word in the visual-only condition. All possible events are separately characterized with respect to A and V. The overlap between P_A and P_V (which is the same as the overlap between P_V and P_A) is removed once from the sum of P_A and P_V, to yield P_{sum}. It is clear that the total probability of stimulus recognition is increased with respect to the unimodal probabilities according to this model, as long as the overlap is not complete.

True multisensory integration has been shown to depend on three sensory requirements: (i) spatial alignment of the different sensory sources: stimuli are more likely integrated when they come from the same location, (ii) temporal alignment: stimuli are more likely integrated when they occur simultaneously, and (ii) inverse effectiveness: multisensory integration is strongest when the two (aligned) modalities are weak.[19]

The principle of inverse effectiveness in multisensory integration indicates that, as the responsiveness to individual sensory stimuli decreases, the strength of multisensory integration increases.[20] Inverse effectiveness intuitively makes sense: highly salient individual cues will be more easily detected and localized (Fig. 1.6 – *situation 3*). Thus, the multisensory combination has a proportionally modest effect on the total neural activity and behavioral performance due to ceiling effects. By contrast, weak unisensory cues evoke fairly few neural impulses and their responses are potentially subject to substantial enhancement when the stimuli are combined (Fig. 1.6 – *situation 1*). In these cases the multisensory response can exceed the sum of their individual responses (in that case there is true multisensory integration) and can have a significant positive effect on behavioral performance by increasing the speed and likelihood of detecting and localizing an event.

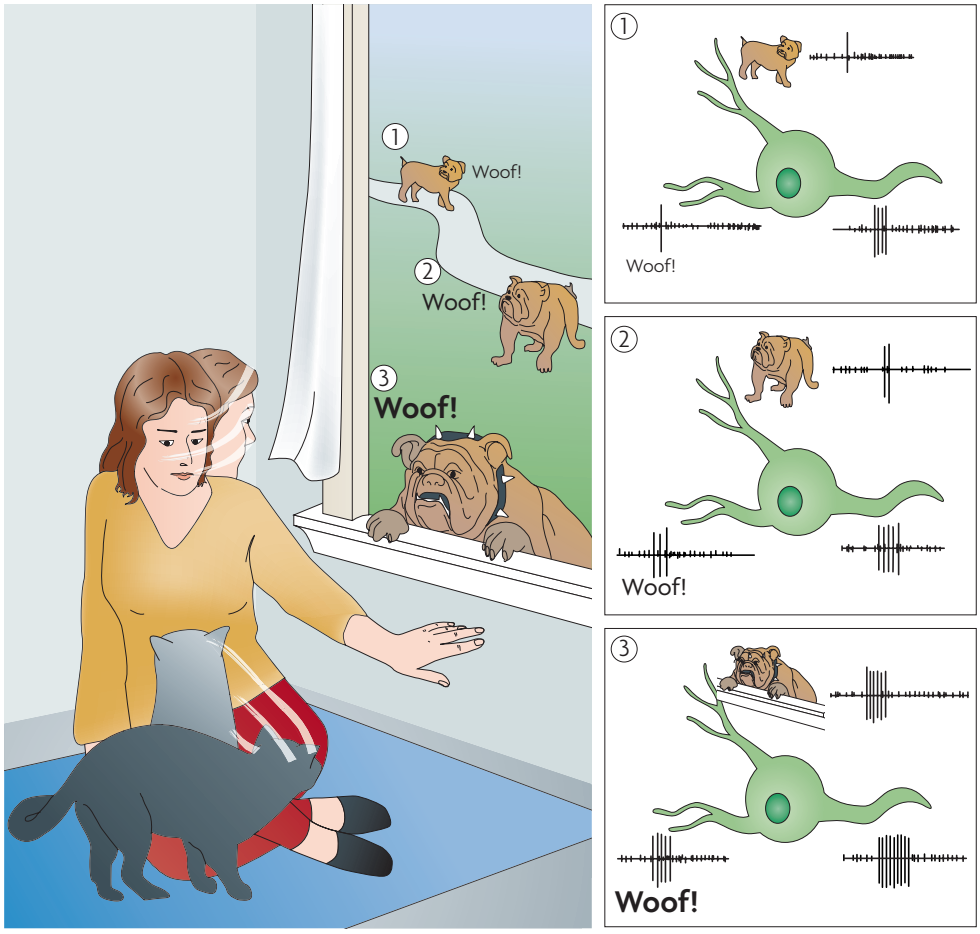


Figure 1.6. Multisensory integration improves detection and speeds responses.

A woman and cat detect the approach of a dog, based on sight and sound (situation 1-3). In each situation a neuron with two input (auditory, left lower corner, and visual, left upper corner) and one output is defined. The number of spikes per modality is correlated with the saliency of the stimulus (increasing from subfigures 1 to 3). When these cues are weak (when the dog is far away), the neural computation involved in their integration is super-additive, such that the response not only exceeds the most vigorous component response, but also exceeds their sum (top). As the dog gets closer, the cues become more effective, unisensory component responses become more vigorous, and integrated responses become proportionately smaller. The computation now becomes additive (middle) and then sub-additive (bottom). Although both the additive and the sub-additive computations also produce responses that exceed the most vigorous component response (that is, they all exhibit multisensory integration), their enhancements are proportionately less than the one shown at the top. All enhancements increase the probability of orientation, but the benefits of multisensory integration are proportionately greatest when cross-modal cues are weakest (illustration used with permission from the author).[*7]

[*7] Stein BE, Stanford TR. Multisensory integration: current issues from the perspective of the single neuron. *Nat Rev Neurosci*. 2008;9(4):255-266. doi:10.1038/nrn2331

The main topic of the thesis is how normal-hearing subjects and CI users integrate audiovisual speech. In this introductory chapter, audiovisual speech perception on a (general) behavioral level has been discussed in order to first explain the basics. In Chapters 2 and 3, audiovisual speech perception will be studied in depth for both normal-hearing listeners and CI-users on a behavioral level.

In the subsequent chapters, to better understand the behavioral differences between normal hearing listeners and CI-users, we try to unravel what occurs in the brain when audiovisual speech is presented. For that, we need a functional neuroimaging technique that can be used for both subject groups.

NEUROIMAGING AND CI-USERS

As referred to before, sensorineural hearing loss and deafness lead to a gradual loss of activity in the auditory nerve.[21] Long-term hearing loss (i.e. duration of deafness) induces structural and functional changes at all levels of the auditory pathways.[22] The loss of one sensory modality can lead to neural plasticity of cortical areas associated with the remaining modalities. Functional neuroimaging studies suggest that long duration of deafness decreases the cortical metabolic activity and can modify the cortical activation patterns involved in audiovisual speech processing.[23,24] Functional neuroimaging techniques have improved our understanding of central auditory pathways, and the cerebral changes induced by sensory deprivation in adults. Unfortunately, several of the neuroimaging methods (functional magnetic-resonance imaging (fMRI), magneto-encephalography (MEG)) are severely limited in their usefulness in CI users, as they are affected by and/or affect the CI.[25] On the other hand, positron emission tomography (PET), which does not affect the CI, is an invasive imaging technique, not suitable for repetitive experiments.[26] Therefore, we used functional near-infrared spectroscopy (fNIRS), a non-invasive, minimally-restrictive and quiet (in contrast to fMRI) optical neuroimaging technique to study cortical activity of CI users and, for reference purposes, normal-hearing subjects.[27] Further explanation on fNIRS can be found in Chapter 4.

OUTLINE OF THIS THESIS

The first part of the thesis focuses on audiovisual speech processing. In **Chapter 2** we focus on inverse effectiveness in normal hearing subjects. This concept holds that multisensory enhancement increases for poorly perceptible unisensory signals, for example in the presence of acoustic background noise or visual distracters.[28] This is not a trivial extension of the classical audiovisual integration studies, as the underlying speech-related sensory signals are complex

and dynamic signals, requiring advanced (top-down) neural processing within the auditory and visual systems. As a follow-up on this study, audiovisual speech perception in CI-users has been studied (**Chapter 3**). We investigated whether CI-users are better able to integrate auditory and visual cues compared to normal-hearing subjects in difficult listening situations.

The aim of the second part is to introduce the neuroimaging technique of fNIRS (**Chapter 4**) that is a non-invasive optical neuroimaging technique, suitable to study cortical activity in CI users. FNIRS is used to study correlates of audiovisual speech perception in CI users and normal-hearing subjects (**Chapter 5**).

REFERENCES

1. Kral A, Dorman MF, Wilson BS. Annual Review of Neuroscience Neuronal Development of Hearing and Language: Cochlear Implants and Critical Periods. *Annu Rev Neurosci*. 2019. doi:10.1146/annurev-neuro-080317
2. Lazard DS, Lee HJ, Gaebler M, Kell CA, Truy E, Giraud AL. Phonological processing in post-lingual deafness and cochlear implant outcome. *Neuroimage*. 2010;49(4):3443-3451. doi:10.1016/j.neuroimage.2009.11.013
3. Blamey P, Artieres F, Başkent D, et al. Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: an update with 2251 patients. *Audiol Neurootol*. 2013;18(1):36-47. doi:10.1159/000343189
4. Lazard DS, Vincent C, Venail F, et al. Pre-, Per- and Postoperative Factors Affecting Performance of Postlinguistically Deaf Adults Using Cochlear Implants: A New Conceptual Model over Time. *PLoS One*. 2012;7(11):e48739. doi:10.1371/journal.pone.0048739
5. Damen GWJA, Beynon AJ, Krabbe PFM, Mulder JJS, Mylanus EAM. Cochlear implantation and quality of life in postlingually deaf adults: Long-term follow-up. *Otolaryngol - Head Neck Surg*. 2007;136(4):597-604. doi:10.1016/j.otohns.2006.11.044
6. Lazard DS, Bordure P, Lina-Granade G, et al. Speech perception performance for 100 post-lingually deaf adults fitted with Neurelec cochlear implants: Comparison between Digisonic® Convex and Digisonic® SP devices after a 1-year follow-up. *Acta Otolaryngol*. 2010;130(11):1267-1273. doi:10.3109/00016481003769972
7. Holden LK, Finley CC, Firszt JB, et al. Factors affecting open-set word recognition in adults with cochlear implants. *Ear Hear*. 2013;34(3):342-360. doi:10.1097/AUD.0b013e3182741aa7
8. Green KMJ, Bhatt YM, Mawman DJ, et al. Predictors of audiological outcome following cochlear implantation in adults. In: *Cochlear Implants International*. Vol 8. ; 2007:1-11. doi:10.1002/cii.326
9. Summerfield AQ, Marshall DH. Preoperative predictors of outcomes from cochlear implantation in adults: Performance and quality of life. In: *Annals of Otology, Rhinology and Laryngology*. Vol 104. ; 1995:105-108. http://www.ncbi.nlm.nih.gov/pubmed/7668594. Accessed February 11, 2018.
10. Gomaa NA, Rubinstein JT, Lowder MW, Tyler RS, Gantz BJ. Residual Speech Perception and Cochlear Implant Performance in Postlingually Deafened Adults. *Ear Hear*. 2003;24(6):539-544. doi:10.1097/01.AUD.0000100208.26628.2D
11. Adunka OF, Buss E, Clark MS, Pillsbury HC, Buchman CA. Effect of preoperative residual hearing on speech perception after cochlear implantation. *Laryngoscope*. 2008;118(11):2044-2049. doi:10.1097/MLG.0b013e3181820900
12. Van Der Marel KS, Briaire JJ, Verbist BM, Muurling TJ, Frijns JHM. The influence of cochlear implant electrode position on performance. *Audiol Neurotol*. 2015;20(3):202-211. doi:10.1159/000377616
13. Rouger J, Lagleyre S, Fraysse B, Deneve S, Deguine O, Barone P. Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proc Natl Acad Sci U S A*. 2007;104(17):7295-7300. doi:10.1073/pnas.0609419104
14. Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb Cortex*. 2007;17(5):1147-1153. doi:10.1093/cercor/bhl024
15. Sumby WH, Pollack I. Visual Contribution to Speech Intelligibility in Noise. *J Acoust Soc Am*. 1954;26(2):212-215. doi:10.1121/1.1907309
16. Auer ET. Investigating speechreading and deafness. *J Am Acad Audiol*. 2010;21(3):163-168. doi:10.3766/jaaa.21.3.4
17. Auer ET, Bernstein LE. Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. *J Speech, Lang Hear Res*. 2007;50(5):1157-1165. doi:10.1044/1092-4388(2007/080)
18. Bernstein LE, Demorest ME, Tucker PE. Speech perception without hearing. *Percept Psychophys*. 2000;62(2):233-252. doi:10.3758/BF03205546

19. Meredith MA, Stein BE. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Res*. 1986;365(2):350-354. doi:10.1016/0006-8993(86)91648-3

20. Meredith M, Stein B. Interactions among converging sensory inputs in the superior colliculus. *Science (80-)*. 1983;221(4608):389-391. doi:10.1126/science.6867718

21. Tucci DL, Born DE, Rubel EW. Changes in spontaneous activity and CNS morphology associated with conductive and sensorineural hearing loss in chickens. *Ann Otol Rhinol Laryngol*. 1987;96(3 Pt 1):343-350. doi:10.1177/000348948709600321

22. Kral A, Sharma A. Developmental neuroplasticity after cochlear implantation. *Trends Neurosci*. 2012;35(2):111-122. doi:10.1016/j.tins.2011.09.004

23. Green KMJ, Julyan PJ, Hastings DL, Ramsden RT. Auditory cortical activation and speech perception in cochlear implant users: Effects of implant experience and duration of deafness. *Hear Res*. 2005;205(1-2):184-192. doi:10.1016/j.heares.2005.03.016

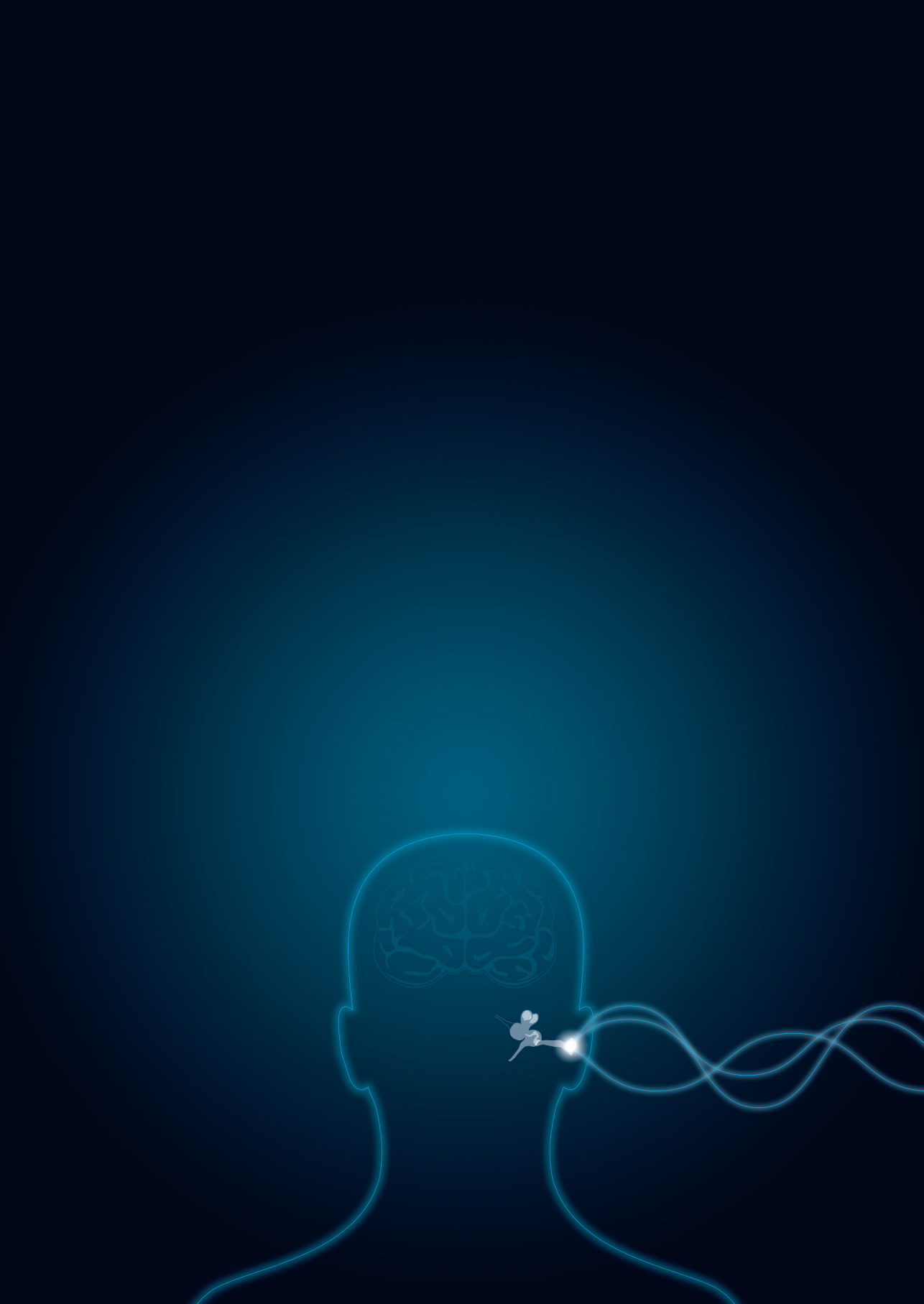
24. Giraud A, Lee H. Predicting cochlear implant outcome from brain organisation in the deaf. *Restor Neurol Neurosci*. 2007;25:381-390. doi:Article

25. Hall DA, Haggard MP, Akeroyd MA, et al. Modulation and task effects in auditory processing measured using fMRI. *Hum Brain Mapp*. 2000;10(3):107-119.

26. Johnsrude IS, Giraud AL, Frackowiak RSJ. Functional Imaging of the Auditory System: The Use of Positron Emission Tomography. *Audiol Neuro-Otology*. 2002;7(5):251-276. doi:10.1159/000064446

27. van de Rijt LPH, van Opstal AJ, Mylanus EAM, et al. Temporal Cortex Activation to Audiovisual Speech in Normal-Hearing and Cochlear Implant Users Measured with Functional Near-Infrared Spectroscopy. *Front Hum Neurosci*. 2016;10(February):48. doi:10.3389/fnhum.2016.00048

28. Stein BE, Meredith MA. The Merging of the Senses. *Cambridge, MA MIT Press*. 1993:xv-211. doi:10.3389/neuro.01.019.2008



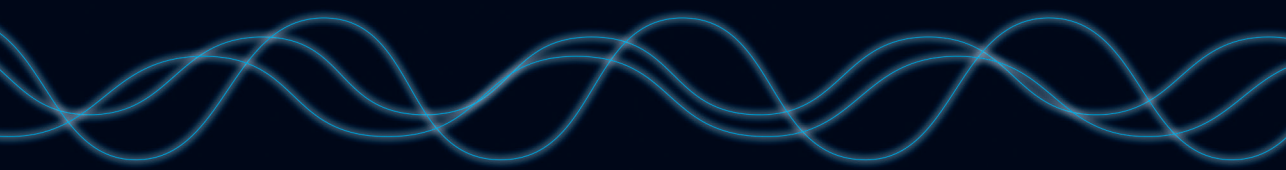
CHAPTER 2

The principle of inverse effectiveness in audiovisual speech perception

Luuk P.H. van de Rijt, Anja Roye, Emmanuel A.M. Mylanus, A. John van Opstal, Marc M. van Wanrooij

Keywords: multisensory, lipreading, listening, hearing, speech recognition in noise

Frontiers in Human Neuroscience. 13:335.
DOI: https://10.3389/fnhum.2019.00335



ABSTRACT

We assessed how synchronous speech listening and lipreading affects speech recognition in acoustic noise. In simple audiovisual perceptual tasks, inverse effectiveness is often observed, which holds that the weaker the unimodal stimuli, or the poorer their signal-to-noise ratio, the stronger the audiovisual benefit. So far, however, inverse effectiveness has not been demonstrated for complex audiovisual speech stimuli. Here we assess whether this multisensory integration effect can also be observed for the recognizability of spoken words.

To that end, we presented audiovisual sentences to 18 native-Dutch normal-hearing participants, who had to identify the spoken words from a finite list. Speech-recognition performance was determined for auditory-only, visual-only (lipreading) and auditory-visual conditions. To modulate acoustic task difficulty, we systematically varied the auditory signal-to-noise ratio. In line with a commonly-observed multisensory enhancement on speech recognition, audiovisual words were more easily recognized than auditory-only words (recognition thresholds of -15 dB and -12 dB, respectively).

We here show that the difficulty of recognizing a particular word, either acoustically or visually, determines the occurrence of inverse effectiveness in audiovisual word integration. Thus, words that are better heard or recognized through lipreading, benefit less from bimodal presentation.

Audiovisual performance at the lowest acoustic signal-to-noise ratios (45%) fell below the visual recognition rates (60%), reflecting an actual deterioration of lipreading in the presence of excessive acoustic noise. This suggests that the brain may adopt a strategy in which attention has to be divided between listening and lipreading.

INTRODUCTION

Speech is a complex, dynamic multisensory stimulus, characterized by both an auditory and a visual information stream. Congruent information of the sensory modalities (i.e. spatial and temporal coincidence of the sensory streams, and their meanings) is integrated in the brain to form a coherent, often enhanced, percept of the common underlying source.[1–3] Indeed, additional synchronous visual information (i.e. speech-reading / lipreading) has a positive impact on speech perception, and audiovisual speech recognition in acoustic noise is substantially better than for auditory speech alone.[4–15]

Audiovisual integration in general, has been the topic of a variety of behavioral and electrophysiological studies, involving rapid eye-orienting to simple peripheral stimuli, spatial and temporal discrimination of audiovisual objects, and the integrative responses of single neurons in cats and monkeys.[16–23] Three main principles have been shown to govern the mechanisms of multisensory integration: i. spatial alignment of the different sources, ii. temporal (near-)synchrony, and iii. inverse effectiveness. The latter holds that multisensory enhancement strongly increases for poorly perceptible unisensory signals, for example in the presence of acoustic background noise or visual distracters.[3] Although these principles have mostly been demonstrated at the neurophysiological level of anesthetized experimental animals (for review, see Stein and Meredith 1993), several studies on audiovisual saccadic eye movements in humans or on manual reaction times in macaques and humans, have revealed systematic modulations of the effects of audiovisual congruency and inverse effectiveness that corroborate the neurophysiological data.[16,24-26]

In this study, we focus on whether the phenomenon of inverse effectiveness can also be applied to speech perception. This is not a trivial extension of the classical audiovisual integration studies, as the underlying speech-related sensory signals are complex and dynamic signals, requiring advanced (top-down) neural processing within the auditory and visual systems. One way of studying the presence of inverse effectiveness in the perception of audiovisual speech stimuli is by adding background noise, which effectively changes the saliency of the auditory stimulus.[11,15,27] By doing so, earlier studies have suggested an absence of inverse effectiveness, as at low unimodal performance scores, the audiovisual enhancement decreases. The principle of inverse effectiveness has also been studied by quantifying the differences in unimodal word-recognition performance scores across (groups of) subjects, however, outcomes were not consistent.[7,15,28,29] To our knowledge, the effect of the visual or auditory recognizability of words (irrespective of background noise) on the presence or absence of inverse effectiveness has not been studied. For example, words that contain more spectral-temporal information, or are articulated more pronouncedly, will likely be better heard or visually recognized over a large range of noise levels. If the principle of inverse effectiveness would hold at the word level,

highly-informative words should benefit less from bimodal presentation than less-informative words. To study this possibility, we determined how well words can be recognized by listening and/or lipreading under noisy listening conditions in normal-hearing subjects.

RESULTS

Overview

Eighteen normal-hearing subjects had to identify 50 words (Table 2.1) occurring in 155 unique five-word sentences, by selecting the words they recognized (ten-alternative forced choice) on a screen. The speech material was based on the Dutch version of the speech-in-noise matrix test developed by Houben and colleagues (see Methods on the construction of the speech material, Fig. 2.1).[30] The words were presented in acoustically-only (A-only, e.g. Fig. 2.1A), visual-only (V-only, e.g. Fig. 2.1D) or bimodal (AV, e.g. Fig. 2.1A and D combined) blocks. An acoustic background noise (Fig. 2.1B) was played in the A-only and AV conditions at five signal-to-noise ratios. Note that the words vary substantially in ongoing amplitude and duration (Fig. 2.1A), spectral-temporal dynamics (Fig. 2.1C), and articulation (Fig. 2.1D). This variation will likely affect speech recognition, and is the foundation on which we will test inverse effectiveness. In what follows, we will quantify how well each word is recognized visually and aurally, then how simultaneous audiovisual presentation of a word affects recognition accuracy, and finally we will determine how unimodal recognition accuracy affects audiovisual enhancement.

Table 2.1 Words of the Dutch matrix test

Name	Verb	Numeral	Adjective	Object
Anneke	geeft	twee	dure	bloemen
Christien	had	drie	goede	boeken
Heleen	kiest	vier	**groene**	boten
Jan	koopt	vijf	grote	**dozen**
Mark	maakte	**zes**	kleine	fietsen
Monique	tekent	acht	mooie	messen
Pieter	**telde**	negen	nieuwe	munten
Sarah	vond	tien	oranje	ringen
Tom	vroeg	twaalf	vuile	schoenen
Willem	wint	achttien	zware	stenen

Bold words indicate an example sentence: 'Tom telde zes groene dozen' (translation: 'Tom counted six green boxes', see Fig. 2.8)

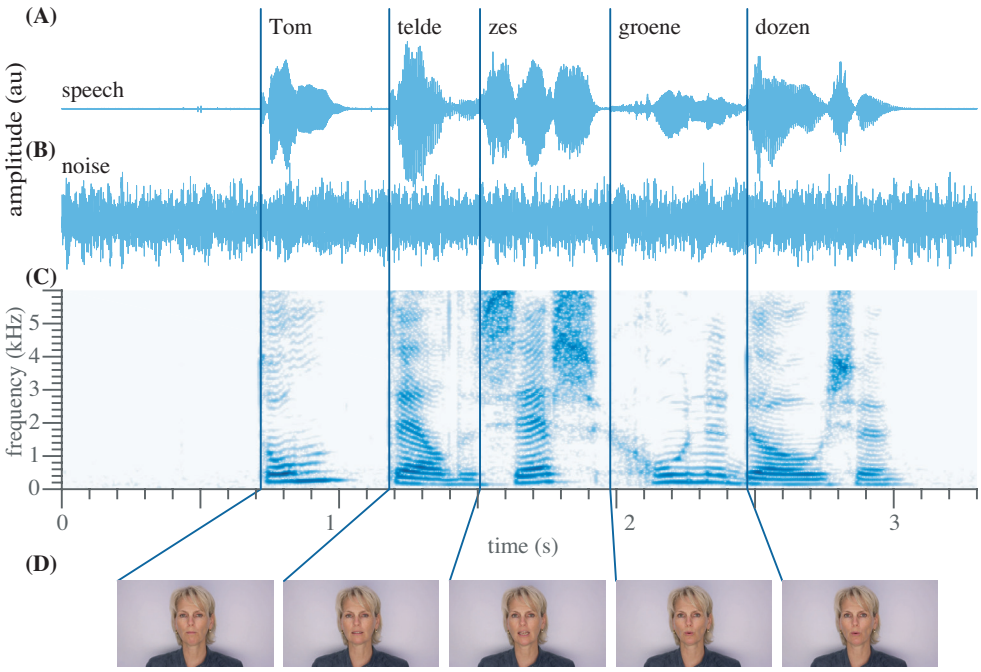


Figure 2.1. Example stimulus.

A) Temporal waveform of the auditory speech signal "Tom telde zes groene dozen" (translation: Tom counted six green boxes. **B)** Waveform of the auditory noise. **C)** Spectrogram of the recorded sentence. **D)** Five videos frames around the onset of the word. Dark blue lines denote the approximate onset of each individual word. Written informed consent for the publication of this image was obtained from the individual shown.

Lipreading

We will first describe the lipreading abilities (V-only). These were quantified for every subject (n=18) and every word (n=50) as the number of correct responses, z, divided by the number of presentations, N(=18), i.e. the correct scores (Fig. 2.2A), in the V-only block. The correct scores varied both across words and subjects from perfect (i.e. 18 correct responses to 18 presentations, e.g. for the word 'vijf' by subject S2), to around chance level (0.1, e.g. a score of 0 correct responses for 18 word presentations for the word 'telde' presented to subject S8). Notably, some words were easily correctly identified by almost all subjects (e.g. 'Mark'), while others were near-never identified ('telde') by anyone. Similarly, some subjects were perfect lip-readers with correct scores for all words near 1.0 (e.g. subject S14), while subject S13, as an extreme case, could hardly identify any words via lipreading.

As the realizations of the visual correct scores were quite noisy (as apparent in the jittery pattern in Fig. 2.2A), the estimates for the proportion of correct scores for each word and subject separately were quite uncertain (average 95%-highest density interval [95%-HDI] was 0.29 [0.14-0.42] across all 900 estimates from 18 subjects and 50 words). We therefore

determined the visual lipreading recognition rates for words, $\rho_{V,w}$, and for each subject, $\rho_{V,s}$ by fitting the following function:

$$F_V(\rho_{V,w}, \rho_{V,s}) = \rho_{V,w} \times \rho_{V,s} \quad (2.1)$$

to the responses from the V-only trials, which are taken to be binomially distributed (see Methods for details on the fitting procedure). This yields 18 visual recognition rates for subjects, $\rho_{V,s}$, and 50 visual recognition rates for words, $\rho_{V,w}$. Multiplication of these rates assumes that they were independent, and thus separable from each other. This assumption seems to hold, at least qualitatively, when looking at the correct scores for each word and subject (cf. Fig. 2.2A and Fig. 2.2B, see also Methods for a more quantitative approach). This procedure smoothened the recognition rate matrix (Fig. 2.2B), and decreased variability in the estimates (as expressed by the small 95%-HDI in Fig. 2.2C/D; average 95%-HDI = 0.09 [0.04-0.14] across 68 parameters). This function also reduced the number of variables from 900 (number of subjects multiplied by number of words) to 68 (number of subjects plus number of words). These features enable a more practical comparison to the other, A-only and AV conditions, to be introduced later on. The model described by eqn. 2.1 is also preferred by having a lower Bayesian Information Criterion (BIC, see Methods) compared to the model that determines recognition rates independently for all subjects and words (5.5k vs 9.0k, respectively).

Moreover, the recognition estimates are in line with the correct-score data (correlation r=0.84, with limited to no discernible bias). Words were generally easily recognized through lipreading (Fig. 2.2D, mean $\rho_{V,w} = 0.77$), but there was considerable variability in visual recognizability across words: many words were identified easily (e.g. mean $\rho_{V,boten} = 0.99$), while others were barely recognizable (e.g. mean $\rho_{V,telde} = 0.03$). Also the ability of subjects to lipread was relatively high on average (Fig. 2.2C, mean $\rho_{V,s} = 0.78$). However, there was a considerable range in lipreading ability. The best lip-readers could recognize ~100% of the easily-identified words (mean $\rho_{V,S14} = 1.00$), while the worst performer could at best recognize ~15% correctly (mean $\rho_{V,S13} = 0.15$). The large variability in visual recognition rates across words and subjects provides a potential way to determine how speech-reading performance affects speech listening, when both auditory and visual speech-recognition cues are presented synchronously.

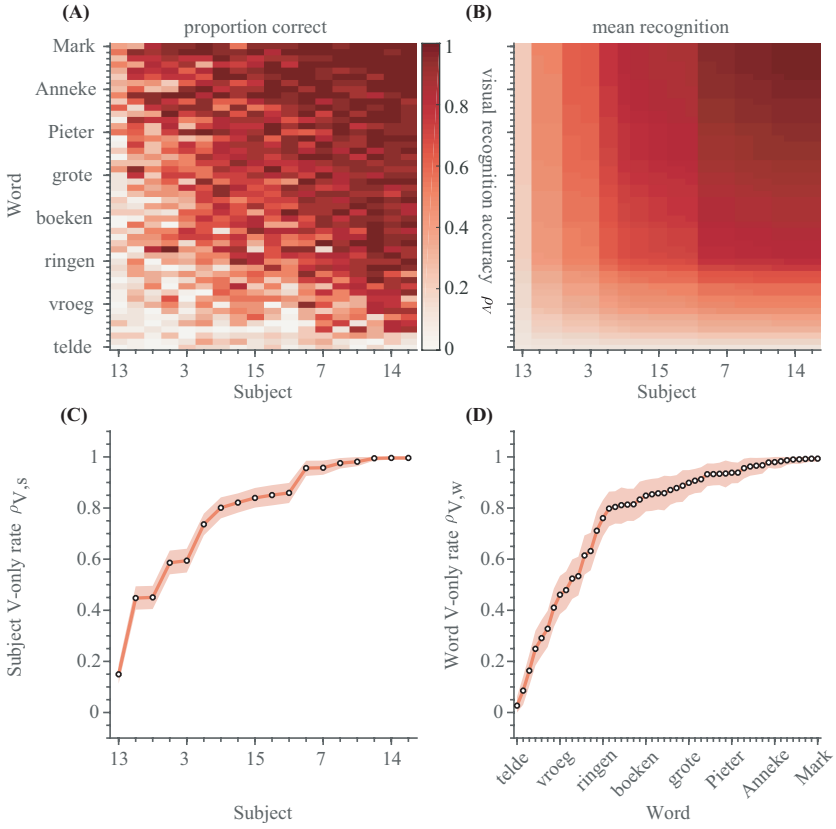


Figure 2.2. Lipreading.

A) Visual recognition scores. The correct score (number of correct responses divided by the number of presentations) is shown separately for every word and subject (900 entries) for the V-only condition. The correct scores and rates have been ordered by the recognition rates of subjects on the abscissa, and of words on the ordinate from low-left to high-right. **B)** The average estimated visual recognition rates (Eqn. 2.1). Same layout as in A. V-only speech recognition rates for **C)** subjects and **D)** words. Rates were ordered from low-left to high-right. Open circles indicate the mean of the estimated rate, colored patch indicates the 95% Highest Density Interval (HDI).

Speech listening

In the A-only block, subjects identified words by listening to the audio recordings of sentences (e.g. Fig. 2.1A, without visual feedback from the lips). A stationary masking noise (e.g. Fig. 2.1B) was played at a constant level of 65 dB SPL, while the sentences were played at an SNR of -21, -16, -13, -10 or -5 dB. In total, the data comprised 4482 different combinations of subject, word, and SNR (not all 250 potential combinations of SNR and word were presented to every one of the 18 subjects). The average word recognition rate was ~50% across all SNRs and subjects (Fig. 2.3A-E). Overall listening performance for SNRs lower than -10 dB was worse than lipreading performance (cf. amount of white in Fig. 2.2A vs. Fig. 2.3A-E). In contrast to lipreading, listening performance was quite similar across subjects (Fig. 2.3A-E). This small

variability across listeners might be expected, as all listeners were normal-hearing, and were therefore likely to understand the speech equally well.

Typically, SNR had a strong influence on the ability to recognize the words through listening (Fig. 2.3A to 3E, from low to high SNR, the correct scores improve from almost 0 to near perfect). To quantify this, we estimated the SNR for which the recognition rate was 50%, i.e. the auditory speech-recognition threshold, θ_A, by fitting the parameters of a logistic psychometric function F_A for every word (with a parametrization as mentioned in [31]):

$$F_A(SNR, \theta_A, \omega_A) = \left(1 + e^{-\frac{2\ln 9}{\omega_A}(\mathrm{SNR} - \theta_A)}\right)^{-1} \quad (2.2)$$

with ω_A the auditory recognition width from 10 to 90% performance (in dB). The width (conversely, the slope) of the psychometric curve, ω_A, did not vary substantially across words or subjects. Therefore, only one value was estimated, which was on average 7.1 dB, 95% HDI: 6.8 - 7.4 dB. As the correct scores did not vary appreciably across subjects, we pooled over subjects, to obtain 50 auditory recognition thresholds, one for each word. To exemplify this, we take a look at the word 'Pieter' (Fig. 2.3K). This word was easily recognized by all subjects at the SNR of -5 dB, leading to a 100% recognition score. In contrast, "Pieter" was almost impossible to identify at the lowest SNR of -21 dB, when subjects identified the word presented in 10% of the cases (chance-level). By fitting a psychometric curve through the data, we obtained a speech listening threshold for this word at -11.5 dB (Fig. 2.3K). Similar to the V-only model (eqn. 2.1), this modeling smoothened the A-only estimates (Fig. 2.3F-J), reduced uncertainty in the parameter estimates (average 95%-HDI from 0.54 [0.35-0.77] to 0.07 [0.00-0.18]), and reduced the number of parameters (from 4482 to 51). The model is (therefore) also favored by the BIC (8.0k vs. 45.3k of a fully-independent model; a model that included a logistic dependence on SNR but allowed for subject and word variability in both the threshold and width had a BIC of 21.2k with 1800 free parameters).

Importantly, auditory speech-recognition thresholds for each word (Fig. 2.3L) varied over a considerable 10-dB range, from the best-recognizable word (mean $\theta_{A,zware}$ = -16.7 dB) to the hardest-to-recognize word (mean $\theta_{A,goede}$ = -6.6 dB), with an average threshold of -12.1 dB.

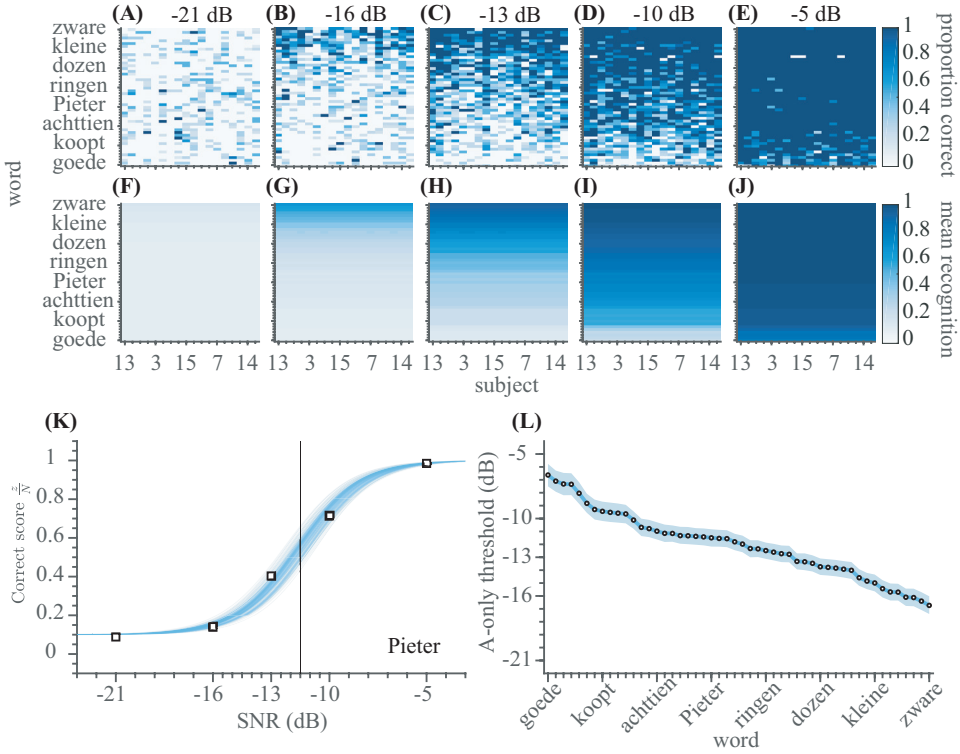


Figure 2.3. Speech listening.

Auditory word-recognition scores. **A-E)** The correct score (number of correct responses divided by the number of presentations) is shown separately for every word and subject (900 entries) for each of the SNRs of -21, -16, -13, -10 and -5 dB. The correct scores have been ordered by the average V-only rates of subjects on the abscissa, and A-only thresholds on the ordinate. **F-J)** The average estimated auditory recognition rates. **K)** Correct scores and psychometric fit for the word 'Pieter' as a function of SNR, averaged across all subjects. Open squares indicate the measured correct scores. Blue shading denotes credible fits (see Methods). Vertical bold grey line indicates the average of likely recognition thresholds. **L)** A-only speech recognition thresholds, ordered from high-left to low-right. Note that a lower threshold indicates better performance. Open circles indicate means of the estimated thresholds, colored patch indicates the 95% HDI.

Audiovisual speech recognition

In the AV-condition, subjects identified words by listening to, and by lipreading, the audiovisual recordings of sentences in the presence of acoustic noise (65 dB SPL, SNR: [-21, -16, -13, -10, -5] dB). The presentation of congruent visual feedback clearly aided recognition performance, as the correct scores (Fig. 2.4A-E) were higher than for the A-only condition (cf. Fig. 2.3A-E). Also, in contrast to the speech listening scores (cf. Fig. 2.3A-E) and more in line with lipreading performance (Fig. 2.2A), the AV scores not only varied over words, but also across subjects (which is visible in the pattern of correct scores in Fig. 2.4A).

We quantified AV performance by fitting a function F_{AV} that combines the characteristics of Eqns. 2.1 and 2.2 for the unimodal performances:

$$F_{AV}(SNR, \theta_{AV}, \omega_{AV}, \rho_{AV,w}, \rho_{AV,s}) = \left(1 - \rho_{AV,w} \times \rho_{AV,s}\right) \times \left(1 + e^{-\frac{2\ln 9}{\omega AV}(SNR - \theta_{AV})}\right)^{-1} + \rho_{AV,w} \times \rho_{AV,s} \quad (2.3)$$

with the audiovisual recognition threshold, θ_{AV} describing the logistic SNR dependence, and two audiovisual recognition rates $\rho_{AV,w}$ and $\rho_{AV,s}$, defining the minimum performance level in the AV condition (i.e. for SNR = -∞) for words and subjects, respectively. Again, the word 'Pieter' is taken as an example to illustrate the fit (Fig. 2.4K, cf. Fig. 2.3K). In contrast to A-only recognition, even at the lowest SNR (-21 dB), this word was easily recognized by all subjects in 75% of the time.

Similar to the V-only and A-only models (eqns. 2.1 and 2.2), this modeling smoothened the AV-only estimates (Fig. 2.4F-J), reduced uncertainty in the parameter estimates (average 95%-HDI from 0.55 [0.35-0.77] to 0.10 [0.00-0.22]), and reduced the number of parameters (from 4482 to 119). Again, the model is favored by the BIC (7.7k vs. 45.2k of a fully-independent model; a model that included a logistic dependence on SNR but allowed for subject and word variability in both the threshold and width had a BIC of 33.1k with 1868 free parameters).

Like for the A-only condition, one value of the width was estimated for all subjects and words (this width was on average 10.5 dB, 95% HDI: 9.5 - 11.4 dB). The audiovisual speech thresholds were determined for words alone (Fig. 2.4L), in line with the auditory speech thresholds (Fig. 2.3L). The thresholds varied over a ~21 dB range (from mean $\theta_{A,\,Tom}$ = -27.6 dB to mean $\theta_{A,goede}$ = -6.4 dB), with an average threshold of -14.7 dB. The subjects' AV recognition rates (Fig. 2.4G) varied from almost negligible (chance) to near-perfect (from mean $\rho_{AV,S13}$ = 0.07 to mean $\rho_{AV,S14}$ = 0.99), with an average rate around 0.63. The AV recognition rates for words (Fig. 2.4H) varied over a similar range (from mean $\rho_{AV,tekent}$ = 0.09 to mean $\rho_{AV,Anneke}$ = 0.98), with an average rate around 0.71. There was considerable uncertainty in the estimation of the word AV rates (e.g., the widest 95%-HDI = 0.02-0.95 for the word 'Tom'), but in general the 95% HDIs for all other parameters were narrow.

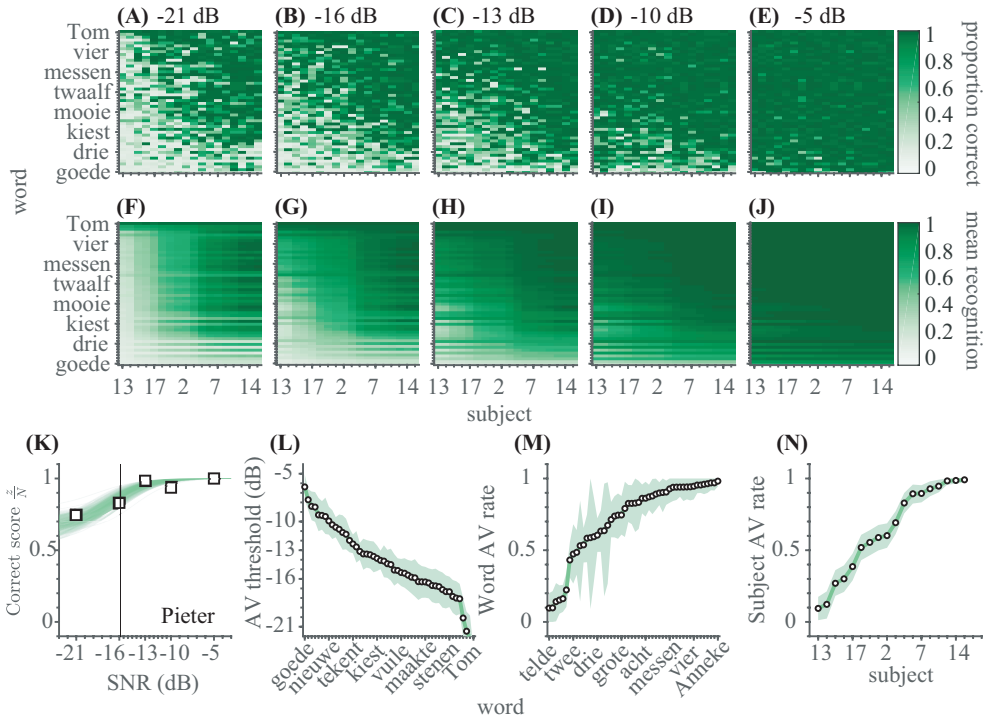


Figure 2.4. Audiovisual speech recognition. A-E)

The audiovisual correct scores are shown separately for every word and subject (900 entries) for each of the SNRs of **A)** -21, **B)** -16, **C)** -13, **D)** -10 and **E)** -5 dB. The correct scores have been ordered by the average AV recognition rates of subjects on the abscissa, and of words on the ordinate. **F-J)** The average estimated audiovisual recognition rates. **K)** Audiovisual correct scores and psychometric fit for the word 'Pieter' as a function of SNR, averaged across all subjects. Open squares indicate the measured correct scores. Green shading denotes credible fits (see Methods). Vertical bold grey line indicates the average of likely recognition thresholds. **L)** AV speech-recognition thresholds, **M,N)** AV recognition rates for words and subjects, ordered from low-left to high-right. Note that a lower threshold indicates better performance. Open circles indicate means of the estimated thresholds, colored patch indicates the 95% HDI.

Audiovisual enhancement

The audiovisual parameters from eqn. 2.3 are considered to be basic descriptors for the audiovisual performance, from which we can derive the audiovisual enhancement by comparing the results to the unimodal parameters from eqns. 2.1 and 2.2. For the audiovisual threshold, the comparison to the auditory threshold indicates how much the SNR can decrease when the visual modality is added, without affecting performance. The change in threshold, $\Delta\theta_{AV}$, relative to the auditory threshold, was thus estimated by rewriting θ_{AV} in eqn. 2.3 as:

$$\theta_{AV} = \theta_A + \Delta\theta_{AV} \tag{2.4}$$

Typically, the audiovisual recognition thresholds were lower (i.e. better) than the auditory recognition thresholds (Fig. 2.5A), by on average -3 dB. This means that the threshold is typically reached at lower SNRs when people speech-read at the same time. The threshold for 35 words improved in the AV condition (95%-HDI lay below 0 dB), while for 15 words there was no difference (95%-HDI included 0 dB).

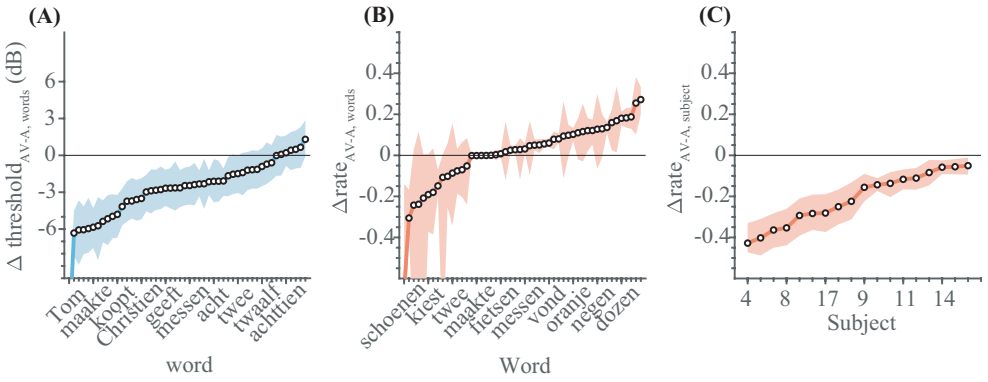


Figure 2.5. Comparison between audiovisual and unimodal conditions.

Change in threshold and rates of AV speech recognition in comparison to unimodal listening conditions. **A)** The change in threshold for each word (eqn. 2.4). Note that a negative change in threshold denotes better performance in AV conditions. **B)** The change in recognition rate for each word (eqn. 2.5). **C)** The change in recognition rate for each subject. For rates, a change larger than 0 denotes better AV performance. Open circles denote the mean of the parameter estimate, colored patches indicate 95% HDI.

Similarly, the minimum performance level in the AV condition is given by multiplying the recognition rates for words and subjects: $\rho_{AV,w} \times \rho_{AV,s}$. This measure quantifies the performance level in the absence of an auditory signal (i.e. when the SNR approaches -∞). In case there really is no auditory signal, one might expect that the minimum audiovisual performance level, given by the rates, would equal the visual performance rate. This, of course, only holds if the stimulus parameters fully determine the subject's performance levels, and if non-stimulus factors, such as task or block design, are irrelevant. We tested this prediction by determining the difference in audiovisual and visual rates for words and subjects:

$$\begin{cases} \rho_{AV,w} = \rho_{V,w} + \Delta\rho_{AV,w} \\ \rho_{AV,s} = \rho_{V,s} + \Delta\rho_{AV,s} \end{cases} \tag{2.5}$$

On average, there was no difference in recognition rates for words (Fig. 2.5B), as the difference values scattered around 0 for most words. In contrast, the subjects' ability to lipread in the AV condition (as reflected by the subjects' recognition rate) was poorer than in the V-only condition (Fig. 2.5C). The rates for all subjects dropped (mean $\Delta\rho = -0.2$, all 95% HDI < 0). This indicates that, on average, audiovisual performance dropped below the V-only performance scores, when poor auditory SNRs caused speech listening to deteriorate completely.

As these last points are important, we will restate them. First, the AV threshold is lowered, making it easier to recognize words at a given SNR. This effectively yields an audiovisual enhancement to speech listening (Fig. 2.5A). Second, words are recognized through lipreading at equal levels in both V-only and AV conditions (Fig. 2.5B). Third, somewhat surprisingly, the lipreading ability of subjects is impoverished in the AV condition (Fig. 2.5C). This suggests that task constraints (i.e. being in an AV condition vs. in a V-only condition) have a significant influence on speech recognition performance, even when stimulus parameters are equivalent (i.e. only a visual, no auditory signal).

Probability summation

Next, we qualitatively compared the AV condition with a model in which audiovisual integration is merely a result of statistical summation rather than of true neural integration. Finding an improved performance (i.e. better speech recognition) in the AV condition is not automatic evidence that the brain integrates the auditory and visual inputs. Indeed, having both modalities available, rather than one, automatically increases the probability of stimulus recognition. In a model of probability summation, participants recognize a word from either the A-only or the V-only condition, which are considered independent processing channels. The probability of word recognition in the presence of the two independent, non-interacting, modalities is given by:

$$P_{sum} = 1 - P_{fail} = P_A + P_V - P_A \times P_V \tag{2.6}$$

where P_{sum} is the probability to successfully recognize a word according to the summation model, P_A is the probability to recognize a word in the A-only condition, and P_V is the probability of recognizing a word in the V-only condition. Both P_A and P_V were estimated according to eqns. 2.1 and 2.2, but there were no additional free parameters to fit for the probability summation model. In order to demonstrate how well this model performs for various unimodal stimulus strengths, we split the data in four groups (Fig. 2.6), as a first, simple approximation, consisting of poor or good V-only lipreading or average A-only listening accuracy (estimated recognition rate below or above 0.55, respectively; for A-only, recognition rates are averaged across SNR; as shown in Fig. 2.1B and Fig. 2.2F-J). Note that there is a weak, negative correlation between the speech listening threshold and lipreading recognition rate at the word level; r = -0.39, 95%-HDI = -0.63 to -0.15, so that each group contains a slightly different number of subject-word combinations.

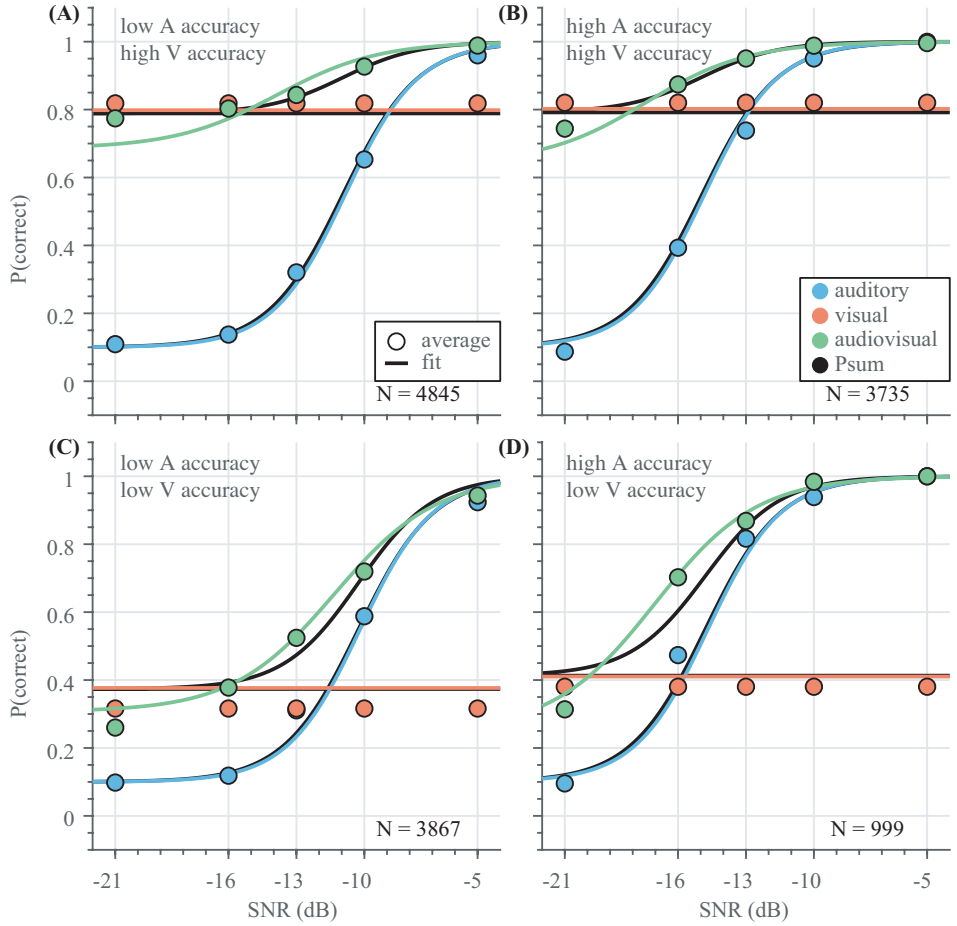


Figure 2.6. Audiovisual speech recognition varies with unimodal information.

Psychometric curves were determined (eqn. 2.1-2.3) from all data divided across 4 groups differing in unimodal performances: visual recognition rate **A,B)** larger and **C,D)** smaller than 0.55; and an auditory recognition rate **A,C)** larger than and **B,D)** smaller than 0.55. Curves indicate the average model estimate, circles denote the average correct score. N is the number of subject-word-SNR combinations for each group.

Despite the differences in unimodal performance, the best-fit performance curves (according to eqn. 2.1-2.3) for each of those groups followed a similar pattern. Auditory performance (Fig. 2.6 – blue) degrades as the signal-to-noise ratio decreases; degradation is worse for words with poor auditory thresholds (Fig. 2.6A,C). Visual performance (Fig. 2.6 – red) is better than auditory performance for a larger range of SNRs if the visual word recognition rate is better (Fig. 2.6A,B). Notably, for all groups, audiovisual performance (Fig. 2.6 – green) is never worse than auditory performance; a clear audiovisual enhancement relative to auditory performance alone is present for a large range of SNRs. While audiovisual performance is typically also better

than visual performance, at very low acoustic SNRs, the multisensory performance tends to be worse than lipreading performance (Fig. 2.6, the green curves and circles drop below the red lines and circles). Overall, the fits to eqns. 2.1-2.3 followed the average correct scores nicely, although the AV fit (green) slightly under- and overshot the correct score at the lowest SNR for the high-accuracy and low-accuracy V-only data, respectively. The V-only fit (red) indicated slightly better performance than the average correct score for low-accuracy V-only data (Fig. 2.6C,D).

Notably, the benchmark probability summation model can describe the audiovisual data quite well, at least qualitatively (Fig. 2.6 – black). This model exhibits unimodal-like performance whenever either unimodal recognition abilities vastly outperforms the other, and shows maximum enhancement when the visual and auditory performances are equal.

We also fitted two other models that can exhibit (supra-additive) enhancements in audiovisual speech perception.[27,29] While qualitatively similar, our version of these models (that also include word and subject variability in the model parameters) performed worse than the probability-summation model (both in terms of how well the fit curves approximated the correct scores, and in terms of the BIC). We will not elaborate on these models here, but would like to note that neither these two models nor the probability-summation model allow for audiovisual performance to drop below visual performance.

Inverse effectiveness – noise level

To test whether the multisensory data adhered to the principle of inverse effectiveness, we first determined the influence of SNR, as a measure of auditory stimulus intensity, on the magnitude of the audiovisual enhancement. For this purpose, we determined the audiovisual enhancement as the difference between the average audiovisual and auditory model fits and correct scores (Fig. 2.6, green and blue, curves and circles). The shape of audiovisual enhancement is largely similar across the four groups. (Fig. 2.7, blue), and indicates 1) that auditory recognition performance improves by adding the visual information especially for low SNRs, and 2) the highest enhancement occurs at high to intermediate noise levels (SNR between -13 and -20 dB). For the lowest SNR of -21 dB, enhancement saturates or decreases slightly (for the correct scores only when A-only and V-only accuracy is low in Fig. 2.7C). So, the principle of inverse effectiveness seems to apply to a large extent, when auditory SNR is considered as the measure of unimodal reliability.

We can also express the audiovisual enhancement relative to the benchmark model of statistical summation. For all 4 groups, the probability-summation model resembles AV speech recognition quite well (Fig. 2.7; black lines close to 0). However, there is a slight deterioration at the lowest SNRs (maximum deterioration of -0.04 to -0.10 at an SNR of -21 dB).

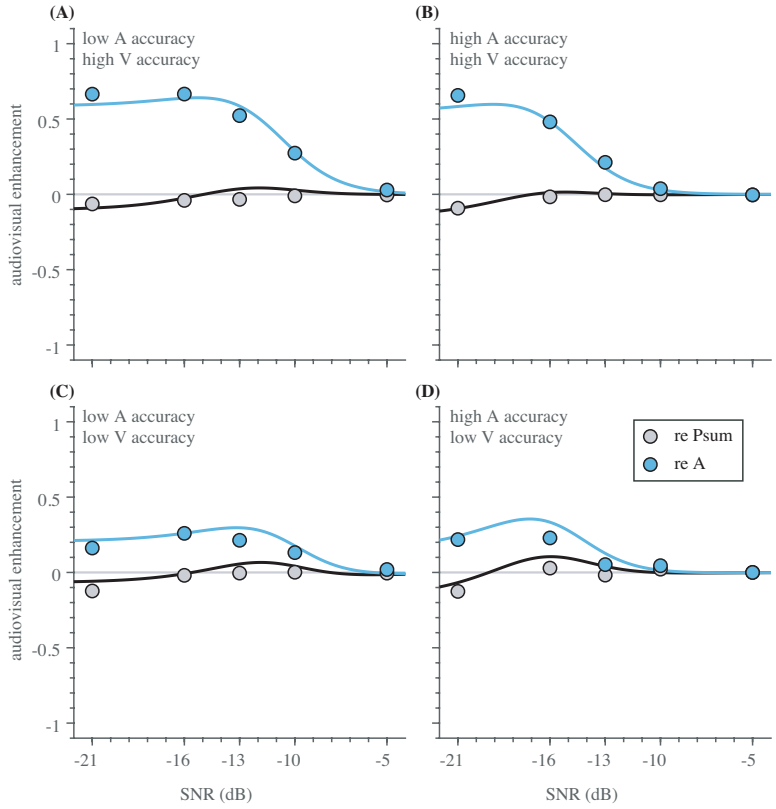


Figure 2.7. Audiovisual enhancement as a function of SNR.

A-D) The average audiovisual enhancement, expressed as proportion correct, as a function of SNR, compared to speech listening only (blue) and the proportion summation model (black). Curves (circles) indicate the enhancement calculated from the average model estimate (average correct score).

Inverse effectiveness – word and subject accuracy

Finally, we tested whether multisensory enhancement correlates negatively with unisensory responsiveness (i.e. A-only thresholds, V-only word and subject recognition rates; rather than stimulus intensity, i.e. SNR), as predicted by the principle of inverse effectiveness. To that end, we determined the multisensory enhancement as the difference in correct scores between the audiovisual and either the auditory, E_{AV-A}, or visual, E_{AV-V}, stimulus, for every word, subject and SNR combination. The slope of the relationship between multisensory enhancement and auditory thresholds or visual recognition rates, respectively, was determined through multiple linear regression analysis:

$$\begin{cases} E_{AV\text{-}A} = \beta_0 - \beta_1\theta_A + \beta_2\rho_{V,w} + \beta_3\rho_{V,s} \\ E_{AV\text{-}V} = \beta_5 - \beta_6\theta_A + \beta_7\rho_{V,w} + \beta_8\rho_{V,s} \end{cases} \quad (2.7)$$

with β_1 the parameter of interest to infer effectiveness of the auditory response, and β_7 and β_8 of the visual response for words and subjects. The other parameters are included to account for confounds such as the effect of the other modality (e.g. the audiovisual enhancement over the auditory response will be negligible if the visual response is minimal). These parameters are an offset to the intercept and reflect the type of integration as shown by the audiovisual data (i.e. super-additive, additive, sub-additive). Note that for the auditory thresholds, the signs are inverted. This ensures that a negative slope would actually indicate inverse effectiveness, even though higher thresholds indicate a worse response.

The audiovisual enhancement over the auditory response (E_{AV-A}, Fig. 2.8A) is larger for words with higher auditory thresholds, with an effectiveness slope $\beta_1 = -0.031$ (95%-HDI: -0.035 to -0.027). The negative slope suggests that the auditory response to each word is inversely effective in driving the multisensory response. The magnitude of the enhancement over the auditory response increases when a word can be more easily recognized through lipreading (i.e. high visual word recognition rate, dark filled dots). This is in line with the observation that the multisensory data follow probability summation quite well, reflecting an additive type of integration (Fig. 2.6 and 2.7). Importantly, the observed inverse effectiveness is not an artefact due to a ceiling effect, as the auditory response allowed for a larger performance benefit (Fig. 2.8A, dotted line).

Multisensory enhancement over the visual response follows the same principles. Words with a low visual recognition rate were more effective at improving the AV response (Fig. 2.8B), with an effectiveness slope $\beta_7 = -0.33$ (95%-HDI: -0.38 to -0.29). Notably, even across subjects, the poorer lipreaders benefit more from audiovisual presentation than excellent lipreaders (Fig. 2.8C), with an effectiveness slope $\beta_8 = -0.42$ (95%-HDI: -0.46 to -0.38).

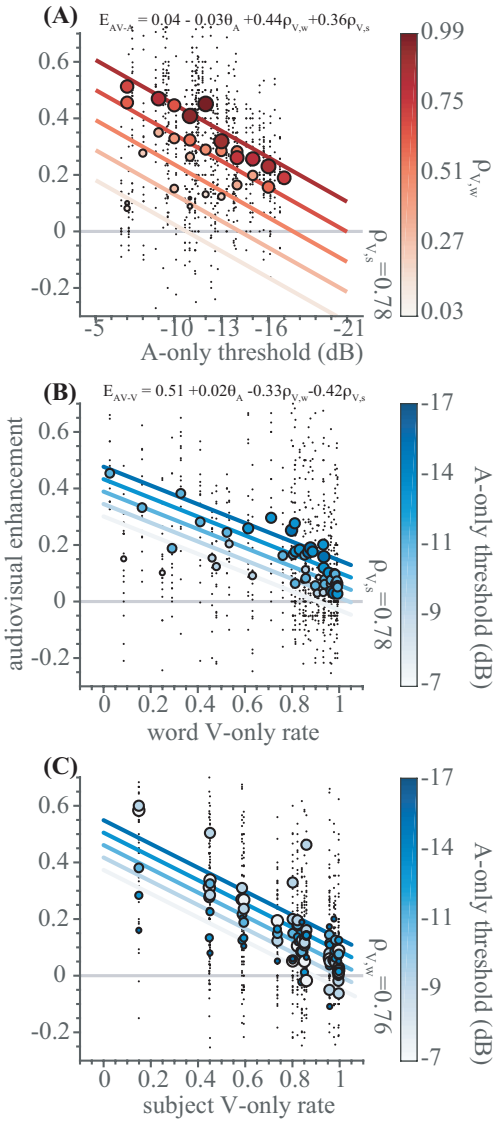


Figure 2.8. Inverse effectiveness.

The audiovisual enhancement over unisensory responses (as defined in the text) as a function of the independent variables **A)** auditory threshold, **B)** visual word recognition rate, **C)** visual subject recognition rate. Note that the x-axis is inverted in **A)**. Black dots indicate the enhancement in correct score for every subject-word-SNR combination. To visualize the effects of the three independent variables on the dependent variable, we binned the variables as follows. The two-dimensional bins were centered on rounded threshold values and for five visual word recognition rates (from the minimum to the maximum rates in equidistant steps) in **A),** and on five auditory thresholds (from the minimum to the maximum thresholds in equidistant steps) and all visual word recognition rate values in **B)** and visual subject recognition rates in **C)**. Circles denote binned average correct scores. Lines indicate the best-fit multiple regression lines for the independent variable of interest (on the abscissa), with intercepts determined by the second, binned variable (indicated by the color bar) and the mean of the third variable (indicated by text). Dot size (color) denotes the cross-sensory performance level (as indicated by the color bars).

DISCUSSION

Overview

This paper reports the occurrence of inverse effectiveness on the recognizability – visually or auditory - of individual words. We determined how well words presented in sentences can be recognized by normal-hearing subjects through listening and/or lipreading under noisy listening conditions. In line with previous research, we found that lipreading improves speech recognition by listening alone (Fig. 2.5A, Fig. 2.6). [4–7] However, we also observed that audiovisual performance levels fall below lipreading performance for the lowest SNR (Fig. 2.5C, Fig. 2.6). Furthermore, we found that the improvements typically saturated at intermediate SNRs, which is largely in line with the principle of inverse effectiveness. We also observed inverse effectiveness across individual words and subjects (Fig. 2.8): the data show that the benefit of adding cross-modal information increased when a word was poorly heard (Fig. 2.8A), when a word was poorly seen (Fig.2.8B), or when the subject was a poor lipreader (Fig. 2.8C).

Performance in lipreading

Our data demonstrate considerable variability in lipreading performance (Fig. 2.2), which has been reported and discussed earlier in the literature.[32] The average performance levels from the current study are relatively high, especially considering that the normal-hearing subjects were not specifically trained to lipread. This is consistent with earlier findings on word and sentence recognition tasks, although more recent papers have reported lower values.[11,27,29,32] One possible explanation for the high lipreading performance might be the use of the closed-set speech-recognition task (i.e. a limited set of words used in a forced-choice behavioral task).

Performance in speech listening

The auditory scores varied mainly across words; subjects could all recognize words through listening at an almost equal performance level (Fig. 2.3). Since all participants had normal hearing, and could therefore be expected to understand speech equally well, the limited variability between subjects corroborated that expectation. The analysis of speech recognition performance in the auditory-only condition revealed speech reception thresholds of -12.1 dB, which is lower than the threshold of -8.4 dB obtained from the original version of the Dutch Matrix test [30].

Models for audiovisual enhancement

The behavioral improvement of audiovisual speech perception can be modeled in various ways. Typically, AV data are compared to the benchmark probability-summation model, in which the auditory and visual channels are considered independent, without true multisensory neural interactions. This model (Eqn. 2.6) matched the data closely (Figs. 2.5 and 2.6).

Rouger and colleagues[29] found that an alternative, optimal-integration model could better describe their data. In their model, spectral-temporal audiovisual cues merge across modalities to optimize the amount of information required for word recognition. Our audiovisual data in poor lipreading conditions (i.e. visual recognition rate for a word is lower than 0.55) compares quite well to the speech-recognition abilities of the normal-hearing subjects of Rouger et al. in the presence of a masking noise (cf. Rouger et al. 2007 - their Fig. 3D).[29]

A third model was proposed by Ma and colleagues, in which words were regarded as points in a multidimensional space, and word recognition becomes a probabilistic inference process.[27] This Bayesian model assumes that certain words occur more frequently than other words (and are more easily recognized), and it uses this pre-knowledge (i.e. priors) to explain the recognition scores for all words.
It is hard to reconcile any of the three models with our observation that in low-SNR conditions, multisensory speech recognition is actually degraded compared to unimodal lipreading without accounting for non-stimulus factors affecting audiovisual speech recognition (Figs. 2.4C and 2.5). The aforementioned models do not include a mechanism for divided attention between the two modalities.[33,34] In such a scheme, the two separate information streams could actually lead to impaired performance in conditions in which either of the two signals may be ambiguous or weak. Thus, even though lipreading might provide sufficient information to recognize words, people are not able to divert their attention away from the auditory stream, despite the absence of a potential signal in that information stream.

Inverse effectiveness
We tested whether the principle of inverse effectiveness also holds in audiovisual speech recognition by: i) modulating the acoustic signals related to background noise, ii) by investigating each subject's lipreading ability, and iii) by comparing to auditory and/or visual recognizability of words.

First, in line with several laboratory studies of multisensory integration using simple sensory stimuli (e.g. white noise bursts and LED flashes), a lower auditory SNR typically induced stronger multisensory enhancement.[16-25] However, here we report that for the lowest SNRs (-21 dB) the enhancement saturated, or even slightly dropped (Fig. 2.7C). This differs quantitatively with the data from Ma et al., who found a significant enhancement drop for low SNRs.[27] Notably, however, Bayesian modelling of audiovisual enhancement in the study by Ma et al. suggested that the largest enhancement shifted to lower SNRs with decreasing vocabulary size. As the vocabulary size in the current experiment was limited to only 50 words (with only 10 possible choices per word category), the model by Ma et al. would also predict the largest enhancement at the lowest SNRs.

Secondly, evidence for inverse effectiveness can be found for individual lipreading abilities; worse lipreaders benefited more from the additional auditory information for the audiovisually presented sentences (Fig. 2.8C). Finally, inverse effectiveness also plays a role at word-level performance, both for vision and for hearing: the hardest to-recognize words exhibited the strongest audiovisual enhancements relative to the unimodal condition (Fig. 2.8). As such, this type of inverse effectiveness found is in line with basic multisensory integration results from earlier studies using stimuli with low-level features (simple noise bursts and LED flashes) and for studies using slightly more complex, spectro-temporally modulating stimuli, but likely also involves a wide network of high-level feature processing (features such as word frequency, familiarity, audiovisual co-occurrence, task constraints; see also the limitation of this study in determining these effects in the following section).[26]

Matrix test

The audiovisual speech material is based on an existing auditory-only matrix sentence test for Dutch native speakers.[30,35] It is not immediately clear whether the observed results hold specifically for the Dutch language, or whether it is immaterial for which language this test has been developed. Numerous audiovisual speech recognition tests have been developed for the English language, with exceptions for native French and Dutch speakers.[4,9,11,13,27,29,36-38] Detailed comparisons are difficult also because the stimuli (monosyllables vs words vs sentences) and the subject populations (normal-hearing vs hearing-impaired) differ. The use of a standardized test, such as the Matrix test, might facilitate comparisons, especially between normal-hearing and hearing-impaired listeners, since the Matrix test is also well-suited to test the hearing-impaired. Comparisons across languages might still be difficult, as, even though an auditory Matrix test is available in many languages, the words may vary in their spectro-temporal properties and thresholds between languages.[30,39-41]

Note that the use of this standardized Matrix test, that was constructed with the intention to evaluate hearing-impaired, includes words that are quite common and that are familiar to the subjects. The dependence of word recognition on higher-level factors beyond the low-level processing of spectro-temporal or articulatory stimulus representation is therefore hard – if not impossible - to determine with these speech materials.

Conclusion

To conclude, lipreading enhances speech recognition (in line with earlier studies); this visual enhancement, however, is affected by the acoustic properties of the audiovisual scene. Visual enhancement for words that are easily recognized by vision alone is impoverished in high acoustic noise conditions. Audiovisual enhancements were highest for intermediate signal-to-noise ratios. Inverse effectiveness holds for words and subjects, for which the poorest visually/auditory-recognizable words underwent the strongest cross-modal enhancements.

MATERIALS AND METHODS

Participants

Eighteen native Dutch-speaking adults (mean age = 26 years, range = 21-40) participated in this study. All gave their informed consent. They were screened for normal-hearing (within 20 dB HL range 0.5 - 8 kHz), and had normal or corrected-to-normal vision. The experiments were carried out in accordance with the relevant institutional and national regulations and with the World Medical Association Helsinki Declaration as revised in March 2017 (https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects). The experiments were approved by the Ethics Committee of Arnhem-Nijmegen (project number NL24364.091.08, October 18, 2011). Written informed consent was obtained before each experiment.

Audiovisual material

The speech material was based on the Dutch version of the speech-in-noise matrix test developed by Houben and colleagues in analogy to a Swedish test.[30-39] In general, a matrix test uses complete sentences that are composed from a fixed matrix of words (Table 2.1). All created sentences shared the same grammatical structure (name, verb, numeral, adjective, object), but were semantically unpredictable. In principle, a set of 10^5 different sentences could be created. Therefore, the test suffered little from potential training confounds when participants were tested multiple times. Houben et al., ensured that the occurrence of phonemes in their test was similar to standard Dutch.[30] For the audiovisual version of the test reported here, we selected a subset of 180 (155 unique) sentences that were grouped into 9 lists of 20 sentences each. In every list, each of the 50 words from the matrix occurred twice, once in the first ten sentences and once in the second ten sentences.

The audio-video material was recorded in a sound-attenuated, semi-anechoic room, using an Olympus LS-5 audio recorder (24-bit/44.1 kHz sampling rate), and a Canon 60D video camera (1280 x 720, 720p HD at 50 frames per second), respectively. All sentences were spoken by a Dutch female speech therapist. If a sentence was not articulated clearly, or if there was a sudden movement of the face or eyes, the sentence was re-recorded. The audio and video recordings were combined off-line using Final Cut Pro X (Mac App OS X Yosemite), and saved in MPEG-4 format, in H.264 codec.

Experimental setup

Audiovisual testing was carried out in the same room in which the material had been recorded. Stimulus presentation was controlled by a Dell PC (Dell Inc., Round Rock, TX, USA) running Matlab version 2014b (The Mathworks, Natick, MA, USA). Participants were seated at a table, 1.0 m in front of a PC screen (Dell LCD monitor, model: E2314Hf, Dell Inc., Texas,

USA). Sounds were played through an external PC sound card (Babyface, RME, Germany) and presented over one speaker (Control Model Series, model number: Control One, JBL, California, USA) placed 1.0 m in front of the participant, immediately above the screen (30° above the interaural plane). Speaker output was calibrated with an ISO-TECH Sound Level Meter (type SLM 1352P) at the position of the listener's head, on the basis of the stationary masking noise.

Stimuli

The stimuli contained digital video recordings of a female speaker reading aloud the sentences in Dutch (Fig. 2.1). In the auditory-only presentation (A-only), the voice was presented without visual input (i.e. black screen, Fig. 2.1A,C) with added background acoustic noise (Fig. 2.1B). In the visual-only presentation (V-only) the video fragments of the female speaker were shown on the screen without an auditory speech signal and noise (Fig. 2.1D). In the audiovisual presentation (AV), the video was presented with the corresponding auditory signal and the masking noise.

The masking noise was created following the procedure reported by Wagener et al.[42] To that end, the 180 sentences were overlaid by applying a random circular shift. Repeating that procedure five times resulted in a stationary masking noise with the same spectral characteristics as the original speech material.

Paradigm

All participants were tested in a closed-set speech-recognition test in A-only, V-only and AV conditions. Prior to the experiment, all participants familiarized themselves with the matrix of 50 words (10 words for each of the 5 categories, Table 2.1) and by practicing the task on 10 randomly selected AV sentences. No improvement in speech recognition was observed during the experimental sessions, which indicates that there was no recognition effect of procedural learning.

The masking noise started and ended 500 ms before and after the sentence presentation. The noise onset and offset included 250 ms ($\sin^2$, $\cos^2$) ramps. In the A-only and AV conditions, the masking noise was fixed at 65 dB SPL (A-weighted), with the speech sound presented at 44, 49, 52, 55, or 60 dB SPL (A-weighted) to obtain signal-to-noise ratios (SNRs) of -21, -16, -13, -10, and -5 dB, respectively. After presentation of the sentence and the end of the noise, the matrix of 50 words was shown on the screen (Table 2.1). Participants were instructed to choose one word from each of the 5 categories (10-alternative forced-choice task). Participants initiated the next trial by pressing the mouse-button.

For each of the sensory modalities (A-only, V-only, and AV), participants were tested in separate sessions on different days. In this way, fatigue and repetitive stimulus presentation

were avoided. In each session, the nine lists of 20 sentences were presented. In the A-only and AV sessions, each sentence was assigned one of the five SNRs pseudo-randomly (each SNR was presented equally often as the others, i.e. 36 times in each session).

Data analysis

For every word (w=1:50), subject (s=1:18), SNR (n=1:5) and sensory modality (m=1:3), we determined the correct score, defined as the number of correct responses, z, divided by the number of presentations, N. The correct score, P(correct), is assumed to be binomially distributed, in which the probability of a success is given by:

$$P(\text{correct}) \sim \text{Binomial}((1-\gamma) \times F(\Psi) + \gamma, N) \tag{2.8}$$

where $F(\Psi)$ is a function that characterizes the recognition performance for the particular stimulus and subject parameters (subject parameters such as SNR and visual recognition rate), described by Ψ; γ is the probability that the subject gives the correct answer, irrespective of the stimulus (the 'guess rate'); $(1-\gamma) \times F(\Psi)+\gamma$ is the probability of success; N is the number of trials; and Binomial denotes the binomial distribution. Here, γ was set to 10% (0.1), as there were ten word alternatives per category. We estimated model parameters Ψ, e.g. the recognition rates, ρ (i.e. how often words were recognized correctly at a given SNR) and the recognition thresholds, θ (i.e. the SNR at which words were recognized in 50% of the presentations), as described in the Results section (eqn. 2.1-2.3).

Statistical Analysis

Parameter estimation of Eqns. 2.1-2.8 was performed using a Bayesian statistical analysis. This analysis requires the definition of priors over the parameters. As a prior for the auditory thresholds, we chose the Gaussian distribution with mean 0 and standard deviation 100, and for the visual recognition rates we took a positive-only beta distribution, for which both shape parameters were set to 1. The audiovisual rate differences (Eqn. 2.5) were modeled as Gaussian distributions with the rates transformed to probit scale (see e.g. (Lee and Wagenmakers 2014, Chapter 9.3).[44] For the multiple linear regression (eqn. 2.7), the data was modeled according to a t-distribution.[45] For the priors on the parameters, Gaussian distributions with a mean of 0 and a standard deviation of 2 were chosen, after normalization of the data.

The estimation procedure relied on Markov Chain Monte Carlo (MCMC) techniques. The estimation algorithms were implemented in JAGS through matJAGS.[46,47] Three MCMC chains of 10,000 samples were generated. The first 10,000 samples were discarded as burn-in. Convergence of the chains was determined visually, by checking that the shrink factor $\hat{R} < 1.1$, and by checking that the effective sample size >1000.[48-50]

From these samples of the posterior distributions, we determined the mean and the 95% highest density interval (95%-HDI) as a centroid and uncertainty estimate of the parameters, respectively.

Model Selection

To test for the appropriateness of the models in eqns. 2.1-2.3, we compared them against less-restrictive models, including fully independent models. To that end, we determined the BIC for the models:

$$BIC = k \ln(n) - 2 \ln(\hat{L}) \quad (2.9)$$

where k denotes the number of parameters of the model (e.g. 68 for eqn. 2.1 and 900 for a fully-independent V-only model), n the number of samples (e.g. 900 for the V-only data), and $\hat{L}$ the maximized value of the binomial likelihood function (e.g. for those $\rho_{V,w}$, and $\rho_{V,s}$. that maximize the likelihood function for the V-only data at hand). The model with the lowest BIC is the preferred model. An alternative model-selection criterion, the Akaike Information Criterion (which contains a smaller penalty term for the number of parameters) yielded the same model selections.

ACKNOWLEDGMENTS

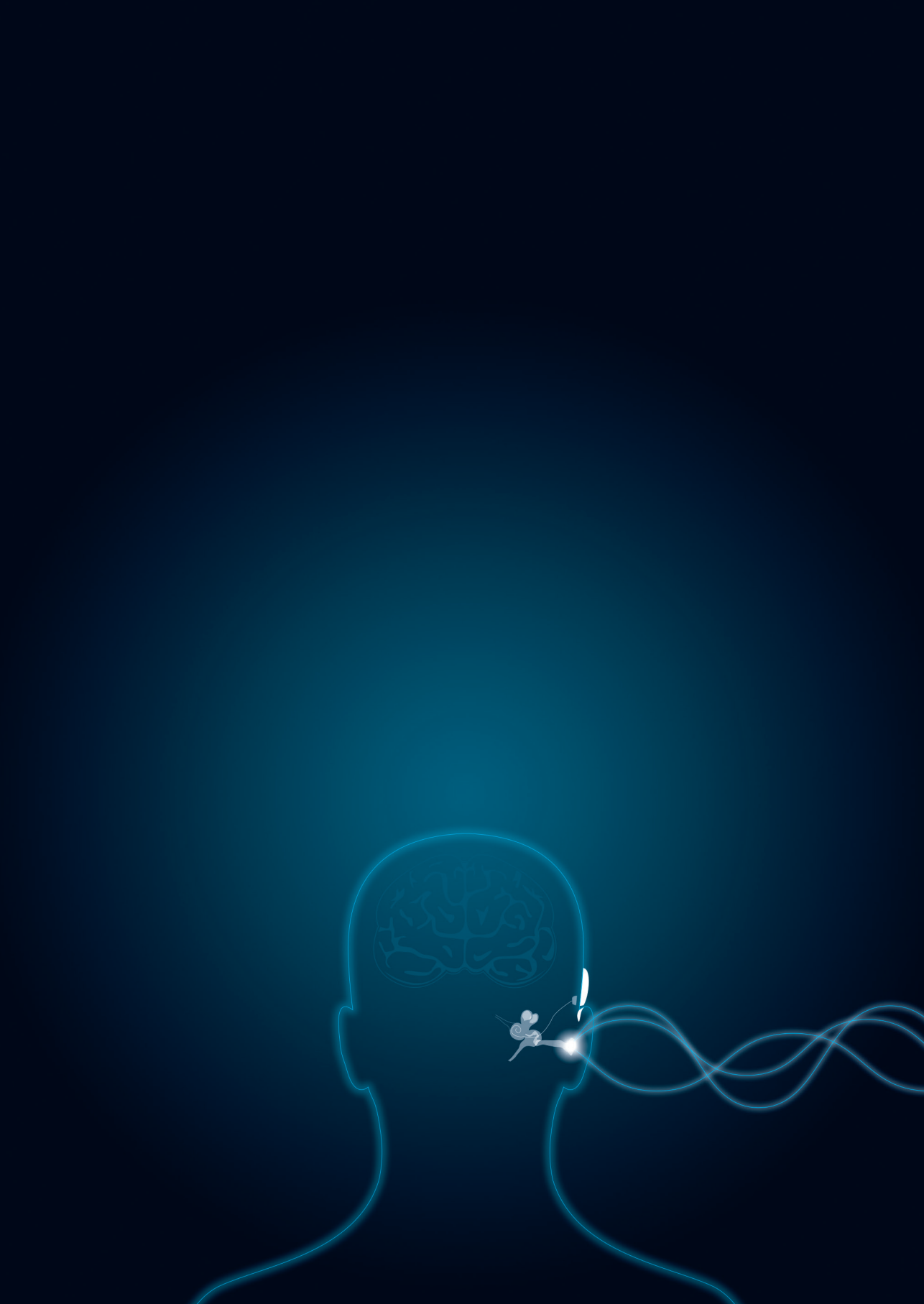
We thank Günther Windau, Ruurd Lof, Stijn Martens, and Chris-Jan Beerendonck for their valuable technical assistance, speech-therapist Jeanne van der Stappen for providing the audiovisual material, and Eefke Lemmen for editing. We are grateful to Ad Snik for providing valuable comments on earlier versions of this manuscript.

REFERENCES

1. van de Rijt LPH, van Opstal AJ, Mylanus EAM, Straatman LV, Hu HY, Snik AFM, van Wanrooij MM. Temporal Cortex Activation to Audiovisual Speech in Normal-Hearing and Cochlear Implant Users Measured with Functional Near-Infrared Spectroscopy. Front Hum Neurosci. 2016;10: 48. doi:10.3389/fnhum.2016.00048
2. Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol. 2000;10: 649–57. doi:10.1016/S0960-9822(00)00513-3
3. Stein BE, Meredith MA. The Merging of the Senses. Cambridge, MA, US: The MIT Press.; 1993.
4. Bernstein LE, Auer ET, Takayanagi S. Auditory speech detection in noise enhanced by lipreading. Speech Commun. 2004;44: 5–18. doi:10.1016/j.specom.2004.10.011
5. Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am. 2000;108: 1197–208. doi:10.1121/1.422512
6. Helfer KS. Auditory and auditory-visual perception of clear and conversational speech. J speech, Lang Hear Res. 1997;40: 432–43.
7. Winn MB, Rhone AE, Chatterjee M, Idsardi WJ. The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. Front Psychol. Frontiers; 2013;4: 824. doi:10.3389/fpsyg.2013.00824
8. MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. Br J Audiol. 1987;21: 131–41.
9. MacLeod A, Summerfield Q. A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. Br J Audiol. 1990;24: 29–43. doi:10.3109/03005369009077840
10. O'Neill JJ. Contributions of the visual components of oral symbols to speech comprehension. J Speech Hear Disord. American Speech-Language-Hearing Association; 1954;19: 429–439. doi:10.1044/jshd.1904.429
11. Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex. 2007;17: 1147–53. doi:10.1093/cercor/bhl024
12. Sommers MS, Tye-Murray N, Spehar B. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. Ear Hear. 2005;26: 263–75. doi:10.1097/00003446-200506000-00003
13. Sumby WH, Pollack I. Visual Contribution to Speech Intelligibility in Noise. J Acoust Soc Am. 1954;26: 212–215. doi:10.1121/1.1907309
14. Tye-Murray N, Sommers MS, Spehar B. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. Ear Hear. 2007;28: 656–68. doi:10.1097/AUD.0b013e31812f7185
15. Tye-Murray N, Sommers M, Spehar B, Myerson J, Hale S. Aging, Audiovisual Integration, and the Principle of Inverse Effectiveness. Ear Hear. 2010;31: 1. doi:10.1097/AUD.0b013e3181ddf7ff
16. Corneil BD, van Wanrooij MM, Munoz DP, van Opstal AJ. Auditory-visual interactions subserving goal-directed saccades in a complex scene. J Neurophysiol. 2002;88: 438–54. doi:10.1152/jn.2002.88.1.438
17. van Barneveld DCPBM, van Wanrooij MM. The influence of static eye and head position on the ventriloquist effect. Eur J Neurosci. 2013;37: 1501–10. doi:10.1111/ejn.12176
18. Alais D, Burr D. No direction-specific bimodal facilitation for audiovisual motion detection. Brain Res Cogn Brain Res. 2004;19: 185–94. doi:10.1016/j.cogbrainres.2003.11.011
19. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. Causal inference in multisensory perception. Sporns O, editor. PLoS One. Public Library of Science; 2007;2: e943. doi:10.1371/journal.pone.0000943

20. Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA. Unifying multisensory signals across time and space. Exp brain Res. 2004;158: 252–8. doi:10.1007/s00221-004-1899-9
21. Bell AH, Meredith MA, van Opstal AJ, Munoz DP. Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. J Neurophysiol. 2005;93: 3659–73. doi:10.1152/jn.01214.2004
22. Meredith MA, Stein BE. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Res. 1986;365: 350–4. doi:10.1016/0006-8993(86)91648-3
23. Wallace MT, Meredith MA, Stein BE. Multisensory integration in the superior colliculus of the alert cat. J Neurophysiol. 1998;80: 1006–10. doi:10.1152/jn.1998.80.2.1006
24. Bremen P, Massoudi R, van Wanrooij MM, van Opstal AJ. Audio-Visual Integration in a Redundant Target Paradigm: A Comparison between Rhesus Macaque and Man. Front Syst Neurosci. Frontiers; 2017;11: 89. doi:10.3389/fnsys.2017.00089
25. Frens MA, van Opstal AJ, van der Willigen RF. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Percept Psychophys. 1995;57: 802–16. doi:10.3758/BF03206796
26. van Wanrooij MM, Bell AH, Munoz DP, van Opstal AJ. The effect of spatial-temporal audiovisual disparities on saccades in a complex scene. Exp brain Res. Springer-Verlag; 2009;198: 425–437. doi:10.1007/s00221-009-1815-4
27. Ma WJ, Zhou X, Ross LA, Foxe JJ, Parra LC. Lipreading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. PLoS One. 2009;4: e4638. doi:10.1371/journal.pone.0004638
28. Tye-Mmurray N, Spehar B, Myerson J, Hale S, Sommers M. Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. Psychol Aging. 2016;31: 380–389. doi:10.1037/pag0000094
29. Rouger J, Lagleyre S, Fraysse B, Deneve S, Deguine O, Barone P. Evidence that cochlear-implanted deaf patients are better multisensory integrators. Proc Natl Acad Sci U S A. 2007;104: 7295–300. doi:10.1073/pnas.0609419104
30. Houben R, Koopman J, Luts H, Wagener KC, van Wieringen A, Verschuure H, et al. Development of a Dutch matrix sentence test to assess speech intelligibility in noise. Int J Audiol. Taylor & Francis; 2014;53: 760–3. doi:10.3109/14992027.2014.920111
31. Kuss M, Jäkel F, Wichmann FA. Bayesian inference for psychometric functions. J Vis. 2005;5: 478–92. doi:10.1167/5.5.8
32. Bernstein LE, Demorest ME, Tucker PE. Speech perception without hearing. Percept Psychophys. 2000;62: 233–52. doi:10.3758/BF03205546
33. Alsius A, Navarra J, Campbell R, Soto-Faraco S. Audiovisual Integration of Speech Falters under High Attention Demands. Curr Biol. Elsevier; 2005;15: 839–843. doi:10.1016/j.cub.2005.03.046
34. Bonnel AM, Hafter ER. Divided attention between simultaneous auditory and visual signals. Percept Psychophys. Springer-Verlag; 1998;60: 179–90. doi:10.3758/BF03206027
35. Stein BE, Stanford TR, Ramachandran R, Perrault TJ, Rowland BA. Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. Exp Brain Res. 2009;198: 113–126. doi:10.1007/s00221-009-1880-8
36. Holmes NP. The principle of inverse effectiveness in multisensory integration: Some statistical considerations. Brain Topography. Springer US; 2009. pp. 168–176. doi:10.1007/s10548-009-0097-2
37. Houben R, Dreschler WA. Optimization of the Dutch matrix test by random selection of sentences from a preselected subset. Trends Hear. 2015;19: 233121651558313. doi:10.1177/2331216515583138
38. Stevenson RA, Nelms CE, Baum SH, Zurkovsky L, Barense MD, Newhouse PA, et al. Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. Neurobiol Aging. Elsevier Ltd; 2015;36: 283–91. doi:10.1016/j.neurobiolaging.2014.08.003

39. Anderson Gosselin P, Gagné J. Older adults expend more listening effort than young adults recognizing speech in noise. J Speech Lang Hear Res. 2011;54: 944–58. doi:10.1044/1092-4388(2010/10-0069)
40. Middelweerd MJ, Plomp R. The effect of speechreading on the speech-reception threshold of sentences in noise. J Acoust Soc Am. 1987;82: 2145–7. doi:10.1121/1.395659
41. Hagerman B. Sentences for testing speech intelligibility in noise. Scand Audiol. Taylor & Francis; 1982;11: 79–87. doi:10.3109/01050398209076203
42. Hochmuth S, Brand T, Zokoll MA, Castro FZ, Wardenga N, Kollmeier B. A Spanish matrix sentence test for assessing speech reception thresholds in noise. Int J Audiol. 2012;51: 536–44. doi:10.3109/14992027.2012.670731
43. Ozimek E, Warzybok A, Kutzner D. Polish sentence matrix test for speech intelligibility measurement in noise. Int J Audiol. 2010;49: 444–454. doi:10.3109/14992021003681030
44. Wagener K, Josvassen JL, Ardenkjaer R. Design, optimization and evaluation of a Danish sentence test in noise. Int J Audiol. Taylor & Francis; 2003;42: 10–7. doi:10.3109/14992020309056080
45. Lee MD, Wagenmakers E-J. Bayesian cognitive modeling: A practical course. Cambridge University Press, 2014. Cambridge University Press, New York; 2014.
46. Kruschke JK. Doing Bayesian Data Analysis. 2nd ed. Elsevier; 2015. doi:10.1016/C2012-0-00477-2
47. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Hornik K, Leisch F, Zeileis A, editors. Proceedings of the 3rd Internaitional Workshop on Disbtributed Statistical Computing. Vienna, Austria; 2003. Available: http://mcmc-jags.sourceforge.net
48. Turner BM, Forstmann BU, Wagenmakers E-J, Brown SD, Sederberg PB, Steyvers M. A Bayesian framework for simultaneously modeling neural and behavioral data. Neuroimage. Elsevier Inc.; 2013;72: 193–206. doi:10.1016/j.neuroimage.2013.01.048
49. Brooks SP, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. J Comput Graph Stat. Taylor & Francis; 1998;7: 434–455. doi:10.1080/10618600.1998.10474787
50. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis, Third Edition (Texts in Statistical Science). Book. Chapman and Hall/CRC; 2013.
51. Kass RE, Carlin BP, Gelman A, Neal RM. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. Am Stat. 1998;52: 93. doi:10.2307/2685466

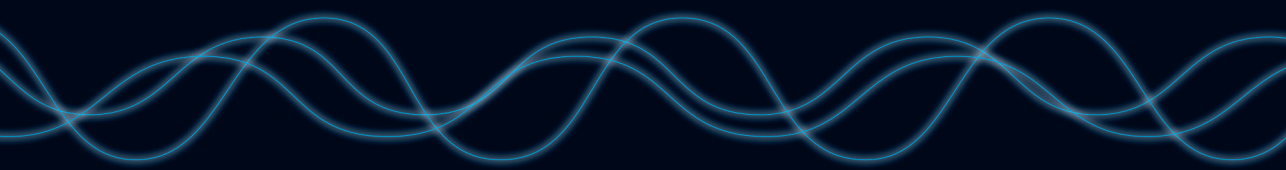


CHAPTER 3

Multisensory integration-attention trade-off in cochlear-implanted deaf individuals

Luuk P.H. van de Rijt, A. John van Opstal[b], Marc M. van Wanrooij[b,1]

Key words: cochlear implant, multisensory integration, speech perception, focused attention, divided attention



ABSTRACT

The cochlear implant (CI) allows profoundly deaf individuals to partially recover hearing. Still, due to the coarse acoustic information provided by the implant, CI users have considerable difficulties in recognizing speech, especially in noisy environments. CI users therefore rely heavily on visual cues to augment speech comprehension, more so than normal-hearing individuals. However, it is unknown how attention to one (focused) or both (divided) modalities plays a role in multisensory speech recognition. Here we show that unisensory speech listening and reading were negatively impacted in divided-attention tasks for CI users - but not for normal-hearing individuals. Our psychophysical experiments revealed that, as expected, listening thresholds were consistently better for the normal-hearing, while lipreading thresholds were largely similar for the two groups. Moreover, audiovisual speech recognition for normal-hearing individuals could be described well by probabilistic summation of auditory and visual speech recognition, while CI users were better integrators than expected from statistical facilitation alone. Our results suggest that this benefit in integration comes at a cost. Unisensory speech recognition is degraded for CI users when attention needs to be divided across modalities. We conjecture that CI users exhibit an integration-attention trade-off. They focus solely on a single modality during focused-attention tasks, but need to divide their limited attentional resources in situations with uncertainty about the upcoming stimulus modality. We argue that in order to determine the benefit of a CI for speech comprehension, situational factors need to be discounted by presenting speech in realistic or complex audiovisual environments.

SIGNIFICANCE STATEMENT

Deaf individuals using a cochlear implant require significant amounts of effort to listen in noisy environments due to their impoverished hearing. Lipreading can benefit them and reduce the burden of listening by providing an additional source of information. Here we show that the improved speech recognition for audiovisual stimulation comes at a cost, however, as the cochlear-implant users now need to listen and speech-read simultaneously, paying attention to both modalities. The data suggests that cochlear-implant users run into the limits of their attentional resources, and we argue that they, unlike normal-hearing individuals, always need to consider whether a multisensory benefit outweighs the unisensory cost in everyday environments.

INTRODUCTION

Speech comprehension is a challenging task. First, the speech signal itself might be hard to recognize due to poor pronunciation, semantic ambiguities and highly variable and rapid articulation rates (>200 words/min[1]). Second, in common everyday environments, even highly salient speech signals are frequently embedded in acoustic background noise and are masked by other talkers. During face-to-face conversation, non-acoustic cues from seeing a talker's mouth can improve speech recognition in those situations, through the integration of visual and auditory information[2–6].

Multisensory integration is beneficial for normal-hearing and normally sighted individuals, whenever multisensory stimuli are in spatial-temporal congruence. The effects of audiovisual integration include behavioral benefits such as shorter response-reaction times[7–9], increased accuracy and precision[7,10], better selection, and reduced ambiguity[11]. At the neuronal level, these effects are typically reflected by enhanced activity[9,12,13]. This also applies to more complex auditory stimuli; supplemental visual input enhances speech perception, and audiovisual speech recognition embedded in noise is considerably better than for auditory speech alone[14–17]. The necessity to integrate non-acoustic information to improve performance becomes especially clear for individuals with hearing impairments, such as profoundly deaf individuals using a cochlear implant (CI). The CI typically recovers hearing to an extent that allows the CI user to understand speech in quiet situations, yet creates significant problems under more challenging listening conditions (e.g., noisy surroundings). In these cases, the CI user should rely more on the information obtained from lip reading. Evidence suggests that CI users are indeed better able to integrate visual information with the perturbed acoustic information than normal-hearing individuals[18,19].

Due to all the observed benefits of multisensory integration, one may forget that it requires paying attention to multiple sensory modalities at the same time. Attention is a neural mechanism by which the brain is able to effectively select a relevant signal from a multitude of competing sources (e.g., finding someone with a red coat in a busy street). When attention is fully focused on a particular sensory modality, say auditory, performance in auditory selection tasks will markedly increase, but visual stimuli will likely be missed, because attention has limited capacity. The opposite occurs when attention is focused on vision. In natural environments, however, the most relevant sensory modality of a potential target may not be known in advance, and therefore focusing attention on a single sensory modality may not be an optimal strategy to maximize perceptual performance. Instead, in such cases, attention should be divided across the relevant modalities. In case of speech perception, these modalities are auditory (listening) and visual (lipreading) signals. Dividing attention

across modalities will allow the brain to integrate the multimodal signals when they originate from the same source, and filter out the perturbing background from unrelated sources.

However, because of its limited capacity, dividing attention in an uncertain sensory environment may lead to decreased performance for stimuli that happen to be unisensory, as each modality will receive less attentional amplification than during a fully focused attention task. Here we compared word-recognition performance during focused and divided attention tasks for CI users and normal-hearing individuals, by presenting unisensory and/or bi-sensory spoken sentences in different sensory-noise regimes. Because CI users have more difficulty to process the perturbed auditory input, more effort (i.e., more attention) will be required to understand auditory speech. Therefore, we reasoned that in a divided-attention task, the lapse in attention to audition (and vision) may lead to poorer unisensory performance scores in CI users. In principle, the same reasoning may hold for normal-hearing participants. So far, it remains unclear from the literature whether CI users can successfully divide their attention across modalities, and whether divided attention affects their speech-comprehension abilities.

RESULTS

Overview

Fourteen normal-hearing participants and seven post-lingually deaf unilateral implanted CI users had to identify 50 words (see Methods), presented in 155 unique five-word sentences, by selecting the words they recognized (10-alternative, open-ended choice) on a screen. The speech material has been used in a previous study[20], in which further details about the material can be found. The stimuli were either presented in two separate unisensory, focused-attention blocks, or in one divided-attention block. We varied task difficulty in both experiments, by blurring the video, and by presenting acoustic background noise at several levels.

In the focused-attention experiment (Fig. 3.1; purple), the sentences were either presented in an acoustic-only block (Fig. 3.1A,C, purple circles), or in a visual-only block (Fig. 3.1B,D, purple circles), in which the participant could focus solely on listening or lipreading, respectively. In the divided-attention experiment auditory (Fig. 3.1A,C, green diamonds), visual (Fig. 3.1B,D, green diamonds) and audiovisual (Fig. 3.2) sentences were presented in pseudo-random order, all interleaved in one block. In this task, participants were free to focus on one modality, or to divide attention across both modalities.

To estimate parameters of interest, such as the signal-to-noise ratio and blur at which performance level was 50% and (attentional) lapse probabilities, we fitted psychophysical-function curves through the data (as fully explained in the Methods). We report on the mean

and 95%-highest-density interval (HDI) of the fitted estimate distributions of the group-level parameters, and show both the fitted curves for each group and the data averaged across participants in the figures.

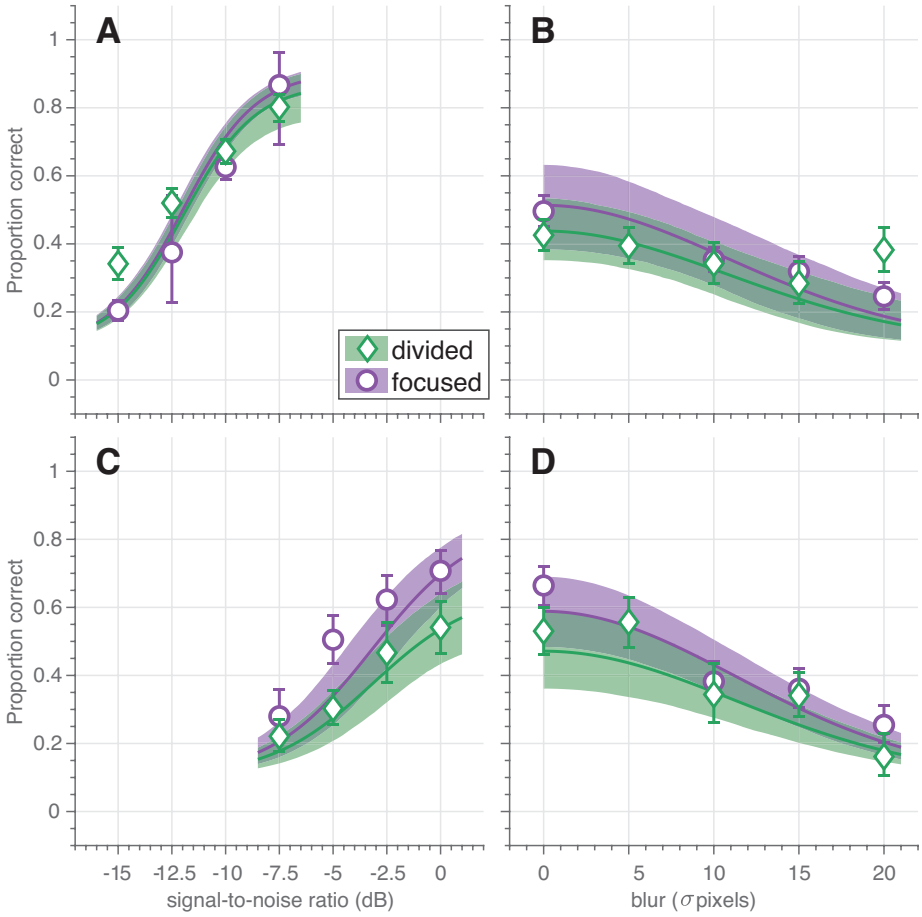


Figure 3.1. Unisensory speech recognition.

A,C) Auditory-only speech recognition (proportion correct) as a function of signal-to-noise ratio (dB) for **A)** normal-hearing participants (n=14) and **C)** CI users (n=7) in the focused- (purple circles) and divided-attention (green diamonds) tasks. Note that although the unisensory stimuli were the same for both tasks, CI users recognized more auditory words correctly in the focused-attention task (purple) than in the divided-attention task (green). This effect was absent for the normal-hearing participants. **B,D)** Visual-only speech recognition as a function of spatial blur (in units of pixel standard deviations) for **B)** normal-hearing participants and **D)** CI users in the focused- (purple circles) and divided-attention (green diamonds) tasks. Note that due to the large similarity in visual recognition scores for both tasks, a psychometric curve was fitted through the combined data (black curve and patch). Symbols and bars indicate mean and 95%-confidence intervals, respectively, of the raw data (proportion correct) pooled across participants. Curves and patches indicate means and 95%-HDI, respectively, of the psychophysical-function group-level fits.

Unisensory Speech Perception

When sentences were presented only acoustically (Fig. 3.1A,C), the two groups clearly differed in their ability to recognize words, as expected. Typically, the normal-hearing participants (Fig. 3.1A) recognized 50% of the words correctly in the unisensory hearing condition at a signal-to-noise ratio (auditory threshold, Eqn. 3.1) of -12 dB (HDI = [-12.4, -11.5] dB) vs. -3.1 dB for the CI users (HDI = [-4.4, -1.7] dB, Fig. 3.1C) for either of the tasks (green and purple). For both groups, the proportion of correctly recognized words strongly depended on the actual signal-to-noise ratio; to increase performance levels from 5%- to 95%-word recognition (psychometric curve width), the signal-to-noise ratio needed to be increased by 7.4 dB on average for the normal-hearing participants (HDI = [6.5, 8.5] dB; Fig. 3.1A) and slightly more for CI users by on average 10.4 dB (HDI = [8.8, 12.2 dB; Fig. 3.1C). As expected, both these results confirm that listening for CI users is considerably more difficult than for normal-hearing participants.

The parameter of main interest in this study is the lapse probability (Eqn. 3.3), i.e. the probability of not recognizing words even at the highest signal-to-noise ratio and without blur. Lapses occurred even in the focused-attention task as evidenced by the non-perfect performance at the highest signal-to-noise ratios; the average performance of normal-hearing participants and CI users saturated at around 90 and 84% correct, respectively (Fig. 3.1A,C, purple; HDI = [85, 94] and [74, 92%). A larger lapse probability for the CI users compared to the lapse probability for the normal-hearing participants may be expected due to technical limits of the cochlear implant and the maximal comfortable loudness levels experienced by the CI users, but note that evidence for any difference was actually small (mean 5%, HDI = [-5, 17]%).

More importantly and more clearly, in the divided-attention task the CI users recognized 22% (HDI = [6.7, 38]%) fewer words than in the focused attention task (Fig. 3.1C, green vs purple). This difference was not clearly evident for the normal-hearing participants (mean difference 3.9%, HDI = [-4.0, 14] %; Fig. 3.1A, green vs purple). Evidence for group differences in auditory lapse probability during the divided-attention experiment was substantial (on average, the lapse probability for normal-hearing participants was 24% lower than for the CI users, HDI = [8, 41] %).

When sentences were presented only visually (Fig. 3.1B,D), the proportion of correctly recognized words depended on the amount of blur, and were largely similar for both groups; the visual threshold (i.e. the blur at 36% of the maximal lipreading performance, Eqn. 3.2 was on average 17.7 and 18.3 pixels for CI users (Fig. 3.1D) and normal-hearing (Fig. 3.1B) participants, respectively (HDI = [16.0,19.7] and [15.2, 21.8] pixels, respectively) for both tasks. Of course, lipreading abilities were far from perfect even without blurring.

No major difference in lipreading performance was observed for the visual lapse probability, so we pooled the data from both tasks to estimate this parameter. Normal-hearing participants (Fig. 3.1B) had a lapse in word recognition in 54% of the cases (HDI = [42, 65]%), while CI users (Fig. 3.1D) incorrectly recognized unblurred visual words in 46% of the cases (HDI = [36, 56]%). While one may expect CI users to be better lip-readers than normal-hearing participants, differences between groups were actually small (on average 8%, HDI = [-8, 23]%).

In summary, largely in contrast to the normal-hearing participants, the CI users experienced more speech-recognition problems when attention had to be divided between more than one sensory modality. These problems were especially conspicuous for listening, the sensory modality that faced the largest difficulties for the CI users.

Multisensory Integration

We next analyzed whether speech perception of audiovisual stimuli would be enhanced for both groups of participants in the divided-attention task (Fig. 3.2). Figs. 3.2A and B show examples of individual participants (NH3 and CI4) in the divided-attention task at a visual blur of 10 pixels. The unisensory data and fits for these two participants (Figs. 3.2A,B brown and green for lipreading and listening, respectively) are in line with the group-level data and fits as described in the previous section (cf. Fig. 3.1, green). The audiovisual speech recognition (Fig. 3.2A, blue and Fig. 3.2B red for NH3 and CI4, respectively) outperforms or equals either unimodal speech recognition; for very low and high signal-to-noise ratios, audiovisual performance tends to equal visual or auditory performance. For intermediate signal-to-noise ratios, audiovisual performance is clearly enhanced. Such an enhancement of multisensory performance could potentially be due to mere statistical facilitation, if the participants would recognize a word by using either the available auditory, or visual information, without actually integrating both inputs. The percept is then determined by whichever sensory channel wins the race (probability summation)[9,13,20]. The audiovisual enhancement would then be fully determined by the unisensory auditory and visual recognition performance during the divided-attention task. To check for this possibility, we compared the data to the prediction from this probability-summation model (Fig. 3.2A,B, black curve, see Methods). For the normal-hearing participant (Fig. 3.2A; cf. black markers and blue curve), the model's prediction corresponded quite well to the data. Hence, despite the improvement in audiovisual recognition rates, the normal-hearing participant did not seem to benefit from multisensory integration. In contrast, although the CI user evidently had difficulty to recognize a pure auditory speech signal in the multisensory divided-attention task (Fig. 3.2B, green; note the increased threshold and the larger lapse probability), they outperformed the probability-summation model for the combined audiovisual speech signals by about 10% at the highest signal-to-noise ratios (Fig. 3.2B, compare red vs black curves).

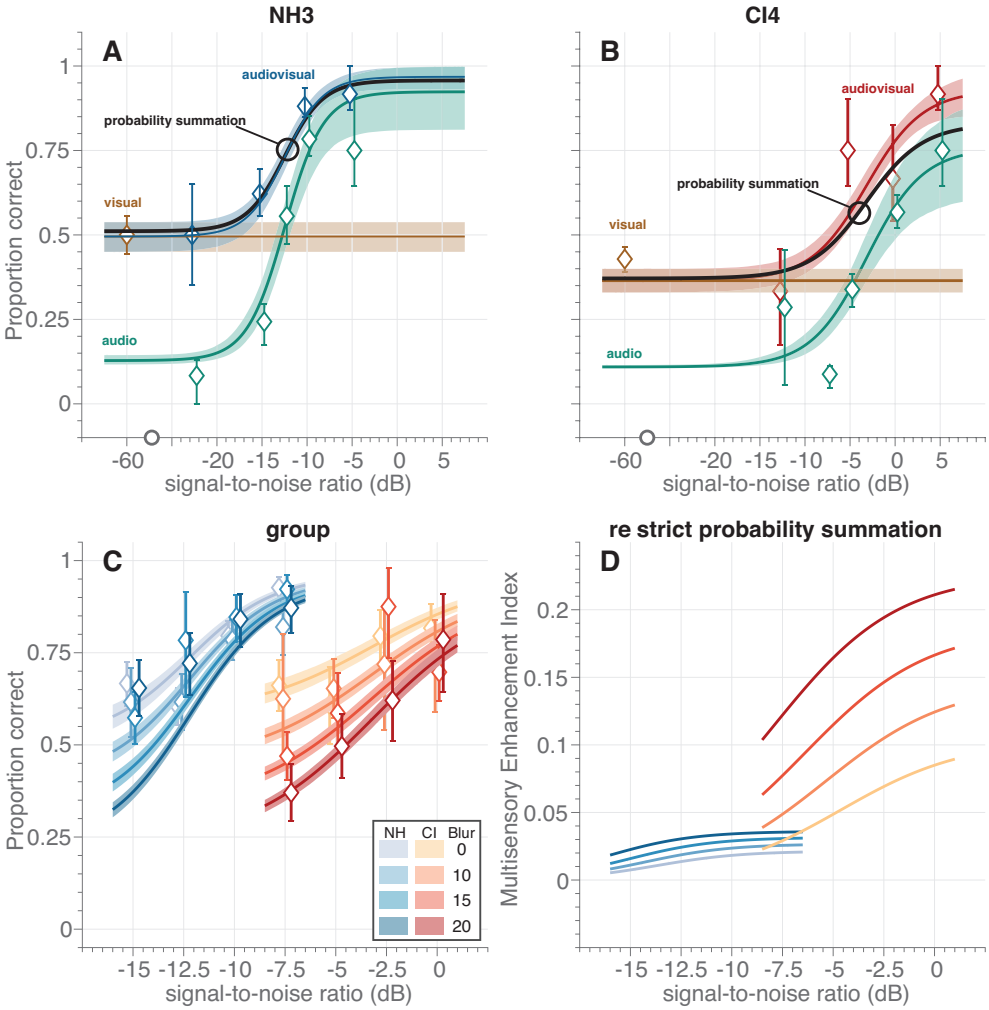


Figure 3.2. Multisensory speech recognition.

Individual data and fit for **A)** normal-hearing (NH) participant NH3 and **B)** CI user CI4. **C)** Audiovisual speech recognition scores as a function of acoustic signal-to-noise ratio (dB) for normal-hearing participants (blueish diamonds) and CI users (reddish diamonds) for four blur values (as indicated by contrast). Symbols and bars indicate mean and 95\%-confidence intervals, respectively, of the raw data (proportion correct) pooled across participants. The data was obtained (by definition) from the divided-attention task. Curves and patches indicate means and 95\%-HDI, respectively, of the psychophysical-function population fits. **D)** Multisensory enhancement index as a function of acoustic signal-to-noise ratio (dB) for normal-hearing participants (blue colors) and CI users (red colors) for four blur values (as indicated by contrast). The multisensory enhancement index quantifies the multisensory enhancement of the trade-off model over strict probability summation.

We quantified the audiovisual performance for all participants of both groups (visualized as a function of the acoustic signal-to-noise ratio for four different magnitudes of visual

blur, Fig. 3.2C) by fitting a probability-summation model that was fully determined by the unisensory auditory and visual recognition performance (Eqns. 3.1-3.4). Typically, the observed multisensory enhancement should be compared to probability-summation of unisensory performance obtained from the same experimental regime, which in the current experiment would be from the divided-attention task. We term this model the strict probability-summation model. In Fig. 3.2C, we show the results of an alternative model, which we designate the trade-off model, that actually captures the multisensory enhancement by using the unisensory data obtained during the focused-attention task. We did this because the increased lapse probability for listening by the CI users in the divided-attention task (Fig. 3.1C) appeared to equal the multisensory enhancement over the strict probability-summation model (e.g. Fig. 3.2B, compare the red fit curve to the black curve). In essence, the difference in recognition scores between the two tasks was captured by the difference in auditory lapse probability, the single trade-off model parameter free to vary between tasks.

Nevertheless, the trade-off model describes the data for both tasks quite well (Table 1, see Methods, and Figs. 3.2A,B). Note that the pooled data generally appear to be at higher performance levels than the group-level fits of the trade-off model, at least for the normal-hearing participants (Fig. 3.2C, blue). This follows from the fact that we individualized the stimulus parameters for each participant; the data was obtained at lower signal-to-noise ratios and higher blurs more often for the better performers. The group-level fits better describe the expected overall group performance through extrapolation to a larger range of signal-to-noise ratios and blurs. By comparing the fits to the audiovisual data (Fig. 3.2C) to the unisensory fits (cf. Fig. 3.1), one can observe that audiovisual speech recognition is better than unisensory speech recognition; even at a blur of 20 pixels and a signal-to-noise ratio of -15 dB for the normal-hearing and of -7.5 dB for the CI users (around 0.2 vs 0.35 for unisensory and multisensory stimulation, respectively).

Table 3.1. Model comparison.

	ΔBIC Normal-hearing	**ΔBIC CI users**
Trade-off	0	0
Strict	12	35
R^2 (trade-off)	0.89	0.78
mean signed error (trade-off)	0.00	+0.01

To illustrate the benefits of multisensory stimulation more clearly, we determined the multisensory enhancement index (MEI, Eqn. 3.5). This index quantifies by how much multisensory performance of the trade-off model was improved over the strict probability-summation model (Fig. 3.2D). A multisensory enhancement index close to zero is in line with

strict statistical facilitation, while positive values are evidence for audiovisual enhancement due to multisensory integration. The index shows marginal improvement for the normal-hearing group (between 0.005-0.036, depending on signal-to-noise ratio and blur, Fig. 3.2D), and a far more prominent benefit for CI users that was about 4-6 times larger (0.023-0.22). A larger multisensory enhancement index for lower-informative stimuli or poorer-performing individuals would be evidence for inverse effectiveness[8,12]. This effect seemed to occur for the groups and the blurs; CI users exhibited more enhancement than the normal-hearing participants (Fig. 3.2D, red vs blue) and the relative multisensory improvements were largest for the highest blurs (Fig. 3.2D, e.g. the multisensory enhancement index for the 0-pixel blur was lower than for the 20-pixel blur, especially for the CI users). In contrast, for acoustic information a direct, rather than inverse, relationship was observed: the lowest signal-to-noise ratios elicited the smallest enhancements (Fig. 3.2D, the MEI curves all decline for lower signal-to-noise ratios).

DISCUSSION

Summary

Results show that CI users benefit from multisensory integration in a divided-attention task (Fig. 3.2), but that their unisensory performance under such conditions deteriorates when compared to listening under focused attention (Fig. 3.1). Interestingly, their multisensory benefit matches the prediction obtained from probability summation of their (better) focused-attention performance (Fig. 3.2). In contrast, the normal-hearing participants do not have poorer unisensory performance in a divided-attention task, and their multisensory scores are accounted for by strict probability summation. Normal-hearing participants reached higher auditory recognition scores than the CI users. As expected, these results confirm the well-known fact that listening for CI users is considerably more difficult. Factors that likely contribute to the difficulties in understanding auditory speech in noise environments are the lack of access to finely-detailed spectral information and a limited dynamic range[21]. In contrast, CI users and normal-hearing participants had similar lipreading skills (Fig. 3.1B,D). This was slightly unexpected, as others have reported better lipreading abilities by CI users[18,22]. The current experiment, however, entailed recognition of a limited closed-set matrix of only 50 words. This potentially makes lipreading for normal-hearing individuals, who might be unaccustomed to lipreading in general, easier than in open sets with many more alternatives. Also, both the CI users and normal-hearing participants do have normal vision. As such, one might perhaps expect similar visual, lipreading skills.

Attentional lapse in unisensory performance

CI users missed fewer words when they could focus on listening alone (in the focused-attention task, Fig. 3.1C) than in situations with uncertainty about the modality of the upcoming stimulus (in the divided-attention task). Note that this is precisely the sensory condition of every-day life. This may suggest that due to impoverished sensory information more effort is required by CI users to be able recognize speech at higher performance levels. However, the extra effort cannot be maintained by CI users if attention has to be spread out across multiple, potentially-informative sensory modalities. The CI users seem to have reached the limits of attentional resources in the divided-attention task. These limits are not reached when sensory information is not impoverished, i.e. for normal-hearing individuals and for lipreading (Fig. 3.1A, B, D; lapse probabilities are similar across tasks).

Multisensory integration

Following this line of reasoning, one may wonder why CI users attempt to lipread at all. Barring any other benefits, the optimal decision would be to focus on the most-informative sensory modality, and ignoring the other. Even for CI users, listening is generally (i.e. in quiet environments) the far better modality for the purposes of speech recognition. Probabilistic, uninformed switching between listening and lipreading would lead to an overall worse performance[23]. One benefit to offset this drawback could be that switching enables individuals to scan the specific environment and determine whether listening or lipreading would be the most informative modality for the given situation[24,25]. Obviously from the current experiments, another benefit could be that the detriment in listening is accompanied by an enhancement of speech recognition for multisensory stimuli. Indeed, although CI users had poorer unisensory recognition scores in the divided-attention task than in the focused attention task (Fig. 3.1), they outperformed the strict probability-summation model (Fig. 3.2D). Conversely, the normal-hearing individuals do follow strict probability summation[20]. Because of this, CI users appear to be better multisensory integrators than the normal-hearing individuals[18] (Fig. 3.2D).

Integration-attention trade-off

Intriguingly, the trade-off model suggests that the exact compensation of the listening decline (Fig. 3.1C) by multisensory enhancement (Fig. 3.2D) may be explained by an integration-attention trade-off mechanism for CI users. To benefit from multisensory integration, attention needs to be divided across all relevant signals. Only then will integration be able to enhance source identification and selection by filtering out irrelevant noise sources. The cost of this benefit is the decline in attentional amplification of unisensory signals. In our model, this is fully and solely captured by the change in auditory lapse probability (Eqn. 3.3), which amounted to be about 22% on average for CI users. The multisensory enhancement follows directly from this increase in lapses (through the trade-off probability-summation model, Eqns. 3.4 and 3.5);

the multisensory enhancement should equal this in magnitude for the weakest visual signals and strongest auditory signals (note that the multisensory enhancement index is 0.22 for the highest blur at a signal-to-noise ratio of 0 dB), and be less for stronger visual signals and weaker acoustic signals (Fig. 3.2D).

Conclusion

Normal-hearing participants can attend extensively on auditory and visual cues, while CI users need to divide their limited attentional resources across modalities to improve multisensory speech recognition - even though this leads to a degradation in unisensory speech recognition. We argue that in order to determine the acoustic benefit of a CI towards speech comprehension per se, situational factors need to be discounted by presenting speech in realistic audiovisual environments.

METHODS

Participants

Fourteen native Dutch-speaking, normal-hearing participants (mean age: 22.3 years ± 1.8, 10 female) and 7 native Dutch-speaking, post-lingually deaf unilaterally implanted CI users (mean age 64.1 years ± 5.3, 3 female) were recruited to participate in this study. All CI users had at least one year of experience with their CI, with a mean of 3.6 years ± 1.8. Five CI users were implanted on the left. The cause of deafness was progressive sensorineural hearing loss for all but three CI users (Ménière's disease, sudden deafness and hereditary hearing loss). Additional contralateral hearing aids were turned off during the experiment. The unaided pure tone average (range 1-4 kHz) of the non-implanted ear ranged between 70 and >120 dB Hearing Loss. However, no CI users had any speech intelligibility for words in quiet with their non-implanted ear at levels < 90 dB Sound Pressure Level (SPL). All normal-hearing participants were screened for normal hearing (within 20 dB HL range 0.5 - 8 kHz). All participants reported normal or corrected-to-normal vision. All participants gave written informed consent before taking part in the study. The experiments were carried out in accordance with the relevant institutional and national regulations and with the World Medical Association Helsinki Declaration as revised in October 2013 (Declaration). The experiments were approved by the Ethics Committee of Arnhem-Nijmegen (project number NL24364.091.08, October 18, 2011).

Stimuli

The audiovisual material was based on the Dutch version of the speech-in-noise matrix test developed by Houben et al[26]. In general, a matrix test uses sentences of identical grammatical structure in which all available words are taken from a closed set of alternatives. The sentences are syntactically fixed (subject, verb, numeral, adjective, object), but semantically unpredictable.

The audiovisual material (Fig. 3.3) including the masking speech noise are reported previously[20]. Briefly, the stimulus material consisted of digital video recordings of a female speaker reading aloud the sentences in Dutch. Auditory speech (Fig. 3.3A,C) was presented with varying levels of acoustic background noise (Fig. 3.3B). Visual speech consisted of the video fragments of the female speaker (Fig. 3.3D). Saliency of the visual speech was altered through blurring, by filtering every image of the video with a 2-D Gaussian smoothing kernel at several pixel standard deviations.

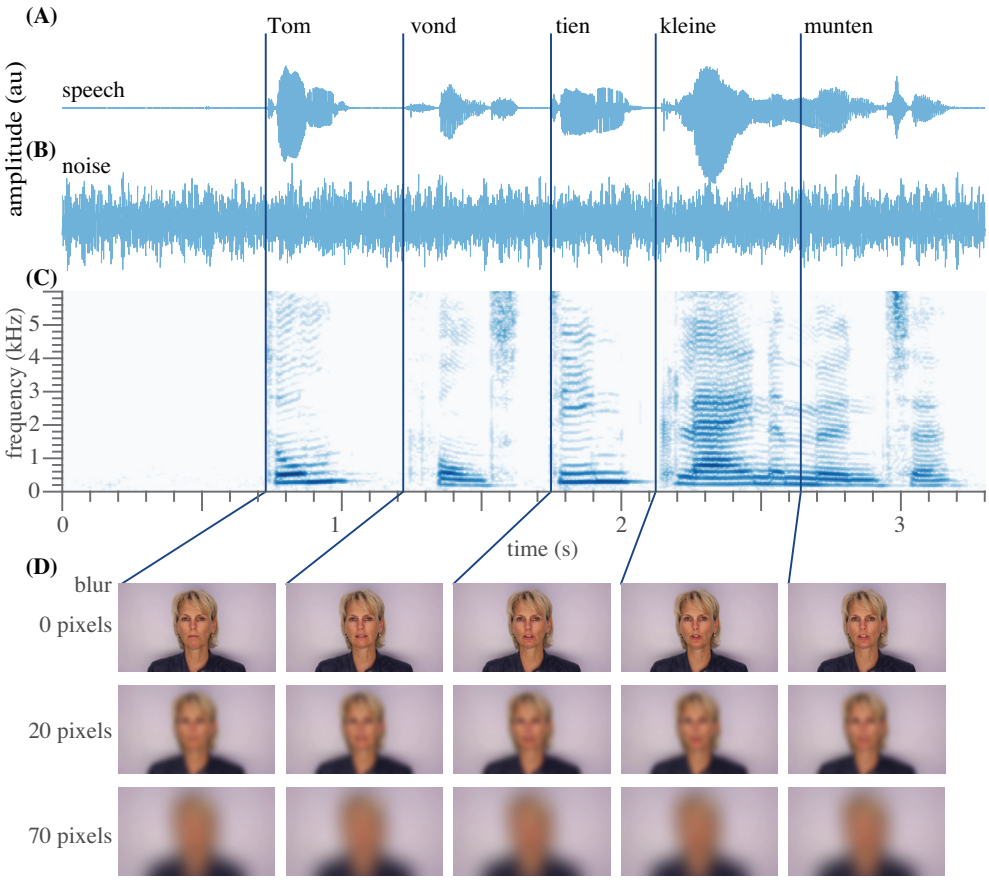


Figure 3.3. Example sentence.

A) Temporal waveform of the auditory speech signal "Tom vond tien kleine munten" (translation: Tom found ten little coins.) **B)** Waveform of the auditory noise. **C)** Spectrogram of the recorded sentence. **D)** Five video frames around the onset of the word, untouched (top), moderately blurred (middle, 20 pixels), and extensively blurred (bottom, 70 pixels, used as a unisensory auditory condition in the divided-attention task). Dark blue lines denote the approximate onset of each individual word. Written informed consent for the publication of this image was obtained from the individual shown.

Set-up

The experiments were performed in an experimental room, in which the walls and ceiling were covered with black acoustic foam that eliminated echoes for sound frequencies >500 Hz[27]. Stimulus presentation was controlled by a Dell PC (Dell Inc., Round Rock, TX, USA) running Matlab version 2014b (The Mathworks, Natick, MA, USA). were seated in a chair 1 m in front of a PC screen (Dell LCD monitor, model: E2314Hf). Sounds were played through an external PC sound card (Babyface, RME, Germany) and presented through one speaker (Tannoy, model Reveal 502) placed above the PC screen, 1 m in front of the participant (30º above the interaural plane). Speaker level was measured with an ISO-TECH Sound Level Meter, type SLM 1352P at the position of the participant's head, using the masking noise.

Paradigm

All participants were tested on a closed-set recognition of six Matrix lists of 20 sentences each (180 words). Participants were instructed to select words from the Matrix list which they recognized.

Familiarization

To familiarize participants with the Matrix test procedure and to obtain an initial estimate for the auditory threshold, 40 unique auditory-only sentences were presented. The signal-to-noise ratio varied adaptively in accordance with the Brand and Kollmeier procedure[28], and the auditory 50% speech recognition threshold was calculated as the average signal-to-noise ratio of the last nine sentences. This threshold was used to individualize the signal-to-noise ratios in focused-attention experiment. For normal-hearing participants, the speech level was fixed at 60 dB SPL, while for the CI users the noise level was fixed at 60 dB SPL. This was also true for both experiments.

Focused-attention task: unisensory speech listening or reading

In this experiment participants listened to auditory-only sentences in one block and viewed visual-only sentences in another block. The participants were asked to accurately indicate the words (10-alternative open-ended choice per word) after each sentence. Each trial was self-paced. Participants either heard 40 or 60 unique sentences in each block.

In the auditory-only block, the auditory speech was presented in acoustic background noise with uninformative visual input (i.e. a black screen for 6 normal-hearing participants; or a heavily blurred video (70-pixel blur) for 8 normal-hearing participants and all CI users). For each sentence, the signal-to-noise ratio was pseudo-randomly picked from 4 to 12 values, that were selected individually based on the results from the adaptive tracking procedure.

In the visual-only block, the video fragments of the female speaker were shown on the screen together with the acoustic background noise and without auditory speech signal. For each sentence, the standard deviation of the Gaussian blurring kernel of the video images was pseudo-randomly picked from 5 to 10 values; the 5 most common values were 0, 6, 12, 16, and 20 pixels both for normal-hearing participants and CI users.

To avoid priming effects of sentence content (but not word content), a sentence was never repeated within a block. For each participant a different set of random signal-to-noise ratios, spatial blurs, and sentence permutations were selected. Importantly in this experiment, participants should focus on one sensory modality, and ignore the other, in order to reach maximum performance.

Divided-attention task: multisensory speech listening and reading

In this experiment, audiovisual sentences (80 to 120 trials) were presented in one block. This experiment was conducted on another day than the focused-attention experiment. For each sentence, a visual blur and an auditory signal-to-noise ratio were chosen in pseudo-random order from five values, yielding 25 audiovisual stimulus combinations, selected in pseudo-random order. These values were selected individually based on the performance in the focused-attention experiment. We aimed for a unisensory speech-recognition performance of 0, 25, 50 and 75% for each participant, but as the maximum performance did not always reach 75%, other values were then chosen by the experimenter. The most common values were the same as for the previous experiment. In the unisensory trials of this task, the visual blur was extreme with a standard deviation of 70 pixels for the acoustic-only trials, and the auditory signal-to-noise ratio was -60 dB for the visual-only trials. Importantly, in contrast to the focused-attention task, participants could use information from both the auditory and visual modality in order to recognize words throughout most of the experiment, although some sentences were only informative in one sensory modality, but not in the other due to either extreme visual blurring (70-pixel blur) or an extremely poor acoustic signal-to-noise ratio (-60 dB signal-to-noise ratio).

Data analysis

For graphical purposes, the proportion of words correct responses are plotted in raw form pooled across participants for each group as mean and 95%-HDI in Figs. 3.1 and 3.2 for the most common signal-to-noise ratios and blurs.

Unisensory psychometric functions

To relate each participant's responses to the intensity of the unisensory stimuli (i.e. auditory signal-to-noise ratio or visual blur), x, we fitted a psychometric function $F(x, \theta)$ to the unisensory

data, the shape of which depended on the sensory modality, *m*. For the auditory-only data, a logistic function was fitted[20,29]:

$$F_A(x_A; \theta_A; \omega_A) = \left(1 + e^{\left(-\frac{2\ln 9}{\omega_A}(x_A - \theta_A)\right)}\right)^{-1} \quad (3.1)$$

where $F_A(x_A; \theta_A; \omega_A)$ characterizes the change in auditory word recognition rate as a function of the auditory signal-to-noise ratio, x_A; θ_A is the auditory recognition threshold for which holds $F_A^{-1}(0.5)$ and ω_A is the auditory recognition width, the stimulus-level range in which F_A ranges from 0.1 to 0.9.

For the visual-only data, an exponential function F_V was taken with only a single parameter:

$$F_V(x_V; \theta_V) = e^{-\frac{x_V^2}{\theta_V^2}} \quad (3.2)$$

where $F_V(x_V; \theta_V)$ characterizes the change in visual word recognition rate as a function of the visual blur, x_V ; θ_V is the visual recognition threshold for which holds $F_V^{-1}(0.3679)$, i.e. for $x_V = \theta_V$. Both functions (Eqns. 3.1 and 3.2) have a sigmoidal shape and fitted the data well (i.e. Fig 3.1).

Lapse

To infer the probability of correct-word recognition Ψ, we included a lapse probability, λ, to the psychometric function *F* for both modalities *m*:

$$\Psi_{m,e} = (1 - \lambda_{m,e}) F_m \quad (3.3)$$

The lapse probability, λ, accounted for the less-than-perfect recognition probability for visual words without blurring and for auditory words at the highest signal-to-noise ratios, both for the CI users and the normal-hearing participants. With probability $\lambda_{m,e}$ a participant has a momentary lapse (i.e. makes a choice independent of stimulus intensity) for modality *m* during experiment *e*. With probability $(1 - \lambda)$ the participant does not have a lapse and has a chance of F_m to give the correct answer. The lapse probability could reflect several issues: e.g. a momentary lapse of attention, blinking during the visual trials, or the lack of increase in information with increasing stimulus intensity due to for example processing issues of the cochlear implant.

Crucially, the estimate for the lapse probability was, at first, inferred separately for the experimental tasks (focused-attention vs divided), as we hypothesized that the separate tasks could differentially affect attentional demands, potentially leading to observed differences in attentional lapses.

We modified this slightly, as we observed no significant differences in the visual lapse probability between experimental tasks (Fig. 3.1). Thus, the final fitted model (Eqns. 3.1-3.3), as reported here, included the auditory lapse probability as the only parameter that was free to vary between experimental tasks. Constraining the model in such a way had no effect on the conclusions.

Multisensory psychometric function defined by probability summation

We modelled the audiovisual speech recognition as a mere statistical-summation effect that is distinct from true neural audiovisual integration. In this model of probability summation (see Introduction), participants recognize a word from either the auditory-only or the visual-only condition, which are considered independent processing channels. Thus, if a subject fails to recognize a word from either one of the modalities, the probability of failure is $(1 - \Psi_A) \times (1 - \Psi_V)$. It then follows that the probability of word recognition in the presence of the two modalities without integration is given by:

$$\Psi_{sum} = \Psi_A + \Psi_V - \Psi_A \times \Psi_V \tag{3.4}$$

where Ψ_{sum} is the probability to successfully recognize a word according to the summation model, Ψ_A is the probability to recognize an auditory word in the auditory-only condition, and Ψ_V is the probability of recognizing a visual word. From this, one can observe that having both modalities available, rather than one, automatically increases the probability of stimulus recognition.

We chose to fit this model because previous evidence[20] showed that speech recognition of the audiovisual materials could be described well by probability summation. Importantly, the data was accurately fitted by this model (see also the section on Model Selection), with one caveat: the fit was better if the lapse probabilities for the audiovisual stimuli (by definition, only presented in the divided-attention task) was set to equal the lapse probabilities as found in the focused-attention task.

This meant that model could only predict an enhancement of speech recognition for multisensory stimuli through a combination of mere statistical facilitation and a change in auditory lapse probability across experimental tasks. To visualize this (Fig. 3.2D), we determined the multisensory enhancement index, MEI:

$$\text{MEI} = \frac{\Psi_{trade\text{-}off}}{\Psi_{strict}} - 1 \tag{3.5}$$

with Ψ_{strict} and $\Psi_{trade-off}$ the probability to successfully recognize a word according to the summation model with an auditory lapse probability taken from the divided-attention (strict)

and focused-attention (trade-off) tasks, respectively. An MEI close to zero is in line with statistical facilitation, and no change in lapse probability. Positive values are evidence for an observed multisensory enhancement and an increased auditory lapse probability.

Guess probability

We also included a guess rate of 10% that accounts for a fixed probability of 0.1 of correctly choosing one of the ten alternatives by chance alone ($0.9\Psi + 0.1$). This was the same for every participant, modality and experimental task, as it depended on the design of the Matrix test itself.

Approximate Bayesian inference

Parameter estimation was performed using approximate Bayesian inference. The models described by eqns. 3.1-3.4 were fitted on all data simultaneously. The parameters were estimated for every participant, which depended on the estimation of overarching group parameters, separately for the normal-hearing participants and CI users, in a hierarchical fashion.

The estimation procedure relied on Markov Chain Monte Carlo (MCMC) techniques. The estimation algorithms were implemented in JAGS[30] through matJAGS[31]. Three MCMC chains of 10,000 samples were generated. The first 10,000 samples were discarded as burn-in. Convergence of the chains was determined visually, by checking that the shrink factor $\hat{R}$ is less than 1.1 and by checking that the effective sample size is larger than 1000[32]. From these samples of the posterior distributions, we determined the mean and the 95%-HDI as a centroid and uncertainty estimate of the parameters, respectively.

Model Selection

To test for the appropriateness of the models in eqns. 3.1-3.4, we compared them against less-restrictive models. To that end, we performed a qualitative check via visual inspection (c.f. Figs. 3.1 and 3.2), but we also quantitatively determined the Bayesian Information Criterion (BIC) for each model:

$$\mathrm{BIC} = k \ln(n) - 2 \ln(\hat{L}) \tag{3.6}$$

where k denotes the number of parameters of the model, n the number of samples and $\hat{L}$ the maximized value of the binomial likelihood function.

Author contributions

LR, JO and MW designed research; LR performed research; MW analyzed data; LR, JO and MW wrote the paper; LR and MW drafted the initial research concept.

Acknowledgments

We thank Günther Windau, Ruurd Lof, Stijn Martens, and Chris-Jan Beerendonck for their valuable technical assistance, speech-therapist Jeanne van der Stappen for being the speaker in the audiovisual material, Eefke Lemmen for video editing, and Roos Cartignij for help in data acquisition. We are grateful to Ad Snik and Emmanuel Mylanus for providing valuable comments on earlier versions of this manuscript.

Funding

This research was supported by EU Horizon 2020 ERC Advanced Grant ORIENT (grant 693400, JO), Cochlear Benelux NV (LR, MW), the Radboud University Medical Center (LR), and Radboud University (MW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author declaration

The authors declare no competing interest.

Data deposition

The data have been deposited in the Donders Institute for Brain, Cognition and Behavior Data Repository at https://doi.org/10.34973/jy8p-dw52.

REFERENCES

1. Miller, J. Divided attention: Evidence for coactivation with redundant signals. *Cogn. Psychol.* 14, 247–279 (1982).
2. Sumby, W. H. & Pollack, I. Visual Contribution to Speech Intelligibility in Noise. *J. Acoust. Soc. Am.* 26, 212–215 (1954).
3. Bernstein, L. E., Auer, E. T. & Takayanagi, S. Auditory speech detection in noise enhanced by lipreading. *Speech Commun.* 44, 5–18 (2004).
4. Helfer, K. S. & Freyman, R. L. The role of visual speech cues in reducing energetic and informational masking. *J. Acoust. Soc. Am.* 117, 842–849 (2005).
5. Summerfield, Q. Lipreading and audio-visual speech perception. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 335, 71–8 (1992).
6. Peelle, J. E. & Sommers, M. S. Prediction and constraint in audiovisual speech perception. *Cortex* 68, 169–181 (2015).
7. Corneil, B. D., Van Wanrooij, M., Munoz, D. P. & Van Opstal, A. J. Auditory-visual interactions subserving goal-directed saccades in a complex scene. *J. Neurophysiol.* 88, 438–54 (2002).
8. Bremen, P., Massoudi, R., Wanrooij, M. M. Van & Opstal, A. J. Van. Audio-Visual Integration in a Redundant Target Paradigm: A Comparison between Rhesus Macaque and Man. *Front. Syst. Neurosci.* (2017).
9. Colonius, H. & Diederich, A. Measuring multisensory integration: from reaction times to spike counts. *Sci. Rep.* 7, 3023 (2017).
10. Alais, D. & Burr, D. The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–62 (2004).
11. McDonald, J. J., Teder-Sälejärvi, W. A. & Hillyard, S. A. Involuntary orienting to sound improves visual perception. *Nature* 407, 906–908 (2000).
12. Stein, B. B. E. & Meredith, M. A. *The merging of the senses. Cambridge, MA: The MIT Press* (1993).
13. van de Rijt, L. P. H. *et al.* Temporal Cortex Activation to Audiovisual Speech in Normal-Hearing and Cochlear Implant Users Measured with Functional Near-Infrared Spectroscopy. *Front. Hum. Neurosci.* 10, (2016).
14. Bosen, A. K., Fleming, J. T., Allen, P. D., O'Neill, W. E. & Paige, G. D. Multiple time scales of the ventriloquism aftereffect. *PLoS One* 13, e0200930 (2018).
15. MacLeod, A. & Summerfield, Q. Quantifying the contribution of vision to speech perception in noise. *Br. J. Audiol.* 21, 131–41 (1987).
16. Sommers, M. S., Tye-Murray, N. & Spehar, B. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear.* 26, 263–75 (2005).
17. Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. & Foxe, J. J. Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cereb. Cortex* 17, 1147–1153 (2006).
18. Rouger, J. *et al.* Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proc. Natl. Acad. Sci. U. S. A.* 104, 7295–300 (2007).
19. Schorr, E. A., Fox, N. A., Van Wassenhove, V. & Knudsen, E. I. Auditory-visual fusion in speech perception in children with cochlear implants. *Proc. Natl. Acad. Sci. U. S. A.* 102, 18748–18750 (2005).
20. van de Rijt, L. P. H., Roye, A., Mylanus, E. A. M., van Opstal, A. J. & van Wanrooij, M. M. The principle of inverse effectiveness in audiovisual speech perception. *Front. Hum. Neurosci.* 13:335 (2019)
21. Friesen, L. M., Shannon, R. V, Baskent, D. & Wang, X. Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *J. Acoust. Soc. Am.* 110, 1150–1163 (2001).
22. Bernstein, L. E., Demorest, M. E. & Tucker, P. E. Speech perception without hearing. *Percept. Psychophys.* 62, 233–52 (2000).
23. Ege, R., Opstal, A. J. Van & Van Wanrooij, M. M. Accuracy-Precision Trade-off in Human Sound Localisation. *Sci. Rep.* 8, 16399 (2018).

24. Ege, R., Van Opstal, A. J. & Van Wanrooij, M. M. Perceived Target Range Shapes Human Sound-Localization Behavior. *eNeuro* 6, (2019).
25. Berniker, M., Voss, M. & Kording, K. Learning priors for Bayesian computations in the nervous system. *PLoS One* 5, e12686 (2010).
26. Houben, R. *et al.* Development of a Dutch matrix sentence test to assess speech intelligibility in noise. *Int. J. Audiol.* 53, 760–3 (2014).
27. Agterberg, M. J. H. *et al.* Improved horizontal directional hearing in bone conduction device users with acquired unilateral conductive hearing loss. *J. Assoc. Res. Otolaryngol.* 12, 1–11 (2011).
28. Brand, T. & Kollmeier, B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J. Acoust. Soc. Am.* 111, 2801–2810 (2002).
29. Kuss, M., Jäkel, F. & Wichmann, F. A. Bayesian inference for psychometric functions. *J. Vis.* 5, 478–92 (2005).
30. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd Internaitional Workshop on Disbtributed Statistical Computing* (2003).
31. Steyvers, M. matJAGS. http://psiexp.ss.uci.edu/research/programs_data/jags/ (2011).
32. Gelman, A. *et al. Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. (Chapman and Hall/CRC, 2013).

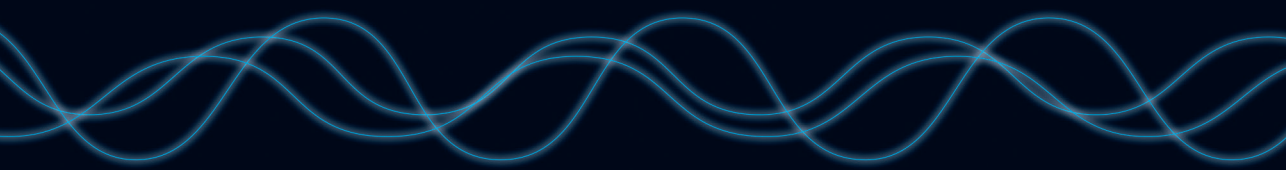
CHAPTER 4

Measuring cortical activity during auditory processing with functional near-infrared spectroscopy

Luuk P.H. van de Rijt, Marc M. van Wanrooij, Ad F.M. Snik, Emmanuel A.M. Mylanus, A. John van Opstal, Anja Roye

Keywords: near-infrared spectroscopy (NIRS), functional near-infrared spectroscopy (fNIRS), brain activity, auditory cortex

J Hear Sci 2018;8(4):9-18.
DOI: https://doi.org/10.17430/1003278



ABSTRACT

Functional near-infrared spectroscopy (fNIRS) is an optical, non-invasive neuroimaging technique that investigates human brain activity by calculating concentrations of oxy- and deoxyhemoglobin. The aim of this publication is to review the current state of the art as to how fNIRS has been used to study auditory function. We address temporal and spatial characteristics of the hemodynamic response to auditory stimulation as well as experimental factors that affect fNIRS data such as acoustic and stimulus-driven effects. The rising importance that fNIRS is generating in auditory neuroscience underlines the strong potential of the technology, and it seems likely that fNIRS will become a useful clinical tool.

BACKGROUND

Functional near-infrared spectroscopy (fNIRS) is an optical neuroimaging technique that assesses cerebral activity based on hemodynamics, which is associated with changes in the transmission of low power near-infrared light directed through the scalp and skull into the brain [1]. A variety of alternative terms have been used for the near-infrared spectroscopy (NIRS) technique, such as diffuse optical topography or tomography (DOT), diffuse optical imaging (DOI), and near infrared imaging (NIRI), although the underlying concept and physiological underpinnings remain similar (for detailed general reviews see e.g.[2–4]).

Brain activity leads to an increase in oxygen consumption, which is accompanied by an increase in cerebral blood flow due to neurovascular coupling.[5] This induces a change in the local oxygenated (HbO_2) and deoxygenated hemoglobin (HbR) concentrations. Given the different absorption coefficients of specific wavelengths of near-infrared light (600–900 nm) by HbO_2 and HbR, changes in the concentration of each of these chromophores can be extracted by measuring changes in the amount of light reflected over time.[6] Due to the relatively low absorbance of near-infrared wavelengths by biological tissue, the cerebral cortex can thus be imaged. Specific parameters of the hemodynamic response observed with fNIRS hence reflect the spatial and temporal characteristics of changes in HbO_2 and HbR, which may be manipulated by experimental paradigms and sensory stimuli (see below).

FNIRS is perfectly suited to the study of auditory processing in human subjects of all ages,[7,8] since fNIRS is a non-invasive and silent brain-imaging technique, as opposed to PET[9] and fMRI.[10] Further, the technique does not interfere with electromagnetic bionic devices such as cochlear implants.[8,11] Since the technique is silent (as opposed to fMRI), subjects can be seated in a normal (laboratory) environment, in which they can readily perform real-world psychophysical tasks, and the technique can be easily coupled with simultaneous EEG recordings. Because of these advantages, an increasing number of researchers are seeing the potential of fNIRS in auditory research for both normal-hearing and hearing-impaired listeners.[12]

The objective of this article is to review the current state of the art as to how fNIRS has been employed to evaluate auditory function, such as in speech, non-speech processing, and auditory attention in adults. In general, obtaining an optimal and stable setup and design for adequate hypothesis testing with fNIRS still remains a challenge. To test hypotheses of auditory processing requires a thorough understanding of the cortical hemodynamic response to acoustic stimuli, and how this response may be modulated by stimulus presentation rate, duration, sound level, and attention. Identifying the experimental factors that might affect the hemodynamic response is paramount for acquiring reliable and valid data.

The specific objectives of this paper are as follows: (i) to introduce the temporal and spatial characteristics of hemodynamic changes to auditory stimulation in general; (ii) to identify experimental factors that affect hemodynamic changes measured with fNIRS; (iii) to obtain insights into common experimental paradigms; and finally (iv) to summarize the contributions fNIRS has made so far to the study of auditory functioning.

Temporal and spatial characteristics of the hemodynamic response

Temporal characteristics of the hemodynamic response

FNIRS should be regarded as an indirect measure of neural activity, as it only measures vascular changes. The hemodynamic response to cortical neural activity relies on the fact that neuronal firing and the associated vascular response are strongly coupled (cf. neurovascular coupling; for a review see[13]).

Although crucial to this neuroimaging method, the mechanisms of neurovascular coupling are still not fully understood. It is clear that active neuronal tissue consumes energy for which the required inflow of oxygen and glucose will be accompanied by a local increase of cerebral blood flow, resulting in a local excess of oxygen in that particular area. This local increase of cerebral blood flow is associated with an increase of HbO_2 and a decrease of HbR (see Figure 4.1 for an example). This characteristic behaviour is usually described as the hemodynamic response function (HRF), and is well characterized for adults.[14,15] The characteristic HRF is related to the blood oxygenation level-dependent (BOLD) response that is also measured with fMRI,[11,16] although the BOLD signal proper is assumed to reflect changes in HbR only (for a review on hemodynamic changes measured with fMRI see[17]).

In general, the onset of the hemodynamic response lags the much faster electrical neural response to sensory stimulation by about 2 s. The changes in HbO_2 and HbR start with a steep increase, which rises to a plateau about 6–10 s after stimulus onset. The recovery time for the HbO_2 and HbR responses to return to baseline is only infrequently reported,[18] and is about 9–10 s.[14] While both hemoglobin species (HbO_2 and HbR) are well correlated regarding their temporal characteristics and shape during the steady state of the stimulus, sometimes an initial overshoot and a post-stimulus undershoot may be observed for both chromophores.[19] These are assumed to be a specific characteristic of neurovascular coupling.[20]

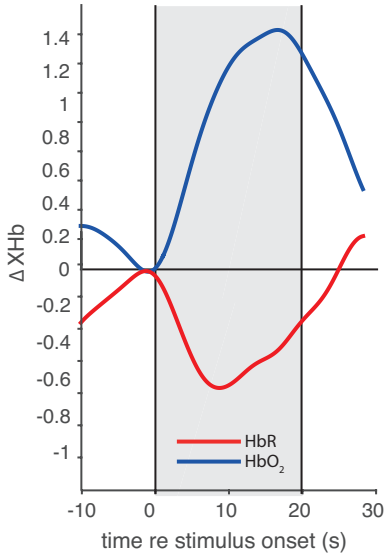


Figure 4.1. Hemodynamic response to auditory stimulation in temporal cortex.

The blue line illustrates the increase of HbO_2 and the red line the decrease HbR in response to the presentation of a speech stimulus (grey patch, 20 s). The sources and detectors were positioned over the left temporal hemisphere. Image adapted from Van de Rijt et al., 2016[27].

Besides the general characteristics of the hemodynamic response, an important question is to what degree it is linearly related to the underlying neural activity, and hence whether it scales with stimulus input strength and obeys the superposition principle to multiple stimuli (on model linearity see e.g.[21]). For example, Soltysik et al.[22] reported that the auditory response obeys linearity for stimuli of a relatively long duration, but reveals nonlinear properties for short-duration stimuli (<10 s). It has also been suggested that responses become non-linear at higher stimulus presentation rates.[23–25] Although, general aspects of hemodynamics might be partly responsible for non-linear response behaviour (e.g. saturation), another contribution could be due to the underlying neuronal responses, which can be enhanced by changes in the acoustic input, but will be suppressed for ongoing, tonic inputs (e.g., due to neural adaptation; for a review see[26]).

Spatial information obtained with NIRS

Figure 4.2 shows the probe template for two optical sources (S) and one photodetector (D) using source–detector distances of 25 and 35 mm, respectively (termed reference or shallow, and deep channel, respectively).[27] In this figure, the detector records the transmitted light coming from two sources, and each source–detector combination is defined as a channel. The sources transmit their light at unique frequencies in order to distinguish, using a lock-in amplifier, which source transmitted the light.

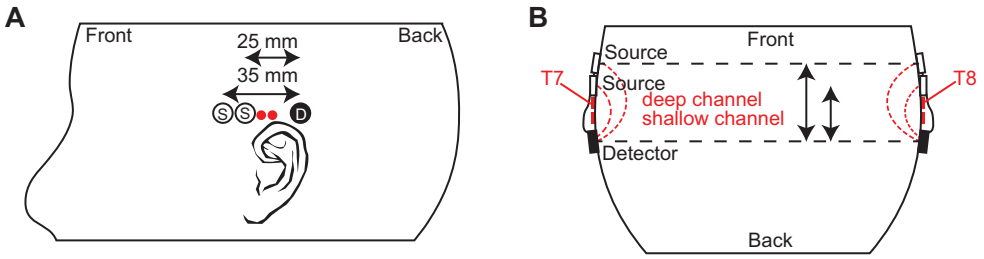


Figure 4.2. Positioning of the optodes.

A) Layout of optical sources (open circles) and photodetectors (filled circles) on the left hemisphere; **B)** schematic top view of probe layout. The estimated T7 and T8 positions of a 10/20 system are also indicated, as these are the supposed superficial centers of the deep and shallow channels (red filled circles). Red dotted lines denote the average path from source to detector, estimated to be part of an ellipsoid with a penetration depth of approximately 2–3 cm. Image adapted from Van de Rijt et al., 2016[27].

The first fNIRS measurements were carried out at only one or a few locations on the skull.[11,28] Since stimulus-evoked brain activity occurs at restricted regions in the brain, one might miss the activation of interest when measuring just one brain area. Hence, a major step was to utilize multi-channel fNIRS systems which allow the possibility of measuring cortical hemodynamics from several cortical locations and construct topographic activity maps.[29–32] Recently, researchers have developed a 140-channel fNIRS system to enhance local sensitivity – measured with several source–detector distances over overlapping regions to enable three-dimensional image reconstructions.[33,34] The method resembles the topographic mapping techniques familiar with fMRI measurements.

Multi-channel measurements can certainly be regarded an important development towards establishing fNIRS as a neuroimaging method that allows neuronal activity mapping over wide or distributed brain areas. However, unlike fMRI, NIRS does not allow structural imaging of the brain, and so several refinements have to be made to overcome this limitation and allow reliable measurements and valid conclusions: 1) Positioning should be accurate and reproducible to guarantee that recordings are taken from the same location; 2) Valid inferences on targeted brain areas recorded with different channels should be possible. 3) One should remove systemic noise from cortical brain activity.

Reliable positioning and valid inferences about underlying sources

Most researchers align the fNIRS channels (area between source and detector) with selected electrode positions of the well-established international 10-20 system.[35–37] Although this procedure secures reliable positioning in general, conclusions about underlying cortical regions can only be drawn in a probabilistic manner.[38] Obviously, some variance of the data will be attributable to the variability of defining the positions based on the 10-20 system across subjects and sessions.[39]

Another option to enhance reliability and validity, and to avoid the variance induced by the 10-20 system, is to align recorded optode locations with anatomical positions of the channels by using magnetic resonance (MR) structural images. Investigators have used markers (e.g. alfacalcidol beads/ vitamin D or E) to determine which cortical structures were measured by fNIRS in the studied participant.[40,41] This procedure ensures that the data were obtained from the region of interest, and therefore 'auditory channels' can be defined *a priori*.[41,42]

A third way to improve reliability is by demonstrating spatial similarity in functional data obtained with alternative neuroimaging methods. Some research groups have used both fNIRS and fMRI to compare cortical measurements of speech-evoked activity.[11] Others have used magneto-encephalography (MEG) and application of a 1000 Hz tone to determine the active region of the auditory cortex and so model the electric source of the N1m response.[28]

Finally, implementing a localizer task into the experimental protocol of the fNIRS recordings itself, besides the experimental contrast, can also be a valuable method to determine regions of interest. For auditory experiments this could be a standard auditory stimulus (tones or noise), or the average response to all experimental stimuli used, which is then compared to a silent baseline period. Channels that exhibit maximal hemodynamic changes may then be followed up in further steps of the analysis.[43] Alternatively, Kennan et al.[44] implemented a motor task (i.e. finger-tapping) within an auditory oddball task to localize the relative position of activation in primary motor cortex. These different approaches may contribute to improved inferences about target areas within and between studies.

Distinguishing physiological noise from cortical brain signals

When looking at the raw fNIRS signals recorded from NIRS channels, which are supposed to target certain brain areas, systemic or physiological noise often pollutes the hemodynamic responses of interest. These physiological sources of noise, such as heartbeat, respiration, or Mayer waves[45] may hide experimental effects which are usually of much smaller amplitude, and it may require sophisticated methods to identify the latter. Using a 'reference channel' offers a possible way to increase the reliability of estimating the hemodynamic response from fNIRS signals.[27,45,46] A reference channel is characterized by a short source–detector distance (range of 1–2 cm, see Figure 4.2), and makes use of the direct relation between source–detector distance and depth reached by photons in tissues underlying the scalp.[47–49] Due to the short distance of the reference channel, it is likely to reflect hemodynamic activity that is taking place within superficial tissues rather than stimulus-evoked brain activity. Signals derived from the reference channel seem to be perfectly suited for subtraction of physiological noise from the measured NIRS signal (i.e. reference channel subtraction (RCS)), and has been demonstrated to facilitate the estimation of evoked cortical hemodynamic responses[50–52] (Figure 4.2). In Figure 4.3 an example is shown of how, at the single-subject level, RCS affects

the average response during auditory stimulation. In general, it improves the signal response (Figure 4.3B; for further explanation see[27]).

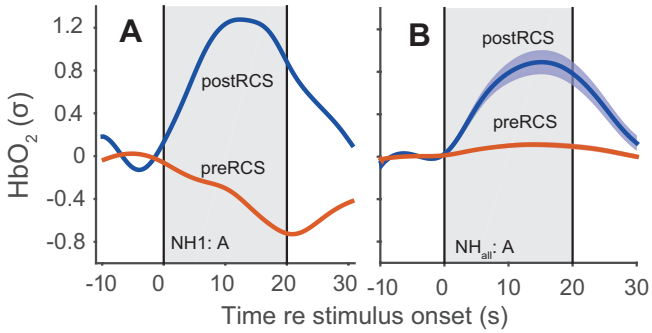


Figure 4.3. Reference channel subtraction.

The red lines depict pre–reference-channel subtraction and the blue lines depict post–reference-channel subtraction. Grey patches indicate auditory stimulus presentation. Stimulus presentation was 20 s. **A)** Averaged normalized HbO_2 data for 12 auditory stimuli of a normal-hearing subject (NH1); **B)** the same for a normal-hearing cohort (n = 33). Image adapted from Van de Rijt et al., 2016[27].

Choosing the most appropriate experimental paradigm

Besides potential methodological difficulties in placing the optodes and removing physiological noise, a further important consideration for optimising data quality is the experimental paradigm used in an fNIRS study. With a few exceptions,[37,44] the majority of NIRS studies employ a block design. In this approach, the different experimental conditions are presented separately within relatively long blocks (4–30 s) of stimulation. Within each block, tokens of the same stimulus type are presented repetitively, or in an ongoing manner. Stimulation blocks are followed by a control condition to allow for the HRF to return to baseline. These periods are usually filled with silence or some kind of unrelated stimulation during the rest period to reduce movement artefacts and keep participants attentive to the experiment.[11]

The general benefit of a block design is reflected in the robustness of the obtained hemodynamic signal. Due to repetitive presentation of a stimulus condition within a block, the captured HRF of the entire block is acquired as a superposition of the individual HRFs to each stimulus presentation. However, this design also has its shortcomings. The effects of individual stimuli within a block cannot be obtained (e.g. different responses to different words within sentences). Further, due to relatively long blocks of stimulation, the obtained responses might be influenced by effects of arousal, selective attention, or other cognitive effects that may vary between blocks and hence confound the actual effect of interest.

As an alternative to the block design, an event-related design[37,44,53] can overcome these attention- or task-related effects. In this case, relatively short stimuli (1–4 s) are presented in much faster succession than the different blocks in a block paradigm. Faster stimulation reduces data acquisition time and hence the total number of epochs (events) can be increased compared to the block design. For the design of the experiment, it is important to consider that the time between two successive stimuli can be short, but should be long enough to allow the HRF to partially return to baseline in order to avoid saturation of the hemodynamic signal. Jittering the inter-stimulus interval may also contribute to reducing random physiological noise in the data. However, due to overlapping HRFs, statistical analysis of the data requires more sophisticated approaches than does a block paradigm (e.g. a general linear model (GLM), see[54]; for a review, see[55]).

Modulating the hemodynamic response by experimental variations

Stimulus-specific and area-specific activations

While NIRS may be considered a reliable and valid tool to study stimulus-driven, bottom-up visual processing,[56] clear evidence that NIRS reflects stimulus-specific and modality-specific activations to acoustic stimuli still needs to be established. The lack of clear evidence is partially due to the use of only a limited number of optodes, and hence *a priori* areas of interest, but also to a lack of systematic experimental designs that target modality and stimulus specificity. The first limitation is overcome by using multi-channel fNIRS that allows spatial brain mapping. It has been shown that maximal hemodynamic changes are indeed measured when channels are centered on the auditory cortex, whereas the optical signal diminishes or disappears for locations away from auditory cortex.[30,43,57,58] This regional specificity of activations is further supported by studies which have demonstrated differential activations at the expected occipital (V1), auditory (A1), and sensorimotor cortical regions for visual stimuli (e.g. checkerboard stimulation), motor tasks (e.g. finger-tapping), and auditory stimulation (tones), respectively.[43,59] More specifically, a recent study by Chen et al.,[59] which measured auditory and visual areas in response to stimuli of both modalities, appeared capable of dissociating auditory from visual activations by showing maximal responses in the associated modality-specific areas. Prior to this certainly necessary systematic experiment, several prior studies had already demonstrated that the hemodynamic response to auditory stimuli can be altered by varying basic, as well as higher level, sound characteristics (bottom-up effects), and also by including top-down task characteristics within the same modality.

Acoustic stimulus driven effects on the hemodynamic response

Loudness modulation. Most studies performed with fMRI have demonstrated that the auditory hemodynamic response is sensitive to variations in sound level.[60,61] Some authors have indicated a positive, nearly linear relationship between the strength of the BOLD signal and sound intensity [62]. It appears that auditory cortical responses measured with fNIRS show such a linear relationship for *perceived* loudness, rather than for the (physical) intensity of the sound.[59] This potential discrepancy between intensity vs. loudness might suggest that fNIRS does not primarily target primary auditory cortex, where intensity effects seem more clear, but mainly relate to activity generated in secondary auditory areas (see discussion in[59]; also on fMRI[63,64]).

Presentation and repetition rate modulation. A difference between block and event-related designs is the interval between consecutive stimuli, which is longer for a block design (3–25 s) and relatively short in an event-related design (1–4 s). When experiments discuss the interval between two stimuli, a clear distinction needs to be made as to whether one is referring to the inter-stimulus interval (ISI) between consecutive stimuli (usually referred to as the presentation rate) or to the inter-stimulus interval between identical stimuli (called the repetition rate). Generally, most studies indicate a nonlinear, inverse relationship between the cortical response and the stimulus presentation rate. As stimuli are presented in fast succession, the cortical response reaches a plateau and may even decrease (evidence from fMRI,[25,65–68]; see also section on the temporal characteristics above). With fNIRS, the effect of sound presentation rate on cortical activation has been investigated by Weiss et al.[69] These authors systematically looked at presentation rates of trains of noise bursts at 2, 10, and 35 Hz. The study confirmed an inverse relationship between HbR concentration change and presentation rate.

In addition, there is the phenomenon of stimulus-specific neural adaptation (for a review see[70]), which holds that responses to an immediately following stimulus (i.e., at short ISI) can be influenced by the response to an immediately preceding stimulus. The size of the response will be reduced if specific stimulus characteristics are repeated. That is why it is useful to distinguish the presentation rate (which concerns different stimuli) and the repetition rate (which refers to identical stimuli). If sufficient time has elapsed before the *same* stimulus is repeated, suppression of the hemodynamic response to the latter may be absent.

A nice illustration of this phenomenon is the 'oddball paradigm' (see e.g.[44]). In a standard oddball paradigm, the subject is presented with a series of repetitive or 'standard' stimuli that are randomly and infrequently replaced with a distinctly different or 'deviant' stimulus. When an identical stimulus (usually called the standard stimulus) is presented several times, the neural system will adapt, leading to reduced neuronal activity for the consecutive stimuli. As a result, the hemodynamic response may saturate.[71] This has been shown by Kennan et al.,[44]

who used a classical auditory oddball design. In the same way, continuous tones do not produce an ongoing hemodynamic response. However, their study also showed that even if the presentation rate is quite high and hence the ISI is short (1.5 s), a low repetition rate of the rare stimuli (which deviate from the repeated standard stimulus) can result in clear responses to these stimuli, even when presented within generally fast sequences of other stimuli. By observing these responses, it makes the technique suitable for experiments which do not last long (e.g. for children).

Stimulus complexity and impact of higher order stimulus categories. Based on fMRI, PET, and animal studies, it can be hypothesized that acoustic complexity can modulate hemodynamic responses. Simple acoustic stimuli (e.g., pure tones) primarily activate the core of the primary auditory cortex, whereas spectrally more complex sounds (e.g., complex noise, vocalizations, music, speech) also activate the surrounding higher order areas (e.g.[72,73]; for a review see[74]). So far, only one study used both simple tones and more complex frequency-modulated sounds within the same fNIRS study.[59]

Besides acoustically driven effects, some research groups have also investigated whether fNIRS shows sensitivity to higher order stimulus features. For example, Pollonini et al.[33] varied intelligibility of auditory stimuli using sounds with otherwise comparable acoustic features (frequency content, spectro-temporal modulations, intensity). They showed that meaningful and intelligible auditory inputs led to a broader area of activation within temporal cortices. The activation decreased for distorted sounds or for non-speech environmental sounds. Bembich et al. reported fNIRS activation only for meaningful words, when compared to meaningless vowel-consonant-vowel syllables.[36] Further, several studies by Minagawa-Kwai et al.[30,57,58] suggested that fNIRS is sensitive to language-specific speech contrasts. They demonstrated that there were left hemispheric hemoglobin changes to phoneme contrasts within the listener's native language that was not present for phoneme contrasts measured in non-native listeners. This left side functional lateralization seems to be driven by the phonemic contrast of the speech, since Sato and colleagues[42] demonstrated that a *prosodic* contrast led to right-sided dominance.

That the emotional valence of non-speech sounds can also yield differences in hemoglobin changes has been shown by Plichta et al.,[43] who reported that both pleasant and unpleasant sounds led to significantly enhanced hemoglobin changes in auditory cortex when compared to neutral sounds. Another group looked into the effects of fear and disgust,[75] and showed that sounds that were associated with fear elicited increased hemoglobin changes within temporal–parietal regions, while disgusting sounds elicited smaller changes. Taken together, these findings underscore that internal representations such as language-specific experiences, and emotional or motivational relevance, can lead to hemoglobin changes that are measurable with fNIRS.

Top-down effects on the hemodynamic response

As described above, auditory cortical responses measured with fNIRS depend on many stimulus-driven factors such as presentation rate, loudness, complexity, intelligibility, experience, and emotional valence. Only a few studies have systematically looked into the effects and response dependencies of attention and task-demands, although it has been suggested by other recording methods that the attentional focus can influence auditory cortical responses (for a meta-analysis on fMRI data see e.g.[76]; for a general review see [77]). Often, fNIRS studies do not really control for attentional effects and simply require the subject to listen without giving a certain response.[11,27,75,78] A notable exception is the fNIRS study of Kojima and Suzuk,[79] which utilized visual stimuli to show that hemodynamic responses in visual cortex are enhanced when participants are asked to perform a visual search task (compared to the inattentional condition).

For auditory stimulation the fNIRS study of Remijn and Kojima[80] assessed auditory-cortical responses within a streaming paradigm. Their results showed that performing a task of actively judging a perceived acoustic rhythm caused significantly larger HbO_2 responses compared to the passive listening condition. In summary, several studies suggest that hemodynamic responses driven by auditory stimulation can be enhanced through auditory attentional engagement.

Reproducibility of fNIRS measurements

A potential advantage of the NIRS technique, compared to other neuroimaging methods, is that the brain activity of patients wearing hearing aids or implants, and also of children, may be measured in a clinical setting. However, a prerequisite for using the technique is to assess its general *reproducibility* or retest reliability. To our knowledge, no study has formally evaluated the reproducibility of different aspects (size, location, amplitude, temporal behaviour) of hemodynamic responses elicited by auditory stimulation. For other modalities, some multichannel fNIRS studies have been carried out to evaluate retest reliability (in the motor cortex, see[39,81]; in the occipital cortex to visual stimulation, see[53]) and they suggest that reliability at the group level exists.

So far, two studies have looked at the reliability of cortical activation in an event-related design,[53,81] while Sato et al.[39] has looked into data reproducibility using a block design. The authors demonstrated that absolute signal amplitudes may vary between sessions, but that the time courses of the signal are highly correlated between sessions ($r > 0.8$). To address the level of reproducibility of fNIRS in occipital cortex, Plichta et al.[53] presented periodic checkerboard stimuli and measured them at a retest interval of 3 weeks, focusing on three different aspects. First, the reproducibility of a number of activated channels over the two sessions was moderate. Second, in a single channel comparison the reproducibility was generally low,

but this improved when channels were clustered (significant activations at first and second session). As a last step, they looked at topographic map activation (*t*-values) within their pre-defined region of interest, and this showed that the fNIRS group activation maps were highly reproducible.

These outcomes show that, on a group level, fNIRS is reliable and trustworthy for fundamental research looking into effects on subjects. However, at this point, reproducibility in single subjects seems to be lacking.[53,81–83] Different causes may underlie this problem. As mentioned before, often only a very limited set of fNIRS optodes is measured, and even if the researcher increases their number, makes exact and reliable positioning, uses data-driven channel selection, and analyses signals over broader areas of interest, these refinements do not always reduce between- and within-subject variance. Some authors suggest implementing MRI-guided techniques[84] to improve within-subject reliability. However, since fNIRS is intended to be used on subjects for whom fMRI scans are to be avoided (children, auditory research, participants with bionic devices), the alignment of fNIRS outcomes with structural and/or functional MRI scans is not an ideal solution.

CONCLUSIONS

This review has aimed to summarize the state of the art of how fNIRS can be used to study auditory central processing. This review indicates that increasing numbers of auditory neuroscience researchers are now readily using fNIRS to measure hemodynamic responses to a range of experimental stimuli and response conditions. Yet, despite the promising results of fNIRS, developing an ideal and stable setup and experimental design for adequate hypothesis testing still remains a challenge. By incorporating some of the aspects reviewed here – for example, details of how the cortical hemodynamic response to acoustic stimuli is modulated by stimulus presentation and repetition rates, sound duration, sound level, and attention – one might be able to acquire reliable and valid fNIRS data.

For further details on the underlying physiological principles,[85,86] available analysis methods, and technological advancements in fNIRS (aspects which lie outside the scope of this review), we suggest reading existing reviews.[2,3,55]

An important asset of fNIRS is that it can be readily combined with other neuroimaging modalities such as fMRI, EEG, PET, and MEG. Evidence comes from the increasing number of publications on multimodal imaging systems.[28,37,44,87,88]

FNIRS is becoming increasingly recognised as a powerful neuroimaging tool to reveal cortical activity in different patient groups of all ages. Typically, this neuroimaging method is silent and non-invasive, as opposed to fMRI and PET respectively. Furthermore, the technique is not impeded by electromagnetic bionic devices, such as a cochlear implant (CI). Anderson et al.[12] has recently shown the potential importance of applying fNIRS for longitudinal studies of cortical auditory function in CI users, giving insights into the correlation between audio-visual interactions and cortical reorganization, before and after cochlear implantation. Their results provide evidence of cortical plasticity within the bilateral superior temporal cortex (STC), suggesting how these effects may potentially explain the considerable variability in CI outcome measures.

REFERENCES

1. Jobsis F. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science (80-)*. 1977;198(4323):1264-1267. doi:10.1126/science.929199
2. Lloyd-Fox S, Blasi a, Elwell CE. Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy. *Neurosci Biobehav Rev*. 2010;34(3):269-284. doi:10.1016/j.neubiorev.2009.07.008
3. Scholkmann F, Kleiser S, Metz AJ, et al. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *Neuroimage*. 2014;85:6-27. doi:10.1016/j.neuroimage.2013.05.004
4. Strangman G, Boas D a, Sutton JP. Non-invasive neuroimaging using near-infrared light. *Biol Psychiatry*. 2002;52(7):679-693.
5. Logothetis NK, Wandell B a. Interpreting the BOLD signal. *Annu Rev Physiol*. 2004;66:735-769. doi:10.1146/annurev.physiol.66.082602.092845
6. Cope M, Delpy DT. System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. *Med Biol Eng Comput*. 1988;26(3):289-294. doi:10.1007/BF02447083
7. Bortfeld H, Wruck E, Boas D a. Assessing infants' cortical response to speech using near-infrared spectroscopy. *Neuroimage*. 2007;34(1):407-415. doi:10.1016/j.neuroimage.2006.08.010
8. Dewey RS, Hartley DEH. Cortical cross-modal plasticity following deafness measured using functional near-infrared spectroscopy. *Hear Res*. 2015;325:55-63. doi:10.1016/j.heares.2015.03.007
9. Johnsrude IS, Giraud AL, Frackowiak RSJ. Functional Imaging of the Auditory System: The Use of Positron Emission Tomography. *Audiol Neuro-Otology*. 2002;7(5):251-276. doi:10.1159/000064446
10. Hall DA, Haggard MP, Akeroyd MA, et al. Modulation and task effects in auditory processing measured using fMRI. *Hum Brain Mapp*. 2000;10(3):107-119.
11. Sevy ABG, Bortfeld H, Huppert TJ, Beauchamp MS, Tonini RE, Oghalai JS. Neuroimaging with near-infrared spectroscopy demonstrates speech-evoked activity in the auditory cortex of deaf children following cochlear implantation. *Hear Res*. 2010;270(1-2):39-47. doi:10.1016/j.heares.2010.09.010
12. Anderson CA, Lazard DS, Hartley DEH. Plasticity in bilateral superior temporal cortex: Effects of deafness and cochlear implantation on auditory and visual speech processing. *Hear Res*. 2017;343:138-149. doi:10.1016/j.heares.2016.07.013
13. Villringer A, Dirnagl U. Coupling of brain activity and cerebral blood flow: basis of functional neuroimaging. *Cerebrovasc Brain Metab Rev*. 1995;7(3):240-276.
14. Boden S, Obrig H, Köhncke C, Benav H, Koch SP, Steinbrink J. The oxygenation response to functional stimulation: is there a physiological meaning to the lag between parameters? *Neuroimage*. 2007;36(1):100-107. doi:10.1016/j.neuroimage.2007.01.045
15. Obrig H, Wenzel R, Kohl M, et al. Near-infrared spectroscopy: Does it function in functional activation studies of the adult brain? *Int J Psychophysiol*. 2000;35(2-3):125-142. doi:10.1016/S0167-8760(99)00048-3
16. Steinbrink J, Villringer A, Kempf F, Haux D, Boden S, Obrig H. Illuminating the BOLD signal: combined fMRI–fNIRS studies. *Magn Reson Imaging*. 2006;24(4):495-505. doi:10.1016/j.mri.2005.12.034
17. Hillman EMC. Coupling mechanism and significance of the BOLD signal: a status report. *Annu Rev Neurosci*. 2014;37(1):161-181. doi:10.1146/annurev-neuro-071013-014111
18. Toronov VY, Zhang X, Webb AG. A spatial and temporal comparison of hemodynamic signals measured using optical and functional magnetic resonance imaging during activation in the human primary visual cortex. *Neuroimage*. 2007;34(3):1136-1148. doi:10.1016/j.neuroimage.2006.08.048

19. Jones M, Berwick J, Johnston D, Mayhew J. Concurrent optical imaging spectroscopy and laser-Doppler flowmetry: the relationship between blood flow, oxygenation, and volume in rodent barrel cortex. *Neuroimage*. 2001;13(6 Pt 1):1002-1015. doi:10.1006/nimg.2001.0808
20. Buxton RB, Uludağ K, Dubowitz DJ, Liu TT. Modeling the hemodynamic response to brain activation. *Neuroimage*. 2004;23 Suppl 1:S220-33. doi:10.1016/j.neuroimage.2004.07.013
21. Horner AJ, Andrews TJ. Linearity of the fMRI response in category-selective regions of human visual cortex. *Hum Brain Mapp*. 2009;30(8):2628-2640. doi:10.1002/hbm.20694
22. Soltysik DA, Peck KK, White KD, Crosson B, Briggs RW. Comparison of hemodynamic response nonlinearity across primary cortical areas. *Neuroimage*. 2004;22(3):1117-1127. doi:10.1016/j.neuroimage.2004.03.024
23. Friston KJ, Fletcher P, Josephs O, Holmes A, Rugg MD, Turner R. Event-related fMRI: characterizing differential responses. *Neuroimage*. 1998;7(1):30-40. doi:10.1006/nimg.1997.0306
24. Rees G, Howseman A, Josephs O, et al. Characterizing the relationship between BOLD contrast and regional cerebral blood flow measurements by varying the stimulus presentation rate. *Neuroimage*. 1997;6(4):270-278. doi:10.1006/nimg.1997.0300
25. Binder JR, Rao SM, Hammeke TA, Frost JA, Bandettini PA, Hyde JS. Effects of stimulus rate on signal response during functional magnetic resonance imaging of auditory cortex. *Brain Res Cogn Brain Res*. 1994;2(1):31-38.
26. Pérez-González D, Malmierca MS. Adaptation in the auditory system: an overview. *Front Integr Neurosci*. 2014;8(5):19. doi:10.3389/fnint.2014.00019
27. van de Rijt LPH, van Opstal AJ, Mylanus EAM, et al. Temporal Cortex Activation to Audiovisual Speech in Normal-Hearing and Cochlear Implant Users Measured with Functional Near-Infrared Spectroscopy. *Front Hum Neurosci*. 2016;10(February):48. doi:10.3389/fnhum.2016.00048
28. Ohnishi M, Kusakawa N, Masaki S, et al. Measurement of hemodynamics of auditory cortex using magnetoencephalography and near infrared spectroscopy. *Acta Otolaryngol Suppl*. 1997;532(s532):129-131. doi:10.3109/00016489709126161
29. Minagawa-kawai Y, Mori K, Naoi N, Kojima S. Neural Attunement Processes in Infants during the Acquisition of a Language-Specific Phonemic Contrast. 2007;27(2):315-321. doi:10.1523/JNEUROSCI.1984-06.2007
30. Minagawa-Kawai Y, Mori K, Sato Y, Koizumi T. Differential cortical responses in second language learners to different vowel contrasts. *Neuroreport*. 2004;15(5):899-903. doi:10.1097/01.wnr.00001
31. Sato Y, Utsugi A, Yamane N, Koizumi M, Mazuka R. Dialectal differences in hemispheric specialization for Japanese lexical pitch accent. *Brain Lang*. 2013;127(3):475-483. doi:10.1016/j.bandl.2013.09.008
32. Yoo S, Lee K-M. Articulation-based sound perception in verbal repetition: a functional NIRS study. *Front Hum Neurosci*. 2013;7(September):540. doi:10.3389/fnhum.2013.00540
33. Pollonini L, Olds C, Abaya H, Bortfeld H, Beauchamp MS, Oghalai JS. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hear Res*. 2014;309(December):84-93. doi:10.1016/j.heares.2013.11.007
34. Olds C, Pollonini L, Abaya H, et al. Cortical Activation Patterns Correlate With Speech Understanding After Cochlear Implantation. *Ear Hear*. 2015;37(3):1-13. doi:10.1097/AUD.0000000000000258
35. Abla D, Okanoya K. Statistical segmentation of tone sequences activates the left inferior frontal cortex: a near-infrared spectroscopy study. *Neuropsychologia*. 2008;46(11):2787-2795. doi:10.1016/j.neuropsychologia.2008.05.012
36. Bembich S, Demarini S, Clarici A, Massaccesi S, Grasso DL. Non-invasive assessment of hemispheric language dominance by optical topography during a brief passive listening test: a pilot study. *Med Sci Monit*. 2011;17(12):CR692-7.

37. Ehlis a-C, Ringel TM, Plichta MM, Richter MM, Herrmann MJ, Fallgatter AJ. Cortical correlates of auditory sensory gating: a simultaneous near-infrared spectroscopy event-related potential study. *Neuroscience*. 2009;159(3):1032-1043. doi:10.1016/j.neuroscience.2009.01.015
38. Okamoto M, Dan H, Sakamoto K, et al. Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10–20 system oriented for transcranial functional brain mapping. *Neuroimage*. 2004;21(1):99-111. doi:10.1016/j.neuroimage.2003.08.026
39. Sato H, Kiguchi M, Maki A, et al. Within-subject reproducibility of near-infrared spectroscopy signals in sensorimotor activation after 6 months. *J Biomed Opt*. 2006;11(1):014021. doi:10.1117/1.2166632
40. Noguchi Y, Takeuchi T, Sakai KL. Lateralized activation in the inferior frontal cortex during syntactic processing: event-related optical topography study. *Hum Brain Mapp*. 2002;17(2):89-99. doi:10.1002/hbm.10050
41. Sato H, Takeuchi T, Sakai KL. Temporal cortex activation during speech recognition: an optical topography study. Cognition 73, B55 - B66. *Cognition*. 1999;73(3):55-66.
42. Sato Y, Mori K, Koizumi T, et al. Functional lateralization of speech processing in adults and children who stutter. *Front Psychol*. 2011;2(April):70. doi:10.3389/fpsyg.2011.00070
43. Plichta MM, Gerdes ABM, Alpers GW, et al. Auditory cortex activation is modulated by emotion: A functional near-infrared spectroscopy (fNIRS) study. *Neuroimage*. 2011;55(3):1200-1207. doi:10.1016/j.neuroimage.2011.01.011
44. Kennan RP, Horovitz SG, Maki A, Yamashita Y, Koizumi H, Gore JC. Simultaneous Recording of Event-Related Auditory Oddball Response Using Transcranial Near Infrared Optical Topography and Surface EEG. *Neuroimage*. 2002;16(3):587-592. doi:10.1006/nimg.2002.1060
45. Zhang Q, Strangman GE, Ganis G. Adaptive filtering to reduce global interference in non-invasive NIRS measures of brain activation: how well and when does it work? *Neuroimage*. 2009;45(3):788-794. doi:10.1016/j.neuroimage.2008.12.048
46. Gagnon L, Perdue K, Greve DN, Goldenholz D, Kaskhedikar G, Boas DA. Improved recovery of the hemodynamic response in diffuse optical imaging using short optode separations and state-space modeling. *Neuroimage*. 2011;56(3):1362-1371. doi:10.1016/j.neuroimage.2011.03.001
47. Zee P van der, Arridge SR, Cope M, Delphy DT. The effect of optode positioning on optical pathlength in near infrared spectroscopy of brain. *Adv Exp Med Biol*. 1990;(277):79-84.
48. Okada E, Firbank M, Schweiger M, Arridge SR, Cope M, Delpy DT. Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head. *Appl Opt*. 1997;36(1):21-31. doi:10.1364/AO.36.000021
49. Cui X, Bray S, Bryant DM, Glover GH, Reiss AL. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *Neuroimage*. 2011;54(4):2808-2821. doi:10.1016/j.neuroimage.2010.10.069
50. Strait M, Scheutz M. What we can and cannot (yet) do with functional near infrared spectroscopy. *Front Neurosci*. 2014;8(May):117. doi:10.3389/fnins.2014.00117
51. Scarpa F, Brigadoi S, Cutini S, et al. A reference-channel based methodology to improve estimation of event-related hemodynamic response from fNIRS measurements. *Neuroimage*. 2013;72:106-119. doi:10.1016/j.neuroimage.2013.01.021
52. Fekete T, Rubin D, Carlson JM, Mujica-Parodi LR. The NIRS Analysis Package: Noise Reduction and Statistical Inference. Zuo X-N, ed. *PLoS One*. 2011;6(9):e24322. doi:10.1371/journal.pone.0024322
53. Plichta MM, Herrmann MJ, Baehne CG, et al. Event-related functional near-infrared spectroscopy (fNIRS): are the measurements reliable? *Neuroimage*. 2006;31(1):116-124. doi:10.1016/j.neuroimage.2005.12.008

54. Plichta MM, Heinzel S, Ehlis a-C, Pauli P, Fallgatter a J. Model-based analysis of rapid event-related functional near-infrared spectroscopy (NIRS) data: a parametric validation study. *Neuroimage*. 2007;35(2):625-634. doi:10.1016/j.neuroimage.2006.11.028
55. Tak S, Ye JC. Statistical analysis of fNIRS data: a comprehensive review. *Neuroimage*. 2014;85 Pt 1:72-91. doi:10.1016/j.neuroimage.2013.06.016
56. Plichta MM, Herrmann MJ, Ehlis a-C, Baehne CG, Richter MM, Fallgatter a J. Event-related visual versus blocked motor task: detection of specific cortical activation patterns with functional near-infrared spectroscopy. *Neuropsychobiology*. 2006;53(2):77-82. doi:10.1159/000091723
57. Minagawa-Kawai Y, Mori K, Furuya I, Hayashi R, Sato Y. Assessing cerebral representations of short and long vowel categories by NIRS. *Neuroreport*. 2002;13(5):581-584. doi:10.1097/00001756-200204160-00009
58. Minagawa-Kawai Y, Mori K, Sato Y. Different brain strategies underlie the categorical perception of foreign and native phonemes. *J Cogn Neurosci*. 2005;17(9):1376-1385. doi:10.1162/0898929054985482
59. Chen L-C, Sandmann P, Thorne JD, Herrmann CS, Debener S. Association of Concurrent fNIRS and EEG Signatures in Response to Auditory and Visual Stimuli. *Brain Topogr*. 2015;28(5):710-725. doi:10.1007/s10548-015-0424-8
60. Lockwood AH, Salvi RJ, Coad M Lou, et al. The Functional Anatomy of the Normal Human Auditory System : Responses to 0 . 5 and 4 . 0 kHz Tones at Varied Intensities. 1999:65-76.
61. Jäncke L, Shah NJ, Posse S, Grosse-Ryuken M, Müller-Gärtner HW. Intensity coding of auditory stimuli: an fMRI study. *Neuropsychologia*. 1998;36(9):875-883.
62. Uppenkamp S, Röhl M. Human auditory neuroimaging of intensity and loudness. *Hear Res*. 2014;307:65-73. doi:10.1016/j.heares.2013.08.005
63. Langers DRM, Backes WH, van Dijk P. Representation of lateralization and tonotopy in primary versus secondary human auditory cortex. *Neuroimage*. 2007;34(1):264-273. doi:10.1016/j.neuroimage.2006.09.002
64. Röhl M, Uppenkamp S. Neural coding of sound intensity and loudness in the human auditory system. *J Assoc Res Otolaryngol*. 2012;13(3):369-379. doi:10.1007/s10162-012-0315-6
65. Rinne T, Pekkola J, Degerman A, et al. Modulation of auditory cortex activation by sound presentation rate and attention. *Hum Brain Mapp*. 2005;26(2):94-99. doi:10.1002/hbm.20123
66. Sheth SA, Nemoto M, Guiou M, Walker M, Pouratian N, Toga AW. Linear and nonlinear relationships between neuronal activity, oxygen metabolism, and hemodynamic responses. *Neuron*. 2004;42(2):347-355. doi:10.1016/S0896-6273(04)00221-1
67. Harms MP, Melcher JR. Sound repetition rate in the human auditory pathway: representations in the waveshape and amplitude of fMRI activation. *J Neurophysiol*. 2002;88(3):1433-1450. doi:10.1152/jn.00156.2002
68. Tanaka H, Fujita N, Watanabe Y, et al. Effects of stimulus rate on the auditory cortex using fMRI with "sparse" temporal sampling. *Neuroreport*. 2000;11(9):2045-2049.
69. Weiss AP, Duff M, Roffman JL, Rauch SL, Strangman GE. Auditory stimulus repetition effects on cortical hemoglobin oxygenation: a near-infrared spectroscopy investigation. *Neuroreport*. 2008;19(2):161-165. doi:10.1097/WNR.0b013e3282f4aa2a
70. Grill-Spector K, Henson R, Martin A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci*. 2006;10(1):14-23. doi:10.1016/j.tics.2005.11.006
71. Nelken I. Stimulus-specific adaptation and deviance detection in the auditory system: experiments and models. *Biol Cybern*. 2014;108(5):655-663. doi:10.1007/s00422-014-0585-7
72. Hall DA, Johnsrude IS, Haggard MP, Palmer AR, Akeroyd MA, Summerfield AQ. Spectral and temporal processing in human auditory cortex. *Cereb Cortex*. 2002;12(2):140-149.

73. Strainer JC, Ulmer JL, Yetkin FZ, Haughton VM, Daniels DL, Millen SJ. Functional MR of the primary auditory cortex: an analysis of pure tone activation and tone discrimination. *AJNR Am J Neuroradiol*. 1997;18(4):601-610.
74. Samson F, Zeffiro TA, Toussaint A, Belin P. Stimulus complexity and categorical effects in human auditory cortex: an activation likelihood estimation meta-analysis. *Front Psychol*. 2010;1(January):241. doi:10.3389/fpsyg.2010.00241
75. Köchel A, Schöngassner F, Schienle A. Cortical activation during auditory elicitation of fear and disgust: a near-infrared spectroscopy (NIRS) study. *Neurosci Lett*. 2013;549:197-200. doi:10.1016/j.neulet.2013.06.062
76. Alho K, Rinne T, Herron TJ, Woods DL. Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. *Hear Res*. 2014;307:29-41. doi:10.1016/j.heares.2013.08.001
77. Lee AKC, Larson E, Maddox RK, Shinn-Cunningham BG. Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hear Res*. 2014;307:111-120. doi:10.1016/j.heares.2013.06.010
78. Santosa H, Hong MJ, Hong K-S. Lateralization of music processing with noises in the auditory cortex: an fNIRS study. *Front Behav Neurosci*. 2014;8(December):418. doi:10.3389/fnbeh.2014.00418
79. Kojima H, Suzuki T. Hemodynamic change in occipital lobe during visual search: visual attention allocation measured with NIRS. *Neuropsychologia*. 2010;48(1):349-352. doi:10.1016/j.neuropsychologia.2009.09.028
80. Remijn GB, Kojima H. Active versus passive listening to auditory streaming stimuli: a near-infrared spectroscopy study. *J Biomed Opt*. 2010;15(3):037006. doi:10.1117/1.3431104
81. Plichta MM, Herrmann MJ, Baehne CG, et al. Event-related functional near-infrared spectroscopy (fNIRS) based on craniocerebral correlations: reproducibility of activation? *Hum Brain Mapp*. 2007;28(8):733-741. doi:10.1002/hbm.20303
82. Kono T, Matsuo K, Tsunashima K, et al. Multiple-time replicability of near-infrared spectroscopy recording during prefrontal activation task in healthy men. *Neurosci Res*. 2007;57(4):504-512. doi:10.1016/j.neures.2006.12.007
83. Schecklmann M, Ehlis A-C, Plichta MM, Fallgatter AJ. Functional near-infrared spectroscopy: a long-term reliable tool for measuring brain activity during verbal fluency. *Neuroimage*. 2008;43(1):147-155. doi:10.1016/j.neuroimage.2008.06.032
84. Boas D a, Dale AM, Franceschini MA. Diffuse optical imaging of brain activation: Approaches to optimizing image sensitivity, resolution, and accuracy. In: *NeuroImage*. Vol 23. ; 2004:S275-88. doi:10.1016/j.neuroimage.2004.07.011
85. Jacques SL. Erratum: Optical properties of biological tissues: A review (Physics in Medicine and Biology (2013) 58). *Phys Med Biol*. 2013;58(14):5007-5008. doi:10.1088/0031-9155/58/14/5007
86. Durduran T, Choe R, Baker WB, Yodh AG. Diffuse optics for tissue monitoring and tomography. *Reports Prog Phys*. 2010;73(7):43. doi:10.1088/0034-4885/73/7/076701
87. Mehagnoul-Schipper DJ, van der Kallen BFW, Colier WNJM, et al. Simultaneous measurements of cerebral oxygenation changes during brain activation by near-infrared spectroscopy and functional magnetic resonance imaging in healthy young and elderly subjects. *Hum Brain Mapp*. 2002;16(1):14-23. doi:10.1002/hbm.10026
88. Wallois F, Mahmoudzadeh M, Patil a, Grebe R. Usefulness of simultaneous EEG-NIRS recording in language studies. *Brain Lang*. 2012;121(2):110-123. doi:10.1016/j.bandl.2011.03.010

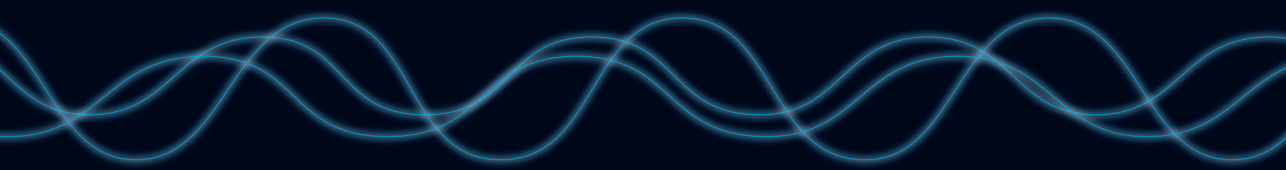
CHAPTER 5

Temporal cortex activation to audiovisual speech in normal-hearing and cochlear implant users measured with functional near-infrared spectroscopy

Luuk P.H. van de Rijt, A. John van Opstal, Emmanuel A.M. Mylanus, Louise V. Straatman, Hai Yin Hu, Ad F.M. Snik, Marc M. van Wanrooij

Keywords: functional near-infrared spectroscopy, fNIRS, audiovisual, auditory cortex, cochlear implant.

Frontiers in Human Neuroscience. 10:48.
DOI: https://10.3389/fnhum.2016.00048



ABSTRACT

Background

Speech understanding may rely not only on auditory, but also on visual information. Non-invasive functional neuroimaging techniques can expose the neural processes underlying the integration of multisensory processes required for speech understanding in humans. Nevertheless, noise (from fMRI) limits the usefulness in auditory experiments, and electromagnetic artefacts caused by electronic implants worn by subjects can severely distort the scans (EEG, fMRI). Therefore, we assessed audio-visual activation of temporal cortex with a silent, optical neuroimaging technique: functional near-infrared spectroscopy (fNIRS).

Methods

We studied temporal cortical activation as represented by concentration changes of oxy- and deoxy-hemoglobin in four, easy-to-apply fNIRS optical channels of 33 normal-hearing adult subjects and 5 post-lingually deaf cochlear implant (CI) users in response to supra-threshold unisensory auditory and visual, as well as to congruent auditory-visual speech stimuli.

Results

Activation effects were not visible from single fNIRS channels. However, by discounting physiological noise through reference channel subtraction, auditory, visual and audiovisual speech stimuli evoked concentration changes for all sensory modalities in both cohorts ($p<0.001$). Auditory stimulation evoked larger concentration changes than visual stimuli ($p<0.001$). A saturation effect was observed for the audiovisual condition.

Conclusions

Physiological, systemic noise can be removed from fNIRS signals by reference channel subtraction. The observed multisensory enhancement of an auditory cortical channel can be plausibly described by a simple addition of the auditory and visual signals with saturation.

INTRODUCTION

Viewing a talking person's face and mouth may enhance speech understanding in noisy environments.[1,2] This effect is due to multisensory integration, in which congruent unisensory signals from multiple modalities are merged to form a coherent and enhanced percept.[3] The mechanisms underlying multisensory integration have been studied extensively at the single-neuron level in animals (review on seminal work in anesthetized cat,[3] and in psychophysical eye movement studies in humans[4,5]). How these mechanisms relate to the neural underpinnings of human speech recognition has been studied with neuroimaging and electrophysiological techniques.[6–8] In individual neurons, the multisensory responses can be much greater than the linear sum of individual unisensory responses. In contrast, for fMRI data, integrating across millions of neurons, super-additivity is typically not found, although multisensory responses are slightly greater than the maximum or mean of the individual unisensory responses[9].

Here, we attempt to characterize multisensory speech processing by applying an alternative, non-invasive method to record neural activity: functional near-infrared spectroscopy (fNIRS). FNIRS assesses cortical hemodynamic changes in blood oxygenation based on changes in the transmission of near-infrared light through biological tissue and its absorption by oxygenated (HbO_2) and deoxygenated (HbR) hemoglobin.[10–14] As fNIRS is a non-invasive, minimally-restrictive and quiet optical technique (as opposed to PET[15] and fMRI[16]), it is ideally suited for auditory studies[17–20] on human subjects of all ages. Furthermore, this technique does not suffer from the severe limitations imposed by electro-magnetic implants (e.g. cochlear implant (CI), [21]). Therefore, it has been successfully used to study human auditory cortex activation by speech stimuli in normal-hearing adults[20] and deaf adults and children using a CI.[22–24]

In this study, we use fNIRS to record supra-threshold auditory, visual and audiovisual speech-evoked activity from temporal cortex of normal-hearing adults and post-lingually deaf unilateral CI users. We use a limited number of fNIRS channels in order to reduce the time and complexity of applying the optodes. Figure 5.1 illustrates the rationale and possible outcomes of our experiments. Pure auditory stimulation is expected to produce a typical hemodynamic response profile (blue[25,26]) in line with the BOLD response (for review, see[27,28]), that reaches its peak at about 6-10 s after a transient stimulus onset. In contrast, pure visual stimulation may produce at best a lower response (red) for a predominantly auditory-responsive area, which could be due to the expectation of a sound being produced by the moving lips.[29] Evidence for clear audiovisual integration would be found if the audiovisual response exceeds mere linear summation of the two unimodal responses, i.e. the additive response, or when it falls below the unisensory auditory response (inhibition).[30] A sub-additive response might be due to either a multisensory or nonlinear saturation effect.

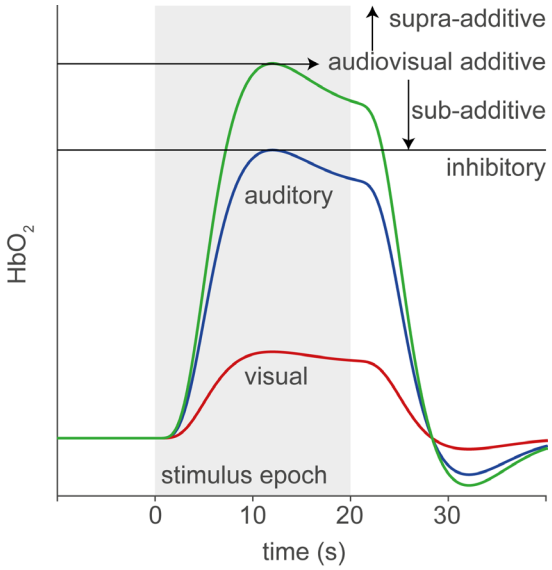


Figure 5.1. Rationale of audiovisual fNIRS experiments.

Hemodynamic responses taken from temporal cortex will differ for the different stimulus modalities, such that to an auditory (blue line) stimulus (grey patch) the response amplitude is larger than to a visual (red) stimulus. We test for potential multisensory integration at the level of temporal cortex, by comparing the hemodynamic response to bimodal stimulus presentation (green) to the linear sum of the visual and auditory responses (additive). Supra- or sub-additive effects on the audiovisual response may be a signature for audiovisual integration.

We also tested a limited number of post-lingually deaf unilateral CI users mainly to examine the feasibility of recording multisensory speech processing at the level of temporal cortex with easily-applied, 4-channel fNIRS in the presence of electrical innervation of the auditory nerve by a CI.

METHODS

Subjects

Thirty-three adult native Dutch-speaking normal-hearing subjects (age: 18-62 years, median 29, 15 male, 18 female) and 5 adult Dutch-speaking post-lingually deaf unilateral CI users (age: 55-59 years, median 57, all female) were recruited to participate in this study. All normal-hearing subjects (within 20 dB of audiometric zero, range 0.5 – 8 kHz) and all CI users had normal or corrected to normal vision. Experiments were conducted after obtaining written consent from the subject. The experiments were approved by the Ethics Committee of Arnhem-Nijmegen (project number NL24364.091.08, October 18, 2011) and were carried out in accordance with the relevant institutional and national regulations and with the World

Medical Association Helsinki Declaration as revised in October 2008 (http://www.wma.net/en/30publications/10policies/b3/).

Table 5.1. Subject demographics of post-lingually deaf cochlear implant users.

CI user	Implanted ear	Etiology	Cochlear implant use (years)	Device
P1	Left	Cogan syndrome	12	C2HighFocus2[1]
P2	Right	Progressive	5	Nucleus24RCA[2]
P3	Left	Progressive	8	C1[1]
P4	Left	Sudden deafness	19	Nucleus 22[2]
P5	Left	Progressive	7	Nucleus24RCS[2]

[1]Advanced Bionics, Stäfa, Switzerland
[2] Cochlear Headquarters, Sydney, Australia

Experimental setup

Subjects sat comfortably in a reclining chair, to reduce head movements and to minimize low-frequency so-called Mayer waves, that are presumably caused by slow variations in blood pressure.[31] The experiment was performed in a darkened experimental room (3.2 × 3.2 × 3.5 m) in which the walls and the ceiling were covered with black acoustic foam that eliminated echoes for sound frequencies > 500 Hz.[32] Background noise level was less than 30 dB, A-weighted (dBA).[33]

Functional near-infrared spectroscopy data were collected with a pulsed continuous-wave NIRS instrument with 4 optical sources and 2 photodetectors (Oxymon MKIII Near-Infrared Spectrophotometer, Artinis Medical Systems BV, Elst, the Netherlands). Each optical source consisted of two lasers with emission wavelengths of 858 or 861 nm and 765 nm. For a comprehensive review of the principles and practicalities of continuous-wave fNIRS, see e.g. Scholkmann.[34]

The fNIRS probe template (Fig. 5.2A and B) consisted of two optical sources and a single detector, typically on both sides of the head (see below), with source-detector distances of 25 and 35 mm, termed reference or shallow and deep channel, respectively. Sources and detectors were embedded in plastic molds, which were secured in place on the skull by adjustable straps. The temporal cortex was located based on the 10-20 system,[35] which roughly estimates its location at T7 for the left hemisphere and T8 for the right hemisphere[36]; Fig. 5.2A and B). As fNIRS measures brain activity over a diffuse area, we did not pinpoint the exact cortical areas per subject: based on Monte Carlo simulations by others,[37–40] the average photon path from source to photodetector is estimated to be an ellipsoid with a penetration depth of approximately 2 to 3 cm. Specifically, the current fNIRS probe template is expected to cover a large area of the temporal cortex.[c.f.22]

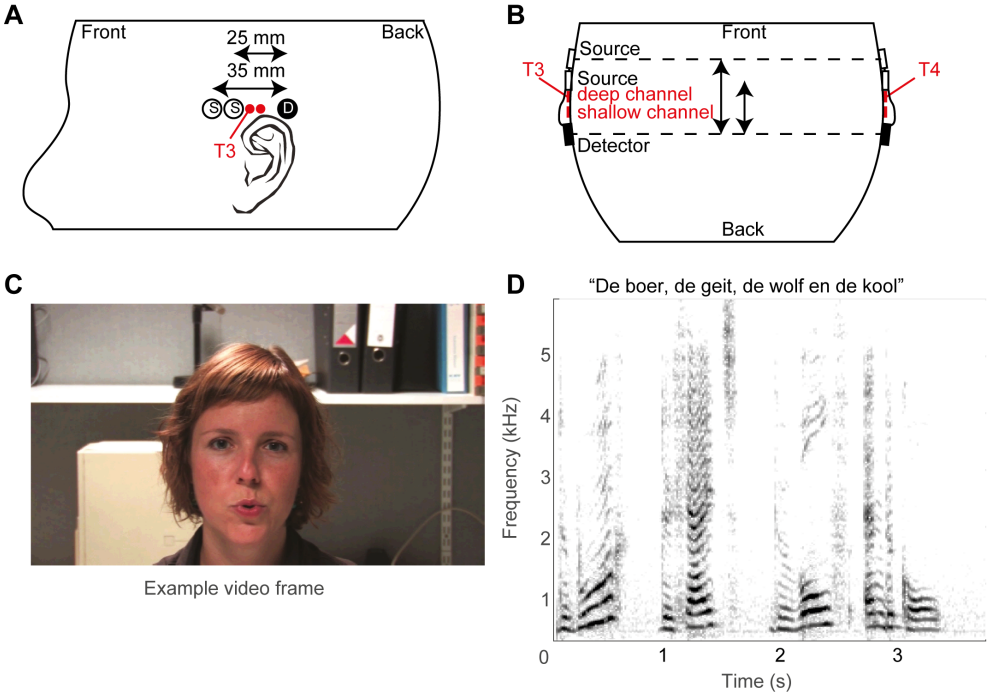


Figure 5.2. Methodological overview.

A) Schematic layout of optical sources (open circles) and photo detectors (filled circles) on the left hemisphere, and **B)** schematic top view of probe layout. The estimated T7 and T8 positions of the 10/20 system are also indicated, as are the supposed superficial centers of the deep and shallow channels (red filled circles). **C)** An example video frame. **D)** A spectrogram of an example sound snippet (the title shows the first words of the story).

For 21 normal-hearing subjects the optodes were positioned by aligning the mid-point of the long-distance (35 mm) source-detector pairs above the preauricular point at the T7 and T8 location of the International 10/20 system on the left and right hemisphere, respectively ([41]). For the other 12 normal-hearing subjects, who were measured prior to the other subjects, only one side was recorded (left hemisphere, T7). For the CI users, only the hemisphere contralateral to the implant was measured with 2 sources and 1 detector, to avoid placement problems of the optodes over the implant. The straps were adjusted to guarantee secure coupling between optodes and scalp at acceptable comfort levels of the subject. Secure coupling was verified online by the presence of a detectable photon count and of a clear cardiac oscillatory response in the raw NIRS trace measured before the experiment. The optodes were connected via optical fibers to the NIRS instrument. The company's software Oxysoft controlled data acquisition, and allowed for online observation of the data. The data were stored at a sampling rate of either 10 (for the early measurements, which included 12 normal-hearing subjects and all 5 CI users) or 250 Hz (for later measurements on 21 normal-hearing subjects). For data analysis, the latter data were downsampled to 10 Hz.

Stimuli

The stimuli were composed of digital video recordings of a female speaker reading aloud children's stories in Dutch (Fig. 5.2C and D). In the auditory-only condition, the voice was presented without visual input (Fig. 5.2D). In the visual-only condition, the video of the woman reading the story was presented on the screen without the auditory signal (Fig. 5.2C). In the auditory-visual condition, the video was presented with the corresponding auditory input. The recordings were digitally edited into 36 20.5-s segments, each consisting of a single vignette from one of three stories (in Dutch: "De boer, de geit, de wolf en de kool", "De professor", and "De prinses"). The three stimulus conditions were presented interleaved in pseudorandom order within a single block. Stimulus generation was controlled by a Dell PC (Dell Inc., Round Rock, TX, USA) running Matlab version 2009b (The Mathworks, Natick, Massachusetts) using Psychophysics Toolbox 3 extensions.[42–44] Sounds were presented through headphones (Sennheiser PCX 350 NoiseGuard, Sennheiser electronic GmbH & CO KG, Wedemark, Lower Saxony, Germany, noise cancellation off) at a comfortable listening volume of 55 dBA, while the video was presented on the Dell PC's monitor. As the implant interfered with placement of headphones for three out of five CI users, the acoustic stimuli to these CI users was alternatively presented via the direct input to the CI or via a free-field speaker.

Paradigm

The 36 segments were played in chronological order, each followed by a silent, dark period ranging from 25 to 50s (randomly drawn from a uniform distribution). Even the shortest intermittent period of 25s allowed the hemodynamic response to return to baseline, while the randomization limited time locking of any periodic physiological signal to stimulus onsets. The segments were presented in three blocks of 12 stimuli each. A single session consisted of these three blocks with intermittent breaks of about 4 to 5 min wherein the light was turned on. Every block started with a baseline measurement (in silence and darkness) of 2 minutes. A session of three blocks (36 segments) took about 45 minutes to complete.

For every block, the 12 segments were pseudo-randomly assigned to an experimental condition (4 segments auditory-only, 4 segments visual-only, 4 segments auditory-visual). Subjects were instructed to pay attention to the segments (watching, listening, both), and were asked afterwards whether they understood the gist of the storyline. Other than that, subjects were not given further task instructions.

Analysis

Signal processing

The optical densities for each channel and wavelength were stored on disk (in the native .oxy-format from the Artinis system) for offline analysis in Matlab (Release, 2014b, the Mathworks, Inc, Natick, MA, USA). Data was read into Matlab via Artinis' proprietary function *oxysoft2matlab*. The 250-Hz sample-rate data were downsampled to 10 Hz (using the *resample* function from Matlab's Signal Processing Toolbox), for computational efficiency.

Physiological noise, such as heart pulsation, respiration, and Mayer waves[45] is mixed with cortical activity in the fNIRS signal. A clear cardiac oscillation is regarded as evidence for a proper contact between the optical probes and the scalp.[46] Therefore, following Pollonini et al. (20), we determined the scalp coupling index (SCI) as the correlation between the two photodetected signals at 765 and ~860 nm, band-pass filtered between 0.5 and 2.5 Hz (typical frequency range for heart rate that excludes low-frequency fNIRS activity), for every optode. Typically, the SCI was highly positive (median 0.98), as expected from physiological signals that have no origin in the neural source,[47] and only 24 out of 354 channels [(21 normal-hearing subjects x 2 hemispheres + 12 normal-hearing subjects x 1 hemisphere + 5 CI users x 1 hemisphere) x 2 channels x 3 recording blocks] had an SCI less than 0.9. These 24 low-SCI channels were rejected from further analysis as we deemed those indicative for poor contact between optode and scalp. Then, to remove cardiac, respiratory, and Mayer wave noise sources, we used the *removeheartbeat* function from the NIRS Analysis Package[48]; in short, this algorithm extracts an oscillatory template from a narrow-frequency filtered average of all channels per subject (with the filter band containing the oscillatory frequencies of interest), and subtracts this from each channel. Then, we band-pass filtered the signals between 0.008 and 0.1 Hz (Fig. 5.3A and B, red and yellow curves). Next, the data was de-trended using a 20^{th}-degree polynomial in order to remove slow temporal drifts (Fig. 5.3A and B, black and purple curves).[49] These processed optical densities were converted to changes in oxygenated and deoxygenated hemoglobin concentration (HbO_2 and HbR, respectively) using the modified Lambert-Beer law.[11,50] Subsequently, the preprocessed data were normalized by the variance in each recorded signal for the entire session.

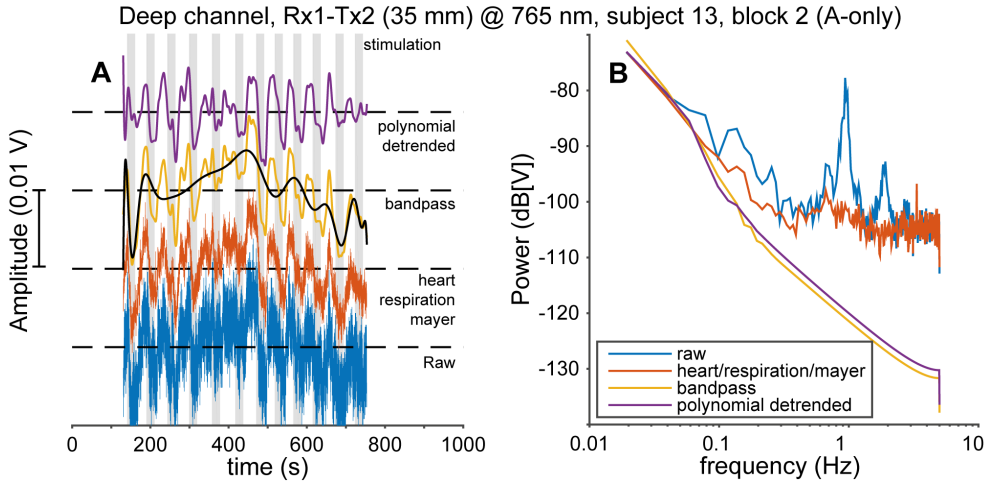


Figure 5.3. Pre-processing.

A) The data are preprocessed in several steps. First, cardiac, respiratory and Mayer oscillations in the raw data (blue, bottom) are removed (red). Then the data are bandpass-filtered between 0.008 and 0.1 Hz (yellow). Subsequently, slow-moving drifts are identified by a polynomial fit (black), which is removed to yield the final signal (purple). **B)** The effects of every pre-processing step on the power spectrum of the data in A.

Despite these filtering procedures, a considerable amount of noise originating from non-cortical physiological processes still remained.[34] To deal with this, we applied *reference channel subtraction* (RCS).[39,51] This assumes that the shallow channel (the signal originating from the shorter-distance optode source-detector) is dominated by non-cortical signals, while the deep channel (the signal arising from the longer-distance optode source-detector) also includes more of the cortical event-related signal of interest. Therefore, we determined the fNIRS signal as the residual signal from a simple linear regression between the deep and shallow channels (Fig. 5.4C). Note that we applied the normalization of data in the graphs both before and after RCS, so that the signals are scaled with respect to the data variance, and are hence dimensionless. An individual trace of HbO_2 for a single normal-hearing subject (NH1) for the shallow (black line), deep channel (red line) and the residual signal, during presentation of auditory snippets, is plotted in Fig. 5.4A. An example of how RCS can affect the average evoked response at the single subject-level is shown in Fig. 5.4B. Even though we can expect that the shallow channel might contain some cortical signal because of the relatively large distance of 25 mm, the RCS procedure in general improved the beta coefficients (Fig. 5.4D; see also section *GLM*) and the signal response (Fig. 5.4E). Because the same (systemic, not event related) noise is present in deep and reference channels, it is successfully removed by RCS. As a result, the variance in the deep channel signal decreases, and hence the signal-to-noise ratio increases. Normalization with a smaller variance leads to an increase of the beta coefficient (Fig. 5.4D), and to the appearance of a clear average activation signal (Fig. 5.4E).

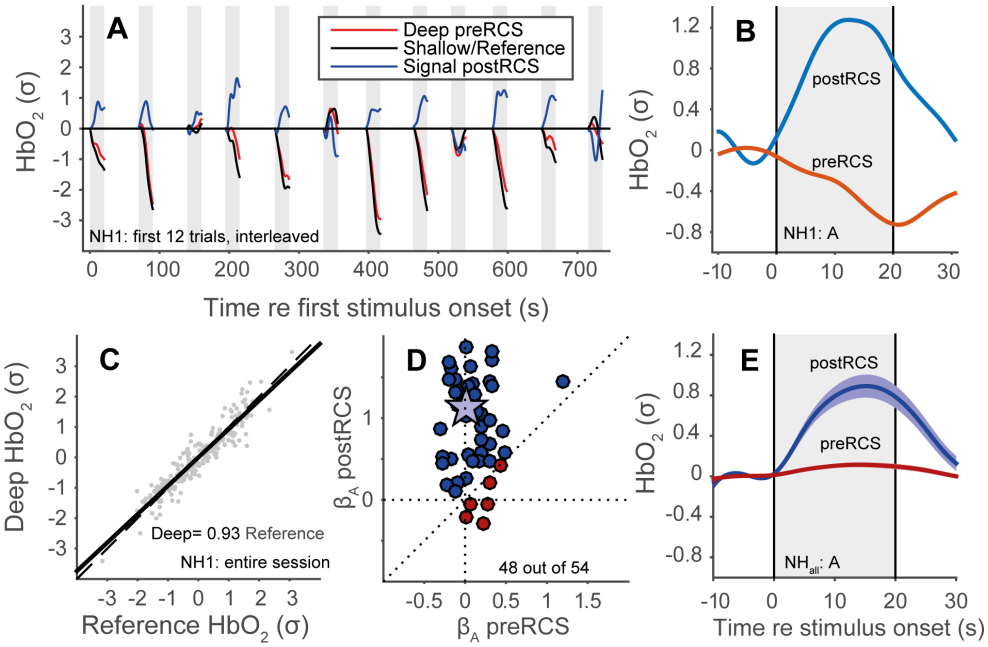


Figure 5.4. Reference channel subtraction.

A) Normalized HbO_2 data for a normal-hearing subject (NH1), 12 auditory trials, colors denote deep pre-RCS channel (red), shallow/reference channel (black) and post-RCS / residual signal (blue). **B)** Averaged normalized HbO_2 data for 12 auditory stimuli of a normal-hearing subject (NH1). Red line represents data before RCS. The blue line represents data after RCS. **C)** Linear regression between the deep and the shallow channel HbO_2 signals. **D)** Regression coefficients after RCS versus before RCS (see Statistics); blue indicates improvement, red inhibition, star subject NH1. **E)** Normal-hearing cohort pre-RCS (red line) and post-RCS (blue line) averages (thick line) and standard error of the means (patch) during auditory stimulation.

STATISTICS

Average

Functional signals were averaged across the twelve repeats of each stimulus modality to calculate the average hemodynamic response for each participant and hemisphere. These traces were averaged across participants and hemispheres (no significant hemispheric differences were observed for the bilaterally-measured 21 normal-hearing subjects according to a Wilcoxon signed-rank test, $p>0.05$ for both HbO_2 and HbR) to determine the mean response for both cohorts.

GLM

We compared both the measured concentration changes of HbO_2 and HbR to a predicted hemodynamic response function (HRF). The HRF consists of a canonical impulse response function h (as used by the SPM toolbox;[52,53]:

$$h(\tau) = \frac{1}{\Gamma(6)}\tau^5 e^{-\tau} - \frac{1}{6\Gamma(16)}\tau^{15} e^{-\tau} \quad (5.1)$$

(with τ time and Γ the gamma function), which peaks at ~5s, convolved with the boxcar function (1 during stimulus presentation, 0 otherwise). After convolution the HRF signal is expected to peak at ~12 s. All pre-processing steps performed on the data were also applied to the HRF signal.

We employed a general linear model (GLM) to quantify the strength between the measured responses to each stimulus condition and the HRF. This model assumes that auditory (β_a) and visual (β_v) inputs independently elicit a hemoglobin concentration change. An extra, third component (β_{av}) is added in this model, which represents the type and amount of multisensory integration during the presentation of audiovisual stimuli:

$$y(t) = X_A(t)\beta_a + X_V(t)\beta_v + X_{AV}(t)\beta_{av} + \varepsilon(t) + C \quad (5.2)$$

with fNIRS data y(t), the explanatory variables X(t), constant regression coefficients β, offset C and Gaussian noise ε(t).

For each GLM fit, we determined the goodness of fit (R^2-value, and the corresponding F and p-values). We took the significance of every regression coefficient as a measure of activation compared to baseline, by determining the corresponding t- and one-sided p-value (larger than 0 for HbO_2 and smaller than 0 for HbR).

Comparisons

To determine whether the beta values differed from a distribution with median 0, the Wilcoxon signed-rank test was applied per cohort. Also, we determined the slope between regression coefficients by determining the optimal fit through simple linear regression. The Wilcoxon Rank Sum test differences in regression coefficients between cohorts were determined. Significance was assessed at the 0.05 alpha level.

RESULTS

Functional NIRS - representative single subject data

We measured fNIRS activity over the temporal cortex of thirty-three normal-hearing subjects and 5 CI users while they were watching and/or listening to auditory, visual and audiovisual speech stimuli (Fig. 5.5). Individual traces of HbO_2 and HbR signals for a single representative normal-hearing subject (NH17) generally increase and decrease, respectively, during the stimulus epochs (Fig. 5.5A and B). Still, despite the extensive pre-processing (see Methods), signal drift, typical for fNIRS measurements,[22] occurs also during the silent dark periods. To deal with this stimulus-independent noise, we averaged the signals over the 12 trials per stimulus modality (Fig. 5.5C and D). The normalized, average HbO_2 over the 12 auditory-only (A) stimuli increases from baseline at sound onset reaching its maximum after about 15 s (Fig. 5.5C, blue), which is slightly more (~9%) than the average for the 12 audiovisual (AV) stimuli (Fig. 5.5C, green). The visual (V) trial average (Fig. 5.5C, red), while also increasing, only reaches a maximum of ~27% of the A maximum. These increases are mirrored in the HbR decreases, albeit with a slightly lower amplitude (Fig. 5.5D). After stimulus offset, HbO_2 and HbR return gradually to baseline (within 10 s). Typically, and as exemplified for this subject, the hemodynamic response corresponds well to the actual signals (cf. Fig. 5.1 and Fig. 5.5B).

Hemodynamic response shapes to auditory, visual and audiovisual stimulation for normal-hearing subjects and CI users

To reveal the shape of the cortical hemodynamic response, we averaged the trial averages over subjects for the A, V, and AV modalities, for the time interval between 10 s before stimulus onset and 10 s after stimulus offset (Fig. 5.6). All modalities demonstrated similar response shapes, albeit with varying amplitudes. The signals changed after stimulus onset (increase for HbO_2 and decrease for HbR) followed by a recovery back to baseline after stimulus offset. The data of the CI users (Fig. 5.6B and 6D) exhibited similar trends (Fig. 5.6A and C), although the standard errors were slightly larger (also due to the lower number of subjects in the CI user cohort). Moreover, the observed response resembled the predicted response shape (cf. Fig. 5.1), at least qualitatively. These similarities in response shapes for all cohorts, modalities and prediction indicate that fNIRS can consistently measure temporal cortical responses to auditory and visual stimuli in normal-hearing and cochlear-implanted adults.

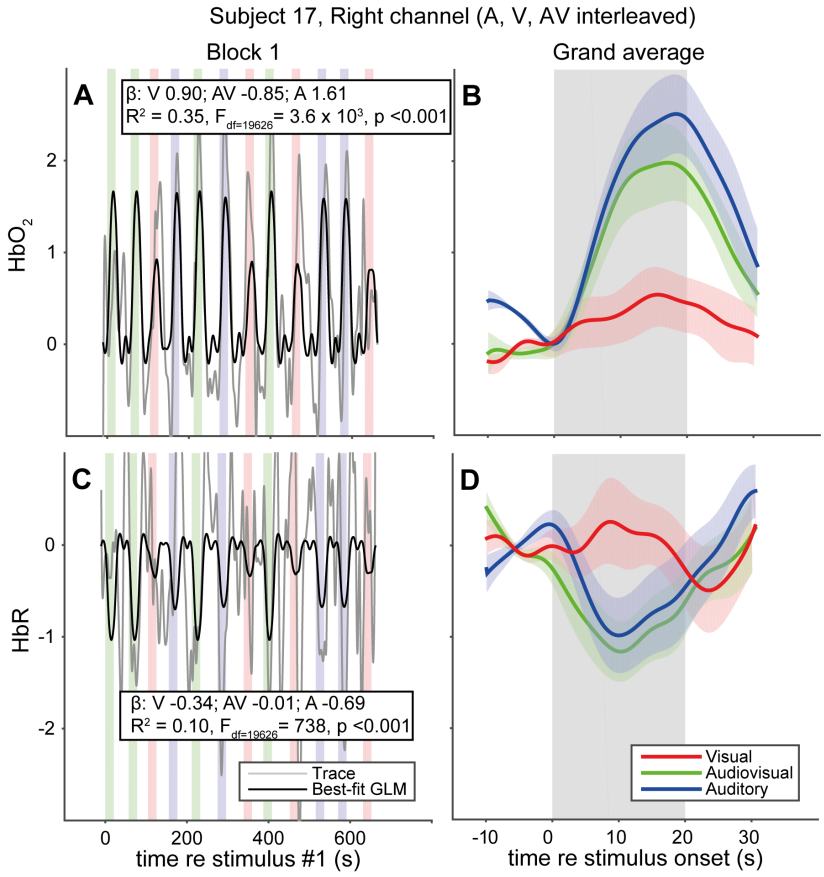


Figure 5.5. fNIRS data of a representative normal-hearing subject.

A) HbO_2 and **C)** HbR traces in a single block of a single normal-hearing subject. Average of the 12 **B)** HbO_2 and **D)** HbR responses measured for A, V and AV stimuli. Colors denote stimulus modality; auditory (blue), visual (red) and audiovisual (green). Rectangular patches denote stimulus activation. The best-fit (predicted) canonical hemodynamic response is shown in (A) and (C) as a black line. Insets in (A) and (C) provide the beta values for individual modalities and the goodness of fit. Shaded areas depict standard error of the mean over trials.

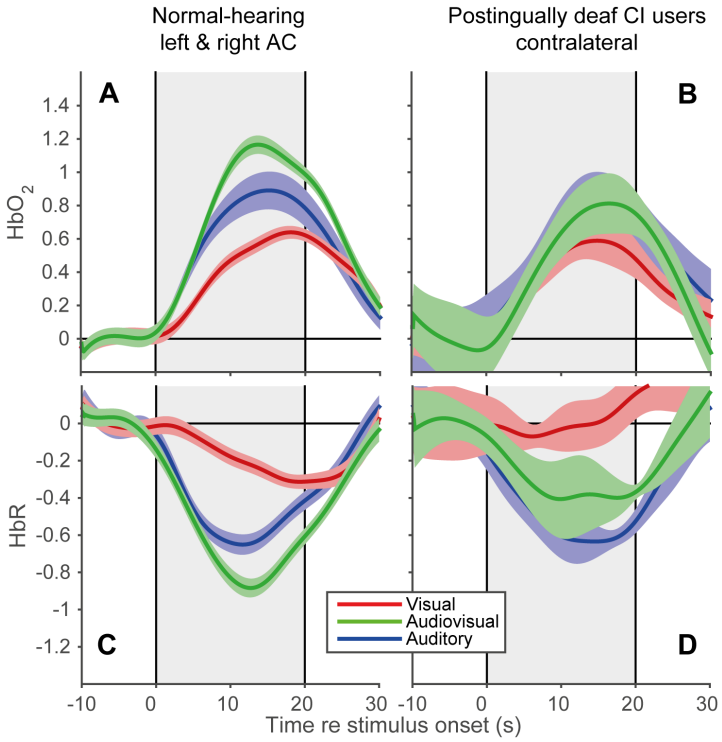


Figure 5.6. Grand average hemodynamic response of normal-hearing subjects and CI users.

Grand average responses for HbO_2 of **A)** normal-hearing subjects and **B)** CI users. Grand average responses for HbR of **C)** normal-hearing subjects and **D)** CI users. For the normal-hearing subjects 54 channels (21 bilateral, 10 unilateral) and for the CI users 5 unilateral channels were recorded. Grey rectangular patch denotes stimulus activation. Colors denote: red – visual; blue – auditory; green – audiovisual stimulation. Shaded areas depict standard error of the mean over subjects.

Cortical hemodynamic amplitude changes reveal saturation

To quantify the evoked responses, we fitted a general linear model (Eqn. 5.1-5.2; *see Methods*) that assumes that auditory and visual inputs independently elicit a hemoglobin amplitude change, also during audiovisual stimulation. An extra, third component is added in this model, which represents the type and amount of multisensory integration during the presentation of audiovisual stimuli. The analysis yields a set of three beta coefficients (Fig. 5.7) representing the modeled amplitude changes for each response component (auditory, visual and audiovisual interaction), for each subject (both cohorts), for both hemispheres (if applicable), and for both HbO_2 and HbR. In line with the grand average response for the normal-hearing cohort (Fig. 5.6A and C), significant activation was observed for the majority of single HbO_2 and HbR channels in normal-hearing subjects by auditory and/or visual stimulation (A: 52/54 and 50/54; V: 47/54 and 38/54 regression coefficients were larger/smaller than 0, for HbO_2 and HbR, respectively). Most channels did not exhibit an additional audiovisual component (AV: 7/54 and 9/54).

In line with the significance of unisensory individual channel activation, the coefficients for both the auditory and visual components reveal a general positive amplitude change for HbO_2 (Wilcoxon signed-rank test; for auditory coefficients: $p<0.001$, $z=6.8$, rank = 1779, and for visual coefficients: $p<0.001$, $z = 6.2$, rank = 1709) on a group-level, although there is a large intra-coefficient variability, with beta values ranging between -0.3 and 1.9. In contrast, comparisons between coefficients show a systematic trend of visual coefficients being smaller than auditory coefficients (Fig. 5.7A open circles; Wilcoxon signed-rank test: $p<0.001$, $z=5.8$, rank = 1657). A similar, opposite pattern arises for HbR (Fig. 5.7D, Wilcoxon signed-rank test: $p<0.001$, $z=-6.6$, rank = 11; $p<0.001$, $z=-5.3$, rank = 182, for A and V, respectively; for comparison between A and V: $p<0.001$, $z=-5.1$, rank = 203). The auditory data signify that we can reliably obtain auditory responses from temporal cortex with fNIRS, and the slightly weaker visual response data arguably imply that cross-modal, visual activation can arise from the same recording site (see also Discussion, *Multisensory integration versus saturation*).

To test for multisensory integration, researchers typically compare the bimodal response to the largest unimodal response[30] (Fig. 5.1). As the far majority of auditory coefficients are larger than the visual coefficients (HbO_2: 49 of 54; HbR: 43 of 54), we chose to compare only the auditory response with the bimodal response for all subjects (Fig. 5.7B and E). Note that the bimodal amplitudes are constituted by the sum of the auditory, visual and audiovisual-interaction coefficients (see Methods). These audiovisual amplitudes are highly similar to the auditory coefficients as all points lie close to the unity line, both for HbO_2 and HbR (Fig. 5.7B and E; Wilcoxon signed-rank test: for HbO_2 $p=0.14$, $z=-1.5$, rank = 688, slope: 0.95; for HbR $p=0.34$, $z=0.95$, rank = 1011, slope = 0.89).

The audiovisual interaction components are almost exactly inversely related to the visual amplitudes (Fig. 5.7C and F; i.e. data points lie close to $y=-x$ line, regression slopes: -0.93 and -0.87 for HbO_2 and HbR, respectively). This might be indicative of a saturation effect as the extra audiovisual interaction effect almost exactly counterbalances any effect a visual component might have (see also Discussion).

Concentration changes evoked in the five CI users (Fig. 5.7, grey squares) resembled those evoked in the normal-hearing subjects. Specifically, the auditory and visual coefficients for the CI users ranged from 0.02 to 1.4 (Fig. 5.7A) and from -1.5 to 0.7 (Fig. 5.7B) for HbO_2 and HbR, respectively. Significant activation from baseline for auditory components was observed for 4 out of 5 and 5 out of 5 subjects for HbO_2 and HbR, respectively. The visual components were significant for 5 out of 5 and 3 out of 5 subjects, respectively. The audiovisual components were significant for 0 out of 5 and 2 out 5 CI users.

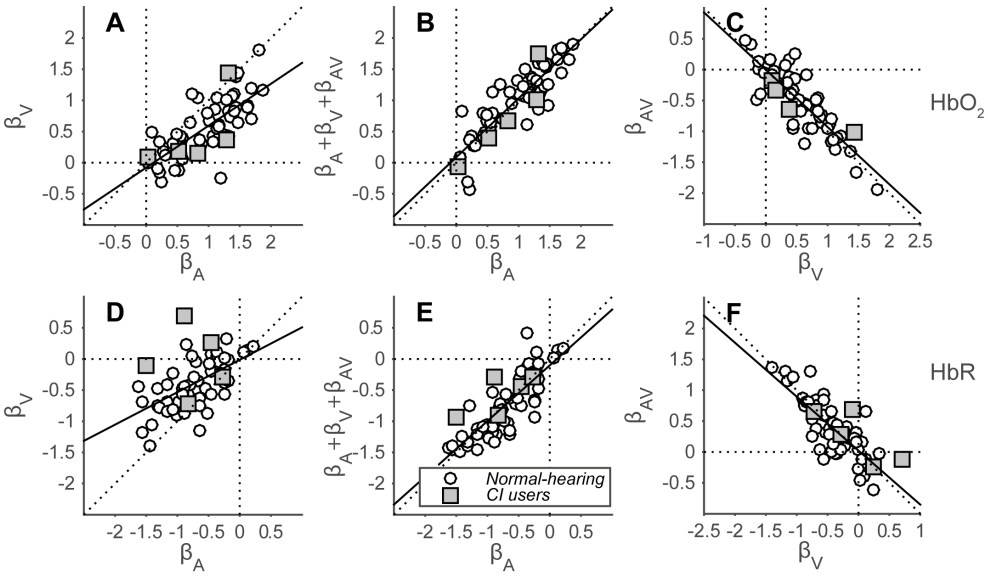


Figure 5.7. Beta coefficients of the GLM.

HbO_2 **(A-C)** and HbR **(D-F)** β-coefficients are shown for all subjects for all stimulus modalities. (A and D): Visual versus auditory regression coefficients. (B and E): summed auditory, visual and audiovisual-interaction (representing the AV response amplitude) versus auditory regression coefficients. (C and F): audiovisual-interaction versus visual regression coefficients. Open circles indicate normal-hearing subjects – filled squares indicate CI users. The black line depicts the best-fit regression line.

DISCUSSION

Overview

In this study, we assessed audio-visual activation in temporal cortex with fNIRS. Specifically, we studied cortical activation as present in concentration changes of oxy- and deoxy-hemoglobin of normal-hearing adult subjects and post-lingually deaf unilateral CI users in response to auditory, visual and auditory-visual speech stimuli. Sounds evoked larger concentration changes than visual stimuli (Fig. 5.7A and D). The audiovisual fNIRS signal resembled the purely auditory response (Fig. 5.7B and E) with the visual component being almost exactly inversely related to the audiovisual component (Fig. 5.7C, F). Interestingly, hemodynamic concentration changes evoked in the CI users strongly resembled those of the normal-hearing subjects (Fig. 5.7).

Feasibility

Since we show robust evoked activity in the temporal cortex for three different sensory conditions in fNIRS data on a group level (Fig. 5.6), fNIRS seems suited to study auditory and visual processing. Furthermore, the responses for the various modalities were consistent when

compared against each other within subjects (Fig. 5.7). Nevertheless, a large idiosyncratic variation on single-modality fNIRS responses (Fig. 5.7) may limit the use of this technique on single subject level. The causes for the observed inter-subject variance might be threefold: 1) methodological, 2) analytical and 3) experimental. First, we will briefly explain and discuss these issues.

A methodological source of inter-subject variability in our data is the placement of the optodes. According to the 10/20 International System, we placed the optodes based on external anatomical landmarks (i.e. nasion and inion).[22,24] Alternatively, one could place the optodes based on functional landmarks, by conducting a short functional localizer experiment, such that the location of the maximal response is searched for in a pilot experiment. For example, tone responsiveness could be determined in order to localize basic auditory-responsive regions.[17] Another optimization of the current 2-channel optode design would be to use multichannel optode arrays,[20,23,24] so that only the channel(s) with the strongest evoked responses are analyzed (as is current practice, e.g.[19]), or to determine a clearly localized response (e.g.[19,20,54]). In addition, one might consider to determine the individual optode positions in such a multi-optode array from anatomical MRI scans per subject.[55–57]

This study reveals that reference channel subtraction (RCS) is a very important factor in the analysis (Fig. 5.4). Typically, this is not performed,[51] although it is considered essential in removing systemic noise.[34] Without RCS, no effect in any of the sensory modalities would be observed in the current data (not shown here, but see Fig. 5.4D and E). A refinement of the current procedure would be to systematically change the inter-optode distances in order to optimally record purely systemic noise (in the reference channel) and the largest evoked hemodynamic signal (in the deep channel). The use of a multichannel optode array with varying optode distances might be ideally suited to disentangle the systemic noise from the evoked signal.

Finally, the experimental paradigm might in itself explain the variability. In this case, it might turn out that idiosyncratic variability is real, and that the amount of neural or hemodynamic activity varies on an individual basis. Variation might then be reduced if the evoked response is maximized for all subjects by specifically tailored experimental paradigms. For example, in the current paradigm, subjects were passively exposed to the stimuli, while active listening typically results in increased cortical activity in humans[58–60] and non-human primates.[61–64] Furthermore, one might refine the stimuli in order to elicit optimal responses from the brain area under consideration. Here, we used speech stimuli, although primary auditory and belt areas might be more responsive to basic acoustic stimuli, such as amplitude-modulated Gaussian white noises, or dynamic spectral-temporal ripples. Yet, higher (belt) auditory cortical regions might respond better to more natural stimuli.

Multisensory integration versus saturation

Our data is in line with cross-sensory influences on neuronal activity, as a clear response was evoked by visual trials over an auditory-responsive, temporal cortical area (Fig. 5.6 and 5.7). This is in line with earlier studies that show a cross-sensory influence on neuronal activity at early cortical areas, which have been traditionally held as unisensory.[65–70] However, we cannot exclude the possibility that recordings may have partially been taken from higher auditory supplementary areas, as fNIRS records signals arising from a large (1-2 cm) diffuse area [71]. As such visual-evoked signals might potentially originate from areas in the superior temporal gyrus that encode for face recognition, lipreading, or other higher-cognitive functions [29,72,73]. The data support the idea that we recorded from predominantly auditory areas, as sounds almost invariantly elicited the largest responses, and the visual activation was nearly completely nulled during audiovisual stimulation (i.e. audiovisual activation was not significantly different from the inverse of visual-only activation, Figs. 5.7C and F).

On a group level, the AV responses hint at auditory dominance (Fig. 5.7), because the visual response, as presented in isolation, does not appear in the AV response. Two distinct mechanisms might underlie this phenomenon. First, a true multisensory integrative effect could have been present ([30], Fig. 5.1), in which the visual component is effectively counterbalanced by an inhibitory audiovisual integration effect (Fig. 5.7C and F). Alternatively, the hemoglobin response might have reached saturation by the supra-threshold, highly intelligible auditory stimulus. Then, adding a visual stimulus will not lead to a stronger response. It seems unlikely that the nearly exact inverse relationship between the audiovisual and visual components in the audiovisual regression model (Fig. 5.7C,F) would be explained by multisensory integration, as it is precisely expected from a saturation effect. To better dissociate these different interpretations, auditory and visual stimuli should be presented in regimes that prevent neural saturation, and/or better characterize the visual response.

Note that with a one-channel setup it is impossible to decide whether the auditory and visual activations originated from the same area, or from spatially separated areas, when the AV response would equal the sum of the A and V responses. However, if the AV response deviates from the purely additive prediction several possibilities may be dissociated, as explained in figure 5.8. Importantly, activation of two distinct, independent brain areas (Fig. 5.8A, E) does not predict the saturation that is observed in our data. The sub-additive AV response observed in our results (a peak activation between the blue and green lines in figure 5.8) allows for two possible scenarios: (i) the signals could have originated from true multisensory neurons (Fig. 5.8C,G), or (ii) from two distinct subpopulations of unisensory-responsive neurons within the recorded area (Fig. 5.8B,F). Note, however, that whenever the peak activation exceeds the additive response (green line), or falls below the strongest unimodal response (blue line), it will be a signature for true multisensory neural integration.

Post-lingually deaf CI users

The brain can reorganize after sensory deprivation, such as caused by deafness.[74,75] The question is whether cross-modal reorganization after deafening might introduce stronger visual effects over auditory cortex in post-lingual deaf subjects. This is not the case in our limited group of CI users (Fig. 5.7A and D), as visual activation was lower than auditory evoked activity.

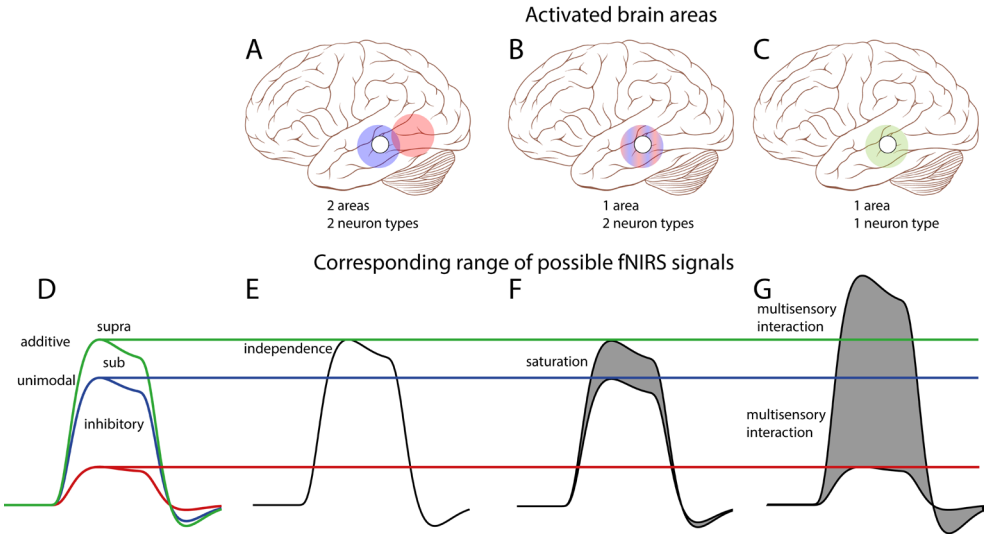


Figure 5.8. Conceptual schematic effects of potential audiovisual interactions on single-channel fNIRS signals. Top. We consider all three possible scenarios: **A)** two spatially separated areas are each activated by either auditory or visual inputs; **B)** neurons are either auditory or visually responsive, but are interspersed within one area; **C)** one area contains bimodal neurons that respond to both auditory and visual stimulation. The open circle on the brain schematically depicts the location of the single T7/8 fNIRS channel, and colored circles depict the activation patterns of indicated areas: blue – auditory, red – visual, green – auditory and visual. Bottom. **D)** Description of potential integrative effects (see also Fig. 5.1). **E)** For two independent areas of unisensory neurons (cf. A), the audiovisual fNIRS signal (black) can only be the sum of the auditory and visual fNIRS signals (and thus equals the additive model – green in D). **F)** For a mix of unisensory neurons in one area (cf. B), both neuron populations will be similarly active for their unisensory-preferred stimulus as for the audiovisual stimulus. The fNIRS signal then equals the additive model, or less (gray area) if saturation of the BOLD response occurs (sub-additive model, between blue and green in D). **G)** For an area with multisensory neurons (cf. C), fNIRS signals could yield any response type. Note that only a multisensory-area can generate multisensory interactions like super-additivity (above green), or inhibition (below blue). (Parts of this image have been taken from https://commons.wikimedia.org/wiki/File:Skull_and_brain_normal_human.svg. Patrick J. Lynch; C. Carl Jaffe; Yale University Center for Advanced Instructional Media; under Creative Commons Attribution 2.5 License 2006.

Our cohort of post-lingual deaf CI users did not differ from the normal-hearing cohort with respect to cortical activation for audio-visual stimuli (Fig. 5.7; grey squares). This is seemingly in contrast to the principle of inverse effectiveness, which suggests that people with sensory impairments might benefit from multisensory integration. Specifically, a larger multisensory

enhancement compared to the purely auditory response would be predicted because of the hearing-impairment of the CI users (and thus the weaker auditory percepts). This is not observed (Fig. 5.7B and E), indicating that either the stimuli were still supra-threshold for these subjects, or that saturation still dominated the audiovisual responses. Both possibilities imply a paradigm that aims at near-threshold stimulation in order to study this principle. Moreover, a larger cohort of CI users is desired when the issues of supra-threshold stimuli and response saturation have been overcome.

CONCLUSION

We found increased activation to auditory, visual and audiovisual stimulation in temporal cortex of normal-hearing subjects and post-lingually deaf CI users using fNIRS. Our findings demonstrate the potential of fNIRS for studying the neural mechanisms of audiovisual integration, both in normal-hearing subjects and in hearing-impaired subjects following cochlear implantation.

ACKNOWLEDGMENTS

We thank Günter Windau and Chris-Jan Beerendonck for their valuable technical assistance.

REFERENCES

1. Helfer KS. Auditory and auditory-visual perception of clear and conversational speech. *J speech, Lang Hear Res*. 1997;40(2):432-443.
2. MacLeod A, Summerfield Q. A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *Br J Audiol*. 1990;24(1):29-43. doi:10.3109/03005369009077840
3. Stein BE, Meredith MA. *The Merging of the Senses*. Cambridge, MA, US: The MIT Press.; 1993.
4. Corneil BD, Van Wanrooij MM, Munoz DP, Van Opstal AJ. Auditory-visual interactions subserving goal-directed saccades in a complex scene. *J Neurophysiol*. 2002;88(1):438-454. doi:10.1152/jn.2002.88.1.438
5. Van Barneveld DCPBM, Van Wanrooij MM. The influence of static eye and head position on the ventriloquist effect. *Eur J Neurosci*. 2013;37(9):1501-1510. doi:10.1111/ejn.12176
6. Beauchamp MS. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr Opin Neurobiol*. 2005;15(2):145-153. doi:10.1016/j.conb.2005.03.011
7. Stein BE. *The New Handbook of Multisensory Processes*.; 2012.
8. Calvert GA, Spence C, Stein BE. The Handbook of Multisensory Processing. 2004.
9. Laurienti PJ, Perrault TJ, Stanford TR, Wallace MT, Stein BE. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp Brain Res*. 2005;166(3-4):289-297. doi:10.1007/s00221-005-2370-2
10. Jobsis F. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science (80-)*. 1977;198(4323):1264-1267. doi:10.1126/science.929199
11. Cope M, Delpy DT. System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. *Med Biol Eng Comput*. 1988;26(3):289-294. doi:10.1007/BF02447083
12. Abdelnour AF, Huppert T. Real-time imaging of human brain function by near-infrared spectroscopy using an adaptive general linear model. *Neuroimage*. 2009;46(1):133-143. doi:10.1016/j.neuroimage.2009.01.033
13. Huppert TJ, Allen MS, Diamond SG, Boas DA. Estimating cerebral oxygen metabolism from fMRI with a dynamic multicompartment Windkessel model. *Hum Brain Mapp*. 2009;30(5):1548-1567. doi:10.1002/hbm.20628
14. Huppert TJ, Hoge RD, Dale AM, Franceschini MA, Boas DA. Quantitative spatial comparison of diffuse optical imaging with blood oxygen level-dependent and arterial spin labeling-based functional magnetic resonance imaging. *J Biomed Opt*. 2006;11(6):064018. doi:10.1117/1.2400910
15. Johnsrude IS, Giraud AL, Frackowiak RSJ. Functional Imaging of the Auditory System: The Use of Positron Emission Tomography. *Audiol Neuro-Otology*. 2002;7(5):251-276. doi:10.1159/000064446
16. Hall DA, Haggard MP, Akeroyd MA, et al. Modulation and task effects in auditory processing measured using fMRI. *Hum Brain Mapp*. 2000;10(3):107-119.
17. Plichta MM, Gerdes ABM, Alpers GW, et al. Auditory cortex activation is modulated by emotion: A functional near-infrared spectroscopy (fNIRS) study. *Neuroimage*. 2011;55(3):1200-1207. doi:10.1016/j.neuroimage.2011.01.011
18. Santosa H, Hong MJ, Hong K-S. Lateralization of music processing with noises in the auditory cortex: an fNIRS study. *Front Behav Neurosci*. 2014;8(December):418. doi:10.3389/fnbeh.2014.00418
19. Chen L-C, Sandmann P, Thorne JD, Herrmann CS, Debener S. Association of Concurrent fNIRS and EEG Signatures in Response to Auditory and Visual Stimuli. *Brain Topogr*. 2015;28(5):710-725. doi:10.1007/s10548-015-0424-8

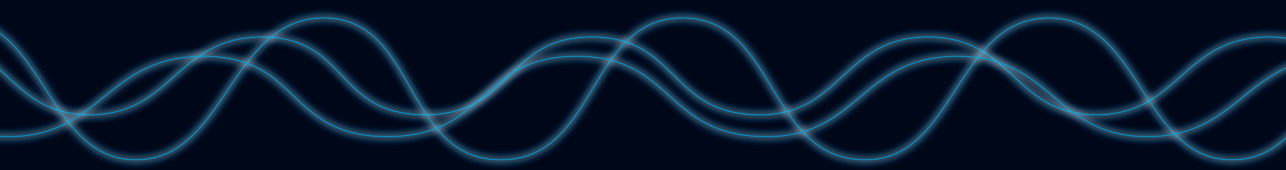
20. Pollonini L, Olds C, Abaya H, Bortfeld H, Beauchamp MS, Oghalai JS. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hear Res*. 2014;309(December):84-93. doi:10.1016/j.heares.2013.11.007
21. Gilley PM, Sharma A, Dorman M, Finley CC, Panch AS, Martin K. Minimization of cochlear implant stimulus artifact in cortical auditory evoked potentials. *Clin Neurophysiol*. 2006;117(8):1772-1782. doi:10.1016/j.clinph.2006.04.018
22. Sevy ABG, Bortfeld H, Huppert TJ, Beauchamp MS, Tonini RE, Oghalai JS. Neuroimaging with near-infrared spectroscopy demonstrates speech-evoked activity in the auditory cortex of deaf children following cochlear implantation. *Hear Res*. 2010;270(1-2):39-47. doi:10.1016/j.heares.2010.09.010
23. Chen L, Sandmann P, Thorne JD, Bleichner MG, Debener S. Cross-Modal Functional Reorganization of Visual and Auditory Cortex in Adult Cochlear Implant Users Identified with fNIRS. *Neural Plast*. 2016;2016:4382656. doi:10.1155/2016/4382656
24. Dewey RS, Hartley DEH. Cortical cross-modal plasticity following deafness measured using functional near-infrared spectroscopy. *Hear Res*. 2015;325:55-63. doi:10.1016/j.heares.2015.03.007
25. Smith SM. Overview of fMRI analysis. *Br J Radiol*. 2004;77(suppl_2):S167-S175. doi:10.1259/bjr/33553595
26. Malinen S, Hlushchuk Y, Hari R. Towards natural stimulation in fMRI—Issues of data analysis. *Neuroimage*. 2007;35(1):131-139. doi:10.1016/j.neuroimage.2006.11.015
27. Steinbrink J, Villringer A, Kempf F, Haux D, Boden S, Obrig H. Illuminating the BOLD signal: combined fMRI–fNIRS studies. *Magn Reson Imaging*. 2006;24(4):495-505. doi:10.1016/j.mri.2005.12.034
28. Cui X, Bray S, Bryant DM, Glover GH, Reiss AL. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *Neuroimage*. 2011;54(4):2808-2821. doi:10.1016/j.neuroimage.2010.10.069
29. Calvert GA, Bullmore ET, Brammer MJ, et al. Activation of auditory cortex during silent lipreading. *Science*. 1997;276(5312):593-596. doi:10.1126/science.276.5312.593
30. Stein BE, Stanford TR, Ramachandran R, Perrault TJ, Rowland B a. Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. *Exp Brain Res*. 2009;198(2-3):113-126. doi:10.1007/s00221-009-1880-8
31. Julien C. The enigma of Mayer waves: Facts and models. *Cardiovasc Res*. 2006;70(1):12-21. doi:10.1016/j.cardiores.2005.11.008
32. Agterberg MJH, Snik AFMM, Hol MKS, et al. Improved Horizontal Directional Hearing in Bone Conduction Device Users with Acquired Unilateral Conductive Hearing Loss. *J Assoc Res Otolaryngol*. 2011;12(1):1-11. doi:10.1007/s10162-010-0235-2
33. Bremen P, van Wanrooij MM, van Opstal AJ. Pinna cues determine orienting response modes to synchronous sounds in elevation. *J Neurosci*. 2010;30(1):194-204. doi:10.1523/JNEUROSCI.2982-09.2010
34. Scholkmann F, Kleiser S, Metz AJ, et al. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *Neuroimage*. 2014;85:6-27. doi:10.1016/j.neuroimage.2013.05.004
35. Jasper HH. Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol*. 1958;10(2):370-375. doi:10.1016/0013-4694(58)90053-1
36. Herwig U, Satrapi P, Schönfeldt-Lecuona C. Using the International 10-20 EEG System for Positioning of Transcranial Magnetic Stimulation. *Brain Topogr*. 2003;16(2):95-99. doi:10.1023/B:BRAT.0000006333.93597.9d
37. Fukui Y, Ajichi Y, Okada E. Monte Carlo prediction of near-infrared light propagation in realistic adult and neonatal head models. *Appl Opt*. 2003;42(16):2881-2887. doi:10.1364/AO.42.002881
38. Strangman GE, Zhang Q, Li Z. Scalp and skull influence on near infrared photon propagation in the Colin27 brain template. *Neuroimage*. 2014;85 Pt 1:136-149. doi:10.1016/j.neuroimage.2013.04.090

39. Brigadoi S, Cooper RJ. How short is short? Optimum source–detector distance for short-separation channels in functional near-infrared spectroscopy. *Neurophotonics*. 2015;2(2):025005. doi:10.1117/1.NPh.2.2.025005
40. Haeussinger FB, Heinzel S, Hahn T, Schecklmann M, Ehlis A-C, Fallgatter AJ. Simulation of Near-Infrared Light Absorption Considering Individual Head and Prefrontal Cortex Anatomy: Implications for Optical Neuroimaging. Hashimoto K, ed. *PLoS One*. 2011;6(10):e26377. doi:10.1371/journal.pone.0026377
41. Niedermeyer E, Lopes da Silva F. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins; 2005.
42. Brainard DH. The Psychophysics Toolbox. *Spat Vis*. 1997;10(4):433-436. doi:10.1163/156856897X00357
43. Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis*. 1997;10(4):437-442. doi:10.1163/156856897X00366
44. Kleiner M, Brainard D, Pelli D. What's new in Psychtoolbox-3. *Perception, 36 ECVP Abstr Suppl*. 2007.
45. Huppert TJ, Diamond SG, Franceschini M a, Boas D a. HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Appl Opt*. 2009;48(10):280-298. doi:10.1364/AO.48.00D280
46. Themelis G, Selb J, Thaker S, et al. Depth of arterial oscillation resolved with NIRS time and frequency domain. In: *Biomedical Topical Meeting*. Washington, D.C.: OSA; 2004:WF2. doi:10.1364/BIO.2004.WF2
47. Yamada T, Umeyama S, Matsuda K. Separation of fNIRS Signals into Functional and Systemic Components Based on Differences in Hemodynamic Modalities. Baron J-C, ed. *PLoS One*. 2012;7(11):e50271. doi:10.1371/journal.pone.0050271
48. Fekete T, Rubin D, Carlson JM, Mujica-Parodi LR. The NIRS Analysis Package: Noise Reduction and Statistical Inference. Zuo X-N, ed. *PLoS One*. 2011;6(9):e24322. doi:10.1371/journal.pone.0024322
49. Pei Y, Wang Z, Barbour RL. NAVI-SciPort solution: a problem solving environment (PSE) for NIRS data analysis. In: *Human Brain Mapping*. Chicago, IL; 2007.
50. Kocsis L, Herman P, Eke A. The modified Beer–Lambert law revisited. *Phys Med Biol*. 2006;51(5):N91-N98. doi:10.1088/0031-9155/51/5/N02
51. Scarpa F, Brigadoi S, Cutini S, et al. A reference-channel based methodology to improve estimation of event-related hemodynamic response from fNIRS measurements. *Neuroimage*. 2013;72:106-119. doi:10.1016/j.neuroimage.2013.01.021
52. Henson R, Friston K. Convolution models for fMRI. In: *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. ; 2007:178-192. doi:10.1016/B978-012372560-8/50014-0
53. Lindquist MA, Meng Loh J, Atlas LY, Wager TD. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*. 2009;45(1 Suppl):S187-98. doi:10.1016/j.neuroimage.2008.10.065
54. Kennan RP, Horovitz SG, Maki A, Yamashita Y, Koizumi H, Gore JC. Simultaneous Recording of Event-Related Auditory Oddball Response Using Transcranial Near Infrared Optical Topography and Surface EEG. *Neuroimage*. 2002;16(3):587-592. doi:10.1006/nimg.2002.1060
55. Barbour RL, Graber HL, Jenghwa Chang, Barbour S-LS, Koo PC, Aronson R. MRI-guided optical tomography: prospects and computation for a new imaging method. *IEEE Comput Sci Eng*. 1995;2(4):63-77. doi:10.1109/99.476370
56. Barnett AH, Culver JP, Sorensen AG, Dale A, Boas DA. Robust Inference of Baseline Optical Properties of the Human Head with Three-Dimensional Segmentation from Magnetic Resonance Imaging. *Appl Opt*. 2003;42(16):3095. doi:10.1364/AO.42.003095
57. Pogue BW, Paulsen KD. High-resolution near-infrared tomographic imaging simulations of the rat cranium by use of a priori magnetic resonance imaging structural information. *Opt Lett*. 1998;23(21):1716. doi:10.1364/OL.23.001716

58. Turner BM, Forstmann BU, Wagenmakers E-J, Brown SD, Sederberg PB, Steyvers M. A Bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage*. 2013;72:193-206. doi:10.1016/j.neuroimage.2013.01.048
59. Vannest JJ, Karunanayaka PR, Altaye M, et al. Comparison of fMRI data from passive listening and active-response story processing tasks in children. *J Magn Reson Imaging*. 2009;29(4):971-976. doi:10.1002/jmri.21694
60. Grady CL, Van Meter JW, Maisog JM, Pietrini P, Krasuski J, Rauschecker JP. Attention-related modulation of activity in primary and secondary auditory cortex. *Neuroreport*. 1997;8(11):2511-2516.
61. Massoudi R, Van Wanrooij MM, Van Wetter SMCI, Versnel H, Van Opstal AJ. Stable bottom-up processing during dynamic top-down modulations in monkey auditory cortex. *Eur J Neurosci*. 2013;37(11):1830-1842. doi:10.1111/ejn.12180
62. Massoudi R, Van Wanrooij MM, Van Wetter SMCI, Versnel H, Van Opstal AJ. Task-related preparatory modulations multiply with acoustic processing in monkey auditory cortex. *Eur J Neurosci*. 2014;39(9):1538-1550. doi:10.1111/ejn.12532
63. Wang X, Lu T, Snider RK, Liang L. Sustained firing in auditory cortex evoked by preferred stimuli. *Nature*. 2005;435(7040):341-346. doi:10.1038/nature03565
64. Osmanski MS, Wang X. Behavioral Dependence of Auditory Cortical Responses. *Brain Topogr*. 2015;28(3):365-378. doi:10.1007/s10548-015-0428-4
65. Foxe JJ, Schroeder CE. The case for feedforward multisensory convergence during early cortical processing. *Neuroreport*. 2005;16(5):419-423. doi:10.1097/00001756-200504040-00001
66. Schroeder CE, Foxe J. Multisensory contributions to low-level, "unisensory" processing. *Curr Opin Neurobiol*. 2005;15(4):454-458. doi:10.1016/j.conb.2005.06.008
67. Ghazanfar AA, Neuhoff JG, Logothetis NK. Auditory looming perception in rhesus monkeys. *Proc Natl Acad Sci U S A*. 2002;99(24):15755-15757. doi:10.1073/pnas.242469699
68. Kayser C, Petkov CI, Augath M, Logothetis NK. Functional imaging reveals visual modulation of specific fields in auditory cortex. *J Neurosci*. 2007;27(8):1824-1835. doi:10.1523/JNEUROSCI.4737-06.2007
69. Kayser C, Logothetis NK, Panzeri S. Visual Enhancement of the Information Representation in Auditory Cortex. *Curr Biol*. 2010;20(1):19-24. doi:10.1016/j.cub.2009.10.068
70. Koelewijn T, Bronkhorst A, Theeuwes J. Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychol (Amst)*. 2010;134(3):372-384. doi:10.1016/j.actpsy.2010.03.010
71. Boas D a, Dale AM, Franceschini MA. Diffuse optical imaging of brain activation: Approaches to optimizing image sensitivity, resolution, and accuracy. In: *NeuroImage*. Vol 23. ; 2004:S275-88. doi:10.1016/j.neuroimage.2004.07.011
72. Sams M, Aulanko R, Hamalainen M, et al. Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neurosci Lett*. 1991;127(1):141-145. doi:10.1016/0304-3940(91)90914-F
73. MacSweeney M, Amaro E, Calvert GA, et al. Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport*. 2000;11(8):1729-1733. doi:10.1097/00001756-200006050-00026
74. Rauschecker JP. Compensatory plasticity and sensory substitution in the cerebral cortex. *Trends Neurosci*. 1995;18(1):36-43. doi:10.1016/0166-2236(95)93948-W
75. Lee DS, Lee JS, Oh SH, et al. Cross-modal plasticity and cochlear implants. *Nature*. 2001;409(6817):149-150. doi:10.1038/35051653

CHAPTER 6

Discussion



The aim of this thesis was to examine whether lipreading in the hearing impaired may provide a useful visual information stream that can be integrated with the – degraded – auditory information stream to improve speech understanding. We tested audiovisual speech perception behaviorally and studied the neural correlates of this through the use of functional near-infrared spectroscopy. In the following sections, the main results are discussed.

AUDIOVISUAL SPEECH PERCEPTION

We determined how well words presented in (nonsense) sentences are recognized by normal-hearing subjects through listening (auditory presentation) and/or lipreading (visual presentation) under noisy listening conditions (both auditory noisy and visual blur; **chapter 2/3**). In line with previous research, we found that listening and lipreading (audiovisual presentation) improves the speech recognition scores compared to auditory only.[1–4] However, in contrast to earlier reports, we observed that audiovisual performance levels fell below visual performance when the acoustic signal-to-noise ratios (SNRs) were low. We also found that the improvements typically saturated at intermediate SNRs, which is expected from the principle of inverse effectiveness. Inverse effectiveness was found at (individual) word-level and subject-level: the data showed that the benefit of adding cross-modal information increased when a word was poorly heard, when a word was poorly seen, or when the participant was a poor lipreader. The improvement of audiovisual speech perception compared to purely auditory speech perception can be modeled in various ways (**chapter 2**). Typically, audiovisual data are compared to the benchmark probability-summation model (*see Introduction for an explanation*), in which the auditory and visual channels are assumed independent, without interaction. The percept is then determined by whichever modality wins the 'race'. This non-integration model matched the data closely. Yet, Rouger et al. reported that an alternative integration model could better describe their data.[5] In their model, spectral-temporal audiovisual cues merge across modalities to optimize the amount of information required for word recognition. Our audiovisual data obtained under poor lipreading conditions (i.e., when the visual recognition rate for a word is lower than 0.55) compared quite well to the data of Rouger et al. (2007 – *their Figure 3D*).

A third model was proposed by Ma et al. (2009) in which this Bayesian model assumes that certain words occur more frequently than other words and are more easily recognized, and their model uses this prior knowledge to explain the recognition scores for all words.[6]

It is hard to reconcile any of the three models with our observation that for low-SNR conditions, multisensory speech recognition is actually degraded compared to unimodal. Note that none of these three models considered non-stimulus factors that may affect audiovisual speech

recognition, such as attention. The aforementioned models did not include a mechanism for dividing attention[7,8] between the two modalities, as we propose in **chapter 2**. In such a mechanism, the two separate information streams could actually lead to impaired performance for conditions in which either of the two signals may be ambiguous, or weak. Thus, even if lipreading might provide sufficient information to recognize words, subjects do not seem to be able to divert their attention away from the auditory stream, despite the absence of a potential signal (e.g., merely noise or a total blur) in that information stream (audio or video).

The ability to divide attention between the auditory and visual modalities was further studied in **chapter 3**. CI users made less errors in speech recognition when they could focus on listening alone than in situations with uncertainty about the modality of the upcoming stimulus modality. Note that this is precisely the sensory condition of every-day life. This may suggest that due to impoverished sensory information more effort is required by CI users to be able recognize speech at higher performance levels. However, the extra effort cannot be maintained by CI users if attention has to be spread out across multiple, potentially-informative sensory modalities. The CI users seem to have reached the limits of attentional resources in the divided-attention task. These limits are not reached when sensory information is not impoverished, i.e. for normal-hearing individuals and for lipreading (Chapter 5; Figs. 5.1A, B, D; lapse probabilities are similar across tasks).

Following this line of reasoning, one may wonder why CI users attempt to lipread at all. Barring any other benefits, the optimal decision would be to focus on the most-informative sensory modality, and ignoring the other. Even for CI users, listening is generally (i.e. in quiet environments) the far better modality for the purposes of speech recognition. Probabilistic, uninformed switching between listening and lipreading would lead to an overall worse performance.[9] One benefit to offset this drawback could be that switching enables individuals to scan the specific environment and determine whether listening or lipreading would be the most informative modality for the given situation.[10,11]

Obviously from the current experiments, another benefit could be that the detriment in listening is accompanied by an enhancement of speech recognition for multisensory stimuli. Indeed, although CI users had poorer unisensory recognition scores in the divided-attention task than in the focused attention task (Fig. 5.1), they outperformed the strict probability-summation model (Fig. 5.2D). Conversely, the normal-hearing individuals do follow strict probability summation.[12] Because of this, CI users appear to be better multisensory integrators than the normal-hearing individuals[5] (Fig. 5.2D).

CROSS-MODAL ACTIVATION

As stated in the introduction of this thesis, cochlear implantation partially restores hearing of deaf individuals. Owing to a wide, partly unexplained variance in outcomes of cochlear implantation, it might be important to study the effects of brain plasticity on CI outcomes. In deaf individuals in which the auditory system is deprived, certain areas in the brain associated with auditory processing can be taken over by the other intact sensory modalities. Due to cross-sensory plasticity, these originally auditory areas might become more responsive to visual stimuli. While being deaf, in everyday life, this might be beneficial. On the other hand, these processes of adaptation to-being-deaf (such as lipreading) might also limit the ability to acquire auditory speech recognition, once hearing is partially restored with a CI. Rouger et al. demonstrated that cross-sensory activation of auditory brain regions prior to implantation correlates with poor speech outcomes with a CI.[13] The opposite has also been shown for a limited group of CI users in **chapter 5,** in which visual activation of temporal areas was similar to normal-hearing subjects. Strelnikov et al. found that auditory speech recovery positively correlated with visual activity in auditory regions measured with PET.[14] Several potential mechanisms have been suggested to mediate cross-sensory reorganization and outcome of speech recognition performance with a CI (e.g., some individuals may be predisposed to strongly rely on visual modes of communication, or, alternatively, visual takeover of auditory brain areas prevents recovery once a CI is used). Functional near-infrared spectroscopy (fNIRS) could contribute to understand and dissociate different potential mechanisms.

Back in 1977, Frans Jöbsis was the first to report the relatively high degree of transparency of brain tissue when using light in the 600-900 nm near-infrared range. The characteristic hemoglobin (Hb) absorption spectra in this wavelength region enable real-time, non-invasive and local detection of changes in Hb oxygenation, assessing changes in brain activity.[15] The charm of what is called fNIRS has exceeded across disciplines – physics, physiology, psychology, statistics, and neuroscience. fNIRS needs to be further comprehended, the processing methods need to be established, the reliability needs to be enhanced, and finally the clinical purposes need to be established (**chapter 4**). Neuroscientists and clinicians have applied fNIRS to a wide range of research questions regarding the functional organization and physiology of the brain, and how they vary across clinical populations. In **chapter 4**, we reviewed the literature on measuring cortical activity during auditory processing with fNIRS. Yet, despite the promising results of fNIRS, developing an ideal and stable setup and experimental design for adequate hypothesis testing still remains a challenge. By incorporating some of the aspects reviewed – for example, details of how the cortical hemodynamic response to acoustic stimuli is modulated by stimulus presentation and repetition rates, sound duration, sound level, and attention – one might be able to acquire reliable and valid fNIRS data. This type of problems captures the essence of the difficulty associated with many hoped-for clinical

implementations of fNIRS. It is clear that fundamental knowledge is important for clinical applications, even if the applications are not yet clearly defined. Functional near-infrared spectroscopy has certainly contributed to fundamental knowledge.

MULTISENSORY INTEGRATION

We assessed audiovisual activation in the temporal cortex with fNIRS (**chapter 5**). Specifically, we studied cortical activation of normal-hearing adult subjects and post-lingually deaf unilateral CI users in response to auditory, visual and audiovisual speech stimuli. The audiovisual stimuli evoked a hemodynamic signal that resembled the purely auditory evoked signal with the visual component being almost exactly inversely related to the audiovisual component (see Figure 5.7C/F in **chapter 5**). Interestingly, the fNIRS signals evoked in the CI users strongly resembled those of the normal-hearing subjects. We demonstrated cross-sensory influences on cortical activity, as a clear response was evoked by visual-only trials in an auditory-responsive, temporal cortical area (**chapter 5**; Fig. 5.6 and 5.7). This is in line with earlier studies that show a cross-sensory influence on neuronal activity at early cortical areas, which have been traditionally held as unisensory.[16–21] On a group level, the audiovisual responses hint at auditory dominance, because the visual-only response does not affect the audiovisual response. Two distinct mechanisms might underlie this phenomenon. First, a true multisensory integrative effect could have been present, in which the visual component was effectively counterbalanced by an inhibitory audiovisual integration effect. Alternatively, the hemoglobin response might have reached saturation by the supra-threshold, highly intelligible auditory stimulus. Then, adding a visual stimulus will not lead to a stronger response. To better dissociate these different interpretations, auditory and visual stimuli should be presented in regimes that prevent neural saturation, and/or better characterize the visual response. For future studies one could use a set up according to **chapter 3** in which the saliency of the auditory and visual stimuli is systematically altered by using different signal-to-noise ratios and spatial visual blur, respectively.

REFERENCES

1. Bernstein LE, Auer ET, Takayanagi S. Auditory speech detection in noise enhanced by lipreading. *Speech Commun*. 2004;44(1-4):5-18. doi:10.1016/j.specom.2004.10.011
2. Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am*. 2000;108(3 Pt 1):1197-1208. doi:10.1121/1.422512
3. Helfer KS. Auditory and auditory-visual perception of clear and conversational speech. *J speech, Lang Hear Res*. 1997;40(2):432-443.
4. Winn MB, Rhone AE, Chatterjee M, Idsardi WJ. The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Front Psychol*. 2013;4(NOV):824. doi:10.3389/fpsyg.2013.00824
5. Rouger J, Lagleyre S, Fraysse B, Deneve S, Deguine O, Barone P. Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proc Natl Acad Sci U S A*. 2007;104(17):7295-7300. doi:10.1073/pnas.0609419104
6. Ma WJ, Zhou X, Ross LA, Foxe JJ, Parra LC. Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One*. 2009;4(3):e4638. doi:10.1371/journal.pone.0004638
7. Alsius A, Navarra J, Campbell R, Soto-Faraco S. Audiovisual Integration of Speech Falters under High Attention Demands. *Curr Biol*. 2005;15(9):839-843. doi:10.1016/j.cub.2005.03.046
8. Bonnel AM, Hafter ER. Divided attention between simultaneous auditory and visual signals. *Percept Psychophys*. 1998;60(2):179-190. doi:10.3758/BF03206027
9. Ege R, Opstal AJ Van, Van Wanrooij MM. Accuracy-Precision Trade-off in Human Sound Localisation. *Sci Rep*. 2018;8(1):16399. doi:10.1038/s41598-018-34512-6
10. Ege R, Van Opstal AJ, Van Wanrooij MM. Perceived Target Range Shapes Human Sound-Localization Behavior. *eNeuro*. 2019;6(2). doi:10.1523/ENEURO.0111-18.2019
11. Berniker M, Voss M, Kording K. Learning priors for Bayesian computations in the nervous system. Brezina V, ed. *PLoS One*. 2010;5(9):9. doi:10.1371/journal.pone.0012686
12. van de Rijt LPH, Roye A, Mylanus EAM, van Opstal AJ, van Wanrooij MM. The Principle of Inverse Effectiveness in Audiovisual Speech Perception. *Front Hum Neurosci*. 2019;13(September):1-15. doi:10.3389/fnhum.2019.00335
13. Rouger J, Lagleyre S, Démonet J-F, Fraysse B, Deguine O, Barone P. Evolution of crossmodal reorganization of the voice area in cochlear-implanted deaf patients. *Hum Brain Mapp*. 2012;33(8):1929-1940. doi:10.1002/hbm.21331
14. Strelnikov K, Rouger J, Lagleyre S, et al. Increased audiovisual integration in cochlear-implanted deaf patients: Independent components analysis of longitudinal positron emission tomography data. *Eur J Neurosci*. 2015;41(5):677-685. doi:10.1111/ejn.12827
15. Jobsis F. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science (80-)*. 1977;198(4323):1264-1267. doi:10.1126/science.929199
16. Foxe JJ, Schroeder CE. The case for feedforward multisensory convergence during early cortical processing. *Neuroreport*. 2005;16(5):419-423. doi:10.1097/00001756-200504040-00001
17. Schroeder CE, Foxe J. Multisensory contributions to low-level, "unisensory" processing. *Curr Opin Neurobiol*. 2005;15(4):454-458. doi:10.1016/j.conb.2005.06.008
18. Ghazanfar AA, Neuhoff JG, Logothetis NK. Auditory looming perception in rhesus monkeys. *Proc Natl Acad Sci U S A*. 2002;99(24):15755-15757. doi:10.1073/pnas.242469699

19. Kayser C, Petkov CI, Augath M, Logothetis NK. Functional imaging reveals visual modulation of specific fields in auditory cortex. *J Neurosci*. 2007;27(8):1824-1835. doi:10.1523/JNEUROSCI.4737-06.2007

20. Kayser C, Logothetis NK, Panzeri S. Visual Enhancement of the Information Representation in Auditory Cortex. *Curr Biol*. 2010;20(1):19-24. doi:10.1016/j.cub.2009.10.068

21. Koelewijn T, Bronkhorst A, Theeuwes J. Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychol (Amst)*. 2010;134(3):372-384. doi:10.1016/j.actpsy.2010.03.010

CHAPTER 7

Summary

Nederlandse samenvatting

In this thesis is explored how auditory and visual speech is processed by normal-hearing individuals and CI users.

Chapter 1 starts with a general introduction about the anatomy and physiology of the healthy hearing organ. Subsequently, is described how, in case of severe deafness, the cochlear implant (CI) might partially restore hearing. After this general problem statement it is explained how lipreading could potentially provide an additional useful stream of information. Lipreading can be integrated with the - degraded (e.g. by ambient noise) - acoustic information to improve speech understanding.
Our behavioral experiments demonstrate how visual information is incorporated by normal-hearing listeners (**chapter 2**) and CI users (**chapter 3**) in audiovisual speech perception.

We assessed how synchronous speech listening and lipreading affects speech recognition in acoustic noise. In simple audiovisual perceptual tasks, inverse effectiveness is often observed, which holds that the weaker the unimodal stimuli, or the poorer their signal-to-noise ratio, the stronger the audiovisual benefit. So far, however, inverse effectiveness has not been demonstrated for complex audiovisual speech stimuli. Here we assessed whether this multisensory integration effect can also be observed for the recognizability of spoken words (**chapter 2**). Speech-recognition performance was determined for auditory-only, visual-only (lipreading), and auditory-visual conditions. To modulate acoustic task difficulty, we systematically varied the auditory signal-to-noise ratio. In line with a commonly observed multisensory enhancement on speech recognition, audiovisual words were more easily recognized than auditory-only words (recognition thresholds of –15 and –12 dB, respectively). We here show that the difficulty of recognizing a particular word, either acoustically or visually, determines the occurrence of inverse effectiveness in audiovisual word integration. Thus, words that are better heard or recognized through lipreading, benefit less from bimodal presentation. Audiovisual performance at the lowest acoustic signal-to-noise ratios (45%) fell below the visual recognition rates (60%), reflecting an actual deterioration of lipreading in the presence of excessive acoustic noise. This suggests that the brain may adopt a strategy in which attention has to be divided between listening and lipreading.

The CI allows profoundly deaf individuals to recover hearing. Still, due to the coarse acoustic information provided by the implant, CI users have considerable difficulties in recognizing speech, especially in noisy environments, even years after implantation. CI users therefore rely heavily on visual speech to augment speech comprehension, more so than normal-hearing individuals. However, it is unknown how attention to one (focused) or both (divided) modalities plays a role in multisensory speech recognition. Here we show that unisensory speech listening and lip reading are negatively impacted in divided-attention tasks for CI users - but not for normal-hearing individuals (**chapter 3**). Our psychophysical experiments reveal

that, as expected, speech-listening thresholds are consistently better for the normal-hearing, while lipreading thresholds were largely similar between both groups. Moreover, audiovisual speech recognition for normal-hearing individuals can be described well by probabilistic summation of auditory and visual speech recognition, while CI users are better integrators than expected from summation. Our results suggested that this benefit in integration, however, comes at a cost. Unisensory speech recognition is degraded for CI users when attention needs to be divided across modalities, i.e. in situations with uncertainty about the upcoming stimulus modality. We speculate that CI users exhibit an integration-attention trade-off. They focus solely on a single modality during focused-attention tasks, but need to divide their limited attention resources to more modalities during divided-attention tasks. We argue that in order to determine the benefit of a CI towards speech comprehension per se, situational factors need to be discounted by presenting speech in realistic or complex audio-visual environments.

Finally, the neuroimaging technique of functional near-infrared spectroscopy (fNIRS; further elaborated in **chapter 4**) was introduced, that allows for non-invasive brain activity measurements in CI users. The use of this technique is described in **chapter 5** where we studied neural correlates of audiovisual speech perception in CI users and in normal-hearing listeners

fNIRS is an optical, non-invasive neuroimaging technique that investigates human brain activity by calculating concentrations of oxy- and deoxyhemoglobin. The aim of this publication was to review the current state of the art as to how fNIRS has been used to study auditory function. We addressed temporal and spatial characteristics of the hemodynamic response to auditory stimulation as well as experimental factors that affect fNIRS data such as acoustic and stimulus-driven effects. The rising importance that fNIRS is generating in auditory neuroscience underlines the strong potential of the technology, and it seems likely that fNIRS will become a useful clinical tool.
Non-invasive neuroimaging techniques can expose the neural processes underlying the integration of multisensory processes required for speech understanding in humans. Nevertheless, noise (from functional MRI, fMRI) limits the usefulness in auditory experiments, and electromagnetic artifacts caused by electronic implants worn by subjects can severely distort the scans (EEG, fMRI). Therefore, we assessed audiovisual activation of temporal cortex with a silent, optical neuroimaging technique: fNIRS. We studied temporal cortical activation as represented by concentration changes of oxy- and deoxy-hemoglobin in four, easy-to-apply fNIRS optical channels of 33 normal-hearing adult subjects and five post-lingually deaf CI users in response to supra-threshold unisensory auditory and visual, as well as to congruent audiovisual speech stimuli. Activation effects were not visible from single fNIRS channels. However, by discounting physiological noise through reference channel subtraction

(RCS), auditory, visual and AV speech stimuli evoked concentration changes for all sensory modalities in both cohorts. Auditory stimulation evoked larger concentration changes than visual stimuli. Physiological, systemic noise can be removed from fNIRS signals by RCS. The observed multisensory enhancement of an auditory cortical channel can be plausibly described by a simple addition of the auditory and visual signals with saturation.

In dit proefschrift is onderzocht hoe visuele en auditieve informatie met betrekking tot spraak wordt verwerkt door normaalhorende individuen en CI-gebruikers.

In **hoofdstuk 1** wordt een algemene introductie gegeven over de anatomie van het oor en de fysiologie van het gehoor bij gezonde normaalhorende individuen. Vervolgens wordt beschreven hoe in geval van ernstige slechthorendheid, het cochleair implantaat (CI) het gehoor gedeeltelijk kan herstellen.
Na deze algemene introductie wordt besproken hoe liplezen als extra informatiebron kan dienen voor CI-gebruikers om het spraakverstaan te verbeteren. In situaties waarin het geluid niet optimaal is, bijvoorbeeld door omgevingslawaai, kan het lipbeeld worden gecombineerd met dit 'ruizige' geluid om het spraakverstaan te verbeteren. Middels psychofysische experimenten wordt onderzocht hoe visuele informatie door normaalhorende individuen (**hoofdstuk 2**) en CI-gebruikers (**hoofdstuk 3**) wordt geïntegreerd bij het spraakverstaan.

Het spraakverstaan in ruis is bestudeerd. Bij simpele audiovisuele perceptuele taken wordt vaak een fenomeen waargenomen wat het omgekeerd evenredige effect wordt genoemd ('inverse effectiveness'). Hierbij geldt dat, hoe zwakker de unimodale stimuli, oftewel hoe slechter hun signaal-ruisverhouding, hoe sterker de audiovisuele integratie. Tot nog toe is geen omgekeerd evenredig effect aangetoond voor complexe audiovisuele spraakstimuli (zoals woorden/zinnen).
In dit onderzoek is gekeken of dit multisensorische integratie-effect ook kan worden waargenomen voor de herkenbaarheid van gesproken woorden (**hoofdstuk 2**). Het spraakverstaan werd bepaald in verschillende condities; puur auditieve, puur visuele en audiovisuele condities. Om de moeilijkheidsgraad van het luisteren te moduleren, is de signaal-ruisverhouding systematisch gevarieerd. In overeenstemming met de literatuur, werd een multisensorische verbetering van het spraakverstaan voor audiovisuele woorden waargenomen, dat wil zeggen dat deze woorden gemakkelijker herkend werden dan alleen auditieve woorden (drempels van respectievelijk -15 en -12 dB). De moeilijkheid om een bepaald woord te herkennen, zowel auditief als visueel, is bepalend voor het optreden van het omgekeerd evenredig effect in audiovisuele spraakverstaan. Het bleek dat woorden die beter gehoord of visueel herkend worden, minder baat hebben bij een bimodale presentatie. Audiovisuele prestaties bij de laagste auditieve signaal/ruis-verhoudingen (zeer lastig te verstaan) (45%) vielen onder de visuele herkenningspercentages (60%), wat een daadwerkelijke verslechtering van het liplezen, in de aanwezigheid van overmatige akoestische ruis, weerspiegelt. Dit suggereert dat de hersenen een strategie hanteren waarbij de aandacht moet worden verdeeld tussen luisteren en liplezen.

Het cochleaire implantaat maakt het mogelijk dat individuen met ernstige slechthorendheid, weer redelijk kunnen horen. Toch hebben CI-gebruikers door de grove auditieve informatie

van het implantaat (CI's zijn niet in staat om omgevingsgeluid te filteren zoals een normaal gehoor dit wel kan) moeite om spraak te herkennen, vooral in rumoerige omgevingen. Dit blijft bestaan, zelfs jaren na implantatie. CI-gebruikers zijn daarom ook afhankelijk van visuele input om het spraakverstaan te verbeteren, meer dan normaalhorende personen. Het is echter onbekend hoe de aandacht voor één (gerichte) of beide (gedeelde) modaliteiten een rol speelt bij de audiovisueel spraakverstaan. Unisensorisch luisteren naar spraak en liplezen wordt lastiger voor CI-gebruikers indien zij hun aandacht moeten verdelen tussen de modaliteiten (**hoofdstuk 3**). Onze psychofysische experimenten laten zien dat, zoals verwacht, auditieve drempels consistent beter zijn voor normaalhorende individuen, terwijl visuele drempels (liplezen) grotendeels gelijk waren voor beide groepen. Bovendien kan het audiovisuele spraakverstaan voor normaalhorende personen goed worden beschreven door statistische facilitatie van auditieve en visuele spraakverstaan, terwijl CI-gebruikers betere integratoren zijn dan verwacht op basis van sommatie (*statische facilitatie*). Onze resultaten suggereren echter dat dit voordeel in de integratie ten koste gaat van het liplezen of luisteren in een audiovisuele omgeving. Unisensorische spraakverstaan wordt slechter voor CI-gebruikers wanneer de aandacht moet worden verdeeld over verschillende modaliteiten, d.w.z. in situaties met onzekerheid over de aankomende stimulus modaliteit. Men zou kunnen speculeren dat CI-gebruikers een afweging maken tussen integratie en aandacht. Zij richten zich alleen op één enkele modaliteit tijdens gerichte aandachtstaken, maar moeten hun beperkte aandacht verdelen over meer modaliteiten tijdens verdeelde aandachtstaken. Om een realistische indruk te krijgen van het spraakverstaan van een CI-gebruiker, zal men de situationele factoren moeten verdisconteren door spraak te presenteren in realistische en/of complexe audiovisuele omgevingen.

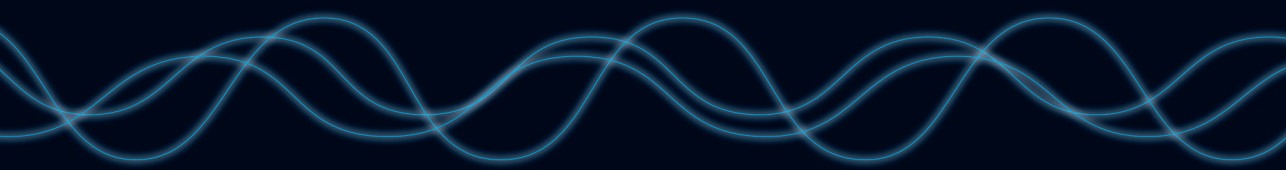
Tot slot is functionele nabij-infrarood spectroscopie (fNIRS; **hoofdstuk 4**) geïntroduceerd, een neuroimaging techniek die het mogelijk maakt om niet-invasieve corticale hersenmetingen te doen. Het gebruik van deze techniek wordt beschreven in **hoofdstuk 5** waar we hersenactiviteit van CI-gebruikers en bij normaalhorende luisteraars hebben bestudeerd.

fNIRS is een optische, non-invasieve techniek die corticale activiteit onderzoekt door middel van concentraties van geoxygeneerd en gedeoxygeneerd hemoglobine te berekenen. Het doel van het review was om de huidige stand van zaken op te maken met betrekking tot de manier waarop fNIRS is gebruikt om de auditieve functie op corticaal niveau te bestuderen. Temporele en spatiële kenmerken van de hemodynamische respons op auditieve stimulatie zijn toegelicht, alsmede experimentele factoren die de fNIRS data kunnen beïnvloeden. Het toenemende belang dat fNIRS genereert in de neurowetenschappen onderstreept het sterke potentieel van de technologie, en het lijkt waarschijnlijk dat fNIRS een nuttig klinisch hulpmiddel kan worden.

Andere non-invasieve technieken, zoals EEG en fMRI (functionele MRI), kunnen ook de neurale processen onderzoeken die ten grondslag liggen aan de audiovisueel spraakverstaan. Toch kan de ruis van de MRI-scanner en de elektromagnetische artefacten veroorzaakt door cochleaire implantaten de data (fMRI, EEG) ernstig verstoren. Daarom hebben we gebruik gemaakt van fNIRS om de corticale activiteit van de temporale cortex te bestuderen. Activatie van de cortex was niet zichtbaar op individuele fNIRS-kanalen. Echter door het verwijderen van fysiologische ruis door middel van een referentiekanaal, werd voor alle 3 de condities (auditief, visueel en audiovisueel) in beide groepen activiteit meetbaar. Auditieve stimulatie zorgde voor grotere concentratieveranderingen in de temporale cortex dan visuele stimuli. De waargenomen multisensorische verbetering van een auditieve corticale kanaal kan worden beschreven door een sommatie van de auditieve en visuele signalen met saturatie (*statische facilitatie)*.

CHAPTER 8

Dankwoord



In 2007 begon ik met de studie geneeskunde en had ik niet gedacht dat ik ooit nog een dankwoord van een proefschrift zou schrijven. Mijn sociale leven, de opleiding tot KNO-arts, tophockey, trouwen, het krijgen van twee kinderen en dat combineren met het schrijven van een proefschrift leek mij bijna onmogelijk. Maar het is gelukt (!) met de hulp van vele mensen. Iedereen die geholpen heeft in enige vorm bij de productie van dit proefschrift wil ik bedanken; een aantal mensen in het bijzonder.

Dr. M. M. van Wanrooij. Beste Marc, ik denk dat iedereen het met mij eens is dat jij de grootste hulp bent geweest in het produceren van dit proefschrift. Samen hebben we ontzettend veel dagdelen doorgebracht, kijkend naar analyse-scripts, manuscripten, figuren en apparatuur in het laboratorium. Ik ben je dankbaar voor je hulp en toewijding. Jouw toewijding maakt dat er wetenschap wordt bedreven van de bovenste plank, maar ook dat je ontzettend veel werk op je neemt. Het bekritiseren van wetenschappelijk literatuur, het bedenken van experimenten en het analyseren van data, zijn een aantal vaardigheden die ik van jou geleerd heb. Ik hoop dat onze wegen zich niet te veel scheiden, zodat we in de toekomst meer onderzoek samen kunnen opzetten.

Prof. dr. ir. A.F.M. Snik. Beste Ad, je bent zeer betrokken geweest bij dit proefschrift, wat onder andere bleek uit je snelle reactie op stukken die ik ter beoordeling stuurde, maar ook uit je interesse in persoonlijke kwesties en je geduld. Je dacht altijd praktisch mee en vooral oplossingsgericht. Jouw feedback en steun zorgden ervoor dat we als team nooit de finish uit het oog verloren. Dank voor je hulp!

Prof. dr. A.J. van Opstal. Beste John, het was fijn dat ik gedurende mijn onderzoek een plek bij jullie op de afdeling Biofysica heb gekregen. De manier hoe jij in de werkgroep, week in week uit, voorzat met veel deskundigheid, werkt inspirerend voor de groep. Je bent erg goed in complexe zaken simpel uitleggen. De gezellige sfeer, ook zeker tijdens kerstdiners met de muziek op de afdeling, zal ik niet snel vergeten. Wat een fijne afdeling om op te werken. Je hulp, de opmerkingen en revisies van dit manuscript hebben mij erg geholpen, waarvoor veel dank!

Prof. dr. E.A.M. Mylanus. Beste Emmanuel, voor mij springen er twee eigenschappen uit als ik aan jou denk: bevlogenheid en enthousiasme. Na onze besprekingen kreeg ik altijd weer veel goede energie om er vol tegen aan te gaan. Met jouw vragen stimuleerde je mij om experimenten op te zetten en onderzoek uit te zetten. Het bespreken van manuscripten zorgde ervoor dat we beiden tot nieuwe of andere ideeën kwamen. Naast je wetenschappelijke kennis ben je ook zeer kundig oorchirurg waar ik nog veel van kan gaan leren!

Dr. A. Roye. Liebe Anja. The moment that I was stuck in research you popped up in the Biophysics department. We designed multiple research proposals, and together we proposed the fundament of this thesis. We spend much time in our underground labs; you were always there when I had questions or concerns regarding any part of the experiment. Unfortunately you had to leave and had to go back to Germany, but by then the blueprint of this thesis was already set up. Thank you for your major support!

Beste staf van de afdeling Keel-Neus-, en Oorheelkunde, dank voor de mogelijkheden om de opleiding onder jullie supervisie te kunnen ontplooien. De sfeer is heerlijk op het werk, maar zeker ook ernaast bij de sociale activiteiten!

Alle proefpersonen die uren beneden in het laboratorium in het donker hebben volgehouden. Ze kwamen overal vandaan; de studenten van biofysica, de arts-onderzoekers, buurtbewoners en familie. Ook dank voor alle technische ondersteuning in het laboratorium, met name Günter en Ruurd. In het bijzonder wil ik Jan Blom bedanken voor het helpen ontwerpen van de helm om de optodes van het NIRS apparaat te fixeren op de proefpersoon.

Lieve Bas, er zat slechts 2,5 week verschil tussen onze start bij de KNO, en sindsdien hebben we veel samengewerkt en bovenal heen en weer gereisd tussen Nijmegen en Amsterdam. Dat er een klik is, bleek vanaf het eerste moment. Fijn is het om jou als collega te hebben, iemand waar je van op aan kunt, maar bovenal als vriend! Laten we samen nog veel genieten van het leven!

Lieve Stijn, ik kan een boekwerk schrijven over hoe betrokken jij bij ons bent, zelfs wanneer je in Amman woonde. Ik ben je erg dankbaar voor onze vriendschap en laten we samen het leven blijven vieren!

Lieve papa, mama, Joline, Liza en Tom, onze onderlinge band is enorm hecht en fijn. Allemaal hebben jullie een bijdrage geleverd aan het ontstaan van dit proefschrift, waarvoor dank, maar bovenal voor het zijn van mijn familie. In het bijzonder wil ik mama bedanken. Jouw organisatie kunsten en het denken in mogelijkheden zijn twee van de honderd eigenschappen die ik van je geleerd heb. Wat een power-vrouw ben je!

Lieve Timme en Max, jullie zijn 2 enorm kanjers waar we enorm trots op zijn! En die me geregeld van mijn proefschrift af hebben gehouden.

Lieve Marijke, je stimuleert mij in de richting waar ik op ga, maar houdt mij ook bij de les. Ik ben erg trots op jou, de manier hoe jij je werkende leven in Zaandam combineert met ons gezinsleven. Samen met jou kan ik de hele wereld aan, ik geniet van onze 2 zonen en van ons heerlijke leven!

CHAPTER 9

Curriculum Vitae

Luuk Pieter Harrie van de Rijt werd op 17 juni 1989 geboren te Nijmegen. Hij groeide op in een gezin met 2 zusjes en 1 broertje in Nijmegen. Als zoon van een tandarts-gnatholoog en een huisarts werd zijn interesse in de mens en het menselijk lichaam al vroeg gewekt. In Nijmegen behaalde hij aan het Nijmeegse Scholen Gemeenschap Groenewoud zijn gymnasium-diploma in 2007. Hij kreeg een aanbod om te gaan hockeyen in de hoofdklasse bij A.M.H.C. Pinoké te Amstelveen en parallel ging Luuk geneeskunde studeren aan de Vrije Universiteit van Amsterdam. Naast de internationale hockeystages richting China en Zuid Afrika, was er ook ruimte binnen de opleiding geneeskunde om buitenlandse ervaring op te doen. In 2008 volgde hij een stage in het St. Maarten Medical Center in Philipsburg te Sint-Maarten en in 2011 het coschap Heelkunde in het Steve Biko Academic Hospital te Pretoria, Zuid Afrika. Tijdens het reguliere coschap KNO in het VUmc werd zijn interesse gewekt voor Keel-, Neus- en Oorheelkunde. Hierop volgde een onderzoek bij prof. dr. Emmanuel Mylanus, wat uiteindelijk uitmondde in een sollicitatie bij de afdeling KNO van het Radboudumc te Nijmegen.

In 2014 startte hij als arts-onderzoeker aan het Radboudumc onder begeleiding van dr. Marc van Wanrooij, prof. dr. Emmanuel Mylanus, prof. dr. John van Opstal en prof. dr. ir. Ad Snik waar de basis voor dit proefschrift werd gelegd. Sinds begin 2016 is hij in opleiding tot Keel-, Neus- en Oorarts in het Radboudumc onder supervisie van prof. dr. Henri Marres en dr. Frank van den Hoogen. Zijn perifere stages volgde hij in het Canisius Wilhelmina Ziekenhuis onder begeleiding van dr. Joost Engel en dr. Bas van den Borne in 2017 en in het Rijnstate Ziekenhuis Arnhem onder begeleiding van dr. Anja Meulenbroeks en dr. Henk Bouman in 2019. In 2020 is hij gestart met de differentiatie otologie in het Radboudumc.

In 2017 is hij getrouwd met Marijke Streng. Zij zijn de trotse ouders van Timme (2018) en Max (2020).

CHAPTER 10

List of publications

van de Rijt LPH, van Opstal AJ, Mylanus EAM, Straatman LV, Hu HY, Snik AFM and van Wanrooij MM. Temporal Cortex Activation to Audiovisual Speech in Normal-Hearing and Cochlear Implant Users Measured with Functional Near-Infrared Spectroscopy. *Frontiers in Human Neuroscience* 2016;10:48: 1-14.

van de Rijt LPH, van Wanrooij MM, Snik AFM, Mylanus EAM, van Opstal AJ and Roye A. Measuring Cortical Activity During Auditory Processing with Functional Near-infrared Spectroscopy. *Journal of Hearing Science* 2018;8(4):9-18

van de Rijt LPH, Roye A, Mylanus EAM, van Opstal AJ and van Wanrooij MM (2019) The Principle of Inverse Effectiveness in Audiovisual Speech Perception. *Frontiers in Human Neuroscience* 2019;13:335:1-15.

van de Rijt LPH, van den Borne SCF, Prinsen CFM, Küsters-Vandevelde HVN. Een zeldzame tumor in de sinus maxillaris: ameloblastoom. *Nederlands Tijdschrift voor Keel-Neus-Oorheelkunde 2020 - 4*

van de Rijt LPH, van Opstal AJ and van Wanrooij MM. Multisensory Integration attention trade-off in cochlear-implanted deaf individuals. *Submitted*.

CHAPTER 11

Research data management

This thesis research has been carried out under the institute research data management policy of the Donders Institute for Brain, Cognition and Behaviour (as of 25-02-2020).

https://hdl.handle.net/2066/231917

FINDABILITY AND ACCESIBILITY

All data are available from the Donders Institute for Brain, Cognition and Behaviour repository at:

Chapter 2:
https://doi.org/10.34973/egzg-gh08

Chapter 3:
https://doi.org/10.34973/87nw-zb11

Chapter 4:
Contains no data.

Chapter 5:
https://doi.org/10.34973/jy8p-dw52