

# AUDIO-VISUAL SPEECH PROCESSING AND EFFECTS OF MULTISENSORY ASYNCRONICITY

Bachelor's Thesis

Aron Petau  
aron@petau.net  
967985  
(Universität Osnabrück)

Supervisors:  
Juliane Schwab  
(Universität Osnabrück)

October 2020



**Abstract:** In the present study I seek to identify possible problems related to learning and speech processing in general when presented with Audiovisual delays. I also examine application specific properties such as the Echo Effect in Smart Hearing Protection Devices. I discuss possible use cases with a focus on Individuals with Autism Spectrum Disorder.

**Keywords:** Multi-sensory Integration, Smart Hearing Protection, Echo Effect, Sensory Asynchrony, Autism Spectrum Disorder, Multi-modal Re-calibration

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Literature Review</b>	<b>3</b>
2.1	Introduction . . . . .	3
2.2	Multi-sensory integration . . . . .	3
2.2.1	Speech and Gestures . . . . .	4
2.2.2	Speech and Visual Lip Movement . . . . .	4
2.2.3	Multi-sensory asynchronies and the Temporal Window of integration . . . . .	4
2.2.4	The McGurk Effect . . . . .	5
2.2.5	Just Noticeable Difference (JND) . . . . .	5
2.3	The Echo Effect . . . . .	6
2.3.1	Delayed Auditory Feedback (DAF) . . . . .	6
2.3.2	Tolerable Delays . . . . .	6
2.3.3	The Special case of Smart Hearing Protection (SHPD) . . . . .	6
2.4	Age Effects . . . . .	6
2.5	Autism Spectrum Disorder . . . . .	6
2.5.1	Possible Differences to Neurotypicals . . . . .	7
2.6	Conclusion . . . . .	7
<b>3</b>	<b>Design</b>	<b>7</b>
3.1	Participants . . . . .	7
3.2	Preparations . . . . .	7
3.3	Assumptions . . . . .	7
3.4	Hypothesis . . . . .	7
3.5	Variables . . . . .	8
3.6	Setup . . . . .	8
3.7	Experiment 1: Delay . . . . .	8
3.8	Experiment 2: Echo Effect . . . . .	8
<b>4</b>	<b>Materials</b>	<b>8</b>
4.1	Content . . . . .	8
4.2	The Hardware . . . . .	8
4.2.1	Screen . . . . .	8
4.2.2	Headphones . . . . .	8
4.3	The Software . . . . .	8
4.4	Environment . . . . .	8
4.4.1	Audio Suppression . . . . .	8
4.4.2	Visual Suppression . . . . .	8
<b>5</b>	<b>Results</b>	<b>8</b>
5.1	Statistical Analysis . . . . .	8
5.1.1	Spectrogram Analysis . . . . .	8
5.1.2	Descriptive Statistics . . . . .	8
5.1.3	Analysis of the Variance . . . . .	8
<b>A</b>	<b>Appendix</b>	<b>11</b>
A.1	Acknowledgements . . . . .	11



# 1 Introduction

It is well known that sensory modalities interact during speech comprehension. I will conduct a Literature Review of prior findings in the field, discuss those and subsequently present my own experiment which aims to observe the effects of delays in auditory speech signals that would occur when utilizing selective digital filtering for background noise or distressing sounds, which could be of great impact, especially for non-neurotypical people diagnosed with Autism Spectrum Disorder.

## 2 Literature Review

### 2.1 Introduction




It has long been known that **correct** and synchronized Visual Input greatly aids peoples ability to perceive audio information and to understand natural language. Seeing the speakers lips especially helps in making sense of what is being talked about. However, this leads to a **bunch** of interesting scientific questions. **How can visual information help us hearing? Are these modalities integrated into one stream of information or are they processed separately? What happens if this process is dysfunctional? What happens when the actual sensory information is corrupted? As in, say, a video call with lagging audio?** I will review these questions and more and discuss why individuals with Autism Spectrum Disorder (ASD) and hearing-impaired individuals can provide special insights into these questions. For that, I will introduce the large research field of Multi-sensory integration, talk about research carried out in different modalities and present the concept of a temporal Window of Integration (TWIN). Then I will continue to deal with questions about the ability to detect asynchronies between modalities and discuss several ideas concerning echos in hearing. Finally, I will have a look at research on individuals with ASD and explain what we know about the differences **towards neurotypicals**, \* concerning multimodal Integration. The goal is to inform about the current state of research and identify possible open questions worth

\* usually called TD - typically developed

more research.

### 2.2 Multi-sensory integration

The idea of multi-sensory integration **goes back a while already**, with a considerable body of research conducted. Why do you have to switch off the radio when you try to park the car? Why is it harder to understand people when you cannot see their face? The most prominent theory **til** date was put forward in 1986 by Meredith and Stein in an article <sup>†</sup>, where they found via observing single cell neurons in several animals that some neurons respond differently to specific sensory inputs. Those neurons that react to input in multiple modalities they called **"multisensory"**, proving that multisensory convergence is a common and essential concept in sensoric processing. Later they refined these findings **in a very popular book called "The Merging of the Senses"** <sup>‡</sup>, putting forward the idea that this convergence is not restricted to a neuronal level, but it is a global concept governing sensory processing in the entire brain. This was called Multi-sensory Integration and is discussed **til** date. Integration was since modeled often via statistical neuronal **networks**. Nakamura (2002) 

Seeing that integration seems to be a common phenomenon, we might ask which purpose it fulfills, whether and how powerfully it can enhance our processing capacity. One paradigmatic study was conducted in **2006 in New York** by Ross et al., where speech processing was observed when **presented** with auditory input alone and contrasted with additional visual information of articulatory movements. They also manipulated the signal-to-Noise Ratio (SNR) in the auditory signals to see whether the quality of the single inputs has any effect. They found a performance increase in understanding up to three times when compared to the uni-modal condition, and also **observed that integration seems to work best with medium SNRs, meaning that our system might be best attuned to only partly corrupted inputs**, corresponding best with a real world scenario, with all kinds of interference noises occurring at

<sup>†</sup>Meredith and Stein (1986)

<sup>‡</sup>Stein and Meredith (1993)

almost all times. Ross et al. (2007)

Much benefiting from technical advancements, a more recent study observes the same phenomenon, while having a closer look at the neuronal behaviour in the actual human brain through scalp recordings via EEG.<sup>§</sup> They extend on the findings by Ross et al. by examining continuous speech versus single syllables, providing a more naturalistic framework. Furthermore, they show an increase in performance even for noise-free congruent situations, once more demonstrating that temporally congruent audiovisual (AV) stimuli (as occurring in natural Face to Face conversation) greatly aid in processing and understanding speech. Crosse et al. (2015)

Related, and similarly striking are our exceptional ability to synchronize to rhythmical stimuli. A 2015 study by Iversen et al. challenged the idea that our timing and synchronization abilities are bound to a specific modality by comparing hearing and deaf individuals. Finding no impairments in rhythmic synchronization in the deaf group, they proposed the existence of an amodal timing system responsible for integration. In support, there was no accuracy difference for the hearing and deaf groups for visual synchronization tasks, hinting towards this timing system being nonexplicit and plastic in nature. Iversen et al. (2015)

### 2.2.1 Speech and Gestures

Another well-established field of research is Audio-gestural integration. Within the framework of Multi-modal integration, the idea that we constantly incorporate information about facial expressions, body language, and hand gestures into our processing of speech and that these not only add to uni-modal processing, but the whole processing pipeline seems to be amodal, or agnostic to the modality. Specifically for speech and gestures, synchronizing effects have again recently been demonstrated by Pouw and Dixon (2019), where again

<sup>§</sup>Referring to an electroencephalogram, a relatively non-invasive method where lots of tiny electrodes are attached to the outside of the human head in order to detect small electrical currents stemming from the outer layers of the brain are measured and then statistically analysed for regional activity.

the benefits of integration were biggest under sub-optimal conditions where subjects heard a slight echo of 150ms as a distraction. Neuronal Syncing happens via visual clues.

In another EEG study by Biau et al. (2015) it has been put forward that rhythmically congruent hand gestures, so-called "beat gestures" have a significant tuning effect on the low frequency oscillatory bands in the brain, which would be a good explanation as to how the integration is realized.

### 2.2.2 Speech and Visual Lip Movement

Another powerful demonstration of Multi-modal integration comes from an oft-cited paper by Calvert et al. (1997), Where they specifically looked at the phenomenon of lip-reading which amounts to trying to assess auditory information visually. In normally hearing participants, lip information being available will lead to a major speech perception increase. The study being conducted with fMRI clearly showed that the Visual Lip-Reading Information only was enough to activate areas in the auditory cortex, suggesting that these stimuli were processed as if they were of auditory nature. Additionally, a counter-check for pseudospeech showed that the activation patterns in the auditory cortex are more than random excitement reactions to face movement, as the activation specifically only occurred when faces actually mouthing real words were presented. For an excellent in-depth review, see Stilp (2020), where several speech configurations are opened up and cases are neatly separated between forward effects, where the context precedes the target, and backwards effects, with the opposite occurring. Especially interesting for us are the backward proximal effects, which would include echo and other typical speech effects.

### 2.2.3 Multi-sensory asynchronies and the Temporal Window of integration

Based on the framework of Multi-sensory integration that was already introduced, a very sensible question might be what the limits of integration are. Some research about properly functioning integration was already presented, but what about

<sup>¶</sup>explain

situations where it does not? In a naturally occurring dialogue that may not be the first thing that comes to mind, but in an ever-increasing digital world of indirectly transmitted speech, we come to note that the temporal alignment of visual information and auditory input is of essence here. Think of the mild annoyance when the subtitles are slightly off, or even gross misunderstandings during an online Video Conference caused by temporal misalignment. A popular term here is the Temporal Window of Integration (TWIN) and it tries to capture some quantity of **minimal synchronicity that needs to be present in order to ensure speech comprehension**. A famous first try at identifying the temporal breaking point of integration **comes from Pollack (1954)**. van Wassenhove et al. (2007) and Team performed a classic simultaneity judgment (SJ) and an Identification Task in a separate experiment, in order to replicate these rather ancient results with more accurate measurements. In a SJ Task, the participant is presented with two stimuli temporally close together and has to decide whether those stimuli occurred simultaneously or not. Their findings are surprisingly **close to the original ones made by Pollack** and conclude that specifically audiovisual (AV) integration successfully occurs within a frame of about 200ms, making AV bi-modal integration relatively resilient against temporal asynchronies. Another important finding for us is that the Modal Order seems to matter. Integration was overall better when auditory stimuli were **training** the visual stimuli, making sense in so far that hearing the sound before seeing the source is quite an unnatural situation and light can travel quite a bit faster than sound, usually arriving earlier at the individual. <sup>‡</sup> Hay-McCutcheon et al. (2009) Further research suggesting that tolerance for visual-leading asynchronies is bigger can be found in Maier et al. (2011). Humans seem to be much more sensitive overall towards auditory-leading stimuli, which is likely explained by the relative minor statistical occurrence in nature.



<sup>‡</sup>just think of how in an approaching storm the lightning occurs before the thunder.

#### 2.2.4 The McGurk Effect

Also essential in the context of Multi-modal integration is a classical illusion dubbed the McGurk effect after the first team to note its existence. In order to produce the effect, they took a video of a speaker and **replaced the phoneme in the auditory canal of the video with a different one**. If done correctly, an incredibly robust "Fusion" occurs, where the visual information of the speakers lips together with the auditory information of a conflicting phoneme get "merged" and form a third phoneme that can be distinctly heard, without being present in any of the stimuli. The effect persists even when the subject is presented with the uni-modal presentations of the phonemes separately and therefore knows the third phoneme cannot be real. Macdonald and McGurk (1978) This rather astonishing effect has been serving as a paradigmatic test for audiovisual integration, for a compelling analysis of why it has to be considered outdated, see Rosenblum (2019). The case is being made, that the McGurk Effect is not fine-grained enough to properly assess multimodal integration in general and may hinder research regarding automaticity of integration.

**Soto-Faraco et al. (2004)** used the McGurk effect in an interesting manner, where they produced the effect in the Independent Dimension in a speeded classification \*\* task, effectively showing that Multi-sensory Integration happens automatically and we cannot just disregard one modality stream of information in processing.

#### 2.2.5 Just Noticeable Difference (JND)

Attunement? Closely related to the question of how large TWIN is, is the concept of the Just Noticeable Difference (JND). While TWIN looks at the breaking point of successful integration, we now talk about a presumed point where integration is still possible, but we already notice the temporal misalignment. Think again Movie subtitles. how many ms do they have to be off-sync in order for us to realize there might be a problem? This is an interesting topic of research, because this point does not seem to be fix, it can vary, depending on the needs of the situation. This amazing ability

\*\*explain

is called Attunement. What's the minimum delay people can notice? Quené (2007) Do not use! it's for tempo in speech, not for asynchrony. Find other sources

What about Sub-Noticeable Delays? Any Studies?

Klockgether and van de Par (2016)

## 2.3 The Echo Effect

### 2.3.1 Delayed Auditory Feedback (DAF)

Delayed auditory feedback classically occurs when ~~an~~ a speaker hears her own voice in a slightly delayed manner, which has been shown to induce stress. Badian et al. (1979) Usually this occurs when the speaker is wearing hearing aids, but a microphone connected to a speaker with some latency for karaoke is another easy example where DAF could occur.

In a rather recent replication of a classic study McNeill (1992) on Gestural Synchronicity, Pouw and Dixon (2019) found a reliable entrainment effect by introducing a 150ms DAF and analyzing subsequent performance.

### 2.3.2 Tolerable Delays

Also utilizing DAF, Stone and Moore (2002) looked at the permissible delays in hearing aids and identified that for regular speech, no disturbance is noticed under 30 ms. This means that any hearing aid processor, in order to be helpful and not actually detrimental, should ideally relay auditory information faster than this threshold.

### 2.3.3 The Special case of Smart Hearing Protection (SHPD)

This becomes especially interesting when confronted with the emerging option of smart hearing devices. Whereas it is nowadays efficiently and fast possible to filter out auditory frequencies, <sup>††</sup> this does have annoying side effects as filtering by frequency completely disregards the nature of the auditory input. With the use of modern digital microcontrollers it becomes possible to preprocess

---

<sup>††</sup>for example, by applying a high- or low-pass filter to make the mid-range frequencies, which contain speech more present

the audio signal in order to decide before relaying on to the integrated speakers, what type of audio is presented. Based on the result, it would become possible to apply a different set of filters, specifically tailored for the incoming signal. This type of advanced filtering comes with a substantial trade-off. Generally, the more complex and advanced a filter becomes, the more processing time is added, introducing more delay for the hearing individual. For an interesting in-depth discussion of this trade-off see Lezzoum et al. (2016)

## 2.4 Age Effects

How much does development and age influence this integration capability?

Going in the opposite direction, looking at age-related hearing loss, Rosemann and Thiel (2018) brought forward strong fMRI data to suggest that with increased hearing loss, the AV integration gets stronger. This would suggest that there likely is no linear relationship between hearing capacity and integration and it supports other claims discussed earlier that integration works best under moderately adverse conditions (such as mild hearing loss). Du et al. (2016) suggest that increased multimodal integration seems to be a common and effective way to compensate impaired speech perception.

## 2.5 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) often presents itself in social interaction and communication deficits and often goes along with atypical processing of sensory information. See APA (2000). The reason it appears in a literature review about multisensory integration is because there have been established consistent findings from a multitude of studies regarding regularities in the atypical processing across individuals with ASD. In Brandwein et al. (2013) this is discussed and extended to more general, basic nonspeech stimuli, suggesting this to be a rather consistent effect.



### 2.5.1 Possible Differences to Neurotypicals

One rather well-established processing difference lies in re-calibration speed, or maybe even the overall capacity for re-calibration. As very well explained in Turi et al. (2016), TD individuals exhibit rapid re-calibration, often shown via SJ Tasks, where the skew of the preceding runs partially determines the judgement in the current run, the individual gets "attuned" to temporal discrepancies. This finding is particularly well demonstrated in Bertelson et al. (2003), using hearing individuals. This rapid re-calibration is very diminished in ASD individuals, one consequence being a lower susceptibility to the McGurk Effect. Another, probably more important one is the reduced ability to optimise sub-optimal speech perception situations. This would also explain very well, why ASD typically start to speak faster and under-perform in language reproduction. More on a comparison with still developing Children can be found in Noel et al. (2017).

For a concise overview see Stevenson et al. (2014)

## 2.6 Conclusion

As could be seen earlier, some of these phenomena are overwhelmingly well researched, while others are still largely open. Even though we know the noticeable latency boundary for a smart hearing protection device is somewhere around 30ms, this refers to self-reported variables, it does not strictly have to coincide with a latency boundary for good performance. It is also an open question whether these boundaries are generally similar for TD and ASD populations. Furthermore, although the DAF is well represented in the research, other Echo-configurations that are imaginable with a SHPD are critically missing.

Write better (longer) Conclusions, make it a mini-discussion

## 3 Design

My own experiment aims at extending the field of research such that some concrete recommendations for the latency of a smart hearing protection device

can be made, where we can be reasonably sure that the additionally introduced latency will not carry negative consequences for speech recognition. Many interesting questions will remain unaddressed, for example, whether a very small latency affects children differently, for example while learning language capacities in school. The debate of whether this latency affects individuals with ASD in a different fashion is also not the focus here. The intention is to set a baseline with adult TD participants and establish a protocol that is repeatable and comparable, even with different demographics.

## 3.1 Participants

The experiment recruited 50 participants via the university internal mailing list targeting cognitive science students. All participants were native german speaking adults with normal or corrected to normal vision. Mean Age:

Standard Deviation:

All Participants participated in both experiments. Participation was completely voluntary. Written consent was obtained from all participants and they could leave at any time without penalties. Participants received Experiment Hours (VP-Stunden) as compensation, a mandatory part to finish the studies. No other compensation was granted.

## 3.2 Preparations

The participants are asked to complete the Experiment with their dominant right hand, handedness is assessed via the 1971 Handedness test. Oldfield (1971)

## 3.3 Assumptions

There is a universal underlying mechanism of multimodal integration for speech perception.

## 3.4 Hypothesis

When presented with greater multimodal inconsistency, speech perception ability decreases.

### **3.5 Variables**

Dependent Variable (DV)

Independent Variable (IV)

### **3.6 Setup**

2 Experiments:

Echo effect and Latency Effects

Latency Effects:

IV: No latency / Latency, varying degrees

DV: Accuracy in identification, Reaction time possibly?

Conditions: Congruent and Incongruent (Latency in Video)

Task: Identify Subject from array of images?

Harder: Identify Object, mixed entities. Was the picture in the sentence or not?

Task Effects? in asynchrony speech vs non speech noisy vs clean

Metrics: Accuracy, Speed Echo Effect:

IV: No Echo / Echo, varying loudness, varying latency

DV: Accuracy in identification, Reaction time possibly?

Conditions: No echo / Loud fast echo, loud slow echo, small fast echo small slow echo

### **3.7 Experiment 1: Delay**

### **3.8 Experiment 2: Echo Effect**

## **4 Materials**

### **4.1 Content**

Uslar et al. (2013)

### **4.2 The Hardware**

#### **4.2.1 Screen**

#### **4.2.2 Headphones**

### **4.3 The Software**

RStudio Team (2020) Audacity (2020) puredata (2020)

### **4.4 Environment**

#### **4.4.1 Audio Suppression**

#### **4.4.2 Visual Suppression**

## **5 Results**

### **5.1 Statistical Analysis**

#### **5.1.1 Spectrogram Analysis**

#### **5.1.2 Descriptive Statistics**

#### **5.1.3 Analysis of the Variance**



## References

- APA, E. (2000). Diagnostic and statistical manual of mental disorders, text revision (dsm-iv-tr). Washington, DC.
- Audacity (2020). Audacity® software is copyright © 1999-2020 audacity team. web site: <https://audacityteam.org/>. it is free software distributed under the terms of the gnu general public license. the name audacity® is a registered trademark of dominic mazzoni.
- Badian, M., Appel, E., Palm, D., Rupp, W., Sitig, W., and Taeuber, K. (1979). Standardized mental stress in healthy volunteers induced by delayed auditory feedback (daf). *European journal of clinical pharmacology*, 16(3):171–176.
- Bertelson, P., Vroomen, J., and De Gelder, B. (2003). Visual recalibration of auditory speech identification: a mcgurk aftereffect. *Psychological Science*, 14(6):592–597.
- Biau, E., Torralba, M., Fuentemilla, L., de Diego Balaguer, R., and Soto-Faraco, S. (2015). Speaker’s hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex*, 68:76–85.
- Brandwein, A. B., Foxe, J. J., Butler, J. S., Russo, N. N., Altschuler, T. S., Gomes, H., and Molholm, S. (2013). The development of multisensory integration in high-functioning autism: high-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. *Cerebral Cortex*, 23(6):1329–1341.
- Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (1997). Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596.
- Crosse, M. J., Butler, J. S., and Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, 35(42):14195–14204.
- Du, Y., Buchsbaum, B. R., Grady, C. L., and Alain, C. (2016). Increased activity in frontal motor cortex compensates impaired speech perception in older adults. *Nature communications*, 7(1):1–12.
- Hay-McCutcheon, M. J., Pisoni, D. B., and Hunt, K. K. (2009). Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis. *International Journal of Audiology*, 48(6):321–333.
- Iversen, J. R., Patel, A. D., Nicodemus, B., and Emmorey, K. (2015). Synchronization to auditory and visual rhythms in hearing and deaf individuals. *Cognition*, 134:232–244.
- Klockgether, S. and van de Par, S. (2016). Just noticeable differences of spatial cues in echoic and anechoic acoustical environments. *The Journal of the Acoustical Society of America*, 140(4):EL352–EL357.
- Lezzoum, N., Gagnon, G., and Voix, J. (2016). Echo threshold between passive and electro-acoustic transmission paths in digital hearing protection devices. *International Journal of Industrial Ergonomics*, 53:372–379.
- Macdonald, J. and McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3):253–257. cited By 284.
- Maier, J., Di Luca, M., and Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of experimental psychology. Human perception and performance*, 37:245–56.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- Meredith, M. A. and Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology*, 56(3):640–662.
- Nakamura, S. (2002). Statistical multimodal integration for audio-visual speech processing. *IEEE Transactions on Neural Networks*, 13(4):854–866.

- Noel, J.-P., De Nier, M. A., Stevenson, R., Alais, D., and Wallace, M. T. (2017). Atypical rapid audio-visual temporal recalibration in autism spectrum disorders. *Autism Research*, 10(1):121–129.
- Oldfield, R. (1971). The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, 9(1):97–113.
- Pollack, S. W. (1954). I 1954 visual contribution to speech intelligibility in noise. *J Acoust Soc Am*, 26:212215.
- Pouw, W. and Dixon, J. A. (2019). Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cognitive Science*, 43(3):e12721.
- puredata (2020). puredata.info.
- Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3):353–362.
- Rosemann, S. and Thiel, C. M. (2018). Audio-visual speech processing in age-related hearing loss: Stronger integration and increased frontal lobe recruitment. *NeuroImage*, 175:425–437.
- Rosenblum, L. D. (2019). Audiovisual speech perception and the mcgurk effect. In *Oxford Research Encyclopedia of Linguistics*.
- Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what i am saying? exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex*, 17(5):1147–1153.
- RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.
- Soto-Faraco, S., Navarra, J., and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, 92(3):B13–B23.
- Stein, B. E. and Meredith, M. A. (1993). *The merging of the senses*. The MIT Press.
- Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., and Wallace, M. T. (2014). The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. *Frontiers in psychology*, 5:379.
- Stilp, C. (2020). Acoustic context effects in speech perception. *WIREs Cognitive Science*, 11(1):e1517.
- Stone, M. A. and Moore, B. C. (2002). Tolerable hearing aid delays. ii. estimation of limits imposed during speech production. *Ear and Hearing*, 23(4):325–338.
- Turi, M., Karaminis, T., Pellicano, E., and Burr, D. (2016). No rapid audiovisual recalibration in adults on the autism spectrum. *Scientific reports*, 6:21756.
- Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., Ruigendijk, E., Brand, T., and Kollmeier, B. (2013). Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *The Journal of the Acoustical Society of America*, 134(4):3039–3056.
- van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3):598–607. *Advances in Multisensory Processes*.

# A Appendix



## A.1 Acknowledgements

Danielle Benesch  
(NSERC-EERS Industrial Research Chair in In-Ear  
Technologies (CRITIAS),  
Université du Québec (ÉTS))