Keeping track of emotions

de Boer, Minke

DOI:
10.33612/diss.179794205

Publication date:
2021

Keeping track of emotions

Audiovisual integration for emotion recognition and compensation for sensory degradations captured by perceptual strategies

Minke J. de Boer

Colophon

The experiments descibed in this thesis were conducted at the University Medical Center Groningen.

This research was supported by a BCN-BRAIN grant from the Graduate School of Medical Sciences (GSMS) and by the Landelijke Stichting voor Blinden en Slechtzienden that contributed through UitZicht (Grant number: Uitzicht 2014 – 28 – Pilot). The funding organizations had no role in the design or conduct of this research.

PhD training was facilitated by the research institute Behavioural and Cognitive Neurosciences (BCN) of the GSMS.

Publication of this thesis was financially supported by the University of Groningen, BCN, and the Professor Mulder Stichting.

About the cover

The cover figure shows fixation proportions on the actors' mouth over time (0-1000 ms after stimulus start), the eye-tracking data is from the experiment described in Chapter 4. The different colored lines indicate different viewing conditions, the solid lines represent data from younger (18-30) participants, while the dashed lines represent data from older (60-80) participants. It can be seen that older participants fixate the mouth more often than younger participants. Additionally, if macular degeneration is simulated (darker blue, purple, and red lines), there are less fixations on the mouth, especially for younger participants.
All chapter pages are variants of the cover figure; for Chapters 2-4, the eye-tracking data from its respective chapter was used. For Chapter 1, 5, and the Appendices, pilot eye-tracking with the author as participant was used.

The main title was brought to you by Menno Veldman.

Cover and layout:	Minke de Boer
Paranymphs:	Famke Huizinga Inge Krabbendam
Printed by:	Gildeprint – Enschede

rijksuniversiteit groningen

umcg

bcn

uitzicht

rijksuniversiteit
groningen

Keeping track of emotions

Audiovisual integration for emotion recognition and compensation for sensory degradations captured by perceptual strategies

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. C. Wijmenga
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

woensdag 13 oktober 2021 om 16.15 uur

door

Minke Jorien de Boer

geboren op 30 juli 1992
te Utrecht

Promotores

Prof. dr. F.W. Cornelissen
Prof. dr. D. Baskent

Beoordelingscommissie

Prof. dr. R. van Ee
Prof. dr. H.J.M. Janssen
Prof. dr. S. van der Stigchel

Table of contents

Chapter 1 p. 6
General Introduction

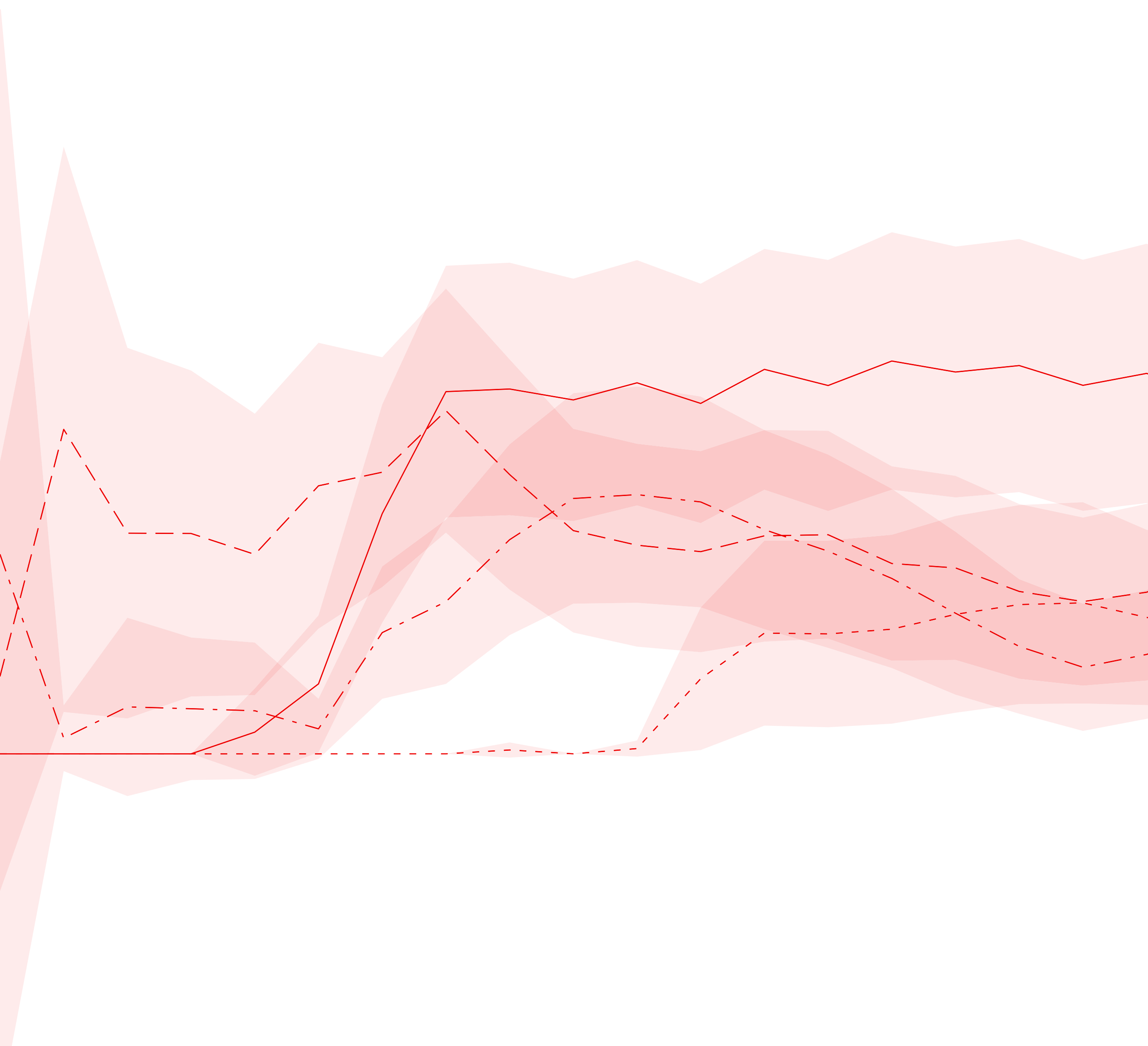
Chapter 2 p. 18
Eyes on emotion: Dynamic gaze allocation during emotion perception from speech-like stimuli

Chapter 3 p. 48
Degraded visual and auditory input individually impair audiovisual emotion recognition from speech-like stimuli, but no evidence for an exacerbated effect from combined degradation

Chapter 4 p. 80
Auditory and visual integration for emotion recognition and compensation for degraded signals are preserved with age

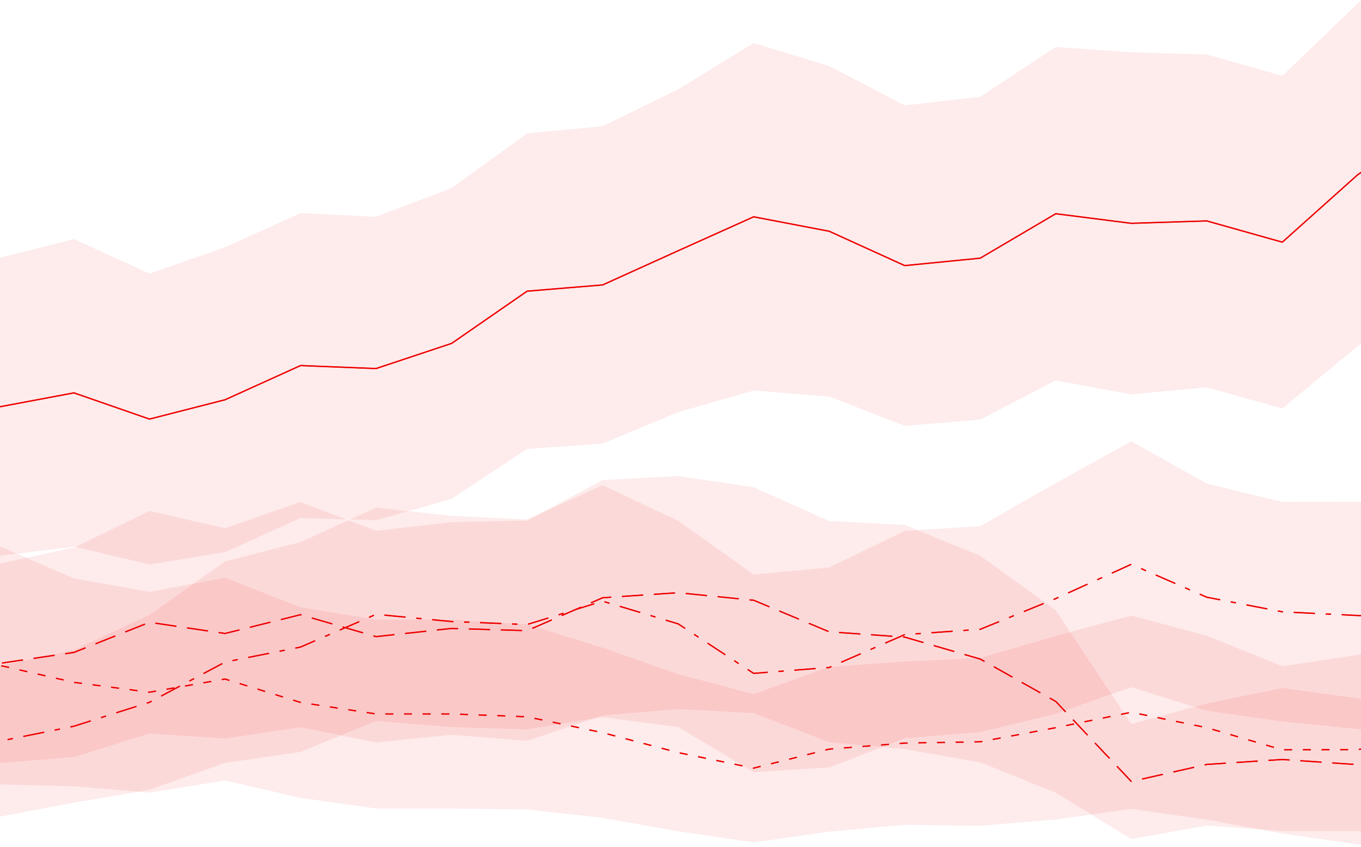
Chapter 5 p. 114
General Discussion

Appendices p. 124
References p. 126
Summary p. 136
Nederlandse samenvatting p. 138
Acknowledgements p. 140
About the author p. 144



Chapter 1

General Introduction



Preface

One of the cornerstones of life for social animals is communication. For humans, the preferred form of communication is through face-to-face speech. Understanding the lexical content of speech is important for communicating successfully. Perhaps equally important, however, is understanding the emotional expressions of one's conversational partner in order to comprehend their intentions and feelings towards you and to respond properly. In our daily lives, social interactions preferably take place as face-to-face communication, in which case emotional content is carried by dynamic acoustic and visual cues. Recognizing emotions is therefore a multimodal and dynamic process, and benefits from integration of the multimodal signals. However, the underlying mechanisms of multimodal integration processes of emotion recognition are unknown. Additionally, it is unknown how these integration processes would change in response to a change in reliability of one or both modalities, for example due to a bad internet connection during an online video call, or when the observer's senses are impaired. Sensory impairments are more common in older adults, and as more people are reaching old age, sensory impairments are becoming increasingly prevalent. When a sensory impairment only affects a single modality, i.e. having either a hearing or vision loss, it is likely that compensation for the impaired modality is possible to a great degree by relying more on the intact modality. However, when both modalities are impaired, especially when the person with the impairment is of older age, where age-related cognitive changes may also be a factor, compensation may be altered or less effective. Therefore, it is increasingly necessary to understand audiovisual integration for emotion recognition in general and with sensory impairments in particular.

The aim of this thesis was to systematically investigate and comprehensively understand audiovisual integration for emotion recognition, focusing on the following questions: 1) In normal vision and hearing, with rich, dynamic emotion cues, how do auditory and visual information contribute to audiovisual integration for emotion recognition? 2) How do real and simulated vision and hearing loss affect emotion recognition and audiovisual integration for emotion recognition? 3) Do healthy observers adapt their perceptual strategies to the presence and absence of video/audio and to simulated impairments? For all these questions, I also investigated how these change with age. To systematically address these questions, I examined how well observers recognize emotions and how recognition and perceptual strategies vary under changing availability of the visual and auditory information. Emotion recognition accuracy was used as a measure of overall performance, and eye-tracking was used to measure perceptual strategies, which in itself can be seen as a measure of information processing.

1.1 Emotion recognition

In psychology, emotion is a concept that has been a challenge to define. During the twentieth century, it has been estimated that over ninety different definitions of 'emotion' have been proposed by psychologists[1]. It is not surprising that there is so much disagreement among emotion theorists about how to conceptualize emotion and interpret its role in life when one considers there is not even consensus on the definition of the term. The definition used in this thesis, which is close to how the term is used in everyday life, is the following: an emotion is a subjective feeling that is directed towards or in response to some object. The object, which can be a person, thing, or event, is perceived as the cause of the emotion. For example, you can be irritated because the bus is late. In this case, the lateness of the bus (object) causes you to be irritated. Generally, when feelings are not attached to a particular object, it is considered a mood, not an emotion. Moods also tend to be longer lasting than emotions.

Just as there are many different definitions of the term emotion, there are many different labels for emotions. The English language contains several hundred labels, although many are synonyms or near synonyms, such as happiness, joy, elation, and cheerfulness. Because so many labels are related, some researchers have proposed that there are several basic or primary emotions, although there is no agreement on how many emotions can be considered primary. This idea of basic emotions was already proposed by Charles Darwin, who argued that specific facial expressions accompany specific emotional states in humans, and furthermore, that these expressions are universal[2]. This work was extended about a century later by Paul Ekman and Wallace Friesen, who created a system that describes the exact facial muscle movements (called action units) that constitute each of six basic emotions: surprise, happiness, fear, anger, sadness, and disgust[3,4]. When Ekman and Friesen showed photographs of these six expressions to a tribe in New Guinea that had had little contact with other cultures, they found that the people of this tribe recognized the same emotions in these facial expressions as Caucasian participants did. The reverse was also true; they took photographs of people of the tribe while they made facial expressions of these six emotions and found that western participants accurately recognized the six emotions. However, more recently, the universality of the recognition of these emotions and the actual existence of basic emotions has been called into question, suggesting that previous findings are mostly due to the use of posed expressions and the use of forced-choice response formats[5–7]. One alternative proposition is that, while emotion as a concept is biologically programmed, the process of learning to express and recognize emotions is largely influenced by cultural factors[8]. Relatedly, it has been suggested that specific emotion categories and labels are culturally specific, but the broad emotional dimensions, such as valence and arousal, are universal[7].

Besides the use of posed expressions and forced-choice response formats, another important factor that may have limited the generalizability of previous findings is the fact that most studies only use a single modality and a very limited set of emotions. In contrast, in daily life cues of emotion can occur in virtually any modality, and modalities are often combined, for example during face-to-face conversation. With technological advances, it has become possible to create a more real-life setting in the lab and doing so can yield rich information that can relate to daily life.

1.2 Audiovisual integration

While all of our senses have their own unique roles and specialties, often information across different senses is integrated for more efficient processing of the sensory input and for preparing the appropriate response to it. One of the most well-known forms of multisensory integration is audiovisual integration. In neuroscience, at the level of single neurons, integration is often measured as a neural response to a multimodal stimulus that is different from the sum of the neural responses to its individual contributors. Generally, this response is a larger increase in activity then expected on the basis of summation (multisensory enhancement), but the response could also be a decrease in activity, called multisensory depression[9]. In behavioral studies, integration is usually seen as more accurate[10,11] and/or faster responses[12,13] to multimodal compared to unimodal stimuli. For integration to occur, the observer must have the assumption (whether true or false) that the auditory and visual signals originate from the same object or event. It is most likely that signals belong to the same object or event, and are integrated, if they co-occur in both space[14,15]and time[16,17], but also other factors, such as having a shared onset or shared physical characteristics (e.g., a high-pitched voice co-occurring with the sight of a woman speaking likely leads to the assumption that the high-pitched voice belongs to the woman). Integration is not likely if there is large spatial discrepancy or large audiovisual asynchrony.

In audiovisual integration the auditory and visual signals are both weighted and merged. The weighting occurs in a statistically optimal fashion: the unimodal signals are weighted according to their reliability such that the variance in the integrated estimate is decreased (in relation to the variance in the unimodal signals) and generates the most reliable estimate possible[18]. Because of this, the multimodal benefit is higher when the unimodal signals are close to threshold, as when the unimodal variance is already low, the multimodal variance cannot decrease much further. Any discrepancies between modalities, for example when a speaker's lip movements are incongruent with the produced sound, are resolved in favor of the more precise or more appropriate modality. Thus, in the case of discrepancies, the more precise modality receives a higher weighting (or is dominant) in the final integrated estimate than the less precise modality. This dominance of one modality over the other can lead to several illusions. For example, in the ventriloquism effect, also called visual capture[19–21], the location of an auditory stimulus is shifted towards the location of the concurring visual stimulus. Ventriloquists use this phenomenon very efficiently, so that it seems to the observer as if the voice of the puppeteer comes from the moving mouth of the puppet. In this illusion, vision dominates over the audition, because of vision's high spatial precision. However, audition can also dominate over vision, such as in the sound-flash illusion, or auditory capture. In this illusion, a single flash of light is accompanied by two auditory beeps, and the single flash is observed as two flashes[22]. Here, audition dominates vision because of its higher temporal precision.

1.3 Speech perception

In speech, the smallest units of sound that allow differentiation between words in any given language are phonemes. An example is the English phoneme /r/ in the word “run”, which allows it to be distinguished from the words “gun” and “fun”. However, it is not simply the case that the brain just pastes the different phonemes together in order to understand what is being said. This is because what an individual phoneme sounds like is strongly influenced by phonemes preceding and following it, overall context of the speech segment or the conversation, and even the voice quality and dialect of the speaker. Speech perception is thus a combination of the bottom-up analysis of the acoustic features of speech, such as individual phonemes and prosody, and the top-down influence of cognitive processing, such as correct interpretation of the words using context provided by other words.

Although we tend to think that speech perception relies solely on auditory information, this is not the case. The visual cues related to the specific placement of the lips and the tongue can transmit speech information (such as a rounded mouth for /o/ and closed lips for /m/) and improve speech perception[23–25]. Lipreading is a strategy used to complement the auditory information, it makes it easier to understand the spoken words and can lower the effort of understanding. Additionally, in noisy environments, especially in situations with multiple talkers, using the lip movements of the speaker can make the difference between understanding the speaker or not. Furthermore, individuals with mild to moderate hearing loss may use lipreading to greatly compensate for their hearing loss[26,27]. Lipreading works as well as it does, mostly because of the natural temporal asynchrony between visual and auditory speech cues[28,29]. Generally, the mouth and lips are first positioned for the right phoneme before sound is produced. Visual speech cues thus often precede auditory speech cues and can likewise give a prediction of what the auditory cue will be. For example, a strong rounding of the lips while keeping the mouth open gives a clear indication that the following word will start with /o/.

This coupling of auditory and visual speech cues can also lead to fascinating illusions. A famous example is the McGurk effect[30]. In the McGurk effect, a listener is presented with an audiovisual stimulus composed of the auditory ‘ba’ and the visual ‘ga’. Strangely, the listener does not notice this incongruency and does not perceive either ‘ba’ or ‘ga’, but instead the auditory and visual information become fused and form the perception of ‘da’. While the McGurk effect will likely not occur in face-to-face conversation, it nonetheless provides convincing evidence for the strong perceptual coupling between auditory and visual speech cues.

1.4 Visual perception

When light enters our eyes, it is refracted by the lens and projected onto the retina. The retina holds the light sensitive cells of our eyes, the photoreceptors. There are two kinds of photoreceptors: rods and cones. The cones allow for the perception of color and details. The rods, on the other hand, do not enable detailed perception, but respond at much lower luminance levels than cones, and are therefore most helpful in dim environments. Cones and rods are not spread evenly across the retina. The density of cones is highest in the macula of the eye, located on the central axis of the eye, thereby allowing detailed color vision. At the center of the macula is a special location, the fovea, where other retinal cells are spread aside so the light falls directly onto the cones. In addition, the absence of rods in the fovea means that the density of cones can be much higher there, explaining why the fovea provides the highest visual acuity. Moving away from the fovea, towards the periphery, the density of cones sharply declines and with it, visual acuity. Because of its small size, the fovea also only samples a small part of the visual field. The normal human visual field is roughly 150 degrees monocularly, and 200 degrees binocularly, whereas the size of the fovea is less than two degrees (about the size of your thumbnail when held at arm's length).

The information of the retina is passed on to the brain (primary visual cortex, V1) via the optic nerve, optic tract, lateral geniculate nucleus (LGN) of the thalamus, and optic radiation. In V1, the location information of the retina is maintained, meaning that a specific location in V1 corresponds to a specific location on the retina (retinotopic representation). And because each location on the retina corresponds to a specific location in the visual field, V1 neurons are sensitive to specific locations in the visual field. However, there is not a one-to-one relation between retinal surface and cortical representations. For example, although the fovea is only a small part of the retina, it takes up over half of the primary visual cortex. This may seem like a waste of brain space, but this is not the case if one considers the importance of information that falls on the fovea. Because the fovea is located on the central axis, it receives information from fixated locations. Generally, we fixate at objects that are of the highest interest to us, therefore, these objects should also receive the most processing power, thus explaining the large coverage of the fovea in the brain. Conversely, as our vision is not very detailed in the periphery, objects located in peripheral vision also do not need to be processed in such detail as objects in the fovea.

1.5 Eye movements

Eye movements are an essential aspect of vision. Without eye movements, the viewed image would quickly fade due to habituation. Several types of eye movements exist, with the role of either stably positioning an image on the retina, or shifting gaze to a different object. Generally, gaze is guided by attention, such that where our attention goes, our gaze follows. The most well-known eye movements, and those most relevant for situations in which an observer is viewing a screen, are fixations, saccades, and smooth pursuit movements. During a fixation, the eye remains relatively still (though never completely still to avoid habituation), allowing the extraction of details of the fixated object, because the fixated object then falls onto the fovea. Fixation durations are highly variable and depend strongly on the amount of information present in the fixated location, and the number of other objects in the scene. Fixations are often alternated with saccades, which are very fast gaze shifts to new locations. When moving objects are present in the scene, smooth pursuit movements may be made, which are slower eye movements aimed at keeping the object on the fovea.

Because we generally move our eyes to locations that are of interest, eye movement measures may be used as a proxy for measures of attention. For example, fixation locations give information about objects of potential interest to the observer, while fixation durations at those locations give information about how relevant those objects were (i.e., longer fixation durations indicate more relevant objects). In addition, the amplitude of saccades inform about how an observer explores an image, whether they jump from one region to another (i.e., large saccades) or explore a smaller region by making smaller saccades (for example, switching gaze between the eyes and mouth when viewing a face).

Eye movements can be measured with eye-tracking. As gaze is a proxy for where an observer is attending and how information is retrieved, eye-tracking measures have been used already since the 19th century. At that time, eye-tracking devices did not exist, so the measurements were done by direct observation. Despite this crude methodology, these early studies led to the, at the time, surprising finding that reading does not involve a smooth pursuit movement across the text, but consists of a series of fixations and saccades (reported in Huey[31]). Nearly a hundred years later, another influential study was published by Yarbus[32], using suction caps placed on the eyes to stably and accurately measure eye movements. Yarbus showed that how an image is viewed, is largely dependent on the task given to the observer, showing for the first time that eye movements are not only guided by the stimulus (bottom-up), but also by the intentions of the observer (top-down).

With the development of video-based eye-trackers, and their relative cheapness, research using eye-tracking has expanded rapidly, owing to its relative ease of use and non-invasiveness. These trackers use a light source, usually infrared, directed at the observer's eye and the light that gets reflected from the eye is sensed by a specialist camera. Depending on the location and angle of the illuminator, the light either gets reflected off the retina and the pupil appears bright (similar to red eyes in photographs), or the illuminator is offset from the optical path and the light does not get reflected off the retina, but off the cornea, giving a dark appearing pupil and a bright corneal reflection. Changes in these reflections can be used to calculate changes in eye rotation, which in turn can be used to calculate changes in point of gaze. In order to calculate points of gaze, an extensive calibration routine is necessary to create a mapping between features in the eye image and the position of gaze in stimulus space.

1.6 Auditory perception

When soundwaves enter our ear, they are passed through to the cochlea, the component of the auditory organ where the soundwaves excite auditory neurons and allow the signal to be transferred to the brain. The basilar membrane resides within the cochlea and resonates with the incoming soundwaves and this resonation leads to movement of the stereocilia of the inner hair cells, which in turn activates the auditory nerve. However, the basilar membrane varies in thickness and stiffness throughout the cochlea, and because of this varying stiffness, different regions of the membrane resonate with different frequencies, there is thus a frequency-to-place mapping (tonotopy), similar to the place-to-place mapping in the retina. Tonotopy is preserved throughout the auditory processing stream and allows us to decompose the incoming sound into its frequency components. In its most basic functionality, tonotopy allows us to hear that a complex tone composed of a 100Hz and a 150Hz tone and a complex tone composed of a 150Hz and a 200Hz tone share the same fundamental frequency of 50Hz. While the usefulness of this ability is perhaps not apparent other than for musicians, in fact, tonotopy can also contribute to our ability of separating sounds from each other, which in turn can help differentiate voices and recognizing that a certain voice belongs to the same person even when they change the pitch of their voice.

Another important ability we have is localizing sounds. Because sounds from both ears reach the auditory cortex, there are small timing, intensity, and frequency differences between the sounds coming from the left and right ear, called interaural differences. These small differences allow us to determine whether a sound is coming from the left or the right (and through somewhat similar mechanisms, from the front or back, and above or below). Once the location of a sound is determined, we generally direct our attention and gaze there. Interaural differences allow us to direct our attention to initially unseen objects (for example, a car coming up behind you) as well as help us distinguish what object is making a specific sound in the soundscape, which is especially helpful, for example, when two people are talking at the same time.

1.7 Vision and hearing loss

Any form of damage to or malformation of the visual or auditory system can lead to temporary or permanent vision or hearing loss, respectively. The causes of damage and malformation can be diverse, with common causes being disease, medication use, genetic abnormalities, and high noise/light exposure. Damage does not have to come from a singular event, but could also be caused by age-related changes in the eye or ear physiology and accumulated noise/light exposure over the life span. Common types of permanent vision loss are glaucoma and macular degeneration, both often occurring in aged individuals. In glaucoma, the optic nerve is damaged, often because of excessive intraocular pressure. Because of this pressure, the optic nerve gets compressed, leading to vision loss starting in the periphery and slowly moving more central and finally leading to complete blindness as the disease progresses. Because the vision loss starts in the periphery, the disease often goes unnoticed for a long time. Treatment to lower intraocular pressure can slow the disease, but the damage to the optic nerve cannot be treated. In macular degeneration, the macula (in the center of which resides the fovea) slowly degenerates and currently this cannot be cured. This degeneration mostly occurs because of an inability to properly remove waste products (for example, remains of dead cells) from the retina, which then build up under the retina and can lead to damage to the photoreceptors. Because the disease does not progress beyond the macula, macular degeneration does not lead to complete blindness. However, as central vision is severely affected, macular degeneration causes severe difficulties with any task that involves detailed perception, such as reading and recognizing faces. Both these types of vision loss arise because of damage to structures of the eye. Another type of vision loss can occur because of brain damage that affects the visual structures, such as in hemianopia, where an entire hemifield is lost after the removal of or severe damage to the V1 in one hemisphere (for example when a tumor residing in V1 is surgically removed).

Hearing loss is generally divided into three types: sensorineural, conductive, and central hearing loss. Of these, sensorineural hearing loss is the most common type related to ageing and leading to permanent hearing loss. In sensorineural hearing loss, one or more of the structures in the inner ear, often the hair cells, are damaged or deficient. The hair cells may be abnormal at birth, or damaged due to age-related physiological changes, or due to prolonged noise exposure. The hair cells are not always uniformly damaged throughout the cochlea, leading to hearing loss at specific frequencies. Sensorineural hearing loss is common in aged individuals, mostly due to loss of hair cells and/or degeneration of auditory neurons. In both cases, the cells at the base of the cochlea, which is sensitive to high frequencies, are often affected first. Age-related hearing loss is thus mostly a loss of high frequency hearing. High frequencies are necessary for proper speech perception, especially consonants in the higher registers (such as the sounds of F, H, and S) and female voices and children's voices may not be heard as well as before the hearing loss. In addition, speech often sounds muffled, further increasing the load on the central auditory system to understand speech.

Sensory loss, at least when the loss is complete, is often followed by cortical reorganization, when the primary sensory cortices of the lost sense are recruited by other senses[33,34]. This cortical reorganization may enhance the functionality of the intact senses that recruit the sensory cortices of the lost sense, although this does not always happen[35]. Individuals with vision or hearing loss may thus be able to compensate well for their loss by relying more on their intact senses, which may have enhanced functionality. However, if someone has both vision and hearing loss, dual sensory loss, it can be expected that the losses amplify each other and the final loss is more severe than only the additive effect of vision and hearing loss. As more people reach old age nowadays, and many vision and hearing impairments are related to age, dual sensory losses are also becoming more common and with it, the need for new treatment and rehabilitation therapies.

As it is difficult to find a group of patients with the same vision and/or hearing loss and patients, especially when they are older, often have comorbidities besides their vision-/hearing loss, simulations can be a useful tool to conduct systematic studies. By simulating sensory impairments, the effects of these impairments can be studied in a homogeneous group of healthy observers. In addition, it is possible to study both intact vision and hearing, as well as well-controlled degraded vision and hearing in the same observer. Found changes in responses (e.g., task performance, eye movements) are then unquestionably caused by the (simulated) impairment, and not confounded by comorbidities. One drawback of the use of simulations is that it is only possible to study acute effects of sensory impairments, whereas individuals with actual sensory impairments often show some long-term adaptation to their impairment.

1.8 Thesis outline

This thesis consists of three experimental chapters, where each chapter builds on the knowledge gained in the previous chapter. In essence, each experiment studies four different aspects: visual, auditory, and audiovisual emotion recognition, and in addition to performance accuracy, the use of eye movements for recognizing emotions as a marker of perceptual strategies. Each chapter examines a different aspect of audiovisual integration for emotion perception. To increase the ecological validity of the studies, I used dynamic, multimodal stimuli in all experiments. The stimuli were audiovisual video expressions of emotions, obtained from the Geneva Multimodal Emotion Portrayals (GEMEP) core set[36]. The greatest strength of this stimulus set is that it contains fairly natural expressions (actors were instructed to imagine scenarios that would produce a specific emotion) and includes a large multitude of emotions (including subtle variations of emotions belonging to the same emotion family, such as anger and irritation). It allows to examine the perception of non-verbal emotion cues due to the use of nonsensical sentences. During each experiment, participants watched videos in which actors expressed one of twelve emotions. The videos were presented audiovisually (AV), video-only (V), or audio-only (A). Emotion recognition accuracy and eye-movements were assessed.

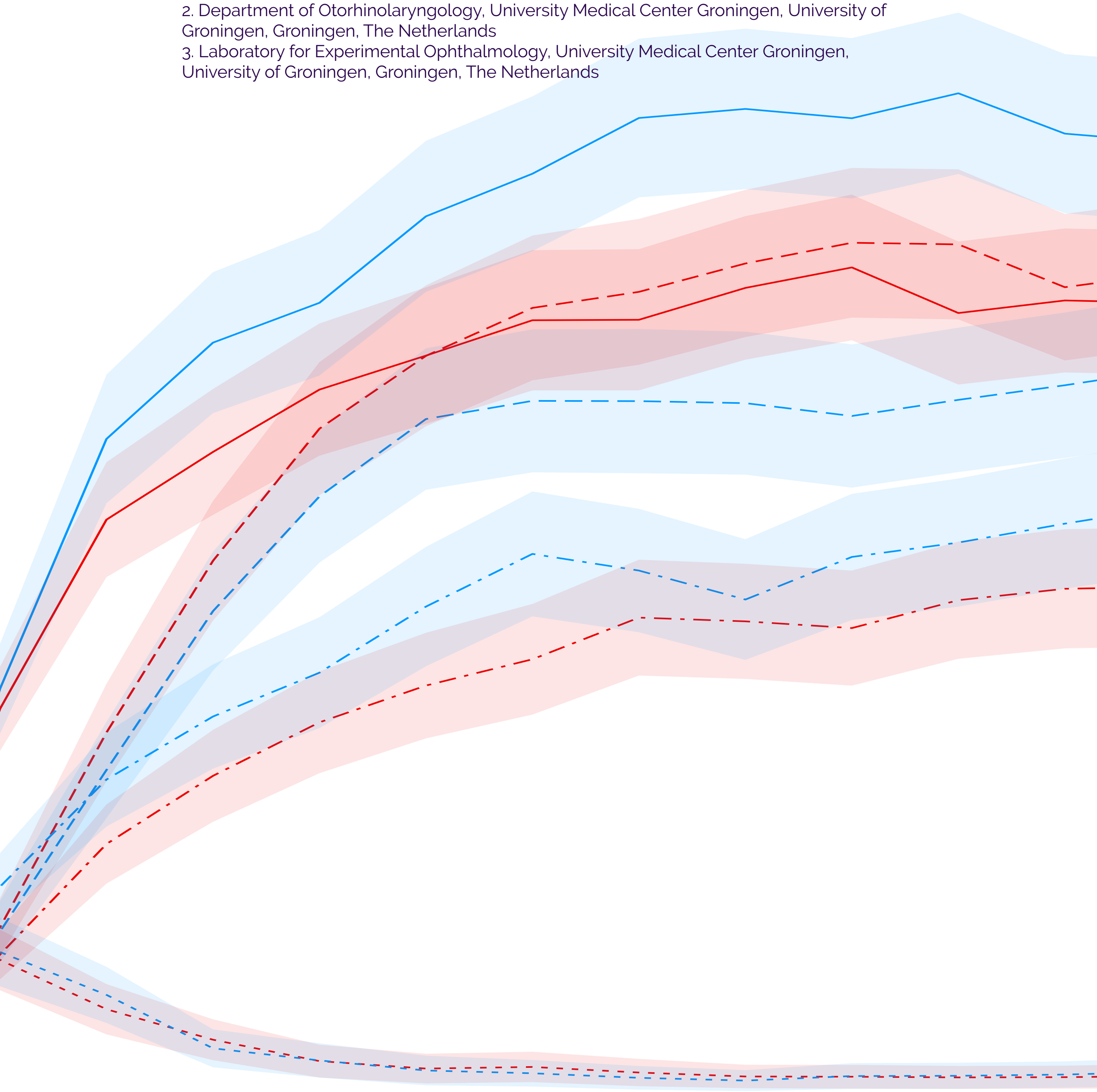
In chapter 2 a baseline understanding of audiovisual integration in emotion perception is created by studying young, healthy observers. More specifically, this chapter addresses the following research questions: 1) How do auditory and visual emotion cues contribute to audiovisual integration for emotion recognition in young, healthy observers and 2) Do these observers adapt their perceptual strategies to the video depending on the presence/absence of audio?

Chapter 3 builds on the knowledge gained from the work in chapter 2 and additionally addresses the effects of degrading visual and auditory information in young, healthy observers. Thus, this chapter focused on the following research questions: 1) How do simulated vision and hearing loss affect emotion recognition and audiovisual integration? and 2) In what manner do young, healthy observers adapt their perceptual strategies to simulated impairments? The visual and auditory degradations were intended to approximate aspects of central vision loss (as occurring in macular degeneration) and age-related hearing loss respectively.

In chapter 4, the effects of healthy ageing in combination with visual and auditory degradations are addressed. In addition, chapter 4 includes a preliminary assessment of the effects of real sensory losses in a small number of patients with varying severities of macular degeneration and age-related hearing loss. This chapter therefore focused on the following research questions: 1) How does healthy ageing affect emotion recognition in general and audiovisual integration for emotion recognition? 2) How does healthy ageing affect the ability to use perceptual strategies to compensate for simulated vision and hearing loss? and 3) How do real vision and hearing loss affect emotion recognition and audiovisual integration?

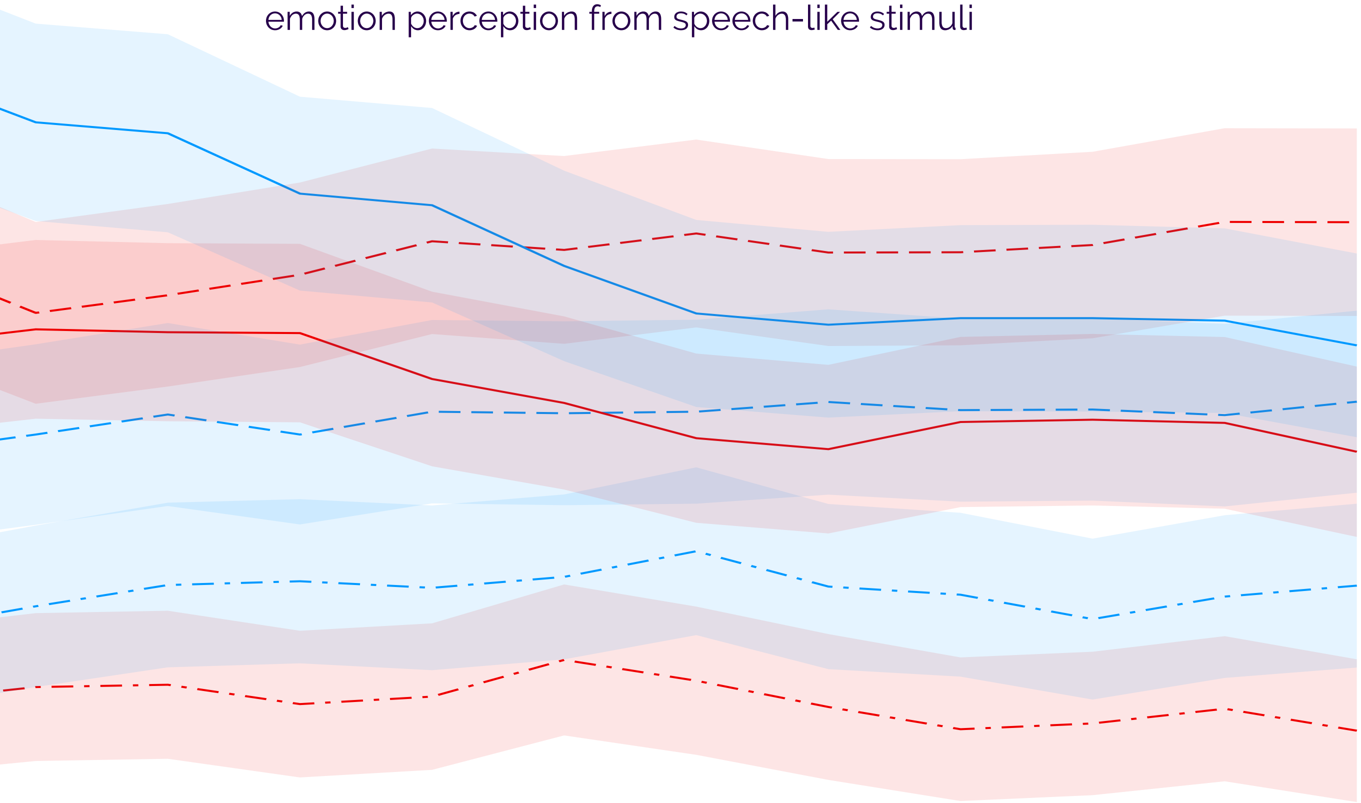
Finally, an overall discussion of the work is provided in chapter 5.

Author affiliations:
1. Research School of Behavioural and Cognitive Neurosciences (BCN), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
2. Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
3. Laboratory for Experimental Ophthalmology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands



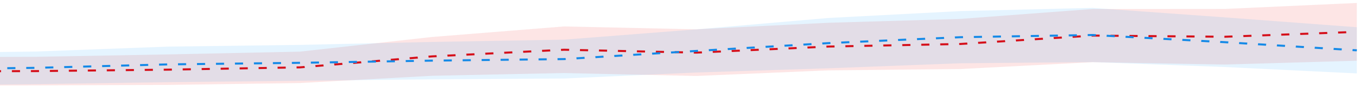
Chapter 2

Eyes on emotion: Dynamic gaze allocation during emotion perception from speech-like stimuli



2020. *Multisensory Research*, 34(1), 17-47

Minke J. de Boer[1,2,3], Deniz Başkent[1,2], Frans W. Cornelissen[1,3]



Abstract

The majority of emotional expressions used in daily communication are multimodal and dynamic in nature. Consequently, one would expect that human observers utilize specific perceptual strategies to process emotions and to handle the multimodal and dynamic nature of emotions. However, our present knowledge on these strategies is scarce, primarily because most studies on emotion perception have not fully covered this variation, and instead used static and/or unimodal stimuli with few emotion categories. To resolve this knowledge gap, the present study examined how dynamic emotional auditory and visual information is integrated into a unified percept. Since there is a broad spectrum of possible forms of integration, both eye movements and accuracy of emotion identification were evaluated while observers performed an emotion identification task in one of three conditions: audio-only, visual-only video, or audiovisual video. In terms of adaptations of perceptual strategies, eye movement results showed a shift in fixations toward the eyes and away from the nose and mouth when audio was added. Notably, in terms of task performance, audio-only performance was mostly significantly worse than video-only and audiovisual performances, but performance in the latter two conditions was often not different. These results suggest that individuals flexibly and momentarily adapt their perceptual strategies to changes in the available information for emotion recognition, and these changes can be comprehensively quantified with eye tracking.

Keywords: emotion perception, perceptual strategies, audiovisual integration, gaze allocation, dynamic, eye movements

2.1 Introduction

Successful social interactions involve not only an understanding of the verbal content of one's conversational partner, but also their emotional expressions. In everyday life, the majority of social interactions takes place as face-to-face communication and emotion perception is thus multimodal and dynamic in nature. Historically, however, emotion perception has been investigated in a single perceptual modality, with static facial emotional expressions being studied most commonly. These unimodal studies have shown one can discriminate between broad emotion categories from visual cues, such as from activations of specific facial muscle configurations[3,37,38] but also from specific body movements and postures[39,40], as well as from auditory cues, such as prosodic speech information[41,42].

The vast amount of literature on multisensory perception in general indicates that observers integrate information in an optimal manner, by weighing the unimodal information based on its reliability prior to linearly combining the now weighted unimodal signals. Because of this, the multimodal benefit, i.e. the strength of the multimodal integration, in perception is largest when the reliability of the unimodal cues is similar and each sense provides unique information. Likewise, when one sense is much more reliable — such as hearing for time interval estimation — this sense will receive a higher weight and the multisensory signal could be roughly equal to the most reliable unisensory signal[10,18,43]. However, while it is well known that observers integrate optimally, it is unknown if they also employ specific perceptual strategies when integrating. For example, how different is the visual exploration of an object when the observer is allowed to touch the object compared to when the observer is not allowed to touch the object? Here, we investigated such multisensory perceptual strategies, and the manner in which they adapt to the presence of multiple modalities, by measuring observers' viewing behavior in the context of emotion perception.

The continual adjustments of weighting unimodal information for multisensory perception make audiovisual integration a flexible process. Consequently, it can be expected that the viewing behavior observers employ also reflects this flexibility. It is long known that people naturally tend to foveate the regions of an image that are of interest[32]. What is of interest in an image is defined by visual saliency[44], but also by the nature of the perceptual task[45]. Võ and colleagues[46] proposed that gaze allocation is a functional, information-seeking process. They performed an eye-tracking study in which participants were asked to rate the likeability of videos featuring pedestrians engaged in interviews. When the video was shown with the corresponding audio, participants mostly looked toward the eyes, nose, and mouth. When the audio signal was removed, there was a decrease in fixations to the face in general, and to the mouth in particular. Thus, despite the fact that the visual signal remained unchanged, the viewing behavior changed, indicating that viewing behavior is not only directed by visual information but also by information in other modalities. These findings led the authors to conclude that gaze is allocated on the basis of information-seeking control processes. On the other hand, one could instead argue that gaze was still mostly guided by saliency. Audiovisual synchrony likely increases the saliency in certain image regions, which are then fixated more often. If the audiovisual synchrony disappears when the video is muted, the saliency of the mouth decreases and it is looked at less. On the other hand, Lansing and McConkie[47], using video recordings of everyday sentences showing only the face of the speaker, found an increase in fixations on the mouth when the video was presented without sound. The participants' task was quite different from that in Võ et al.[46] however, as here participants were required to repeat the spoken sentence. In this study[47], the mouth provides the majority of the information relevant for the task and gaze is thus directed toward it, and even more so when the task is made more difficult by removing the audio. Hence, while both these studies[46,47] used similar stimuli, the findings are drastically different, which would indicate that gaze allocation is indeed a flexible information-seeking process.

While speech sounds are mainly produced with mouth movements, many facial features additionally contribute to emotional expressions. Emotion perception from speech may thus be more complex than speech perception in terms of predicting gaze allocation. Naturally, in face-to-face communication, humans do not observe an isolated face, but a dynamic whole body that contributes with gestures and posture that may be relevant for recognizing emotions. It has been shown that observers can, under some conditions, recognize emotions from bodily expressions equally well as they can from facial expressions (for a review, see de Gelder 2009[39]). Additionally, studies showed that emotional prosody (such as pitch, tempo, and intensity) affects what facial emotion is perceived when the emotion in the voice is incongruent with the emotion in the face[48,49]. It has also been shown that visual attention is guided by emotional prosody, where observers look more often at faces expressing the same emotion than at faces expressing a different emotion[50,51]we evaluated whether emotional prosodic cues in speech have a rapid, mandatory influence on eye movements to an emotionally-related face, and whether these effects persist as semantic information unfolds. Participants viewed an array of six emotional faces while listening to instructions spoken in an emotionally congruent or incongruent prosody (e.g., "Click on the happy face" spoken in a happy or angry voice. However, these studies on the integration of facial expressions with emotional prosody mostly used static images as visual stimuli. It could thus be that observers did not necessarily attribute the face and voice to the same person, or the emotions were not being expressed at the same time. In addition, while vocal emotion always unfolds over time, a static image of a facial expression does not, despite the fact that facial expressions are dynamic in real life.

Therefore, in the present study, aiming for enhanced ecological validity, we presented dynamic multimodal emotional stimuli that always contained congruent emotion cues to express one of twelve different emotions, and also included emotions from the same family, such as anger and irritation. The stimuli were obtained from the Geneva Multimodal Emotion Portrayals (GEMEP) core set[36], which contains audiovisual video recordings of emotional expressions, with actors uttering a short nonsense sentence in an emotional manner. The video recordings show the actor from the waist up and therefore include both facial expressions as well as body, arm, and hand gestures. These stimuli have been shown to be recognizable well above chance level and were rated to be fairly believable and authentic. We used this stimulus set to measure how auditory and visual information is integrated for emotion perception.

For the purpose of this study, we consider information from two modalities as integrated when the addition of a second modality modulates the perception of the first modality[52–54], or vice versa, or when the two modalities are combined into a unified multimodal percept (for similar descriptions, see[55,56]). This combination into a unified percept could be indicated by, e.g., a gain in task performance larger or smaller than expected on the basis of independent summation of auditory and visual information or when an illusory percept arises due to the fusion of incongruent visual and auditory information (McGurk effect[30]). Relevant to our study, one form of integration is when observers alter their viewing strategies under different circumstances and tasks[46,57].

Here, we used eye tracking to gain insight into observers' viewing strategies and in what way they extract and make use of information from the stimuli. Based on previous studies examining viewing behavior during emotion perception, we cannot make a clear prediction about which areas will be fixated on most of the time, as most of these studies used static stimuli. However, two scenarios are likely: either gaze is mostly guided by information-seeking processes, or gaze is mostly guided by saliency. From the information-seeking perspective, when the task is to decode a speaker's emotional state — the focus of the current study — and congruent audio is added to a video signal, the audio signal may help in decoding the emotional information, as the information in the two modalities overlaps to some extent. Hence, auditory information could render certain visual information largely redundant, such as the motion of a speaker's mouth. Therefore, it may no longer be necessary to look at the mouth to retrieve that information and gaze can be directed elsewhere to examine different, potentially more unique, information. Alternatively, emotion recognition may rely mostly on salience, in which case an observer would always look at the most expressive region, such as the mouth for happy expressions and the eyes in angry expressions[58]. In this case we do not expect any changes in viewing behavior in response to the presence or absence of audio. Consequently, a change in viewing behavior in response to a change in modalities available can provide complementary information to task performance as a measure of audiovisual integration. In order to analyze what regions of the stimulus participants were looking at, we employed an Area-of-Interest (AOI) based analysis. Our AOIs were dynamic to capture the dynamic nature of the stimuli. Previous studies have shown that, when observing faces, most fixations are on the eyes, nose, and mouth[59,60]. In addition, it has been shown that hand movements are frequent in emotion expression[61], hence observing these movements might be useful as well for identifying the expressed emotion. Therefore, we focused our analysis on the fixations on the eyes, nose, mouth, and hands, which all could drastically change in location over the time course of the video.

To assess the presence of audiovisual integration, we evaluated whether the accuracy scores for emotion identification differed for audio-only, video-only, and audiovisual stimulus presentation. A difference in accuracy is an indication of integration and the direction this difference is in indicates whether any changes in viewing behavior are indeed functional, i.e. lead to better performance. Several studies have shown that emotion perception improves when participants have access to more than one modality conveying the same emotion[48,49,62]. Conversely, other studies have implied visual information dominates over auditory information and that — consequently — multimodal information may not necessarily improve emotion recognition and the contribution of the audio may be limited[63–65]. These conflicting findings may be the result of differences in the reliability of the auditory and visual information presented in these studies. Collignon and colleagues[55] found visual dominance when the stimuli were presented without any noise, but found robust audiovisual integration when they added noise to the visual stimulus. The visual dominance was found despite the fact that the unimodal emotion recognition performance (correct recognition rate) was the same for the noiseless visual and auditory stimuli. Thus, it appears that in noise-free environments, visual information is often treated as more reliable. Based on this, we hypothesized that we would find visual dominance in participants' accuracy scores.

2.2 Materials and Methods

2.2.1 Participants

In total, 23 young healthy participants volunteered to take part in the experiment (ten male, mean age = 23 ± 2.3 years, range: 20–31). One participant did not pass all screening criteria (described below in subsection 2.2. *Screening*) and was therefore excluded from the experiment before data collection. One other participant was excluded due to severe difficulties in calibrating the eye tracker. Consequently, 21 participants completed the entire experiment (nine male, mean age = 23 ± 2.4, range: 20–31) and were included in the data analysis. The sample size was initially based on similar previous studies on audiovisual emotion perception[55,62,66,67] and was subsequently modified in order to ensure proper counterbalancing of the experimental blocks. All participants were given sufficient information about the nature of the tasks of the experiment, but were otherwise naïve as to the purpose of the study. Written informed consent was collected prior to data collection. The study was carried out in accordance with the Declaration of Helsinki and was approved by the local medical ethics committee (ABR nr: NL60379.042.17).

2.2.2 Screening

Prior to the experiment, potential participants' hearing and eyesight were tested to ensure auditory and (corrected) visual functioning was within the normal range.

Normal auditory functioning was confirmed by measuring auditory thresholds for pure tones at audiometric test frequencies between 125 Hz and 8 kHz. A staircase method, similar to typical audiological procedures, was used to determine the thresholds, in a soundproof booth. Testing was conducted at each ear, always starting with the right ear. In order to participate in the experiment, audiometric thresholds at all test frequencies needed to be as good as or better than 20 dB HL for the better ear. Normal visual functioning was tested with measurements of visual acuity and contrast sensitivity (CS). These tests were performed using the Freiburg Acuity and Visual Contrast Test (FrACT, version 3.9.8)[68,69]. A visual acuity of at least 1.00 and a logCS of at least 1.80 (corresponding roughly to a 1% luminance difference between target and surround) were used as cutoff thresholds to participate in the experiment. Visual tests were performed on the same computer as used in the main experiment.

Additional exclusion criteria were neurological or psychiatric disorders, dyslexia, and the use of medication that can potentially influence normal brain functioning.

2.2.3 Stimuli

The stimuli used in this study were taken from the Geneva Multimodal Emotion Portrayals (GEMEP) core set (for a detailed description, see Bänziger at al.[36]), which consists of 145 audiovisual video recordings (mean duration: 2.5 s, range: 1–7 s) of emotional expressions portrayed by ten professional French-speaking Swiss actors (five female). The vocal content of the expressions were two pseudo-speech sentences with no semantic content but resembling the phonetic sounds in western languages (“nekal ibam soud molen!” and “koun se mina lod belam?”). Out of the total set of 17 emotions, 12 were selected for the main experiment. The selection was made to produce a well-balanced design, such that all actors portrayed the selected emotions, and further, these emotions could be distributed evenly on the quadrants of the valence-arousal scale[70], see Table 1, thereby balancing positive and negative emotions as well as high- and low-arousal emotions within the selected stimulus set. This resulted in a total of 120 stimuli used in our experiments. The five remaining emotions that were excluded from data collection were used as practice material to acquaint participants with the stimulus materials and the task.

Table 1. The selected emotion categories used in the experiment. The emotions for the main experiment are distributed over the quadrants of the valence-arousal scale[70]. The five additional emotions, listed under the table, are used for the practice trials.

		Valence	
		Positive	**Negative**
Arousal	**High**	Amusement Joy Pride	Fear Despair Anger
	Low	Pleasure Relief Interest	Irritation Anxiety Sadness
Additional		Disgust Contempt Surprise Admiration Tenderness	

The audio from all movie files was edited in Audacity (version 2.1.2; http://audacityteam.org/), to remove any audible noise or clipping from the audio recordings, and saved as 16-bit

WAV-files. To do so, in most cases, the editing consisted of using the built-in ‘Noise Reduction’ effect to reduce background noise as much as possible without affecting the speech signal. In rare cases, the files contained clipping, which was removed by manually adjusting the clipped regions of the waveform. Audio recordings were then root-mean-square (RMS)-equalized in intensity level, and re-merged with the corresponding video files (thereby replacing the old audio) using custom-made scripts.

2.2.4 Experimental Setup

Experiments were performed in a silent room, which was dark except for the illumination provided by the screen. Participants were seated in front of a computer screen at a viewing distance of 70 cm with their head in a chin and forehead rest to minimize head movements. Stimuli were displayed and manual responses were recorded using MATLAB (Version R2015b; The Mathworks, Inc., Natick, MA, USA), the Psychophysics Toolbox (Version 3)[71–73] and the Eyelink Toolbox[74] extensions of MATLAB. The stimuli were presented full-screen on a 24.5-inch monitor with a resolution of 1920 × 1080 pixels (43 × 24.8 degrees of visual angle). Average screen luminance was 38 cd/m^2. Stimulus presentation was controlled by an Apple MacBook Pro (early 2015 model). Audio was produced by the internal soundcard of this computer and presented binaurally through Sennheiser HD 600 headphones (Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany). The sound level was calibrated to be at a comfortable and audible level, at a long-term RMS average of 65 dB SPL.

To measure eye movements, an Eyelink 1000 Plus eye tracker, running software version 4.51 (SR Research Ltd., Ottawa, Ontario, Canada), was used. Gaze data were acquired at a sampling frequency of 500 Hz. The eye tracker was mounted on the desk right below the presentation screen. At the start of the experiment, the eye tracker was calibrated using its built-in nine-point calibration routine. Calibration was verified with the validation procedure in which the same nine points were shown again. The experiment was continued if the calibration accuracy was sufficient (average error of less than 0.5° and a maximum error of less than 1.0°). A drift check was performed both at the start of the experiment and after each break. If the drift was too large (i.e., more than 1.0°), the calibration procedure was repeated.

2.2.5 Procedure

In this study, behavioral and eye-tracking data were obtained to identify accuracy and gaze fixation of emotion identification with dynamic stimuli. In each trial, prior to each stimulus presentation, a central fixation cross appeared for a random duration between 500 and 1500 ms. The response screen followed each stimulus presentation after 100 ms and remained on screen until the participant made his or her response. The order of events in a typical trial is shown in Fig. 1.

Participants were asked to identify the emotion presented in one of three stimulus presentation modalities: audio-only (A-only), video-only (V-only), or audio and video combined (AV). They were asked to respond as accurately as possible in a forced-choice discrimination paradigm, by clicking on the label on the response screen corresponding with the identified emotion. Emotion labels were shown and explained before the experiment. Participants were further instructed to blink as little as possible during the trial and maintain careful attention to the stimuli.

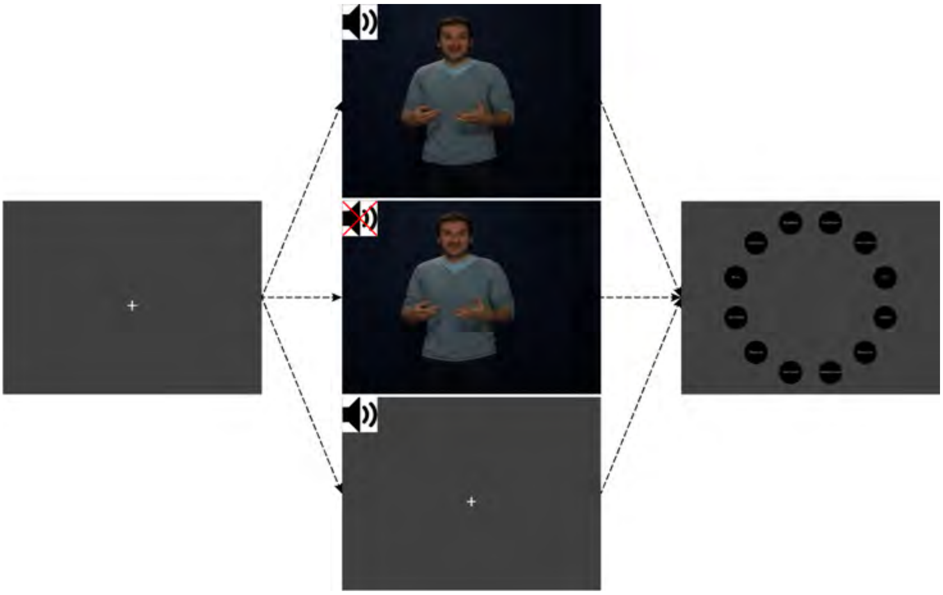


Figure 1. Schematic representation of the events in a single trial. Participants first were shown a fixation cross (left), followed by the stimulus, presented audiovisually (middle top), visually (middle), or aurally (middle bottom). After stimulus presentation, a response screen (right) with labels indicating the possible emotions appeared and remained on screen until the participant made a (forced) response. Emotion labels were in Dutch, from top right going clockwise they are: opgetogen (joy), geamuseerd (amusement), trots (pride), voldaan (pleasure), opgelucht (relief), geïnteresseerd (interest), geïrriteerd (irritation), ongerust (anxiety), verdrietig (sadness), bang (fear), wanhopig (despair), and woedend (anger).

In total, each participant was presented with all 120 stimuli (twelve emotions × ten actors) in all three blocks: an A-only block, a V-only block, and an AV block. Block order was counterbalanced between participants. Stimulus order within each block was randomized. Participants were encouraged to take breaks both within and between blocks (breaks were possible after every 40 trials) to maintain concentration and prevent fatigue. Breaks were self-paced and the experiment continued upon the participant pressing the spacebar. Following each break, a drift correction was applied to the eye-tracking calibration. Fifteen practice trials (five training trials for each modality) preceded the experiment to familiarize participants with the task and stimulus material. In total, the experiment consisted of 375 trials, including the 15 practice trials, and took at most one hour to complete. Feedback on the given responses was provided during the practice trials only.

2.2.6 Analyses of Behavioral Data

To assess the presence of audiovisual integration, we tested whether performance for emotion identification differed for A-only, V-only, and AV stimulus presentation. We additionally employed a measure that quantifies the size of the effect from audiovisual integration, i.e. whether audiovisual integration is sub-additive (i.e., lower than expected based on the simultaneous and independent processing of both unisensory modalities), additive (i.e., equal to a summation of the auditory and visual evidence), or supra-additive. A supra-additive effect would be indicative of a gain in performance beyond what is gained by independently summing the information from both modalities[75,76].

Accuracy scores for each emotion and modality were converted to unbiased hit-rates[77] prior to further analyses. Unbiased hit-rates (H_u) were used to account for response biases. Unbiased hit-rates were then arcsine-transformed to ensure normality and analyzed in R (version 3.6.0; R Foundation for Statistical Computing, Vienna, Austria — https://cran.r-project.org) with repeated-measures ANOVA (*aov_ez* from the *afex* package, version 0.25-1). For the ANOVA, arcsine-transformed H_u was the dependent variable, and *modality* (with three levels; A-only, V-only, and AV) and *emotion* (with 12 levels) the fixed-effects variables. Greenhouse–Geisser correction was performed in cases of a violation of the sphericity assumption. Effect sizes are reported as generalized eta-squared (*ges*). Pairwise comparisons were performed to test main effects (comparing different modalities) and interactions (the effect of modality for each emotion) using *lsmeans* from the *emmeans* package (version 1.4.1). For comparing differences between modalities, the Bonferroni correction was applied to make sure our conclusions were not based on a possibly too liberal adjustment. For comparing modality differences between emotions, we used the False Discovery Rate (FDR) correction in order to ensure no effects were lost due to strict adjustments of *p*-values due to the many pairwise comparisons made.

For a quantitative assessment of the AV integration effect, we tested if the measured performance for AV exceeded the statistical facilitation produced by A+V. To quantify the predicted H_u for the independent summation of A and V we used the following equation[75,76]:

$$H_{u\_pred}(AV) = H_u(A) + H_u(V) - H_u(A) \cdot H_u(V) \tag{1}$$

If the H_u for the AV modality exceeds the predicted H_u, as assessed by a paired *t*-test, this indicates A and V are integrated in a supra-additive manner[78,79]. Paired *t*-tests were only performed when at least the differences between AV and V-only and between AV and A-only were significant.

2.2.7 Analyses of Eye-Tracking Data

Fixations were extracted from the raw eye-tracking data using the built-in data-parsing algorithm of the Eyelink eye tracker. We performed an AOI-based analysis for fixations made during stimulus presentation (only for the AV and V-only modalities as for the A-only modality there is no visual stimulus aside from a fixation cross). Trials with blinks longer than 300 ms during stimulus presentation were discarded. The analysis was restricted to fixations made between 200 ms and 1000 ms after stimulus onset. The first 200 ms were discarded because this is the time needed to plan and execute the first eye movement. No data after 1000 ms were taken into account to limit data analysis to the duration of the shortest movie at 1000 ms.

In the videos, the eyes (left and right), nose, mouth, and hands (left and right) of the speaker were chosen as AOIs. Because the stimuli are dynamic, we created dynamic AOIs. Coordinates of the AOI positions for each movie and each frame were extracted using Adobe® After Effects® CC (Version 15.1.1; Adobe Inc., San Jose, CA, USA). For the face AOIs, these coordinates were obtained by placing an ellipsoid mask on the face area and applying a tracker using the 'Face Tracking (Detailed Features)' method, which automatically tracks many features of the face (see Fig. 2 for an example frame with AOIs drawn in). Face track points were visually inspected and manually edited (i.e. moved into the correct place) whenever the tracking software failed to correctly track them.

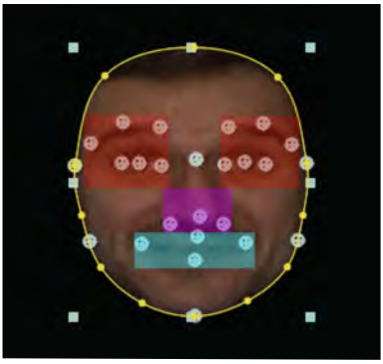


Figure 2. Face tracking in Adobe After Effects CC. The yellow line is the ellipsoid mask after automatic alignment to the contours of the face. Each circled cross is a face track point. The colored rectangles indicate the locations of the different areas of interest (AOIs); the red rectangles denote the right- and left-eye AOIs, the purple rectangle shows the nose AOI, and the blue rectangle specifies the mouth AOI.

Coordinates of all obtained face track points for each movie frame were stored in a text file and used to create rectangular AOIs. For the eyes' AOI we used the coordinates of the following face track points: 'Right/Left Eyebrow Outer' for the *x*-position of the lateral corner, 'Right/Left Eyebrow Inner' for the *x*-position of the medial corner, 'Right/Left Eyebrow Middle' for the top, and the middle between the *y*-positions of 'Left Pupil' and 'Nose tip' for the bottom, indicating the eye–nose border. Two individual AOIs were created for the left and right eye, which were later merged for analyses. For the nose AOI: the eye–nose border as the top, the nose–mouth border (middle between the *y*-positions of 'Right Nostril' and 'Mouth Top'), the *x*-position of 'Right Nostril' for the left corner, and the *x*-position of 'Left Nostril' for the right corner. For the mouth AOI: the *x*-position of 'Mouth Right' for the left corner, the *x*-position of 'Mouth Left' for the right corner, the nose–mouth border for the top, and the *y*-position of 'Mouth Bottom' for the bottom. Each AOI was expanded by 10 pixels on each side (20 pixels across the horizontal and vertical axes), except at the eye–nose and nose–mouth borders. Overlap between AOIs was avoided. The actual size of each AOI varied across actors and frames e.g. due to some actors being closer to the camera.

For the hand AOIs, the 'Track Motion' method was used, in which a single tracker point (per hand) was used to track position. The tracker point was placed approximately in the center of the hand. The track point was manually edited whenever the tracking software failed to correctly track it. This happened often due to the complex movements the hands made in most movies. Figure 3 shows example frames from one movie. After extracting the coordinates, a sphere with a radius of 75 pixels was used to create the AOI.

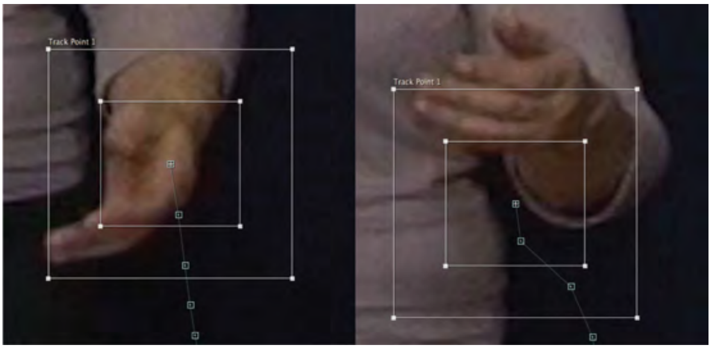


Figure 3. Hand tracking using Adobe After Effects CC. In both images, the attach point is at the center (from which the coordinate is extracted), the inner box is the feature region (i.e., what the tracked region looks like), and the outer box is the search region of the tracker (i.e., the region in which the tracker will search for the feature region). Additionally, the tracked points in previous frames can be seen. As can be seen in the left image, tracking works well early in the movie. As the hand starts to change shape later in the movie, however, the tracker errs. This can be seen on the right image where the tracker loses the hand from sight and tracks the arm and background instead.

Then, for each fixation datapoint we checked whether the fixation was on one of the AOIs (with the coordinates from the movie frame co-occurring with the time of the fixation), leading to one binary vector for each AOI with the same length as the length of the fixation data. These vectors were then averaged per trial, giving a mean fixation proportion on each AOI for each trial. Lastly, the means were arcsine-transformed. A mixed linear regression was performed in *R* (using *lmer* from the *lme4* package, version 1.1-21) on correct trials only, as we were most interested in examining whether changes in viewing behavior due to changes in modality availability were adaptive, leading to good performance. In line with the analyses of unbiased hit-rates, the model included *modality*, *emotion*, and *AOI* as fixed effects, which were allowed to interact with each other. Random intercepts were included for participant and movie and a random slope for modality was included for both participant and movie if the model still converged (otherwise, only a random slope for modality was included for participants). Overall significance of the main effects and interactions was assessed using the *Anova* function from the *car* package (version 3.0-3). Pairwise comparisons were performed to test whether fixation proportions on different AOIs differed for different modalities and emotions using *lsmeans*. As before, for comparing differences between modalities, the Bonferroni correction was applied while for comparing differences between emotions we used the FDR correction.

Lastly, we ran a second model to test whether fatigue or boredom, which may have occurred due to the lengthy duration of the experiment, had an effect on fixation patterns, by adding experimental block to the model. There was no significant effect of block on fixation patterns ($\chi^2_1 = 1.79$, $p = 0.18$), ruling out additional effect from potential boredom and fatigue.

2.3 Results

Participants identified dynamic emotional expressions presented in movies while their eye movements were recorded. The objective of this study was to see if emotions are processed similarly whether conveyed in a unimodal (A-only, V-only) or multimodal (AV) manner, as measured by performance levels and fixation patterns. To achieve this objective, here we present analyses of accuracy and gaze differences for different modalities and emotions. Accuracy and fixation data for individual participants can be found in Supplementary Figs S1, S2, S3, and S4. Confusion matrices for each modality can be found in Fig. S5.

2.3.1 Accuracy across Modalities and Emotions

Accuracy scores in unbiased hit-rate (H_u) and averaged over all participants and testing blocks is shown in Fig. 4. On average, participants performed the task with a mean accuracy of 0.37, well above the chance level of 0.083.

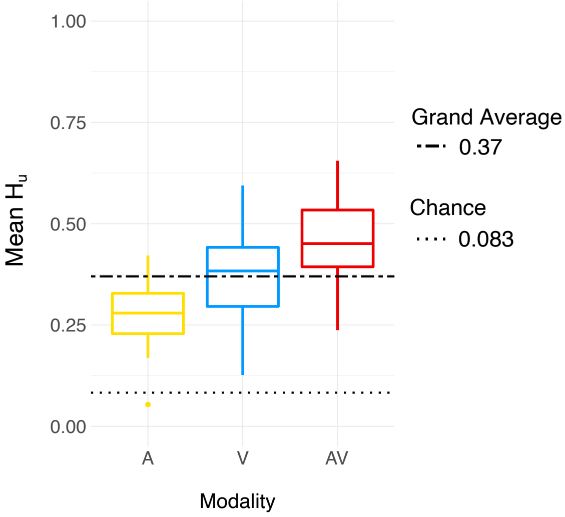


*Figure 4. Task performance for each modality, shown as unbiased hit-rates (H_u) and averaged across all participants and blocks. Each box shows the data between the first and third quartiles. The horizontal solid line in each box denotes the median. The whiskers extend to the lowest/highest value still within 1.5 * interquartile range. Dots are outliers. The black dashed dotted line indicates the grand average performance (0.37). The black dotted horizontal line indicates chance level performance (0.083).*

A visual inspection of Fig. 4 suggests performance is lowest for the A-only modality and highest for the AV modality. This was also confirmed by the ANOVA, which had H_u as the dependent variable, and *modality* and *emotion* as independent variables. The model showed an overall effect of *modality* ($F_{2,40}$ = 42.7, p < 0.001, *ges* = 0.18), a main effect of *emotion* ($F_{11,220}$ = 53.1, p < 0.001, *ges* = 0.48), and a significant interaction between *modality* and *emotion* ($F_{22,440}$ = 5.2, p < 0.001, *ges* = 0.07). Bonferroni-adjusted pairwise comparisons showed performance was significantly different between all modalities (A-only – AV: t_{40} = −9.13, p < 0.001; A-only – V-only: t_{40} = −5.80, p < 0.001; V-only – AV: t_{40} = 3.34, p = 0.006). Therefore, performance was lowest for A-only (mean accuracy = 45%), intermediate for V-only (mean accuracy = 62%), and highest for AV (mean accuracy = 70%), with all differences between modalities being significant.

Further inspection of the *modality*-by-*emotion* interaction showed that, in general, performance was lowest for A-only, intermediate for V-only, and highest for AV, but this was not true for all emotions. In fact, for most emotions (except for Pleasure, Relief and Anxiety), there was no significant difference in performance between V-only and AV. In addition, for some negative valence emotions (Fear and Anger), none of the comparisons between modality pairs produced a significant difference. Lastly, for Pleasure, Relief, and Despair the difference between V-only and A-only was not significant. The complete list of all comparisons is given in Table 2 and further visualized in Fig. 5.

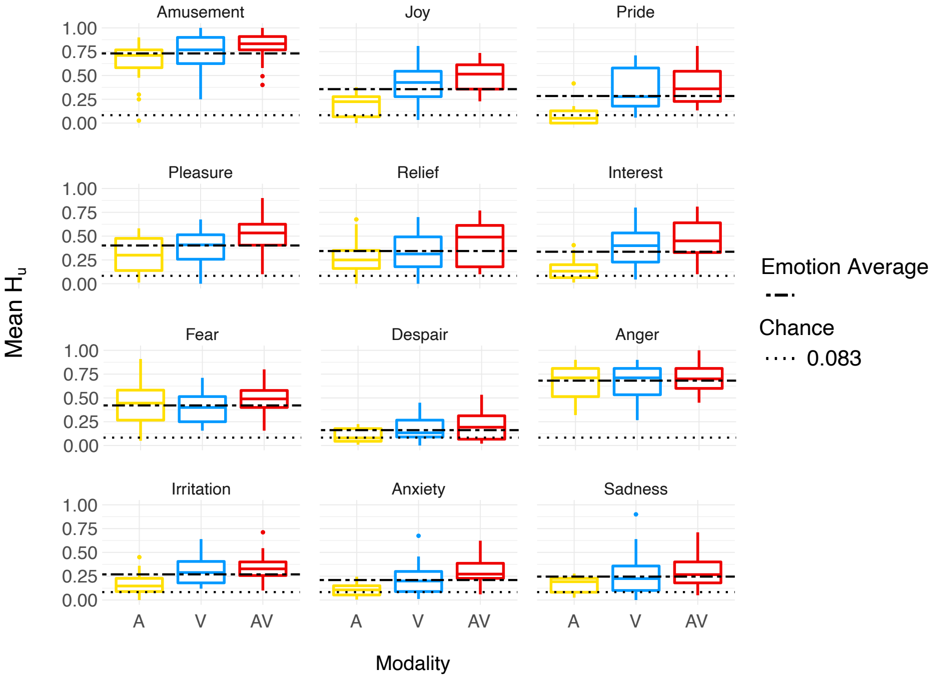


Figure 5. Task performance for each modality, shown as unbiased hit-rates (H_u), similar to Fig. 4, but shown for each emotion. The black dashed dotted line in each panel indicates the average performance for that particular emotion. The black dotted horizontal line indicates chance level performance (0.083).

While AV performance was significantly higher than both A-only and V-only performance, indicating that AV integration took place, the AV integration effect was sub-additive as performance for AV was significantly lower than predicted on the basis of additivity (t_{20} = −3.06, *p* = 0.006; H_{u_pred}(AV): 0.52 ± 0.12, H_{u}(AV): 0.45 ± 0.10). Considering individual emotions, only for anxiety, pleasure, and relief performance differed between both AV and V-only and between AV and A-only, and thus, only for these emotions it was further tested whether AV performance was supra-additive. AV performance was not significantly different from the predicted additive performance for Anxiety (t_{20} = 0.006, *p* = 0.99; H_{u_pred}(AV): 0.30 ± 0.16, H_{u}(AV): 0.30 ± 0.14), for Pleasure (t_{20} = −1.33, *p* = 0.20; H_{u_pred}(AV): 0.56 ± 0.05, H_{u}(AV): 0.51 ± 0.05) or for Relief (t_{20} = −1.54, *p* = 0.14; H_{u_pred}(AV): 0.50 ± 0.05, H_{u}(AV) : 0.43 ± 0.05), indicating that the AV integration effect was additive in all three emotions.

*Table 2. Contrasts for the modality-by-emotion interaction showing the model estimate differences, with the False Discovery Rate (FDR)-adjusted p-values in parentheses. A positive contrast means performance in the first condition was better than in the second of the comparison (and v.v.). Significant differences are indicated in **bold.***

	Contrast		
	AV–V	**AV–A**	**V–A**
Positive valence, high arousal			
Amusement	0.09 (0.09)	**0.23 (<0.001)**	**0.15 (0.005)**
Joy	0.09 (0.07)	**0.37 (<0.001)**	**0.28 (<0.001)**
Pride	0.09 (0.09)	**0.48 (<0.001)**	**0.40 (<0.001)**
Positive valence, low arousal			
Pleasure	**0.15 (0.005)**	**0.23 (<0.001)**	0.08 (0.10)
Relief	**0.12 (0.03)**	**0.19 (<0.001)**	0.07 (0.16)
Interest	0.08 (0.12)	**0.37 (<0.001)**	**0.29 (<0.001)**
Negative valence, high arousal			
Fear	0.09 (0.20)	0.08 (0.20)	−0.02 (0.74)
Despair	0.05 (0.37)	**0.13 (0.02)**	0.09 (0.12)
Anger	0.04 (0.72)	0.04 (0.72)	0.008 (0.87)
Negative valence, low arousal			
Irritation	0.04 (0.42)	**0.21 (<0.001)**	**0.17 (0.001)**
Anxiety	**0.12 (0.02)**	**0.25 (<0.001)**	**0.13 (0.01)**
Sadness	0.05 (0.28)	**0.17 (0.003)**	**0.11 (0.04)**

2.3.2 Fixation Patterns across Modalities and Emotions

Fixation proportions, averaged over all stimuli and participants, are shown for all AOIs in Fig. 6. Figure 6a shows how the fixation proportions change over the analyzed time course, while Fig. 6b shows the fixation proportions averaged over the trial. Figure 6 suggests differences in viewing behavior between modalities.

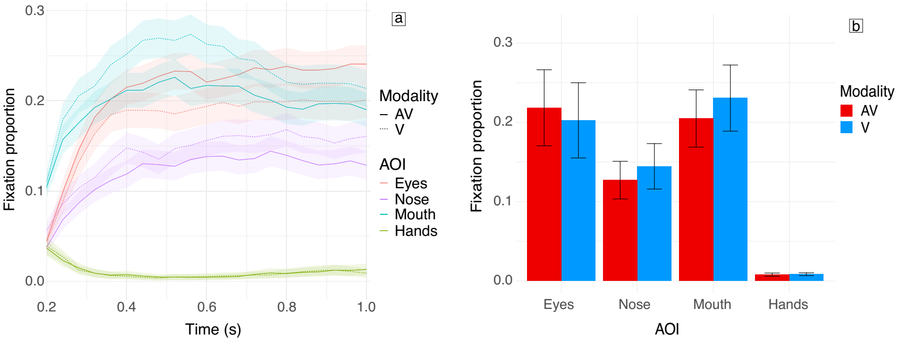


Figure 6. Fixation proportions for correct trials on all areas of interest (AOIs) (face, i.e., eyes, nose, mouth; and hands), across the analyzed time course (a) and averaged over the analyzed time course (b), both averaged over all stimuli and participants. Shaded areas around each line (a) and error bars (b) denote the standard error of the mean (SEM).

The regression model confirmed this. The model included *modality*, *emotion*, and *AOI* as fixed effects (and their interactions). A random intercept was included for both *participant* and *movie*, and a random slope for *modality* for *participant*. There was a main effect of *AOI* (χ^2_3 = 3314.1, p < 0.001), a significant interaction between *modality* and *AOI* (χ^2_3 = 34.2, p < 0.001), and a significant interaction between *emotion* and *AOI* (χ^2_{33} = 184.2, p < 0.001). Significant main effects and interactions were followed up with post-hoc testing, as further detailed below.

Bonferroni-corrected pairwise comparisons showed that, in general, the mouth was fixated more often than the eyes (*z*-ratio = −7.4, p < 0.001) and nose (*z*-ratio = −14.9, p < 0.001), the eyes were fixated more often than the nose (*z*-ratio = 7.5, p < 0.001) and all face AOIs were fixated more than the hands (all p < 0.001). Additionally, participants fixated more on the mouth (*z*-ratio = −3.1, p = 0.002) and nose (*z*-ratio = −2.3, p = 0.02) and less on the eyes (*z*-ratio = 3.08, p = 0.02) in the V-only modality compared to the AV modality. There was no difference in fixation proportions on the hands (*z*-ratio = −0.1, p = 0.92). Lastly, the results of the emotion by AOI interaction can be found in Table 3 and are visualized in Fig. 7. Because fixations on the hands were so scarce, only comparisons between the face AOIs are shown. In general, the same pattern can be seen for each emotion; most fixations are on the mouth, then the eyes, then the nose, and lastly on the hands (not shown in the table). There is only one exception to this: participants fixated on the eyes more often than on the mouth for Anger (*z*-ratio = 2.6, p = 0.01).

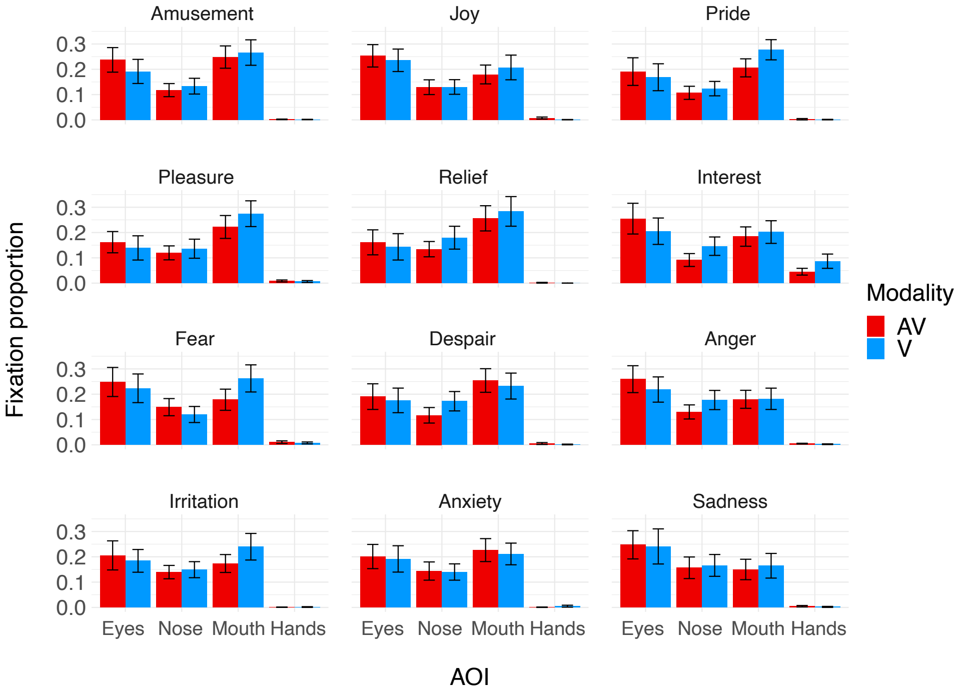


Figure 7. Fixation proportions for correct trials on all areas of interest (AOIs) averaged over the analyzed time course, averaged over participants. The panels show fixation proportions for different emotions. The error bars denote the standard error of the mean (SEM). See Fig. S7 for fixation proportions across the analyzed time course for all emotions.

Table 3. Contrasts for the emotion-by-AOI (area of interest) interaction. The table shows the model estimate difference for the contrasts, with False Discovery Rate (FDR)-adjusted p-values in parentheses. A positive contrast means the first AOI was fixated more than the second of the comparison (and v.v.). Significant differences are indicated in ***bold****.*

	Contrast		
	Eyes – Mouth	**Eyes – Nose**	**Mouth – Nose**
Positive valence, high arousal			
Amusement	**−0.09 (<0.001)**	**0.10 (<0.001)**	**0.20 (<0.001)**
Joy	0.04 (0.13)	**0.12 (<0.001)**	**0.07 (0.01)**
Pride	**−0.18 (<0.001)**	0.03 (0.32)	**0.21 (<0.001)**
Positive valence, low arousal			
Pleasure	**−0.17 (<0.001)**	0.03 (0.22)	**0.20 (<0.001)**
Relief	**−0.20 (<0.001)**	−0.03 (0.25)	**0.17 (<0.001)**
Interest	0.02 (0.55)	**0.12 (<0.001)**	**0.11 (<0.001)**

Negative valence, high arousal			
Fear	−0.04 (0.17)	**0.09 (0.003)**	**0.12 (<0.001)**
Despair	**−0.12 (0.001)**	0.01 (0.74)	**0.14 (<0.001)**
Anger	**0.07 (0.01)**	**0.11 (<0.001)**	0.04 (0.09)
Negative valence, low arousal			
Irritation	**−0.07 (0.048)**	0.04 (0.13)	**0.10 (<0.001)**
Anxiety	−0.02 (0.42)	**0.09 (0.001)**	**0.11 (<0.001)**
Sadness	0.005 (0.89)	0.05 (0.19)	0.05 (0.20)

2.4 Discussion

The present study examined whether observers flexibly adapt their viewing behavior to the presence of audio during the recognition of videos of emotional expressions. We measured audiovisual integration by examining participants' eye movements and emotion identification performance while they viewed video recordings of dynamic emotion expressions with or without the corresponding audio. Our main finding is that there is evidence for integration of auditory and visual information when observers recognize emotions, evident from adapted viewing behavior in response to the changes in modality availability. This adaptation in viewing behavior was present even though there was no evidence for supra-additive integration, as derived from task performance. Moreover, adding audio to the video signal changed observers' viewing behavior, even when the addition of audio did not result in any improvement in identification performance. This implies that auditory signals are used in emotion perception for communication when they are present, and when they are not present people cope well by extracting auditory emotional cues visually, for example by observing mouth movements. Together, our results suggest observers flexibly shape their perceptual strategies based on the audiovisual information available.

2.4.1 Sub-additivity of Audio and Visual Information during Emotion Recognition with Multimodal Stimuli

Firstly, we asked whether our participants would integrate auditory and visual information when performing the emotion identification task or whether visual information alone would mostly be sufficient. When averaged over emotions, task performance was significantly higher in the AV modality than in either of the unimodal modalities, indicating audiovisual integration took place. These findings are in line with studies that compared only two basic emotion categories and used static visual face stimuli combined either with a spoken word[49] or a spoken neutral sentence[48], and with a study that compared all six basic emotions and used short audiovisual videos[62].

These previous studies combined show that emotion recognition improved when information from more than one modality is available, provided the multimodal information is congruent (as was the case in our study). However, we found that the audiovisual integration effect was not particularly strong; performance in the AV modality did not exceed performance gain associated with statistical facilitation as is predicted by 'supra-additivity'[78,79] and was even sub-additive. Our data thus show that audiovisual integration took place, but yet led to a smaller gain in performance than would occur if the auditory and visual evidence would be summed.

However, while some researchers suggested supra-additivity to be the hallmark of multisensory integration[80], originating from the pioneering single-cell electrophysiology of the cat superior colliculus[81], others have argued that many multisensory behaviors do not rely on supra-additivity when the presented stimuli are not close to detection threshold[82,83]. Since we used stimuli with very rich visual and auditory cues, and also in ideal listening and viewing conditions with no distortions, performance in unimodal conditions was already relatively high. As the inverse effectiveness rule states: the strength of multimodal integration is inversely related to the effectiveness of the unimodal stimuli[80]. Therefore, it remains a possibility that the AV integration effect was sub-additive for the specific study conducted here; however, if unimodal performance were lower (i.e., closer to chance level), for example due to decreased auditory and visual signals, the integration effect could be stronger and perhaps become supra-additive.

2.4.2 Multimodal Viewing Does not Always Facilitate Emotion Recognition

In addition to the overall effect of an improvement in performance in the AV modality, we analyzed task performance per emotion. We expected more visual dominance for the basic emotion categories included in the used stimulus set (joy, sadness, fear, and anger) and more integration for the fine-grained emotions (e.g., irritation, despair). Our behavioral data indicated that audiovisual integration – i.e. performance in AV being different from performance in the V-only and A-only conditions – did not occur for all emotions. We found that for many emotions, performance did not differ between AV and V-only, while performance for A-only was mostly lower than in both AV and V-only. Therefore, our behavioral findings would suggest decisions were made primarily on the basis of the visual information and contribution from auditory information was limited. Unlike our expectation, visual dominance was present not only in the basic emotion categories we included, but also in many of the fine-grained emotion categories. The only exceptions were three low-arousal emotions: pleasure, relief, and anxiety (see Table 2). Hence, at least for some fine-grained emotions, combining auditory and visual information increased performance. However, AV performance was never supra-additive for the included emotions.

Our data show a similar pattern to the validation by Bänziger et al.[36] of the stimulus set that was used in the present study. Although these authors did not make all the comparisons we made (in Table 2), their Table 5 (core set rating, 12 repeated emotions only) similarly hints toward visual dominance for many emotion categories investigated. Audiovisual integration, when measured by task performance, thus seems to be the exception, rather than the rule. This is again likely related to inverse effectiveness; performance in the video-only modality was generally higher than performance in the audio-only modality, and hence, the visual dominance observed here could be due to differences in information reliability of this specific stimulus set. This idea is strengthened by findings from Collignon et al.[55], who found visual dominance when audiovisual emotion stimuli were presented without any noise, but evidence for audiovisual integration when they added noise in the visual modality, thus decreasing the reliability of the visual information.

Unreliability of the audio information for emotion recognition may be inherent to this modality. There may be less clear prototypical expressions of specific emotions in the audio (e.g., laughter, crying) than there are in the video (e.g., smile, frown). This could lead to lower reliability for the auditory compared to the visual modality and consequently result in visual dominance. This could explain why visual dominance is commonly found in experiments employing dynamic face/voice stimuli[63,65,67] or dynamic body/voice stimuli[64]. Alternatively, low reliability of auditory information may be inherent to the stimulus material, for example because the use of non-words makes the auditory cues less salient and thus less reliable, which may explain the discrepancy between our data and some other studies[48,49] that did find behavioral evidence for an effect of adding audio to a visual stimulus. However, it should be noted that these studies used a static visual stimulus, which may have decreased its salience and/or reliability and consequently increased the utilization of auditory information. Lastly, some methodological decisions may have affected our participants' ability to integrate the audio with the video; the intensity level of all audio recordings was RMS-equalized, which can take away some of the loudness cues related to emotions that occur in everyday life (e.g., a sad expression is generally quieter than an angry expression). Additionally, the audio was presented over headphones and not via a speaker, which could lead to some spatial disparity between the auditory and visual cues. Although, in principle, audiovisual temporal synchrony should be a stronger cue than the spatial co-location, we cannot exclude if participants experienced spatial disparity and therefore focused less on the auditory cues.

It should be noted that audiovisual integration and visual dominance are not necessarily mutually exclusive. While visual information alone might be sufficient for recognizing emotions, the addition of auditory information could still provide more evidence and allow for faster emotion identification, while accuracy rate remains the same. In complex real-life situations with many interfering audiovisual signals, such added evidence may play a more important role than in the ideal conditions of lab testing. Investigating response times or other measures of cognitive processing could therefore be a beneficial addition. However, for our study, the stimuli used were relatively long and participants were only able to respond after the stimulus ended, and therefore, there is a strong possibility that participants already decided on their answer before the stimulus ended. All these factors, if not controlled for, could make response times unreliable. There may be another method to explore this option, namely, in situations where audiovisual integration may become more important, such as under compromised conditions (e.g., noisy audio or blurred video). A decrease in the reliability of the information in one modality could increase the need for integration, possibly leading to supra-additive integration effects on performance. Furthermore, this could clarify whether there is more visual or more auditory dominance, or whether both channels of information equally contribute to an integrated percept.

2.4.3 There Is No General Tendency to Focus on the Eyes when Recognizing Emotions

In contradiction to popular belief, we did not find a general tendency to fixate on the eyes. There are indications that especially for the recognition of more complex emotions, the eyes are most informative[84]. This view is supported by an ERP study that indicated the eyes as the starting point of emotion recognition. They found that the integration of facial emotional information starts at the eyes, then moves downward across the face, and stops when enough information is integrated to classify an expression[85]. Additionally, it has been shown in both healthy observers as well as in observers with Autism Spectrum Disorders that increased gaze duration to the eyes is correlated with higher emotion recognition performance[86,87]. On the other hand, some eye-tracking studies have indicated that (Western Caucasian) observers distribute fixations evenly across the face[6], whereas other studies have shown observers mostly fixate the areas that are diagnostic for specific emotions (e.g., more fixations on the mouth for happy images and more fixations on the eyes for angry images[88]). Lastly, there is evidence that fixation patterns are perhaps not only specific for different emotions, but also shift when a stimulus is dynamic. Blais and colleagues found that observers more or less equally sampled the eyes and mouth when stimuli were static, but fixated mostly on the center of the face when stimuli were dynamic[89].

Be that as it may, none of these studies used audiovisual stimuli as was done here and the use of audiovisual stimuli seems to greatly impact where an observer will look. Here, we did not find a general tendency to fixate on the eyes, nor on the center of the face. Additionally, as can be seen from Fig. 7, in line with previous studies our participants fixated mostly on the eyes for Anger stimuli[58,90], but in contradiction to previous studies they did not mostly fixate on the mouth for Joy stimuli (the close equivalent to the basic emotion happiness used in other studies), but instead sampled the eyes and mouth equally often. For the majority of other emotions, the mouth was fixated most often, followed by the eyes and nose.

2.4.4 Gaze Behavior During Emotion Perception for Communication Does Not Simply Reflect Visual Saliency

Contrary to our behavioral data, our fixation data suggest clear usage of audio information and thus indicate there is at least an interaction between auditory and visual information. When averaged over emotions, observers viewed the mouth less and the eyes more in the AV modality compared to the V-only modality; there was also an increase in fixations on the nose. These findings suggest observers flexibly adapt viewing behavior to fixate regions that they feel would maximize performance depending on whether audio is present or absent. There are several studies that give indications on why the increased fixations on the nose and mouth in the V-only modality might be beneficial for performance when audio is lacking.

First, the nose has been proposed to be an optimal fixation landmark for global face perception, at least for static facial images[91,92]. After all, from this vantage point, it is possible to both rapidly direct the gaze to either the eyes or the mouth, as these regions are more or less equidistant from the nose. Moreover, it may also be possible to simultaneously gather (crude) visual information from both the eyes and the mouth using lower resolution peripheral vision[93]. Additionally, biological motion can be processed well in the periphery[94], making fixating on the nose a good strategy if one wishes to retrieve dynamic information from both the eyes and the mouth. It is a fair assumption that in the V-only modality, participants tried to gather as much visual information as possible to compensate for the lack of audio signal, and therefore fixated more on the nose in order to also access visual information from both the eyes and mouth.

Second, increasing the proportion of fixations on the mouth could then serve to gather more fine-grained visual emotional information. Such an increase in fixations on the mouth is not commonly reported in the literature and whether or not it is found seems to depend on the task participants performed. For example, while Lansing and McConkie[47] also found an increase in mouth fixations in the V-only modality, Võ and colleagues[46] found a decrease in mouth fixations when sound was muted. However, while their stimuli were similar, their experimental tasks were rather different: both featured videos of people speaking (only face, neck, and shoulders visible) but in the study by Võ *et al.* participants had to rate the likeability of the video, while in the Lansing and McConkie study participants performed a speech identification task. These and our own findings indicate eye gaze reveals how the perceptual strategies flexibly adapt to the available information and the nature of the specific task.

The modality and task dependency of eye gaze indicate that gaze is not simply dictated by visual saliency. If it were, one would expect to always find most fixations on the mouth in dynamic face stimuli. Mouth movements are quite large and thus more salient compared to those of other facial features. Moreover, an increase in fixations on the eyes in the AV compared to the V-only condition is not expected either, as the visual stimulus did not change. Our findings, as those of others[46,47], therefore indicate that gaze is guided by an information-seeking process. Moreover, that V-only performance exceeded A-only performance for most emotions, suggests the visual information provided by the mouth can be a vital substitute for the missing auditory information.

2.4.5 Perceptual Strategies Suggest Auditory Rather than Visual Dominance

While task performance could be taken to indicate visual dominance for many emotions, there is no compelling evidence for visual dominance to be found in the viewing behavior. Although for many emotions, no significant difference in accuracy was found between AV and V-only, there was a clear effect on viewing behavior when adding audio to the video.

Our data suggest that viewing behavior, and by extension the manner in which the task is performed, adapts as a function of both the available information and by the degree to which the information is task-relevant.

That participants' viewing behavior changed depending on the presence of audio, is indicative that gaze is not only guided by the visual information, which remained the same in the AV and V-only modalities, but also by the presence of auditory information. One might therefore even argue for auditory dominance instead of visual dominance for emotion perception. Evidently, when there is audio, one uses it and adapts viewing behavior accordingly, perhaps because some areas do not have to be fixated anymore to obtain the information present in those areas. As an example, the movements of the mouth can provide cues of the expressed emotion, but the audio (produced by those same mouth movements) likely provides the same cues (as well as some unique information), resulting in redundancy in information across the two modalities. It is therefore no longer necessary to look at the mouth when audio is present and one is free to look for cues of the expressed emotion elsewhere, the eyes perhaps. That this adaptation does not always result in improved task performance could be because there simply is not more information in the visuals, wherever one looks. Our present study cannot yet fully confirm or reject whether emotion perception is guided preferentially and perhaps even compulsory by auditory information. To test this idea, one would have to see changes in behavior, be it viewing behavior or otherwise, in the presence of any audio – e.g. noise – compared to the absence of audio. Regardless, our data show that even when audiovisual integration is not apparent from the task performance, from the adaptations in viewing behavior it is clear the two modalities are integrated and shape the decision-making process. This study thus also underlines the need for measuring more than just task performance if one wishes to draw conclusions on audiovisual integration in emotion perception.

2.4.6 Limitations and Future Directions

Due to the many comparisons made in Tables 2 and 3 and the corrections therefore applied to the significance values, the comparisons might be underpowered. Future studies can be designed based on the knowledge produced in this study, where a subset of stimuli or conditions could be selected, producing fewer comparisons, or alternatively use a larger sample size, and better statistical power. Additionally, future studies should explore the integration process further by not only manipulating modality availability, but also manipulating information availability within modalities, for example by blurring (parts of) the image or using speech-shaped noise instead of actual emotional speech. Using stimuli specifically designed for it, measuring response times could also be a good addition, to further explore potential AV integration effects, in addition to accuracy performance. Lastly, though it would decrease the ecological validity of the stimuli, future studies could consider the use of (dynamic) incongruent audiovisual stimuli, possibly with differing reliabilities of the audio and video, to explore whether a continuum from visual to auditory dominance exists.

Our fixation data suggest that while the majority of fixations made were directed to our AOIs, a large part of the fixations was elsewhere on the screen. It can be inferred from Figs 6 and 7 that the fixations captured for each condition add up to roughly half of fixations made, although there are quite large individual differences (see Figs S3 and S4). This could indicate our participants either had an interest also for other areas of the screen, which we would have seen as a clustering of these outside AOI fixations on specific regions, such as the abdomen of the actor, or decided to browse around the screen more, which would be evident by fixations dispersed over the screen. However, an inspection of this with heat maps (see Fig. S6 for fixation heat maps for all modalities) showed that actually most fixations were indeed directed toward the face, with only a minority of the fixations directed elsewhere, mainly on the body and toward the hands. It therefore seems more likely that participants relatively often looked just outside the AOIs. Additionally, it is peculiar that only few of the fixations made were directed toward the hands of the actors, despite our expectation that observers would use the information that can be gathered from hand gestures. Speculating, it is very well possible that observers need not fixate on the hands in order to retrieve the information they convey; viewing hand movements with peripheral vision might give enough information to recognize the emotion that is being expressed by the gestures. Future studies could test these hypotheses for example by removing the face, forcing observers to use other information.

We analyzed fixation data over an 800-ms time window for all stimuli (200 ms after the start of the stimulus until 1000 ms after the start, based on the length of the shortest video clip). Because of this, some gaze data was discarded. We chose not to use the full movie as participants may have decided which emotion was being expressed before the end of the movie clip (which is more likely to occur in long movies) and their gaze data after their decision might therefore reflect task-irrelevant viewing behavior. Nevertheless, we find that the pattern of results does not change if we take the full movie into account (see Fig. S8 for a comparison of average fixation proportions for the full movie and the used time window), confirming that the choice to use an 800-ms time window was an appropriate one.

It should be noted that some noise was present in the audio of the original stimulus materials. While careful consideration had been taken to remove this noise from the original stimulus materials, some noise may have been left which could have made the audio less reliable and may have biased performance to visual dominance. However, since our fixation data argue against visual dominance, it seems unlikely that any potentially remaining noise after pre-processing the audio substantially affected task performance.

Finally, it can be argued that the visual information in the stimuli contained two distinct cues for emotion: facial expressions and body expressions. Since this is not the case in the auditory modality, one could say that in the AV modality, participants had access to three emotion cues (face, body, and voice), in the V-only modality to two emotion cues (face and body), but in the A-only modality to only one emotion cue (voice). Following this line of reasoning, it is thus not surprising that V-only performance was much higher than A-only performance, and that AV and V-only performances did not differ. Future studies should explore this further, for example by comparing performance for face + voice, body + voice, and face + body + voice conditions. In this example, observers always have access to two modalities, but the number of cues – and possibly also the quality of the cues – in the visual modality changes.

2.4.7 Conclusions

For the perception of emotions, observers generally utilize multiple sources of information when these are available. While this was not evident from our behavioral measure of task performance as for many emotions performance on the multimodal task could be quite reliably predicted from performance on the visual task, viewing behavior did change based on information source availability even in the absence of a difference in performance. It can therefore be concluded that people change their perceptual strategies depending on the available information in an attempt to maximize performance. Drawing conclusions about integration of auditory and visual information thus is not only defined by the outcome (i.e., task performance), but also by the process (which can be studied with eye tracking). This study, with the use of dynamic multimodal emotion expressions, has taken a small step toward studying the perception of emotions in an ecologically more valid setting than with simpler materials. Further, it highlights the need for using multiple measures of emotion recognition if one wishes to deduce a comprehensive profile of audiovisual integration in emotion perception.

2.5 Acknowledgements

We gratefully acknowledge Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer for permission for using the GEMEP core set and for publishing sample images from the core set. We also thank Birte Gestefeld, Alessandro Grillini, Rijul Soans, Paolo Toffanin, and Anita Wagner for their help with programming and data analysis.

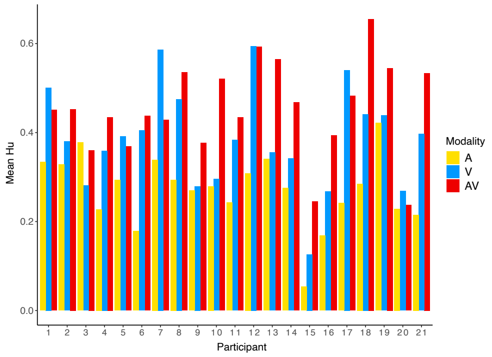
2.6 Funding

The first author was supported by a BCN-BRAIN grant from The Graduate School of Medical Sciences (GSMS), University of Groningen, the Netherlands. This project was supported by the following foundation: the Landelijke Stichting voor Blinden en Slechtzienden that contributed through UitZicht (Grant number: Uitzicht 2014 – 28 – Pilot). The funding organizations had no role in the design or conduct of this research. They provided unrestricted grants.

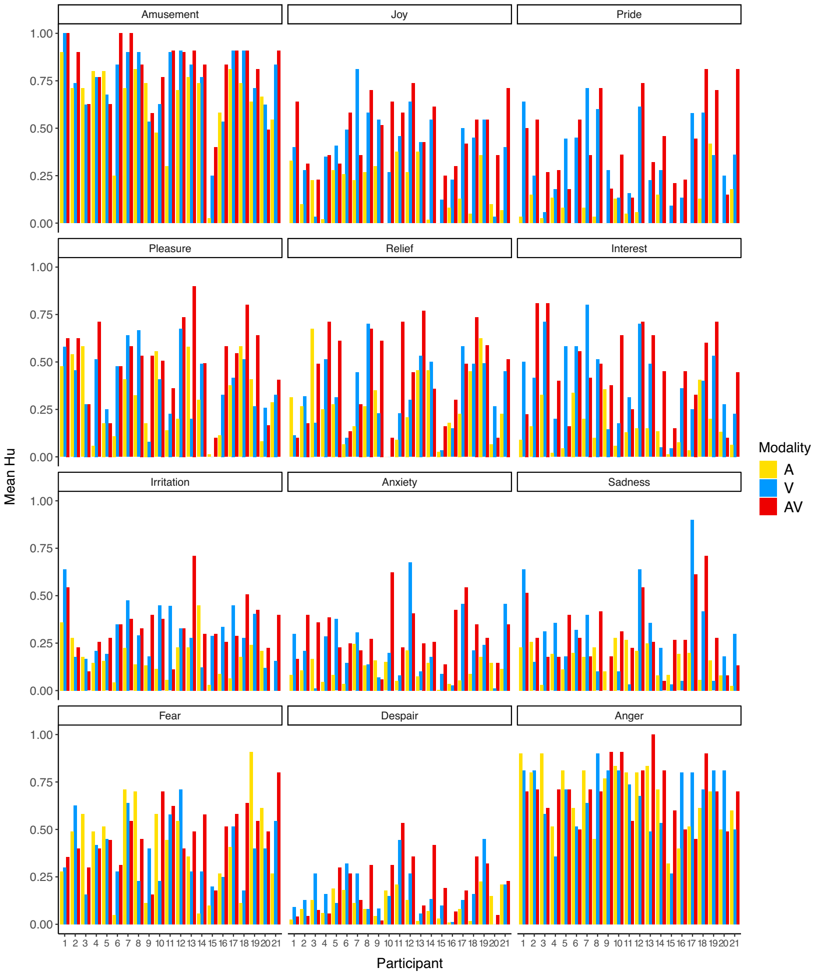
2.7 Data availability

The datasets generated for this study can be found in the DataverseNL repository https://doi.org/10.34894/NXSWFR. All data are publicly available.

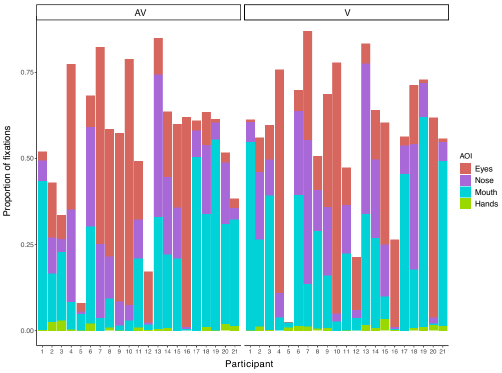
Supplementary Material



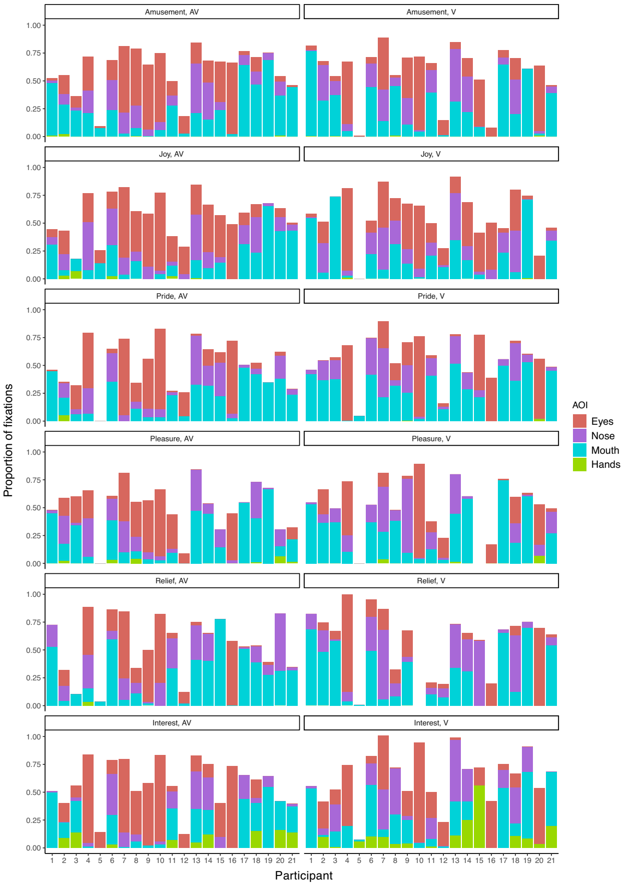
Supplementary Figure S1. Task performance for each modality and participant, shown in unbiased hit-rates (H_u), averaged over trials. Each bar shows an individual participant's mean H_u for a specific modality. A-only performance modality is indicated by yellow bars, V-only performance is indicated by blue bars, and AV performance is indicated by red bars.



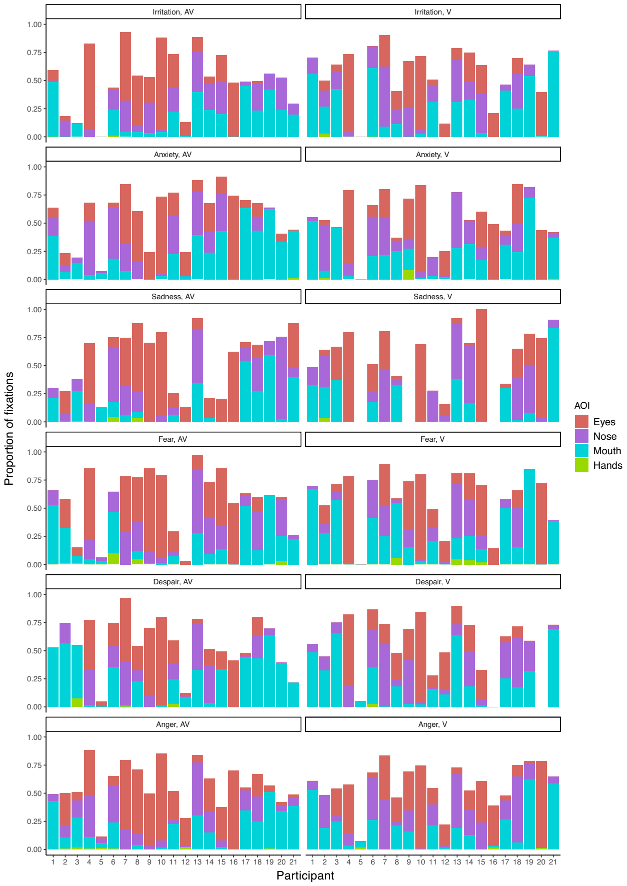
Supplementary Figure S2. Task performance for each modality, emotion, and participant, shown in unbiased hit-rates (H_u), averaged over trials. Each panel shows the result for one of the emotions. Each bar shows an individual participant's mean H_u for a specific modality. A-only performance modality is indicated by yellow bars, V-only performance is indicated by blue bars, and AV performance is indicated by red bars.



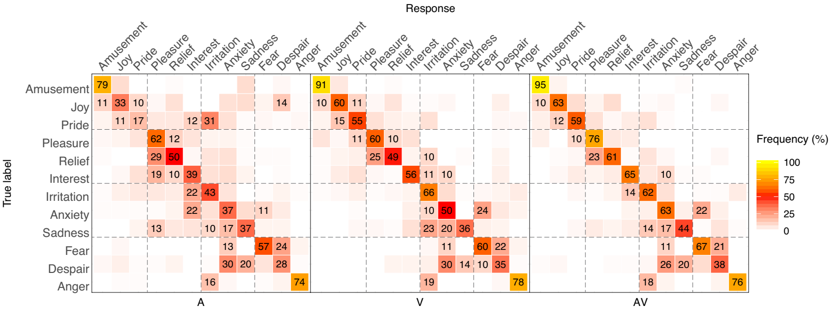
Supplementary Figure S3. Average fixation proportions for each modality and participant, averaged over trials. Only fixation proportions for correct trials are shown. Each panel shows the result for one of the modalities. Each bar shows an individual participant's mean fixation proportions on the different AOIs. Fixations on the eyes are shown in red, fixations on the nose are shown in purple, fixations on the mouth are shown in blue, and fixations on the hands are shown in green.



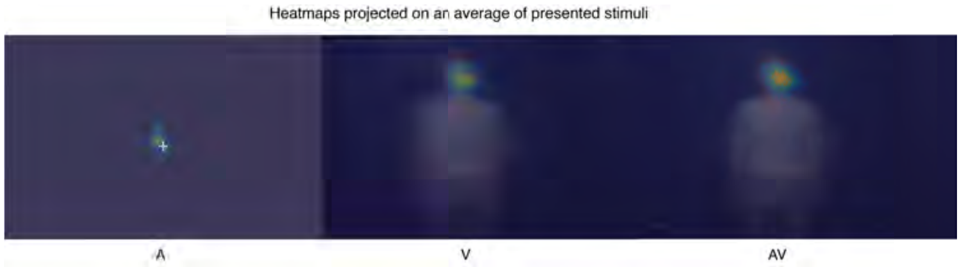
Chapter 2



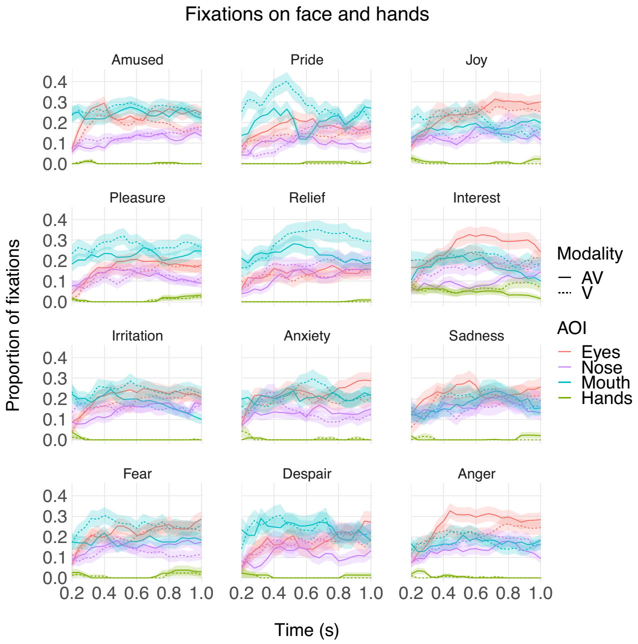
Supplementary Figure S4. Average fixation proportions for each modality, emotion, and participant, averaged over trials. Only fixation proportions for correct trials are shown. Each panel shows the result for one of the modality-emotion combinations. Each bar shows an individual participant's mean fixation proportions on the different AOIs. Fixations on the eyes are shown in red, fixations on the nose are shown in purple, fixations on the mouth are shown in blue, and fixations on the hands are shown in green.



Supplementary Figure S5. Confusion matrices shown for all three modalities. Each row represents the true label for an emotion (i.e., the correct answer), while each column represents the given response. Values on the diagonal indicate correct responses, while values off the diagonal indicate incorrect responses. Values are given as frequency percentages. Only for frequencies > 10% the numeric value is displayed, below 10% only the color is indicative of the frequency.



Supplementary Figure S6. Fixation heat maps shown for all three modalities, averaged over participants and trials. Colors indicate how often a certain area was fixated (yellow: often, dark blue: hardly ever). The heat maps are projected on top of an image averaging the presented stimuli in each modality (i.e., AV and V-only: an average of all movies, A-only: a fixation cross).



Supplementary Figure S7. Fixation patterns for correct trials on all AOIs across the analyzed time-course, averaged over participants. The panels show fixation proportions for different emotions. Shaded areas around each line denote the standard error of the mean (SEM).

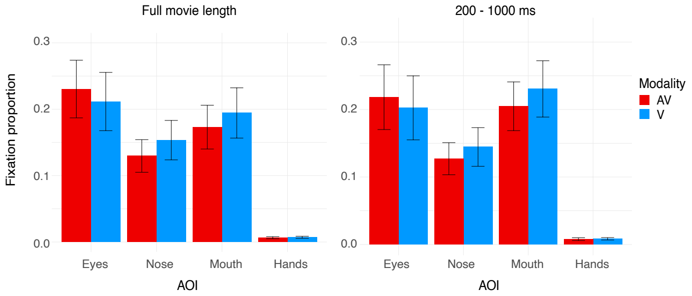
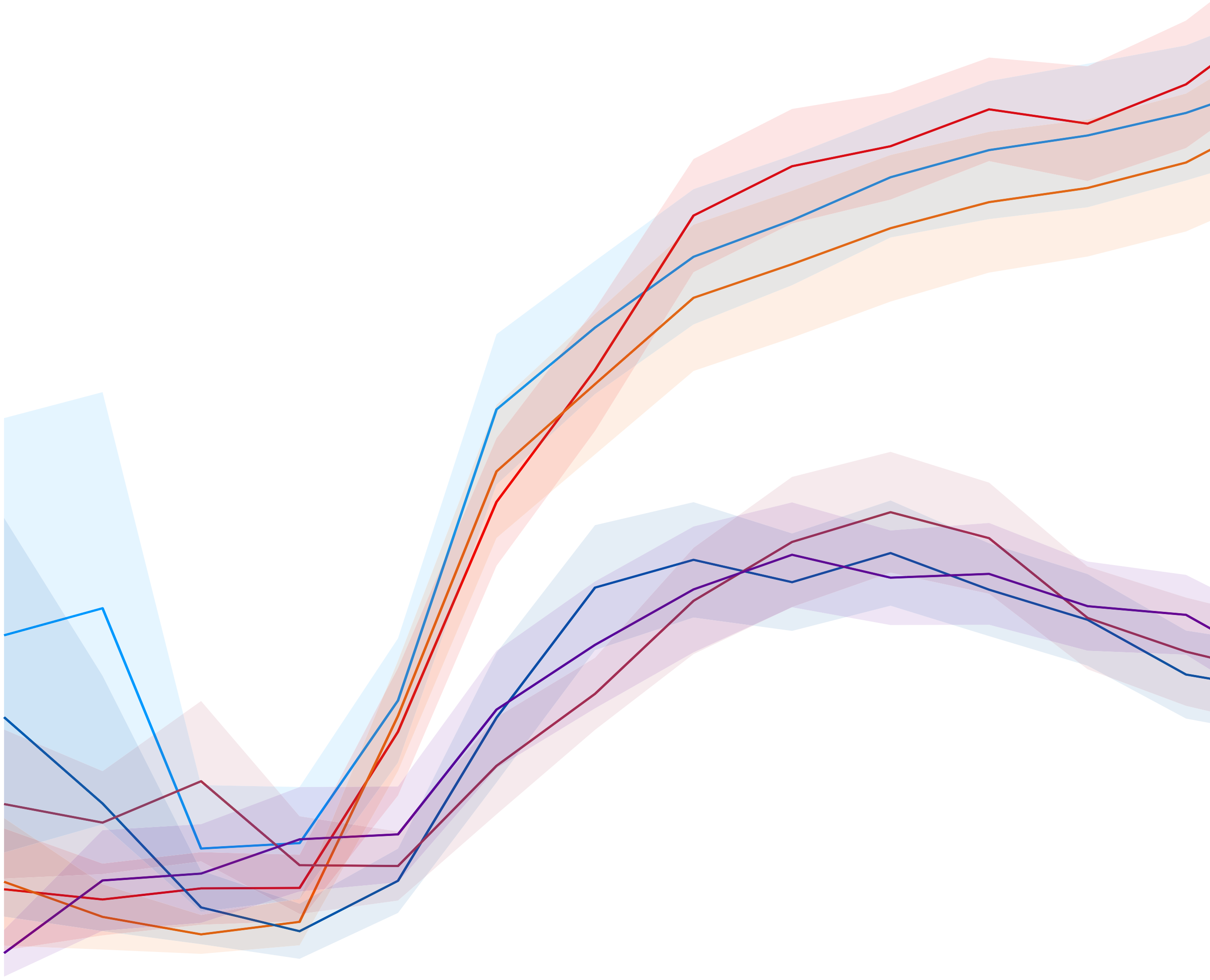


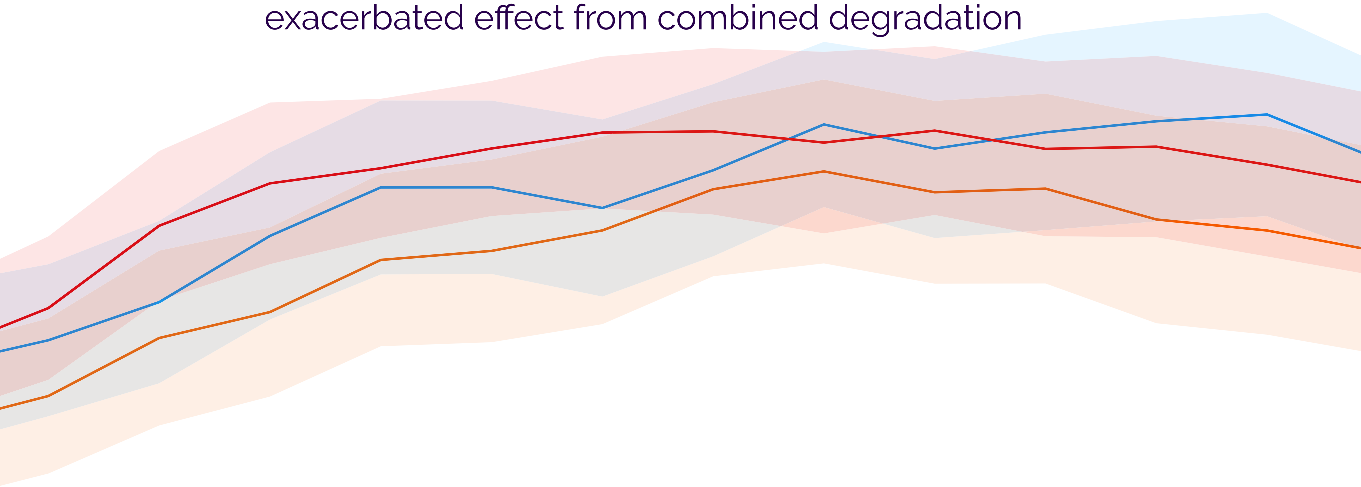
Figure S8. Fixation proportions for correct trials on all AOIs (eyes, nose, mouth, and hands) averaged over: the full movie length (left) and the analyzed time-window of 200 – 1000 ms (right), both averaged over stimuli and participants. Error bars denote the SEM.

Author affiliations:
1. Research School of Behavioural and Cognitive Neuroscience (BCN), University of Groningen, Groningen, the Netherlands
2. Laboratory of Experimental Ophthalmology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
3. Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
4. Institute of Acoustics, Technische Hochschule Lübeck, Lübeck, Germany



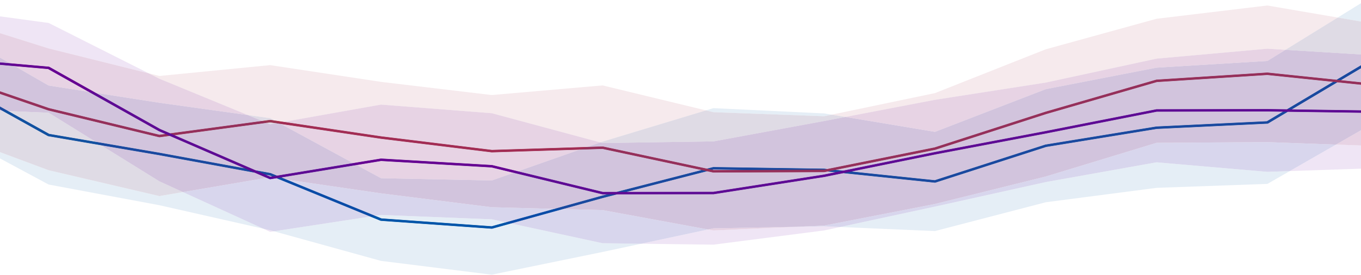
Chapter 3

Degraded visual and auditory input individually impair audiovisual emotion recognition from speech-like stimuli, but no evidence for an exacerbated effect from combined degradation



2021. *Vision Research*, 180, 51-62

Minke J. de Boer[1,2,3], Tim Jürgens[4], Frans W. Cornelissen[1,2], Deniz Başkent[1,3]



Abstract

Emotion recognition requires optimal integration of the multisensory signals from vision and hearing. A sensory loss in either or both modalities can lead to changes in integration and related perceptual strategies. To investigate potential acute effects of combined impairments due to sensory information loss only, we degraded the visual and auditory information in audiovisual video-recordings, and presented these to a group of healthy young volunteers. These degradations intended to approximate some aspects of vision and hearing impairment in simulation. Other aspects, related to advanced age, potential health issues, but also long-term adaptation and cognitive compensation strategies, were not included in the simulations. Besides accuracy of emotion recognition, eye movements were recorded to capture perceptual strategies. Our data show that emotion recognition performance decreases when degraded visual and auditory information are presented in isolation, but simultaneously degrading both modalities does not exacerbate these isolated effects. Moreover, degrading the visual information strongly impacts recognition performance and viewing behavior. In contrast, degrading auditory information alongside normal or degraded video had little (additional) effect on performance or gaze.

Nevertheless, our results hold promise for visually impaired individuals, because the addition of any audio to any video greatly facilitates performance, even though adding audio does not completely compensate for the negative effects of video degradation. Additionally, observers modified their viewing behavior to degraded video in order to maximize their performance. Therefore, optimizing the hearing of visually impaired individuals and teaching them such optimized viewing behavior could be worthwhile endeavors for improving emotion recognition.

Keywords: emotion perception, eye-tracking, central scotoma, age-related hearing loss, audiovisual, dynamic

3.1 Introduction

The perception of another persons' emotional intent is an essential element in human communication. Normally, communication takes place face-to-face, making emotions multimodal and dynamic in nature. Because of this multimodal nature of emotions, proper auditory and visual functioning is required to correctly recognize others' emotions. Currently, it is unknown how effects of vision and hearing loss on emotion perception interact with each other.

With the ageing population, the prevalence of sensory impairments is rising. Difficulties in communication are one of the major problems these individuals face, especially in those impaired in both hearing and vision. For example, it has been shown that individuals with hearing loss exhibit a reduced range in rating non-speech emotional sounds for both valence and arousal compared to hearing controls[95]. The valence and arousal levels of sounds can affect mood, induce or reduce stress[96,97], and the degree to which sounds attracts attention[98]. Consequently, a reduction in the perceived range of valence and arousal levels could negatively affect hearing impaired listeners' emotional responses to sounds. In line with this, in cochlear implant users, vocal emotion recognition accuracy is correlated with quality of life[99].

Multisensory perception studies indicate that observers integrate information in an optimal manner, by weighing unimodal sources based on their reliability prior to linearly combining them. Because of this optimality, multimodal integration is largest when the reliability of the unimodal sources is similar and each provides unique information[10,18,43]. Normally, vision more reliably encodes information in the spatial domain while hearing is better suited towards encoding information in the temporal domain. Yet, despite this specialization, the senses do not uniquely encode this information. For this reason, damage to a sensory organ may affect all of its information encoding, or primarily affect the domain it is specialized for. Consequently, having both vision and hearing loss may have unpredictable consequences. It may either exacerbate the overall effects of the impairments, or, alternatively, domain-specific information necessary for task performance may still be obtained via the other, non-specialized channel.

While studies have been performed that investigate the effects of vision and hearing loss on emotion perception, these were mostly in populations with either only a vision loss or a hearing loss, but not both together. Despite this, results of these studies can still inform about the possible effects that combined vision and hearing loss may have. For example, in age-related macular degeneration (AMD), a common form of vision impairment, it has been shown that visual emotion perception is impaired, although the results are not always consistent. AMD affects up to twenty percent of the elderly population[100] and generally leads to a scotoma (i.e., a region of reduced light sensitivity) in central vision due to a deterioration of the macula. Because of the disease's effect on central vision, it seems likely that AMD would affect emotion recognition, as recognizing most facial expressions requires detecting small, detailed movements[4]. Indeed, as an indirect support of this expectation, face identification is impaired in patients with AMD and their performance is positively correlated with their visual acuity and contrast sensitivity[101], which are both reduced in AMD. Moreover, AMD patients performed near normal levels for facial emotion categorization (i.e., categorize a facial expression as happy, angry, or neutral), but performed much worse when having to decide whether a face was expressive or not[102]. Additionally, Johnson et al.[103] found that eye movements in AMD patients were more randomly distributed over the face, compared to controls, which typically show a T-shape pattern of fixations around the eye and mouth regions.

In the auditory domain, there is some debate on whether hearing loss affects auditory emotion recognition or whether existing results are related to hearing loss per se or to ageing or cognitive decline in addition to hearing loss. Acoustic cues for auditory emotion recognition are mainly conveyed by prosodic features of speech, such as contours of fundamental frequency and its related harmonic structures[104]. To properly perceive these cues, usable hearing in the low frequency range, up to 750 Hz, is necessary[105]. Older individuals with hearing loss generally have hearing loss at higher frequencies, with reasonably preserved hearing at lower frequencies. Therefore, they may recognize acoustic cues related to emotions despite their hearing loss. However, despite preserved hearing in the frequency range required for perceiving acoustic emotion cues, hearing loss, especially at moderate and severe levels, can affect abilities for frequency discrimination and resolution, and temporal resolution. These are all necessary to accurately perceive acoustic cues related to emotional information[106]. Fully in line with this, studies show that both adults and children with hearing loss perform worse in auditory emotion recognition[107,108]. Additionally, Most and Aviner[107] found a lack of performance increase in audiovisual presentation of emotion stimuli compared to visual presentation of emotion stimuli in the children with hearing loss, while this increase was present in the children with normal hearing. This indicates that the children with hearing loss could not adequately use the auditory information present in the audiovisual stimulus. However, the findings in children with hearing loss may be strongly confounded by differences in their development of emotion perception, which is likely also affected by hearing loss and the age at which children receive hearing aids or cochlear implants[109]. The use of hearing aids in older adults seems to slightly increase their emotion recognition performance, but does not fully restore it to the levels of normal hearing older or younger listeners[110].

Consequently, it remains unclear whether existing findings in individuals with unimodal sensory impairments are due to the missing sensory input, i.e., an acute effect, or a general ageing effect, or cognitive impairments brought about by ageing or the sensory impairments, i.e., long-term effects. For example, a study by Orbelo et al.[111] found that impaired vocal emotion recognition in elderly participants with very mild hearing loss was not predicted by their hearing loss, nor by age-related cognitive decline. Their results are indicative that effects found in individuals with hearing loss may be related to general ageing instead of their sensory impairments per se and this may also apply to vision loss. However, in this specific study, with pure-tone hearing thresholds of on average 24 dB HL (± 12 dB), it may be that the hearing loss in the elderly participants was too mild to have a measurable impact on their performance, making it hard to draw definitive conclusions. Furthermore, existing findings do not provide clear predictions on the effects of multimodal sensory impairments.

Therefore, the current study was focused on possible acute effects of sensory impairments on emotion recognition. To additionally be able to investigate the effect of combined impairments across modalities, the present study used modifications of the video and audio signals of movies to degrade visual and auditory information presented to a healthy group of young volunteers. These degradations intended to approximate some aspects of vision and hearing impairment, in simulation. The use of such simulations creates a homogeneous and otherwise healthy fictitious "patient" group, while recruiting healthy young participants ensures that any effects of (simulated) hearing and vision loss will not be due to ageing or cognitive decline. This allows measuring the possible acute effects of sensory impairments while any long-term adaptation that may occur in real sensory impairments is excluded.

In the current study, we degraded the information in such a way to mimic a relative central scotoma in the visual domain and a degradation similar to age-related sensorineural hearing loss in the auditory domain. Because we wanted our visual degradation to be close to the visual experience of AMD individuals, we chose a relative central scotoma, which still provides some visual information, as most AMD individuals are not fully blind in their scotomatic region. Instead, AMD individuals most often experience blurred or hazy vision, followed by distortions, such as straight lines looking crooked[112]. The addition of a moderate level of age-related sensorineural hearing loss creates a hypothetical "typical" elderly AMD individual, as hearing loss is common in the elderly population[113].

In addition to affecting emotion recognition ability, it can be expected that vision and hearing loss change the way in which emotions are perceived and processed. This can be quantified by examining differences in eye movements for individuals with and without vision/hearing loss. Gaze allocation is proposed to be a functional information-seeking process[45,46,114]. Therefore, it can be expected that gaze adapts to the changes in information due to degraded visual and auditory signals. For example, observers generally increase fixation duration as task difficulty increases[115]. Additionally, studies have shown that AMD patients typically develop a preferred retinal locus (PRL)[116,117], a peripheral retinal location that patients use for fixation when the fovea is no longer functional. The PRL is generally located near the border of their scotoma[118,119]. While the location of the PRL could just be determined by spontaneous reorganization in the primary visual cortex, it could also be functional; the closer the PRL is to the original fovea, the higher the visual acuity in that region will be.

In our present study, the acute effects of visual and auditory degradation were tested, using videos that depict different emotions. First, we tested for the "pure" effects of degradation by degrading visual or auditory information while at the same time removing the audio or video, to ensure no cross-modal compensation is possible. In addition, degradation effects were tested both individually and in combination, by degrading only the visual or auditory information and leaving the other modality intact, as well as by simultaneously degrading both the visual and auditory information. By doing this, we could test the possible effects of the degradations in situations where cross-modal compensation is and is not possible. Because observers without sensory impairments seem to rely mostly on visual information in emotion recognition in audiovisual presentation of videos[55,64], we expected that auditory degradation would minimally, or perhaps even not, impact recognition abilities when proper visual information was present. Likewise, it may be expected that visual degradation will impact performance more and possibly increase reliance on the auditory information. Moreover, we expected that combined visual and auditory degradation would impact performance more than only visual degradation, as in this situation an increased reliance on the auditory information provides less benefit. Besides assessing emotion recognition performance, viewing behavior was examined by measuring eye-movements made during stimulus presentation, in an attempt to capture changes in viewing strategies as a result of degraded modalities. Because degradation of information will surely increase emotion recognition difficulty, and higher task difficulty has been shown to increase fixation durations[115], it seems likely that observers will fixate longer under degraded viewing/listening conditions. Increases in fixation duration because of a simulated scotoma have already been found in visual search tasks[120,121]. Furthermore, Cornelissen and colleagues[121] found an increase in saccadic amplitude with a simulated central scotoma, but only when the scotoma was absolute (i.e., complete disappearance of visual input within the scotoma), and not when it was relative (i.e., low contrasts within the scotoma region). Based on this, we expected that fixation durations would be longer under degraded conditions, but that there would be no effect on saccadic amplitude, as the visual impairment simulated in the current study is a relative central scotoma. In addition, we expected that healthy observers would fixate in such a way that the observer's area-of-interest is just outside the border of their artificial scotoma, provided they have at least somewhat adapted to the scotoma. Thus, if the observer would be trying to view someone's face, they would position the scotoma such that the face is adjacent to the scotoma border.

3.2 Methods

The stimuli and methods used in this study are directly based on and modified from previous studies by the authors and by the creators of the stimulus materials[36,114]. In the previous study by de Boer et al. emotion recognition performance and gaze behavior were studied in young, healthy observers that viewed the stimuli audiovisually, only the video, or only the audio. No signal degradation was used in the previous study.

3.2.1 Participants

Twenty-four healthy, native Dutch participants volunteered to take part in the experiment (nine male, mean age = 23 years, *SD* = 2.9, range: 19-29). All participants were given ample information about the nature of the experiment, but were otherwise naïve as to the purpose of the study. Written informed consent was obtained prior to screening and data collection. The study was carried out in accordance to the Declaration of Helsinki and was approved by the local medical ethics committee (ABR nr: NL60379.042.17). Participants received a payment of €8,00 per hour for their participation in accord with departmental guidelines.

3.2.2 Screening

Prior to the experiment, all participants' eyesight and hearing were tested to ensure (corrected) visual and auditory functioning was within the normal range. Normal visual functioning was tested with measurements of visual acuity and contrast sensitivity (CS). Tests were performed using the Freiburg Acuity and Visual Contrast Test (FrACT, version 3.9.8)[68,69]. For inclusion in the experiment, participants needed a visual acuity of at least 1.00 and a logCS of at least 1.80 (corresponding to a luminance difference of approximately 1% between target and surround). Visual tests were performed binocularly and on the same computer and screen as used in the main experiment. Auditory functioning was tested by measuring auditory thresholds for pure tones at audiometric test frequencies between 125 Hz and 8 kHz. For inclusion, audiometric thresholds at all test frequencies had to be as good as or better than 20 dB HL at the better ear. The thresholds were determined using a staircase method based on typical clinical procedures. The participant sat inside a soundproof booth during testing. Testing was conducted on each ear, always starting with the right ear. Additional exclusion criteria were neurological or psychiatric disorders, dyslexia, and the use of medication that could influence normal brain functioning.

3.2.3 Stimuli

The stimuli used in the experiment were taken from the Geneva Multimodal Emotion Portrayals (GEMEP) core set[36], a short demo showing only the face of the actor can be found at the Geneva Emotion Recognition Test (GERT) demo at: https://www.unige.ch/cisa/emotional-competence/home/exploring-your-ec/. This set consists of 145 audiovisual video-recordings (mean duration: 2.5 s, range: 1-7 s) of emotional expressions portrayed by ten professional French-speaking Swiss actors (five male). The vocal content of the expressions was one of two pseudo-speech sentences with no semantic content, but resembling the phonetic sounds in western languages ("nekal ibam soud molen!" and "koun se mina lod belam?"). Out of the 17 emotions present in the set, 12 were selected for the main experiment, see Table 1 for all emotions and how they are distributed over the valence-arousal scale[70].

Table 1. The selected emotion categories used in the experiment. The emotions are distributed over the quadrants of the valence-arousal scale[70].

		Valence	
		Negative	**Positive**
Arousal	**High**	Amusement Joy Pride	Fear Despair Anger
	Low	Pleasure Relief Interest	Irritation Anxiety Sadness

The reason for using many emotions was to avoid any ceiling effects that are often found in emotion research[56,122,123], as changes in performance due to the degradations may not be entirely visible if normal performance is close to ceiling. Portrayals from two actors that were found to be less clearly recognizable in our previous work[114] were used as practice material to acquaint participants with the stimulus materials and the task. Thus, this resulted in a total of 96 unique stimuli used in the main experiment and a total of 24 unique stimuli used in practice trials.

3.2.4 Visual stimulus degradation

Custom MATLAB scripts were used to produce a gaze-contingent relative scotoma. A semi-circular shape, centered on gaze position, was used to mimic an approximate vision loss in an individual with progressed binocular AMD, see Figure 1b-c. The simulated scotoma extended roughly 17° horizontally and 11.5° visual angle vertically (731 x 497 pixels) and had soft edges. Since AMD individuals generally do not perceive a hole in the location of their scotoma, but instead perceive distortions or blur, we decided to blur rather than remove the region in the video that was covered by the simulated scotoma. Additionally, because some information still passes through the scotoma for most AMD individuals, we designed the scotoma in a way that would still allow viewing larger hand and body movements. Further, looking more at the hands may be a compensatory strategy that patients use if they can no longer see facial expressions, and with our design, we aimed to capture these strategies.

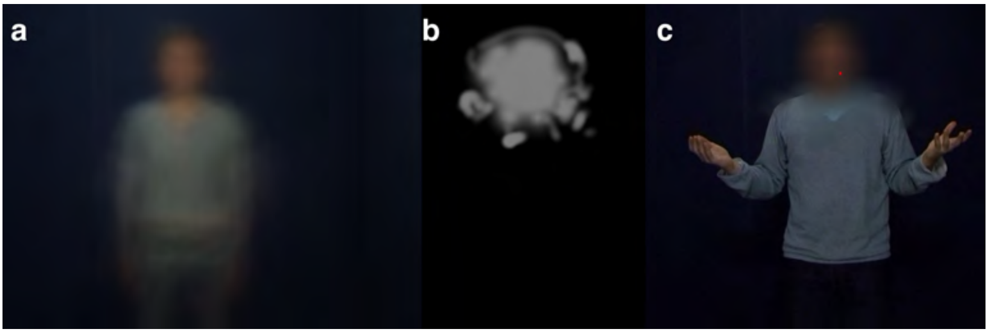


Figure 1. a) Still image created by averaging together all frames of all videos. This image preceded stimulus presentation in all conditions, except in the A and dA conditions. b) Shape of the scotoma mask, drawn approximately to scale. The scotoma was gaze-contingent and the center of the scotoma was positioned on the gaze location. c) Scotoma overlaid on a still image of one video. The scotoma is centered on gaze position, indicated by the red dot. This dot was not visible during the experiment.

A Gaussian low-pass filter (using the MATLAB functions *fspecial* and *imfilter*) with a cut-off (at full width at half maximum, FWHM) of 0.15 cycles/deg was used to create a blurred version of the video. Then, the blurred video was overlaid on the non-blurred video, and the alpha-layer of the scotoma image was used to indicate which region should be blurred and how strongly. Thus, only within the mask the video was blurred, outside the mask the video was not blurred. Four different orientations of the simulated scotoma were created: original (as in Figure 1b), left-right flipped, up-down flipped, and left-right and up-down flipped. Orientation was randomized between trials. While changing the orientation from trial to trial is unlike a real scotoma, this was done to ensure the results would not rely too strongly on the scotoma's shape in a specific orientation, while avoiding a too simplistic simulation. It was found that orientation did not significantly affect recognition performance (F (3, 69) = 0.64, p = 0.589).

Participants were instructed that the scotoma was gaze-contingent and that they could use compensatory eye-movements in order to peripherally look at regions in the video they found interesting or helpful.

3.2.5 Auditory stimulus degradation

The audio signal was degraded in three aspects inspired by three characteristics of sensorineural hearing impairment: increased absolute thresholds, loudness recruitment, and the effects of broader auditory filters on speech envelopes in the auditory system. To implement these degradations the hearing impairment (HI) simulation of Siebe et al.[124] was used, which was inspired by the HI simulation of Nejime and Moore[125]. The degradation consists of two sequential modules: one for sound envelope processing, and one for loudness perception.

The rationale behind the first module, the envelope-processing module, is that envelopes are represented as they are in the impaired auditory system via broader auditory filtering, whereas the fine structure is preserved as in normal hearing. This module processed the input audio signal using a Gammatone filter bank with normal-hearing (NH) bandwidths of one equivalent rectangular bandwidth (ERB) at one ERB spacing of center frequencies between 80 Hz and 10 kHz, and extracts the fine structure using a Hilbert transform. Furthermore, it extracted the Hilbert envelope using a second Gammatone filter bank with one ERB spacing of center frequencies, but with double the bandwidth (i.e., the degraded filters are two ERB wide). This bandwidth was selected to be at the lower edge of the range that was found in hearing impaired (HI) individuals[126]. Hilbert envelopes from broader filters were then multiplied onto Hilbert fine structure signals in each frequency band. Narrowband envelopes can be partially recovered from a NH fine structure signal if they are analyzed using auditory filters of normal bandwidth (which the participants listening to these stimuli have; cf. Ghitza[127]). To minimize this unwanted recovery, i.e., to provide "degraded envelopes" within the auditory system of the NH listeners, an iterative procedure was used whereby the output of the multiplication procedure was passed through a NH Gammatone filter bank and the fine structure extracted using the Hilbert transform was multiplied again with the target impaired envelopes. Ten such iterations were used in the present study, which results in relatively high correlation with the desired speech envelope after modeled NH auditory processing[128].

The subsequent loudness module sets the level in each band such that the perceived loudness for a NH listener was manipulated in a way that resembles the perceived loudness of an (average) HI listener. For this second manipulation, the output signal of the envelope-processing module was fast Fourier transformed (FFT-ed) into six octave-spaced channels with frequencies between 250 Hz and 8 kHz. The level in each channel was extracted and adjusted such that the categorical loudness[129] of an average HI listener was achieved. This procedure was done based on average categorical loudness data[130]. As a last step, the spectral signal was transformed back into the time domain using the inverse FFT. The loudness module therefore also sets the audiometric threshold of the simulation. For the present study these degradations were implemented by taking a moderate hearing impairment as the base (according to Table 2) for the degradation manipulations. The specific values of this audiogram were selected to be similar to the standard audiogram N3 as defined in Bisgaard et al.[131]. Lastly, the sound level was root-mean-square (RMS) equalized to the intact audio, in order to ensure any effects found were not only due to an overall decreased loudness.

Table 2. Audiometric thresholds based on a typical, relatively flat moderate hearing impairment and used for the audio degradation manipulations.

Frequency (Hz)	250	500	1000	2000	4000	8000
Threshold (dB HL)	40	40	45	54	62	70

3.2.6 Experimental set-up

The experiment was performed in a dark and quiet room, the only illumination present was provided by the monitor. The stimuli were presented full-screen on a 24.5-inch monitor with a resolution of 1920 x 1080 pixels (43 x 24.8 degrees of visual angle). Average screen luminance was 38 cd/m^2. Participants were seated in front of the screen at a viewing distance of 70 cm with their head placed in a chin- and forehead rest to minimize head movements. Stimulus display and response recording was controlled using the Psychophysics Toolbox (Version 3)[71–73] and Eyelink Toolbox[74] extensions of MATLAB (The Mathworks, Inc., Version R2017a). An Apple MacBook Pro (mid 2015 model) was connected to the monitor and controlled stimulus presentation. Audio was produced by the internal soundcard of this computer and presented binaurally through Sennheiser HD 600 over-ear headphones (Sennheiser Electronic GmbH & Co. KG). The sound level was calibrated to be at a comfortable and audible level, at a long-term RMS average of 65 dB SPL.

An Eyelink 1000 Plus eye-tracker (SR Research Ltd.), running software version 4.51, was used to measure participants' eye movements. Monocular gaze data was acquired at a sampling frequency of 1000 Hz. Due to technical issues, eye-tracking data for the second session of participant 11 and the first session of participant 12 were recorded at 250 Hz instead of 1000 Hz. The eye-tracker was mounted on a desk just below the presentation screen. The eye-tracker was calibrated at the start of the experiment using the built-in 9-point calibration routine. Calibration was verified with the validation procedure in which the same nine points were displayed again. The experiment was continued if the calibration accuracy was sufficient (i.e., average error of less than 0.5° and a maximum error of less than 1°). Drift was checked for after every fourth trial and after each break. The calibration procedure was repeated if the participant moved during breaks and whenever there was more than 1° of drift in more than one consecutive drift check.

3.2.7 Procedure

During the experiment, both behavioral and eye-tracking data were obtained to identify accuracy of emotion identification and gaze patterns during emotion perception with dynamic stimuli, respectively. In each trial, participants were asked to identify the emotion presented in one of the eight stimulus presentation conditions listed in Table 3. For the A and dA conditions, a fixation cross preceded the stimulus presentation for a random duration between 600 and 1600 ms. The fixation cross remained on screen during stimulus presentation in the A and dA conditions. For all other conditions, a full-screen image displaying the averaged frames of all videos (see Figure 1a), presented for a random duration between 600 and 1600 ms, preceded the stimulus. This averaged image was presented instead of the fixation cross so participants could already orient their gaze, which could be especially helpful in the conditions where a scotoma was present.

Table 3. Experimental conditions used in the experiment. Both modalities were either shown as they are (intact), degraded, or absent.

		Video		
		Intact	**Degraded**	**Absent**
Audio	**Intact**	AV	AdV	A
	Degraded	dAV	dAdV	dA
	Absent	V	dV	■

All participants were asked to respond as accurately as possible in a forced-choice discrimination paradigm, by clicking on the label on the response screen that corresponded with the identified emotion. All twelve emotions were always displayed together on the response screen. Participants' response (emotion label) was recorded as well as whether the response was correct or not. Participants were further instructed to blink as little as possible during the trial and maintain careful attention to the stimuli.

In total, each participant was presented with all 96 stimuli (twelve emotions x eight actors) in all eight conditions, each stimulus was thus seen eight times. The experiment was divided into six experimental blocks. In each experimental block all eight conditions were presented in sub-blocks that contained one sixth of the stimuli (i.e., 16 trials per sub-block, 128 trials per experimental block). The order of conditions between experimental blocks was counterbalanced using balanced Latin Squares within and across participants. Stimulus order for each condition was randomized. Participants were able to take a break after every second sub-block (i.e., every 32 trials) and were encouraged to take breaks in order to maintain concentration and prevent fatigue. Breaks were self-paced and the experiment continued upon a mouse-click from the participant. The eye-tracker was recalibrated if the participant moved during the break, otherwise only a drift correction was performed.

The experiment was preceded by 64 practice trials (eight practice trials for each condition) to familiarize the participants with the stimulus material and the task. For the practice trials, block order was fixed in the following order: AV, V, A, AdV, dAV, dV, dA, dAdV. Stimulus order within each practice block was randomized. After each practice trial, participants received minimal feedback on their given response (correct/incorrect), no feedback was given during the experiment.

Overall, the experiment consisted of 832 trials, including the 64 practice trials, and took about 2.5 hours to complete. The experiment was separated over two test sessions performed on separate days to avoid fatigue.

3.2.8 Analyses of behavioral data

Accuracy scores for each condition and emotion were first converted to unbiased hit-rates[77] to account for any response biases. The unbiased hit-rates (H_u) were then arcsine transformed to create a normal distribution and a repeated measures ANOVA was performed in R (version 3.6.0), using function *aov_ez* from the *afex* package (version 0.25-1), with the arcsine transformed H_u as the dependent variable and *condition* (with eight levels), *experimental test session* (first/second), and their interaction as fixed-effects variables. The Greenhouse-Geisser correction was performed in cases of a violation of the sphericity assumption. Effect sizes are reported as generalized eta-squared (*ges*).

Significant main effects were followed up by post-hoc tests to test which conditions were significantly different from each other. Due to many possible comparisons that can be made with eight conditions, we performed separate t-tests to compare conditions we expected to differ beforehand. P-values of the t-tests were Bonferroni corrected. The following comparisons were made:

- AV with AdV, dAV, dAdV, V, and A
- dAdV with AdV and dAV
- V with A and dV
- A with dA

Non-significant t-tests were followed up with Bayesian t-tests using the *ttestBF* function from the *BayesFactor* package (version 0.9.12-4.2).

We additionally performed an exploratory omnibus paired comparisons test, which compared all conditions to each other using *lsmeans* from the *emmeans* package (version 1.4.1). To correct for multiple comparisons, the False Discovery Rate (*FDR*) correction was used.

3.2.9 Analyses of eye-tracking data

The built-in data-parsing algorithm of the Eyelink eye-tracker was used to extract fixations from the raw eye-tracking data. As only a fixation cross was presented during the A and dA conditions, the eye-tracking data from these conditions was not analyzed. Only those conditions in which a video was shown (AV, V, AdV, dAV, dV, and dAdV) were considered for the eye-tracking analyses. For fixation locations, we performed an Area-of-Interest (AOI) based analysis. In addition, we tested for differences between conditions in fixation durations and saccadic amplitudes. The analyses were restricted to fixations made during stimulus presentation, and only those made until 1000 ms after stimulus onset. No fixation data after 1000 ms were considered to limit data analysis to the duration of the shortest movie, which lasted 1000 ms. In addition, this aimed to discard any data that no longer was task-related, i.e. after a participant decided on a response, which is more likely to occur at a longer interval after stimulus onset. Trials with single blinks longer than 300 ms during stimulus presentation were discarded. Additionally, only trials with a correct response were included, as our main interest was in gaze behavior prior to correct recognition. This allowed examining whether changes in gaze behavior due to information degradation and availability of audio were adaptive and led to good performance.

The eyes (left and right), nose, mouth, and hands (left and right) of the actors were chosen as AOIs. Because the stimuli are dynamic, the AOIs were dynamic as well. Coordinates of the AOI positions for each stimulus and each frame were extracted using Adobe After Effects (Version 15.1.1). The coordinates for the face AOIs were obtained by applying the 'Face Tracking (Detailed Features)' method, which automatically tracks many face features. Face track points at each frame were visually inspected and manually edited whenever the tracking software failed to track them correctly. For the hand AOIs, the 'Track Motion' method was used. A single tracker point per hand was used to track position. The tracker point was placed roughly in the center of the hand. Again, tracking was inspected visually and manually edited where needed. Coordinates of all obtained face and hand track points for each stimulus were stored in a text-file and used to create point AOIs. For the eyes we used the coordinates of the left and right pupil, for the nose the coordinates of the nose tip, and for the mouth we used the mean of the y-positions of 'mouth top' and 'mouth bottom' coordinates for the y-coordinate, and the mean of the x-positions of 'mouth left' and 'mouth right' coordinates for the x-coordinate of the AOI. Note that left and right are in reference to the actor, not the observer. So, the left eye and hand are generally on the right side of the screen and vice versa for the right eye and hand.

Then, for each fixation data-point the Euclidean distance between the fixation and each AOI was calculated. To test whether the Euclidean distance to each AOI changed for the different conditions, linear mixed effects regression was carried out in *R* using the *lmer* function from the *lme4* package (version 1.1-21). Euclidean distances were averaged per trial. In the model, the averaged Euclidean distance between the fixation location and each AOI were used as dependent variables, and *AOI* and *condition* (with six levels) were added as fixed effects, *participant* and *movie* were included as random intercepts. No random slopes were added, as the model did not converge when these were added. Overall significance of main effects and interactions was tested with the *Anova* function from the *car* package (version 3.0-3). Pairwise comparisons were performed to test whether fixation proportions on different AOIs were different between conditions, sessions, and response accuracy using *lsmeans* and corrected for multiple comparisons using the *FDR* p-value adjustment.

In addition, we tested whether fixation durations and saccadic amplitudes differed between conditions using linear mixed effects regression (with the *lmer* function). Fixation durations and saccadic amplitudes were extracted from the parsed data file. Saccades with amplitudes larger than the diagonal of the monitor, which was 49.6°, were filtered out, removing less than 1% of saccades. For both analyses, *condition*, *session*, and *response accuracy* were added as fixed effects and allowed to interact with each other. Similar to the AOI analysis, random intercepts for *participant* and *movie* were added, but without random slopes, as the models did not converge when these were added. Again, significance of main effects and interactions was assessed with the *Anova* function and pairwise comparisons with *FDR* correction were performed using *lsmeans*. Non-significant differences were followed up with Bayesian t-tests or ANOVA's (with the *ttestBF* and *anovaBF* functions from the *BayesFactor* package) to assess the amount of evidence for the differences being the same.

3.3 Results

3.3.1 Accuracy across conditions

Overall, participants performed the task with a mean accuracy of 0.41; accuracy scores in unbiased hit-rates (H_u) are shown in Figure 2, averaged over testing blocks and emotions. Because the H_u score is a combined score of the regular hit-rate corrected for misses and false positives, H_u is generally lower than the regular hit-rate, although the scale does not change. Overall, it appears that performance is best in the original AV condition, then decreases for V, and decreases further for A. For conditions where one modality was degraded and the other intact (dAV and AdV) and when both modalities were degraded (dAdV), performance is not severely impacted compared to AV. Lastly, performance for a single degraded modality (dV and dA) is worse than its equivalent single non-degraded modality (V and A).

The ANOVA, which had the arcsine transformed unbiased hit-rate (H_u) as dependent variable and *condition* and *session* as fixed effects, showed a significant main effect of *condition* (F (7, 161) = 95.4, $p < 0.001$, *ges* = 0.49). The main effect of *session* (F (1, 23) = 4.3, $p = 0.05$, *ges* = 0.002) and the interaction between *condition* and *session* (F (7, 161) = 0.5, $p = 0.76$, *ges* = 0.0006) were not significant, indicating that there was no learning effect.

The post-hoc t-tests with Bonferroni corrected p-values showed that AV performance was higher than V ($t(23) = 7.3$, $p < 0.001$) and A ($t(23) = 13.8$, $p < 0.001$), and V was higher than A ($t(23) = 9.6$, $p < 0.001$), thus replicating our previous results[114]. Additionally, AV performance was higher than conditions with degraded visual information (AdV: $t(23) = 3.8$, $p = 0.01$; dAdV: $t(23) = 4.7$, $p = 0.001$), but not with only degraded auditory information (dAV: $t(23) = 0.43$, $p = 1.0$). The Bayesian t-test showed that there was anecdotal evidence for no difference in recognition performance between AV and dAV ($BF_{01} = 2.47$). Additionally, dAdV performance was lower than dAV ($t(23) = 3.7$, $p = 0.01$), but not significantly different from AdV ($t(23) = 0.7$, $p = 1.0$). There was anecdotal evidence for performance being the same in dAdV and AdV ($BF_{01} = 1.59$). Lastly, V performance was higher than dV performance ($t(23) = 5.6$, $p < 0.001$), and A performance was higher than dA performance ($t(23) = 4.3$, $p = 0.003$).

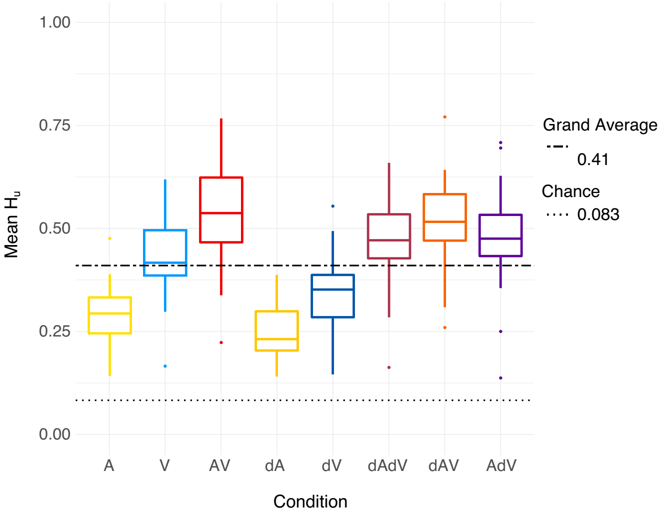


*Figure 2. Task performance for each condition, shown as unbiased hit-rates. Averaged across emotions and blocks. Each box shows the data between the first and third quartiles. The horizontal solid line in each box denotes the median. The whiskers extend to the lowest/highest value still within 1.5 * interquartile range, dots are outliers. The black dotted line indicates chance level performance (0.083). The black dashed-dotted line denotes the grand average accuracy over conditions and participants (0.41). Degraded conditions are shown in darker hues of the intact condition. Colors for AV conditions in which one or more modality is degraded are a mix between the degraded modality and intact AV.*

The results for the exploratory omnibus pairwise comparisons (*FDR* corrected) can be found in Table A.1. Except for the comparisons between AV and dAV and between AdV and dAdV, all comparisons show significant differences. Because we realize that the valence- and arousal level of an emotion may affect which cues (visual or auditory) may be most useful, we reanalyzed the data after combining individual emotions into their respective quadrants (see Table 1). We found that, while the overall performance differs per quadrant, the pattern across conditions stayed the same. That is, for all quadrants, performance is lowest with A, higher with V, and highest with AV. Additionally, performance drops when a degraded modality is presented in isolation (dA, dV), but not much when these are combined (dAdV). See Supplementary Material B for details.

To summarize, we found decreased performance for AdV and dAdV compared to AV, but not for dAV compared to AV, indicating that, at least for the materials used here, participants seem capable of compensating for degraded auditory, but not for degraded visual information. Hence, results show that there could be a hierarchy in the processing of the information in each modality, and this hierarchy can further affect how much degradation in that modality can be compensated for by the other modality.

3.3.2 Saccadic amplitude differences

Saccadic amplitudes, averaged over all stimuli and participants, for each condition are shown in Figure 3. The figure only shows saccadic amplitudes for saccades made during the first 1000 ms of correctly recognized trials. Figure 3 suggests differences in saccadic amplitudes for the different conditions, with larger amplitudes for conditions with degraded visual information.

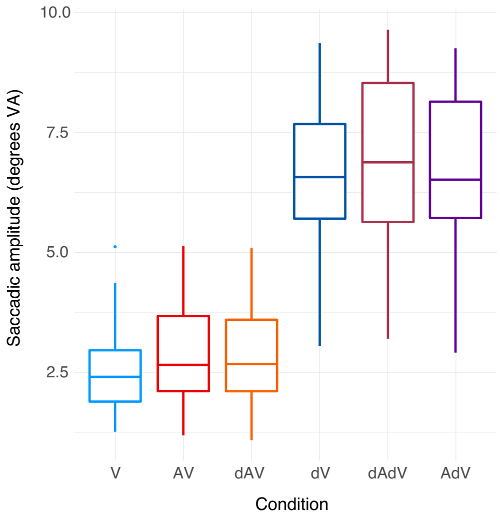


Figure 3. Saccadic amplitude in degrees of visual angle for correct responses in each condition, averaged over stimuli and participants. The horizontal solid line in each box denotes the median. Colors for each condition correspond to the same colors in Figure 2.

The regression model confirmed this. The model included *condition* as a fixed effect and random intercepts for both *participant* and *movie*. There was a significant main effect of *condition* (Chi^2 (5) = 3455.8, p < 0.001). A follow-up on the main effect of *condition* showed that saccades in conditions with intact visual information (AV, V, and dAV) were smaller than in conditions with degraded visual information (AdV, dV, dAdV), all p < 0.001. Additionally, participants made smaller saccades in the V compared to the AV (*z-ratio* = 2.64, p = 0.01) and dAV (*z-ratio* = -2.33, p = 0.02) conditions. Saccadic amplitudes were not significantly different between AV and dAV (*z-ratio* = 0.31, p = 0.76), and the Bayesian t-test indicated substantial evidence for the same saccadic amplitudes in AV and dAV (BF_{01} = 4.21). Lastly, participants made smaller saccades in the dV condition compared to dAdV (*z-ratio* = -3.06, p = 0.003), but not compared to the AdV condition (*z-ratio* = -1.31, p = 0.20), although the evidence for the null hypothesis was anecdotal (BF_{01} = 2.22). Saccadic amplitudes were also not significantly different between AdV and dAdV (*z-ratio* = -1.82, p = 0.08), but again, the evidence for no difference was anecdotal (BF_{01} = 1.46).

Participants thus made larger saccades in conditions with degraded video than in conditions with intact video. Additionally, removing the audio led to somewhat smaller saccadic amplitudes.

3.3.3 Fixation duration differences

Figure 4 shows fixation duration, averaged over all stimuli and participants, for each condition and the two test sessions. As in Figure 3, Figure 4 only shows fixation durations for fixations made during the first 1000ms of correctly recognized trials. Similar to saccadic amplitude, there appears to be a difference between conditions, with shorter fixations for conditions with degraded visual information.

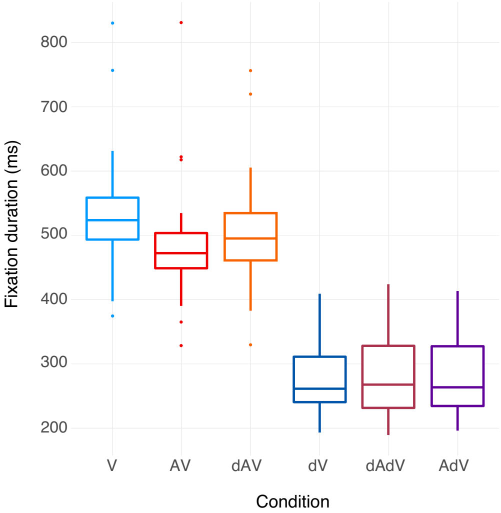


Figure 4. Fixation duration in ms for correct responses in each condition, averaged over stimuli and participants. The horizontal solid line in each box denotes the median. Colors for each condition correspond to the same colors in Figure 2.

The differences were tested with a regression model that included *condition* as a fixed effect, with random intercepts for *participant* and *movie*. There was a significant main effect of *condition* (*Chi*^{2} (5) = 2792.1, *p* < 0.001).

FDR-corrected pairwise comparisons for the main effect of condition showed that participants made longer fixations in the V condition than in the AV (*z-ratio* = -6.01, *p* < 0.001) and in the dAV condition (*z-ratio* = 4.76, *p* < 0.001). The difference between AV and dAV was not significant, but there was only anecdotal evidence for similarity (*z-ratio* = -1.27, *p* = 0.257, BF_{01} = 1.61) In addition, fixation durations were longer in the conditions with intact visual information (AV, V, dAV) than in the conditions with degraded video (AdV, dV, dAdV), all *p* < 0.001. There were no significant differences in fixation duration between conditions with degraded visual information, all *p* > 0.88, the evidence for no difference was substantial (BF_{01} = 7.80). Degrading the visual information thus led to a decrease in fixation durations.

3.3.4 Fixation distance differences between conditions

Fixation heatmaps for the first 1000ms of gaze data for audio-only conditions, conditions with intact video, and conditions with degraded video are shown in Figure 5. The heatmaps are overlaid on a 1000ms window averaged video image. Heatmaps for individual conditions can be found in Figure A.1. Average fixation distance to all AOIs in each condition, averaged over participants is shown in Figure 6. Differently colored bars indicate the different conditions, the x-axis shows the different AOIs. As before, only fixation data for the first 1000ms of correctly recognized trials are included in the figure and analysis. It should be noted that fixation distances in conditions with degraded visual information should be interpreted with the scotoma size in mind; it is expected that the fixation distances would decrease with a smaller scotoma.

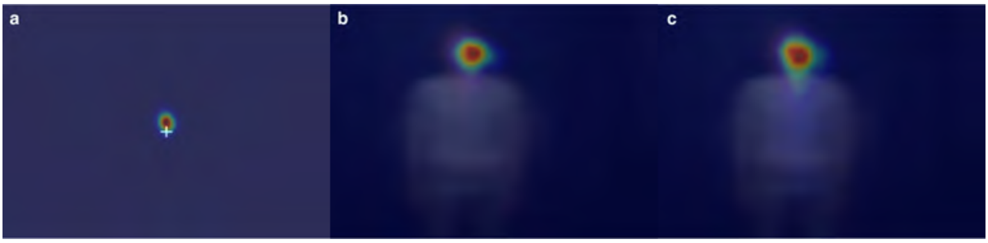


Figure 5. Fixation heatmaps overlaid on a 1000ms window averaged video image. a) Fixation heatmap for the audio-only conditions (A, dA). b) Fixation heatmap for conditions with intact video (V, AV, dAV). c) Fixation heatmap for conditions with degraded video (dV, dAdV, AdV). Heatmaps for individual conditions can be found in Figure A.1.

Figure 6 indicates that under degraded visual information, participants look away from the face AOIs and slightly closer to the hand AOIs, indicating that participants moved their gaze downwards and not solely to the left or right. The regression model also confirmed this pattern. The model included *AOI* and *condition*, and their interaction, as fixed effects. *Participant* and *movie* were added as random intercepts. There were significant main effects of *AOI* (*Chi*^{2} (5) = 73939.4, $p < 0.001$), and *condition* (*Chi*^{2} (5) = 7594.2, $p < 0.001$). Additionally, the interaction between *condition* and *AOI* was significant (*Chi*^{2} (25) = 6514.1, $p < 0.001$).

Overall, participants fixated the face more closely than the hands (all $p < 0.001$). Additionally, the nose and mouth were fixated at a shorter distance than both the left eye (left eye – nose estimate = 0.52, $p < 0.001$; left eye – mouth estimate = 0.60, $p < 0.001$) and the right eye (right eye – nose estimate = 0.45, $p < 0.001$; right eye – mouth estimate = 0.53, $p < 0.001$), there was no significant difference in fixation distance between the nose and mouth (estimate = 0.08, $p = 0.14$) or between the left and right eye (estimate = 0.07, $p = 0.20$). Lastly, there was no significant difference in fixation difference between the left and right hand (estimate = -0.03, $p = 0.57$).

Pairwise comparisons for the *AOI*-by-*condition* interaction, including Bayes factors for non-significant contrasts, are shown in Table A.2. The interaction showed that participants fixated the face AOIs at a further distance for conditions with degraded visual information

(AdV, dV, dAdV) compared to conditions with intact visual information (AV, V, dAV), all $p < 0.001$. Additionally, fixation distances to the hand AOIs were generally smaller for conditions with degraded visual information (p's < 0.03), except for the difference between AdV and AV, V, and dAV for the right hand (p's > 0.08), and between dAV and dAdV for the right hand (estimate 0.22, $p = 0.13$, $BF_{01} = 4.65$). Interestingly, participants fixated more closely to all AOIs for the dV condition compared to both the AdV and dAdV conditions, all $p < 0.05$. The differences between AdV and dAdV were never significant, all $p > 0.21$ and there was generally substantial evidence for similarity (BF_{01} range: 2.85 – 3.69). Lastly, there were no significant differences in fixation distance between conditions with intact video, all $p > 0.12$, although the evidence for similarity was mostly anecdotal for the comparisons between AV and V (BF_{01} range: 0.76 – 4.23) and between V and dAV (BF_{01} range: 0.24 – 3.61), but generally substantial for the comparisons between AV and dAV (BF_{01} range: 2.37 – 4.62).

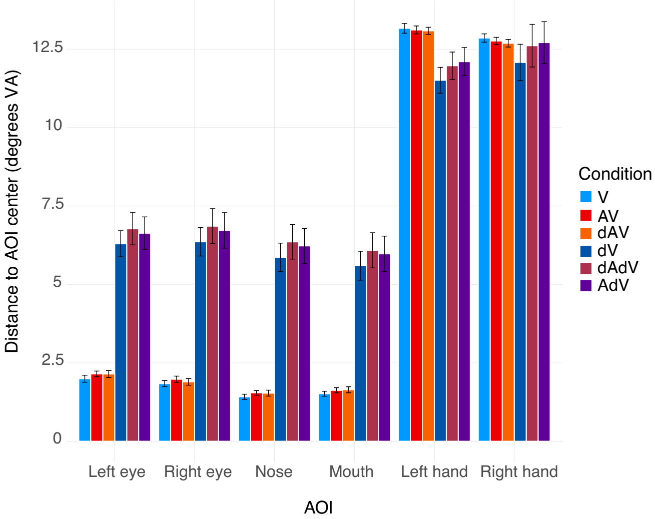


Figure 6. Euclidian fixation distance to AOI center in degrees of visual angle for each condition, averaged over stimuli and participants. Error bars denote the SEM. Colors for each condition correspond to the same colors in Figure 2.

To summarize, participants moved their fixations further from the actor's face and closer to the left hand when the video was degraded. Additionally, participants fixated all AOI's at a slightly closer distance in the dV condition than in the AdV and dAdV conditions. There was evidence that fixation distances were similar for the AdV and dAdV conditions and also for the AV and dAV conditions.

3.4 Discussion

Overall, we find that adding any audio to any video greatly improves emotion recognition. At least for the task and stimulus used here, the addition of either intact or degraded audio to intact or degraded video leads to improvement in emotion recognition. In line with this finding, degrading audio does not seem to impair emotion recognition or affect gaze behavior more than only degrading the video. We found that emotion recognition accuracy and gaze behavior did not significantly differ between the AdV and dAdV conditions, although the evidence for their similarity was generally not substantial. Additionally, degraded auditory information presented alongside intact visual information did not significantly affect performance or gaze behavior compared to intact audiovisual presentation. Moreover, there was some evidence for similarity between the AV and dAV conditions. Lastly, video degradation always impacted both accuracy and gaze behavior, independent of the quality of the audio signal (intact, degraded, or absent).

Our results thus suggest that while audio greatly facilitates emotion recognition, it cannot fully compensate for the negative effects of visual degradation, in line with the low recognition accuracy for audio-only conditions. The asymmetry in compensation may additionally relate to the known asynchrony in visual and auditory perception during speech perception. In audiovisual speech, visual cues may precede auditory cues by several hundred milliseconds[28,29]. Because of this order, visual cues provide information about the onset of the acoustic signal, but also about the amplitude envelope of the speech[28]. Therefore, in speech, early visual cues make auditory cues more predictable, yet auditory cues cannot increase the predictability of visual cues. This natural asynchrony between visual and auditory cues could be one of the reasons for the fact that intact vision can compensate for a degradation in auditory information, while auditory information cannot fully do so for a degradation in visual information.

3.4.1 Combined visual and auditory degradation does not exacerbate isolated effects

For degraded stimuli, we found that our signal degradations had the desired effect of increasing task difficulty and decreasing recognition performance, as was aimed for. This was derived from the pure effects of degradation (i.e., the conditions in which one modality was degraded and the other modality was absent): we found that dV performance was significantly lower than V performance and dA performance was lower than A performance. The isolated effects were not enhanced when combining degraded video and degraded audio in the dAdV condition as the performance level for dAdV was much higher than for dV and dA. Thus, it appears that the addition of any information to a degraded modality increases the amount of information that can be used for emotion recognition and simultaneous degradation in two modalities do not exacerbate their individual effects. In addition, we found that the presence of an additional modality can sometimes completely negate the effect of the degraded modality. Performance for degraded auditory but intact visual information (dAV) was similar to AV performance. However, for degraded visual information, this was not the case; for conditions with degraded visual information and intact or degraded audio (AdV and dAdV respectively), we found decreased performance compared to AV. Moreover, AdV and dAdV performances did not differ significantly, and there was anecdotal Bayesian evidence for similar performance, suggesting that degraded audio on top of degraded video did not decrease performance further. Thus, it appears that, at least for the materials we have used here, participants could fully compensate for the degraded audio by relying more on the intact visual information. In contrast, they could not compensate for the degraded video by relying more on the intact audio. Considering the fact that A performance was much lower than V performance, it might be that the audio did not provide enough or not the right kind of information to compensate for the degraded vision. On the other hand, studies have suggested a dominance of visual over auditory information for emotion perception, at least for similar materials[55,64], thus it could also be that participants relied mostly on the visual information by default, possibly because they were not adapted well enough to the degraded visual signal to shift their attention more to the auditory cues and rely more on them. To discover which of these mechanisms is occurring, further studies would need to be performed in participants that are well adapted to the degradations. This is possible in individuals with hearing and/or vision impairments, or in healthy observers that underwent an extensive adaptation procedure.

3.4.2 Viewing behavior suggests observers use peripheral information to perceive emotional expressions

Our findings for gaze behavior are consistent with the performance results. Viewing behavior was similar for the AV and dAV conditions, at least for the measures examined here. Overall, the biggest differences in gaze behavior were between conditions with and without a degraded visual signal. We found that with degraded video, participants made larger saccades and fixations of shorter duration. Additionally, they moved their fixations away from the face AOIs and somewhat closer to the hand AOIs when video was degraded. There is an indication that participants placed the face AOIs adjacent to the border of their scotoma: the scotoma extended 17 deg x 11.5 deg of visual angle, and participants fixated the face AOIs at distances at roughly half the height of the scotoma (6 deg of visual angle) in visual degradation conditions. This is in line with findings in macular degeneration patients[132] and in control observers with simulated scotoma's[133,134], and suggests that the participants in the current study developed perceptual strategies that are similar to what is seen with a preferred retinal locus (PRL) in patients. In a previous study[114], we have shown that observers generally fixate on the face when identifying emotions. Considering the small fixation distance to the face AOIs for intact visual stimuli and the large fixation distance to the hand AOIs, it can be assumed that participants in the current study also mainly fixated on or near the face. Combining that with the fact that under degraded video, participants' fixations were closer to the hand AOIs than in intact video, and that, in the videos, the hands were generally located inferior to the face, suggests that participants shifted their gaze downwards while using their superior visual field to view the face. While moving gaze down likely makes the scotoma cover the lower body and the hands, which may seem undesirable, it was still possible to view larger movements even when they were covered by the scotoma, due to the relative nature of the scotoma.

3.4.3 Observers increase fixation duration and make larger saccades when viewing degraded video

Our finding that participants' fixation durations were shorter under visual degradation is in contradiction with the idea that observers fixate longer with more difficult tasks[115] and with findings of longer fixation durations with simulated scotoma's for visual search tasks[120,121]. It cannot be that our finding of shorter fixation duration under degraded visual signal is due to the task not being more difficult, as performance always decreased for visual degradation and thus, even though eye-tracking analyses were based on correct responses, we can safely assume that the task was more difficult. Whether fixation durations become longer or shorter might therefore strongly depend on the task and stimulus used. For example, McIlreavy and colleagues[135] used a visual search task with natural images and found that a simulated central scotoma had no effect on mean fixation duration. Henderson et al.[136] used an object identification and recollection task and found a decrease in fixation duration when a central scotoma was present. There is another discrepancy between our and Cornelissen et al.'s[121] findings; they only found an effect on saccadic amplitude for the absolute central scotoma, not the relative central scotoma. The absolute scotoma took on the background color and luminance, while for the relative scotoma the information on the display was shown with very low contrast (3%) within the scotomatic region. Thus, for the relative scotoma, some information was still perceivable, while for the absolute scotoma this was not the case. The scotoma used here was relative as well, as the video within the scotoma was severely blurred and some information could still be perceived (e.g., whether the observer was viewing the face or the body of the actor); yet visual degradation still affected saccadic amplitude. It could be that the blurring was so severe that the scotoma, while technically relative, was effectively perceived as absolute.

One reason for the discrepancies between ours and previous findings might be related to the various types and roles of superior colliculus cells; Walker and colleagues (1997) proposed that there is an ongoing competition in the superior colliculus between cells that stabilize fixation and cells that program saccades. In the presence of peripheral objects, the saccade programming cells increase their firing rate, which increases the probability that a saccade is made. When the presence of peripheral objects is combined with absent foveal information, as in the case of an absolute scotoma, it is even more probable that the balance is shifted more towards saccades. In the materials used here, there was only a single object that was also strongly attention grabbing: the actor. Thus, when it is possible to fixate on the actor (when the video is intact), observers do so, evident by longer fixation durations and small saccades. However, when fixating on the actor leads to not being able to see the actor (when video is degraded by a central scotoma), observers saccade away from the actor in order to see them. At that moment, the actor is located in the periphery, firing rates in the saccade programming cells increase, and saccading back to the actor becomes increasingly probable. Together, this leads to both shorter fixation durations and on average larger saccades (which are needed to move the scotoma away from the actor). In the studies that found longer fixations and no effect on saccadic amplitude[74,120], many objects were present on the display. Thus, when foveal vision was removed by a scotoma, this may have increased saccade generation. However, since it is not immediately obvious towards which object a saccade should be directed, and observers should additionally continuously attempt to process the objects parafoveally/peripherally, which is only possible during fixation, the lack of foveal vision may not necessarily lead to a shortening of fixation durations.

3.4.4 Removing audio affects viewing behavior, degrading audio does not

While we did not find any effects of degraded audio on gaze behavior, a complete absence of audio did affect gaze. In the intact visual, absent audio (V) condition, participants made smaller saccades and fixations with longer durations compared to AV and dAV. In the degraded visual, absent audio (dV) condition, participants made smaller saccades compared to dAdV and fixated all AOIs at the shorter distance than in dAdV and AdV. The fact that the difference in fixation distance for V compared to AV and dAV conditions were not significant (although there were trends in the same direction), might be related to the fact that the fixation distances to AOIs were generally small for V, and thus, differences in fixation distance between V and AV/dAV are then also small and unlikely to reach significance. With respect to differences in saccadic amplitude and fixation duration, the effects for V, compared to AV and dAV, and effects for dV, compared to dAdV and AdV, are not the same. This might be related to the role of the superior colliculus in stabilizing fixations and programming saccades, as discussed above. In the absence of audio, it may be more important to have a stable fixation in order to extract sufficient information and gaze may therefore be placed closer to the AOI. In addition, there is no audio that directs attention to its source, which in this situation is the speaker, so there is less ‘saccade generating’ information in the stimulus. Together, this can explain the longer fixation durations and smaller saccades found in the V condition. In the dV condition however, as explained above, there is very limited foveal information, and with the actor being in peripheral vision (when participants are fixating with their peripheral), more saccades are generated, thus annulling some of the effects that absent audio has on gaze behavior. The need to focus gaze more when audio is absent may explain why participants fixated the AOIs at a closer distance for dV than for dAdV and AdV; this need may have led participants to be more thorough in placing the scotoma, in order to have the border of the scotoma as close to the AOI as possible. As this is likely more effortful, the presence of audio in dAdV and AdV could explain why participants did not use the same care in those conditions.

3.4.5 Limitations and future directions

It should be noted that the fact that we found that observers are affected by degraded visual information, but not by degraded auditory information when it is accompanied by video, may be strongly dependent on the specific materials we used, which had very rich visual cues and possibly less clear auditory cues. On the other hand, the results may also be related to the fact that, generally, observers seem to rely more on visual information than on auditory information for proper perception of emotions[55] and the aforementioned asynchrony between visual and auditory cues. Our results hold the promise that individuals with hearing loss may also be able to compensate for their degraded hearing by relying more on their intact vision. However, there is a chance that cognitive decline due to ageing or the sensory degradations may affect the capacity of (elderly) individuals to compensate.

By design, our study only allowed measuring the possible acute effects of sensory impairments and thus disregards any long-term adaptation that may occur in real sensory impairments. Future studies are needed in individuals with sensory impairments as well as in healthy elderly observers to untangle the effects of general ageing from the effects of sensory impairments.

Studies with different audiovisual emotion materials, for example by including sentences with meaningful semantic content, may shed light on the apparently stronger effects for visual information compared to auditory information.

Lastly, it would be interesting to investigate what specific information in the audio and video signals cause the multimodal facilitation. A likely explanation would be the temporal correlation, as for example speech correlates strongly with the movements the mouth makes. If it is purely related to temporal correlation then replacing the original audio by a tone that fluctuates in fundamental frequency, where these fluctuations represent visual expressions, should already facilitate recognition.

3.4.6 Conclusions

Altogether, the present data show that the combined effects of degraded visual and auditory input do not exacerbate their isolated effects. Thus, there is redundancy in the information relevant to emotion recognition. Such redundancy, which in this study was most notable in vision, can supplement degraded information in another modality, here in audio. It remains an open question whether this redundancy remains still present after long-term central and cognitive changes induced by sensory loss. Additionally, we have shown that observers adapt their viewing behavior to degraded video in order to maximize recognition. Teaching this optimized viewing behavior to visually impaired individuals that do not show this behavior spontaneously could therefore be a starting point for rehabilitation targeted at improved emotion recognition.

3.5 Data availability

The datasets generated for this study can be found in the DataverseNL repository via https://doi.org/10.34894/4XDHZ8. All data is publicly available.

3.6 Acknowledgements

We gratefully acknowledge Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer for permission for using the GEMEP core set and for publishing sample images from the core set. We also thank Ben Williges, Remco Renken and Alessandro Grillini for help with programming and data analysis and Merel de la Rie for help with data collection.

3.7 Funding

The first author was supported by a BCN-BRAIN grant from The Graduate School of Medical Sciences (GSMS), University of Groningen, the Netherlands. This project was supported by the following foundation: the Landelijke Stichting voor Blinden en Slechtzienden, contributing through UitZicht (Grant number: Uitzicht 2014 – 28 – Pilot). The funding organizations had no role in the design or conduct of this research. They provided unrestricted grants.

3.8 Declaration of interest

None

Supplementary Material A

*Table A.1. Contrasts for the omnibus pairwise comparisons of the main effect of condition. It shows the model estimate differences with the FDR adjusted p-values in parentheses. The conditions in the columns are subtracted from the conditions in the rows. A positive contrast means performance in the first condition (rows) was better than in the second of the comparison (columns), and v.v. Significant differences are indicated by **bold** typeface. New comparisons are shown with a white background, while comparisons that were also made in the t-tests are shown with a light grey background.*

	AV	V	A	AdV	dAV	dAdV	dV	dA
AV	■	**0.12 (<0.001)**	**0.31 (<0.001)**	**0.07 (<0.001)**	0.007 (0.73)	**0.09 (<0.001)**	**0.23 (<0.001)**	**0.38 (<0.001)**
V		■	**0.18 (<0.001)**	**-0.06 (0.007)**	**-0.12 (<0.001)**	**-0.04 (0.045)**	**0.10 (<0.001)**	**0.25 (<0.001)**
A			■	**-0.24 (<0.001)**	**-0.31 (<0.001)**	**-0.23 (<0.001)**	**-0.09 (<0.001)**	**0.07 (0.001)**
AdV				■	**-0.06 (0.003)**	0.01 (0.49)	**0.15 (<0.001)**	**0.31 (<0.001)**
dAV					■	**0.08 (<0.001)**	**0.22 (<0.001)**	**0.38 (<0.001)**
dAdV						■	**0.14 (<0.001)**	**0.30 (<0.001)**
dV							■	**0.16 (<0.001)**
dA								■

*Table A.2. Contrasts for the AOI by condition interaction for fixation distance. It shows the model estimate differences with the FDR-adjusted p-values in parentheses. A positive contract indicates participants fixated closer to the AOI in the second condition than in the first condition (and v.v.). Significant differences are indicated by **bold** typeface. Bayes factors for the null hypothesis (BF_{01}) are added to non-significant comparisons in italic typeface.*

	Left Eye	Right Eye	Nose	Mouth	Left Hand	Right Hand
AV – V	0.16 (0.24) *BF_{01}: 0.76*	0.16 (0.24) *BF_{01}: 1.03*	0.15 (0.28) *BF_{01}: 1.11*	0.14 (0.31) *BF_{01}: 1.69*	-0.03 (0.85) *BF_{01}: 4.23*	-0.01 (0.91) *BF_{01}: 2.88*
AV – dAV	-0.05 (0.68) *BF_{01}: 4.66*	0.04 (0.76) *BF_{01}: 2.37*	-0.04 (0.78) *BF_{01}: 4.62*	-0.04 (0.72) *BF_{01}: 4.55*	0.02 (0.85) *BF_{01}: 4.51*	0.09 (0.55) *BF_{01}: 3.85*
AV – AdV	**-4.36 (<0.001)**	**-4.59 (<0.001)**	**-4.53 (<0.001)**	**-4.18 (<0.001)**	**0.99 (<0.001)**	0.25 (0.08) *BF_{01}: 4.65*
AV – dV	**-4.08 (<0.001)**	**-4.29 (<0.001)**	**-4.23 (<0.001)**	**-3.86 (<0.001)**	**1.51 (<0.001)**	**0.76 (<0.001)**
AV – dAdV	**-4.47 (<0.001)**	**-4.70 (<0.001)**	**-4.63 (<0.001)**	**-4.26 (<0.001)**	**1.17 (<0.001)**	**0.31 (0.03)**
V – dAV	-0.21 (0.12) *BF_{01}: 0.24*	-0.13 (0.37) *BF_{01}: 2.97*	-0.18 (0.19) *BF_{01}: 0.62*	-0.18 (0.18) *BF_{01}: 0.43*	0.05 (0.81) *BF_{01}: 3.61*	0.10 (0.53) *BF_{01}: 1.59*

	Left Eye	**Right Eye**	**Nose**	**Mouth**	**Left Hand**	**Right Hand**
V – AdV	**-4.52 (<0.001)**	**-4.75 (<0.001)**	**-4.68 (<0.001)**	**-4.32 (<0.001)**	**1.02 (<0.001)**	0.26 (0.08): BF_{01}: 4.55
V – dV	**-4.24 (<0.001)**	**-4.46 (<0.001)**	**-4.38 (<0.001)**	**-4.0 (<0.001)**	**1.54 (<0.001)**	**0.78 (<0.001)**
V – dAdV	**-4.63 (<0.001)**	**-4.87 (<0.001)**	**-4.78 (<0.001)**	**-4.40 (<0.001)**	**1.19 (<0.001)**	**0.32 (0.03)**
dAV – AdV	**-4.31 (<0.001)**	**-4.63 (<0.001)**	**-4.50 (<0.001)**	**-4.14 (<0.001)**	**0.97 (<0.001)**	0.16 (0.29) BF_{01}: 4.65
dAV – dV	**-4.03 (<0.001)**	**-4.33 (<0.001)**	**-4.20 (<0.001)**	**-3.81 (<0.001)**	**1.49 (<0.001)**	**0.67 (<0.001)**
dAV – dAdV	**-4.42 (<0.001)**	**-4.74 (<0.001)**	**-4.60 (<0.001)**	**-4.21 (<0.001)**	**1.15 (<0.001)**	0.22 (0.13) BF_{01}: 4.65
AdV – dV	**0.29 (0.047)**	**0.30 (0.04)**	**0.30 (0.03)**	**0.32 (0.02)**	**0.52 (<0.001)**	**0.51 (<0.001)**
AdV – dAdV	-0.10 (0.46) BF_{01}: 3.16	-0.11 (0.41) BF_{01}: 3.23	-0.10 (0.49) BF_{01}: 3.33	-0.07 (0.61) BF_{01}: 3.57	0.17 (0.23) BF_{01}: 2.85	0.06 (0.68) BF_{01}: 3.69
dV – dAdV	**-0.39 (0.007)**	**-0.41 (0.004)**	**-0.40 (0.005)**	**-0.40 (0.005)**	**-0.34 (0.02)**	**-0.45 (0.003)**

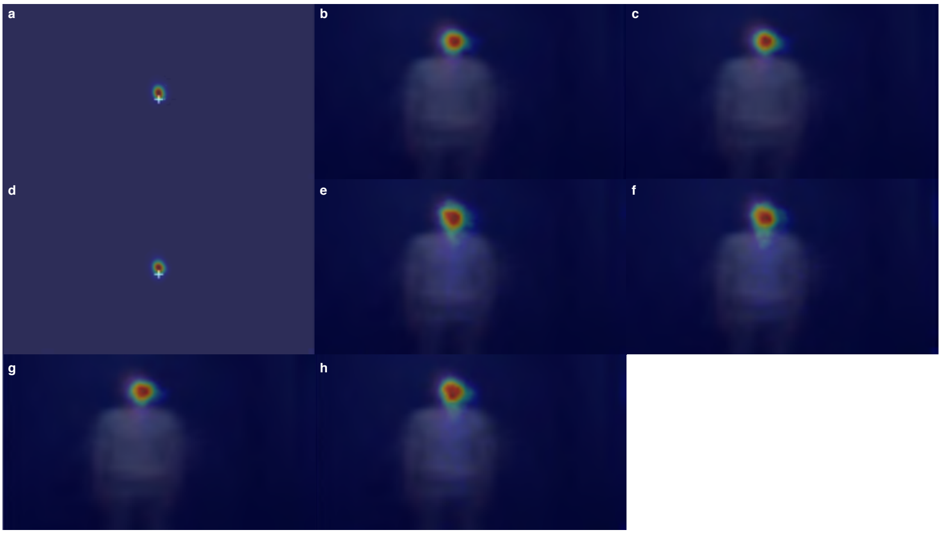


Figure A.1. Fixation heatmaps overlaid on a 1000ms window averaged video image. Different panels show heatmaps for different conditions. a) Fixation heatmap for A. b) Fixation heatmap V. c) Fixation heatmap for AV. d) Fixation heatmap for dA. e) Fixation heatmap dV. f) Fixation heatmap for dAdV. g) Fixation heatmap for dAV. h) Fixation heatmap AdV.

Supplementary Material B

This supplementary analysis investigates the effect of condition per valence-arousal quadrant. For this analysis, participant's responses were first recoded such that incorrect within-quadrant responses (such as responding "Relief" when the expressed emotion was "Interest") are considered correct responses. After recoding, a repeated measures ANOVA was performed in R (version 3.6.0), using the function *aov_ez* from the *afex* package (version 0.25-1), with the arcsine transformed H_u as the dependent variable and *condition* (with eight levels) and *emotional quadrant* (with four levels), as well as their interaction, as independent variables. The Greenhouse-Geisser correction was performed in cases of a violation of the sphericity assumption. Effect sizes are reported as generalized eta-squared (*ges*).

Significant main effects were followed up by post-hoc tests to test which conditions were significantly different from each other by means of an exploratory omnibus paired comparisons test, which compared all conditions to each other using *lsmeans* from the *emmeans* package (version 1.4.1). To correct for multiple comparisons, the False Discovery Rate (*FDR*) correction was used.

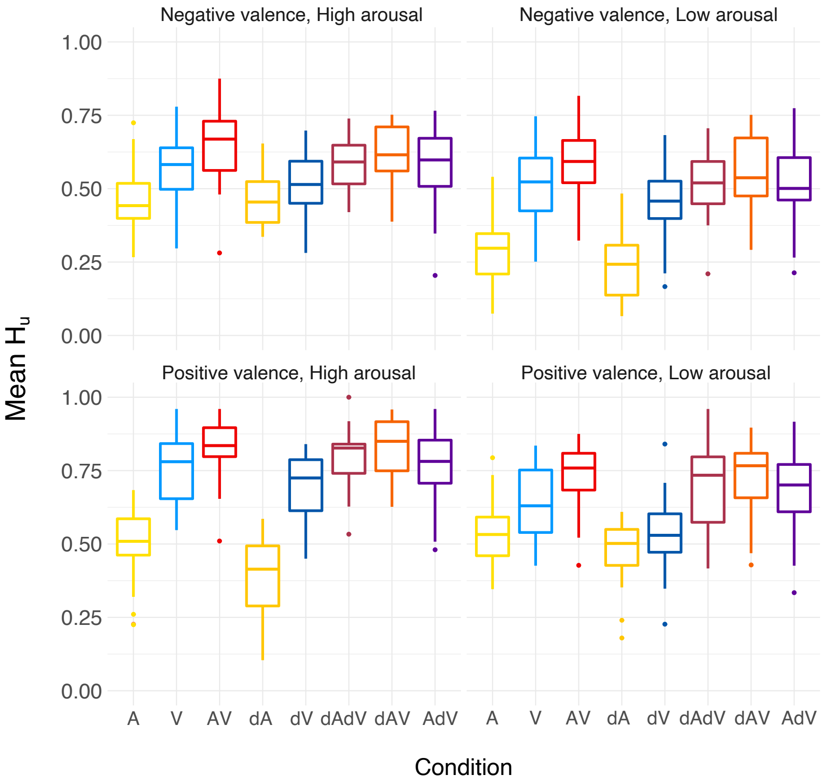


Figure B.1. Performance in mean unbiased hit-rates (H_u) per emotional quadrant (as in Table 1, main manuscript, section 2.3). For this figure, if the response was from the same quadrant as the expressed emotion, the response was considered correct.

The outcome was a significant main effect of *condition* (*F* (7, 161) = 91.4, *p* < 0.001, *ges* = 0.44), significant main effect of *quadrant* (*F* (3, 69) = 81.9, *p* < 0.001, *ges* = 0.35), and a significant interaction between *condition* and *quadrant* (*F* (21, 483) = 12.7, *p* < 0.001, *ges* = 0.11). While from Figure B.1 and Tables B.1-B.4 it is clear that there are differences between quadrants, for example accuracy is much higher for the positive valence, high arousal quadrant, differences between conditions are relatively stable between quadrants. For all quadrants, audio only performance is lower than video only performance, which in turn is lower than audiovisual performance. In addition, performance drops when a degraded modality is presented in isolation (dA, dV), but not much when these are combined (dAdV), similar to the findings presented in section 3.1.

Table B.1. Contrasts for the omnibus pairwise comparisons of the main effect of condition for emotions with negative valence and high arousal. It shows the model estimate differences with the FDR adjusted p-values in parentheses. The conditions in the columns are subtracted from the conditions in the rows. A positive contrast means performance in the first condition (rows) was better than in the second of the comparison (columns), and v.v. Significant differences are indicated by bold typeface.

	AV	**V**	**A**	**AdV**	**dAV**	**dAdV**	**dV**	**dA**
AV		**0.09** **(0.002)**	**0.19** **(<0.001)**	**0.07** **(0.015)**	0.03 (0.33)	**0.06** **(0.030)**	**0.14** **(<0.001)**	**0.20** **(<0.001)**
V			**0.10** **(<0.001)**	-0.02 (0.55)	**-0.06** **(0.04)**	-0.02 (0.40)	0.05 (0.064)	**0.11** **(<0.001)**
A				**-0.12** **(<0.001)**	**-0.16** **(<0.001)**	**-0.12** **(<0.001)**	-0.05 (0.094)	0.008 (0.78)
AdV					-0.04 (0.15)	-0.007 (0.79)	**0.07** **(0.015)**	**0.13** **(<0.001)**
dAV						0.03 (0.24)	**0.11** **(<0.001)**	**0.17** **(<0.001)**
dAdV							**0.08** **(0.007)**	**0.13** **(<0.001)**
dV								0.06 (0.05)
dA								

Table B.2. Contrasts for the omnibus pairwise comparisons of the main effect of condition for emotions with negative valence and low arousal.

	AV	**V**	**A**	**AdV**	**dAV**	**dAdV**	**dV**	**dA**
AV		**0.08** **(0.007)**	**0.33** **(<0.001)**	**0.08** **(0.003)**	0.03 (0.22)	**0.09** **(0.002)**	**0.14** **(<0.001)**	**0.39** **(<0.001)**
V			**0.25** **(<0.001)**	0.007 (0.81)	-0.04 (0.13)	0.01 (0.74)	**0.06** **(0.028)**	**0.31** **(<0.001)**
A				**-0.24** **(<0.001)**	**-0.29** **(<0.001)**	**-0.24** **(<0.001)**	**-0.19** **(<0.001)**	**0.06** **(0.034)**
AdV					-0.05 (0.074)	0.004 (0.89)	0.06 (0.051)	**0.30** **(<0.001)**
dAV						0.05 (0.059)	**0.10** **(<0.001)**	**0.35** **(<0.001)**
dAdV							0.05 (0.064)	**0.30** **(<0.001)**
dV								**0.25** **(<0.001)**
dA								

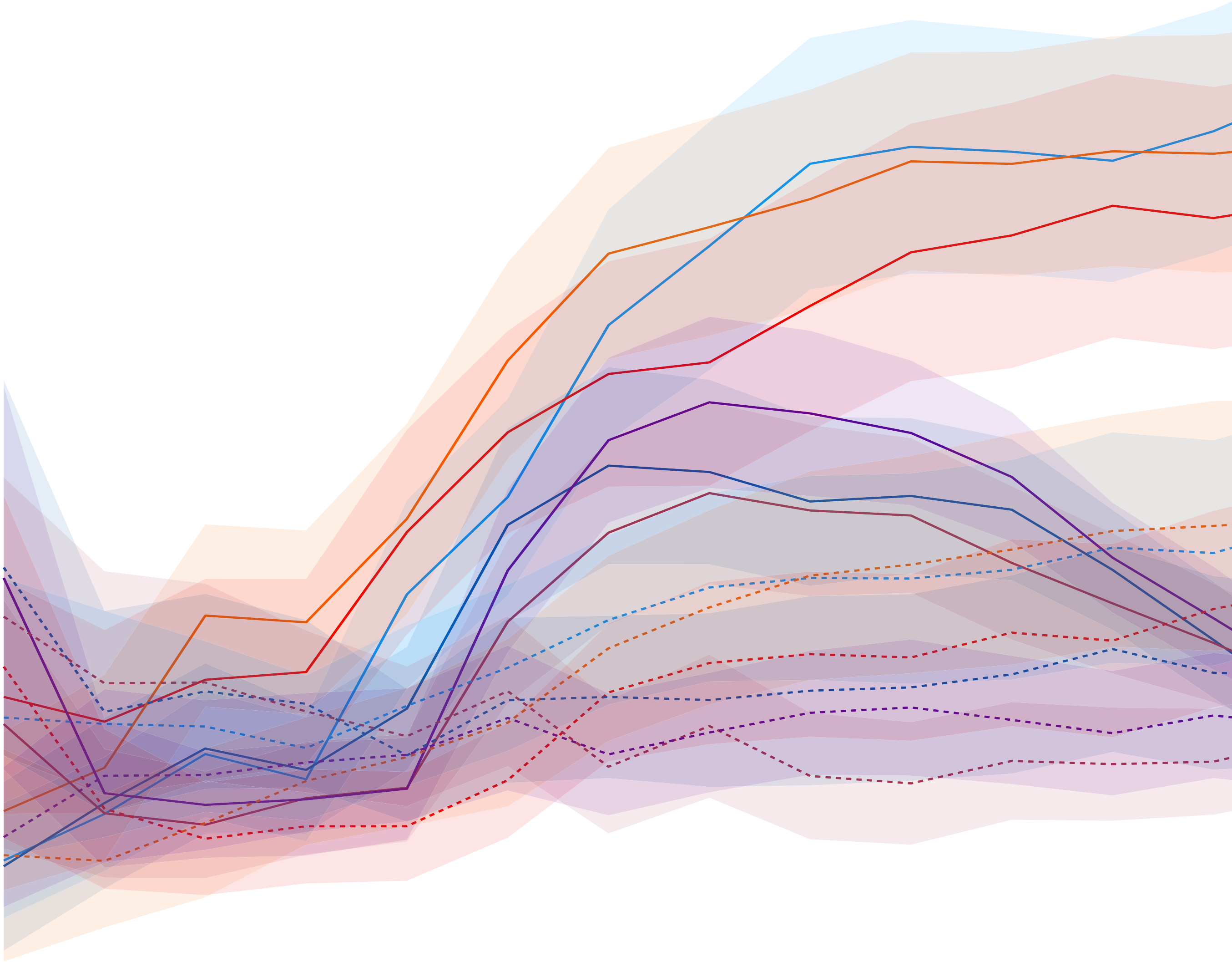
Table B.3. Contrasts for the omnibus pairwise comparisons of the main effect of condition for emotions with positive valence and high arousal.

	AV	**V**	**A**	**AdV**	**dAV**	**dAdV**	**dV**	**dA**
AV		**0.08** **(0.003)**	**0.37** **(<0.001)**	**0.07** **(0.017)**	-0.009 (0.74)	0.035 (0.21)	**0.17** **(<0.001)**	**0.48** **(<0.001)**
V			**0.29** **(<0.001)**	-0.02 (0.56)	**-0.09** **(0.001)**	-0.05 (0.093)	**0.09** **(0.001)**	**0.40** **(<0.001)**
A				**-0.31** **(<0.001)**	**-0.38** **(<0.001)**	**-0.34** **(<0.001)**	**-0.20** **(<0.001)**	**0.11** **(<0.001)**
AdV					**-0.07** **(0.007)**	0.03 (0.27)	**0.10** **(<0.001)**	**0.42** **(<0.001)**
dAV						0.04 (0.12)	**0.18** **(<0.001)**	**0.49** **(<0.001)**
dAdV							**0.14** **(<0.001)**	**0.44** **(<0.001)**
dV								**0.31** **(<0.001)**
dA								

Table B.4. Contrasts for the omnibus pairwise comparisons of the main effect of condition for emotions with positive valence and low arousal.

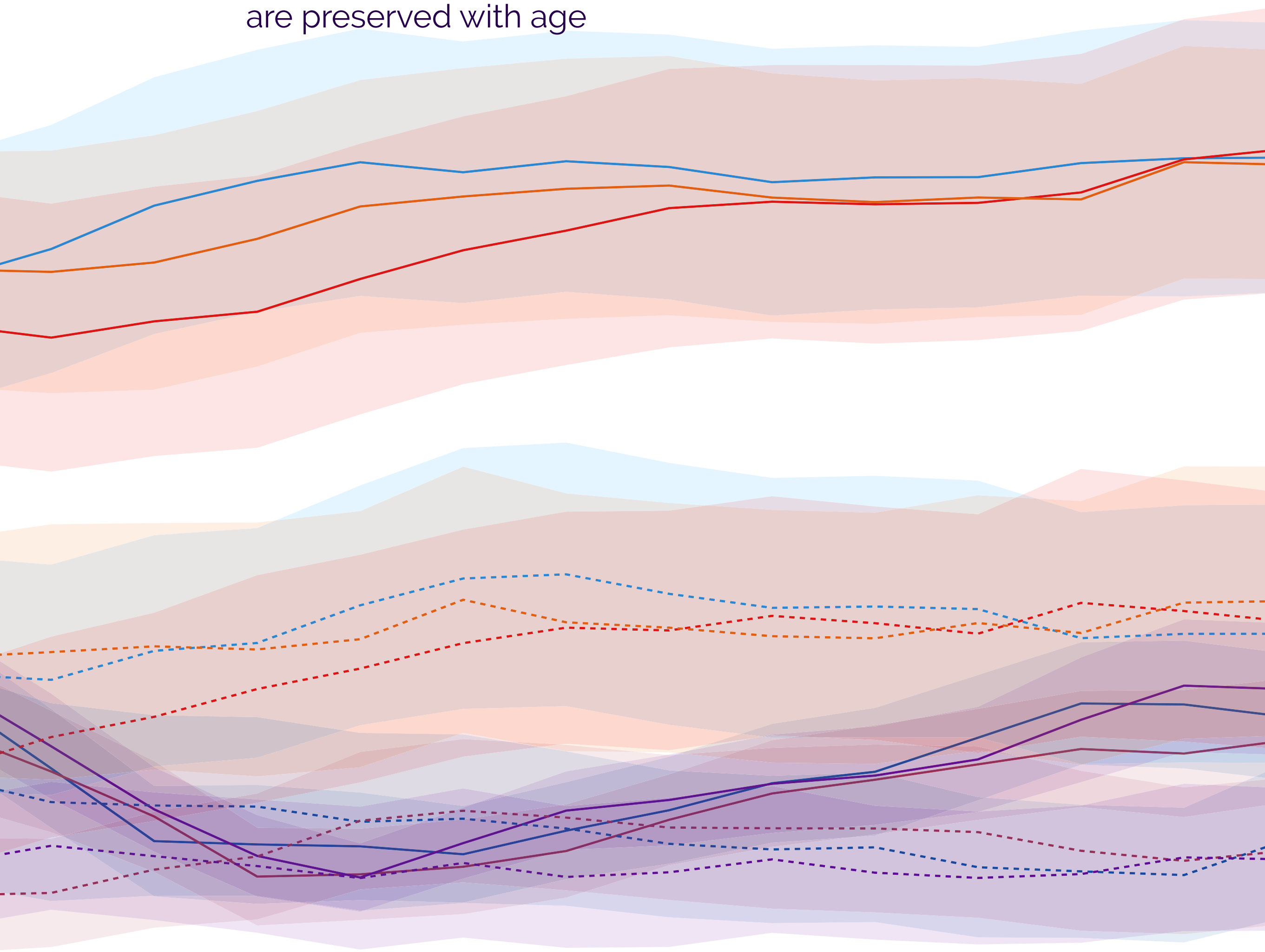
	AV	**V**	**A**	**AdV**	**dAV**	**dAdV**	**dV**	**dA**
AV	■	**0.10** **(<0.001)**	**0.20** **(<0.001)**	0.04 (0.14)	0.002 (0.94)	0.03 (0.23)	**0.21** **(<0.001)**	**0.27** **(<0.001)**
V		■	**0.10** **(<0.001)**	**-0.06** **(0.031)**	**-0.10** **(<0.001)**	**-0.07** **(0.016)**	**0.10** **(<0.001)**	**0.17** **(<0.001)**
A			■	**-0.16** **(<0.001)**	**-0.20** **(<0.001)**	**-0.16** **(<0.001)**	0.008 (0.79)	**0.07** **(0.014)**
AdV				■	-0.04 (0.16)	0.008 (0.80)	**0.16** **(<0.001)**	**0.23** **(<0.001)**
dAV					■	0.03 (0.25)	**0.20** **(<0.001)**	**0.27** **(<0.001)**
dAdV						■	**0.17** **(<0.001)**	**0.23** **(<0.001)**
dV							■	**0.06** **(0.030)**
dA								■

Author affiliations:
1. Research School of Behavioural and Cognitive Neurosciences (BCN), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
2. Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
3. Laboratory for Experimental Ophthalmology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
4. Institute of Acoustics, Technische Hochschule Lübeck, Lübeck, Germany



Chapter 4

Auditory and visual integration for emotion recognition and compensation for degraded signals are preserved with age



In press. *Trends in Hearing*

Minke J. de Boer[1,2,3], Tim Jürgens[4], Deniz Başkent[1,2],
Frans W. Cornelissen[1,3]

Abstract

Since emotion recognition involves integration of the visual and auditory signals, it is likely that sensory impairments worsen emotion recognition. In emotion recognition, young adults can compensate for unimodal sensory degradations if the other modality is intact. However, most sensory impairments occur in the elderly population and it is unknown whether older adults are similarly capable of compensating for signal degradations. As a step towards studying potential effects of real sensory impairments, this study examined how degraded signals affect emotion recognition in older adults with normal hearing and vision. The degradations were designed to approximate some aspects of sensory impairments. Besides emotion recognition accuracy, we recorded eye movements to capture perceptual strategies for emotion recognition. Overall, older adults were as good as younger adults at integrating auditory and visual information and at compensating for degraded signals. However, accuracy was lower overall for older adults, indicating that ageing leads to a general decrease in emotion recognition. In addition to decreased accuracy, older adults showed smaller adaptations of perceptual strategies in response to video degradations. Concluding, this study showed that emotion recognition declines with age, but that integration and compensation abilities are retained. In addition, we speculate that the reduced ability of older adults to adapt their perceptual strategies may be related to the increased time it takes them to direct their attention to scene aspects that are relatively far away from fixation.

Keywords: ageing; eye-tracking; audiovisual; emotion recognition; sensory impairment

4.1 Introduction

A fundamental component of human communication is speech, but to correctly perceive the underlying message the speaker's emotional intent also needs to be correctly perceived and recognized. Emotion recognition in daily life involves optimal integration of the visual and auditory signals conveyed by the speaker. Sensory impairments could thus impair emotion recognition, although it is also possible that any remaining intact senses can, at least partially, compensate for an impaired sense. However, as sensory impairments occur relatively often in older individuals[137], it is unknown whether general ageing or age-related cognitive decline confounds the effects of sensory impairments on emotion recognition, or whether older age could possibly increase the negative effects of sensory impairments by limiting compensatory abilities. As a step towards studying the effects of sensory impairments on emotion recognition in individuals with dual sensory impairments, here we examined the role of stimulus degradations on recognition accuracy and perceptual strategies for emotion recognition in older adults with normal hearing and vision. Additionally, we compared these with previous findings in younger adults with normal hearing and vision[138].

4.1.1 Sensory impairments and their effects on emotion recognition

The most common permanent sensory impairments in the older population that may impact emotion recognition are age-related hearing loss (affecting up to half of the elderly population)[113,137,139], which is generally a sensorineural hearing loss with decreased auditory sensitivity in higher frequencies[113,140], and age-related macular degeneration (AMD), where a deterioration of the macula leads to central vision loss (affecting up to twenty percent of the elderly population)[100,141]. While cataract is technically more common than AMD[142], cataract can be treated quite well and generally does not lead to permanent vision loss. Both age-related hearing loss and AMD can be expected to impact emotion recognition, because of the loss of auditory emotion cues and difficulty of seeing face details clearly, respectively. For hearing loss, it has been shown that both children and older individuals with hearing loss show poorer auditory emotion recognition than normal hearing controls[107–109,143]. While hearing aids were shown to improve emotion recognition marginally, they do not seem to restore emotion recognition to the levels of normal hearing younger or older listeners[110]. Additionally, individuals with AMD show poorer facial emotion recognition than controls[102,103], and this difference remained even when the stimulus was magnified up to twice its original size. In addition, eye movements of AMD individuals were much more variable in position than eye movements of controls[103].

However, it is unclear whether existing findings in individuals with unimodal sensory impairments are mostly due to the effect of degraded sensory input, or confounded by a general ageing effect, a long-term adaptation to the impairment, or cognitive decline brought about by ageing or the impairments. For example, Orbelo et al.[111] found that elderly participants (between 65 and 83 years of age) that had mild hearing loss and did not wear hearing aids showed decreased auditory emotion recognition. However, this decrease could not be explained by their hearing loss, nor by age-related cognitive decline, leading the authors to conclude that the decrease was related to a general ageing effect. It should be noted that in this study, the participants had pure-tone hearing thresholds of on average 24 dB HL (±12 dB, average of both ears) only. Consequently, the participants' hearing loss may have been too mild to measurably affect their performance.

In addition to the possible confounding effect of age, studies on the effects of unimodal sensory impairments on emotion recognition give little to no insight about the possible consequences of dual sensory impairments for emotion recognition, which can occur relatively frequently, i.e., up to thirty percent, in the older population[137,144–147]. Therefore, as a first step, in a previous study[138], we established the individual and combined effects of audio and video degradations on emotion recognition in a healthy group of young volunteers. By means of stimulus degradations, we intended to approximate some of the purely sensory and instantaneous consequences of hearing and vision impairments in simulation, that is, the lack of sensory input only, and not long-term adaptation or cognitive changes that may occur in real sensory impairments. The audio and video signals were degraded to mimic a moderate age-related sensorineural hearing loss and a relative central scotoma (i.e., reduced sensitivity within the scotomatic region, but not a complete loss of perception), respectively. We found that isolated audio and video degradations, that is, presenting degraded audio or video without presenting the corresponding other sense, decreased emotion recognition performance to a similar degree. However, while presenting degraded video alongside normal audio decreased performance, presenting degraded audio alongside normal video did not affect performance. Moreover, degrading both the audio and the video led to a similar performance decrease as only degrading the video. Thus, for dynamic video stimuli at least, the isolated effects of degradation do not necessarily get exacerbated when combined. Moreover, intact vision may compensate for degraded audio, but intact audio cannot compensate for degraded vision. In addition, as evidenced by eye-tracking, we found that participants adapted their perceptual strategies, by making larger saccades and looking away from the face of the actor, in response to video degradations, but not to audio degradations. These adaptations may compensate to a certain degree for the visual degradation, but this is unknown.

4.1.2 The effects of age on emotion recognition

What remains unclear is whether this compensation also occurs in older adults, especially considering that there is evidence from previous research for a global decline in emotion recognition with ageing[148,149]. It has been proposed that audiovisual emotion recognition peaks between 15 and 30 years of age and declines linearly after that[150]. Despite this general decline, there is some evidence that older adults benefit more from multimodal stimuli than younger adults such that no age-related deficits could be established in the multimodal conditions[122,151]. However, others found that older adults show a similar benefit to younger adults from multimodal stimulus presentation, such that the age-related difference observed in unimodal stimuli is preserved for multimodal stimuli[152]. In addition to a possible increased benefit from multimodal stimuli, there is some evidence for preserved or even superior emotion recognition in older adults for positive emotions, especially happiness, the so-called 'positivity effect'[123,153–155], at least for the recognition of facial expressions. The positivity effect is proposed to arise from an attentional bias towards positive and away from negative emotions[156,157]. In summary, it is thus unclear whether an age-related deficit in emotion recognition will be found when using multimodal stimuli, especially for displays of positive emotions.

Besides changes in emotion recognition ability across the lifespan, there is also evidence from eye-tracking studies that older adults view emotional expressions differently from young adults. Studies have found that older adults tend to focus more on the mouth or bottom half of the face than on the eyes or top half, whereas young adults show a reversed tendency[158,159]. In the study by Wong and colleagues[159], looking at the bottom half of the face was negatively correlated with recognition accuracy for the emotions of anger, fear, and sadness, providing a straightforward explanation for why older adults may have impaired recognition of negative emotions.

4.1.3 The current study

Real-life emotion recognition almost always involves dynamic stimuli (i.e., during face-to-face conversations) and both younger and older adults seem to recognize dynamic emotion stimuli somewhat better than static emotion stimuli[160,161]. Therefore, aiming for good ecological validity, in our present study we presented dynamic stimulus presentation, in the form of short movie-clips, to healthy young and older adults. These stimuli and the applied stimulus degradations were the same as in our previous study[138]. The use of these simulations creates a homogeneous fictitious "patient" group, while the use of two age groups allow disentangling the effects of (simulated) hearing and vision impairment from general ageing effects. We used eye-tracking to examine perceptual strategies (i.e., determine when observers look where and what kind of eye-movements they make to achieve this), especially important here as some studies have shown that older adults view emotional expressions differently than younger adults.

Based on existing literature, we expected that older participants would have worse performance on emotion recognition than young participants[148,149]. However, for intact audiovisual stimulus conditions, older adults might recognize the expressed emotions with the same accuracy as younger adults, owing to previous work showing that audiovisual integration provides a larger benefit for older adults[122,151]. Additionally, we expected that older adults would perform as well as, or even better than young adults for positive emotions, both in intact audiovisual and unimodal conditions[123,153–155]. Finally, in our previous work[138] with young normal hearing adults, we found evidence for compensation, as degraded audio did not reduce performance if it was accompanied by any video, regardless of whether the video was intact or degraded. As cognitive functioning declines with age[162], we expected that older adults might not compensate for degraded information as well as young adults.

Since gaze allocation is a flexible information-seeking process[45,46,114], we expected that gaze patterns would differ between conditions as well as between age groups. In line with previous findings[158,159], for all conditions we expected that older adults would fixate more on the lower facial features (specifically, the mouth) than on the upper facial features (the eyes) compared to younger adults. Additionally, older adults' gaze adaptations to degradations could either be similar to those of young adults or these would be less adaptive or even entirely different from those of young adults. Finally, in line with our hypothesis that age effects in performance would be neutralized in multimodal conditions, we expected that this would also hold for gaze such that older participants would attend more to the upper facial features in the multimodal conditions compared to the unimodal conditions.

4.2 Methods

In the present experiment, both performance and eye-tracking data were obtained to identify accuracy of emotion recognition and gaze patterns during emotion perception with dynamic stimuli, respectively. The methods, including stimuli, procedures, and analyses used in this study closely resemble those used in previous studies by the authors[114,138]. The original – unmodified – stimulus materials were first described in[36].

In the study by de Boer et al.[138], emotion recognition performance and gaze behavior were studied in young, healthy observers that viewed the stimuli in three modalities: with audio and video combined, only the video, or only the audio. Their study aimed at understanding basic aspects of audiovisual integration under sensory degradations. The data collected in our present study in healthy older adults is compared to their data[138]. Lastly, for an informal comparison, preliminary data from five individuals with macular degeneration and hearing loss (called patient participants from here on) are included here.

4.2.1 Participants

Twenty-four healthy, native Dutch participants, selected to be over sixty years old and self-reported to have normal vision (or corrected-to-normal vision) and normal hearing, volunteered to take part in the experiment (twelve males, mean age = 66 years, *SD* = 3.2, range: 61-72). All participants were given sufficient information about the nature of the experiment, but were otherwise naïve as to the exact purpose of the study. Two participants did not complete the experiment because their glasses proved incompatible with the eye-tracker. One participant did not complete the experiment because the need to be in the headrest for the eye-tracking measurements made the participant uncomfortable. Therefore, a total of 21 participants completed the entire experiment (ten males, mean age = 66 years, *SD* = 3.4, range: 61-72).

In addition to the data collected here, a previously collected dataset for a different study with similar methods[138] containing data from 24 young, healthy, and native Dutch participants (nine males, mean age = 23 years, *SD* = 2.9, range: 19-29) was used as a control dataset in the present study to test for ageing effects.

Written informed consent was obtained prior to screening and data collection. The study was carried out in accordance to the Declaration of Helsinki and was approved by the local medical ethics committee (ABR nr: NL60379.042.17). All participants received a payment of €8,00 per hour for their participation.

4.2.2 Screening

Participants' eyesight and hearing were tested before the experiment. Normal visual functioning was assessed with measurements of visual acuity and contrast sensitivity (CS), using the Freiburg Acuity and Visual Contrast Test (FrACT, version 3.9.8)[68,69]. Normal vision was considered as a visual acuity (VA) of at least 0.80 and a logCS of at least 1.80 (corresponding to a luminance difference of approximately 1% between target and surround). Visual tests were performed binocularly and on the same computer and screen as used in the main experiment, with participants wearing their regular glasses or contact lenses. Auditory functioning was assessed by measuring auditory thresholds for pure tones at audiometric test frequencies between 125 Hz and 8 kHz. Auditory thresholds were determined using a staircase method based on typical clinical procedures. The participant sat inside a soundproof booth during audiometric testing and testing was conducted on each ear, always starting with the right ear. Since some hearing loss is nearly unavoidable in older populations[113], we have used a somewhat relaxed criterion for normal hearing compared to typical clinical procedures. For older participants, we aimed for the normal hearing definition from the European Working Group on Genetics of Hearing Impairment[163], where the pure-tone average (PTA; the average sensitivity at 500Hz, 1kHz, 2kHz, and 4kHz) is to be as good as or better than 20 dB HL at the better ear.

Four older participants did not have normal vision and five older participants did not have normal hearing according to our criteria (i.e., visual acuity < 0.8 and/or PTA > 20 dB HL). Two participants had both non-normal vision and non-normal hearing. As a result, in total, despite perceiving themselves as normal seeing and normal hearing, seven participants did not have normal vision and/or hearing according to the criteria listed above. We still opted to keep these participants in the experiment to maintain a good number of participants. Additionally, given that they self-reported to have normal vision and hearing, these participants could still be considered representative of the aimed age group. Visual acuity, contrast sensitivity levels, and audiometric thresholds for all participants are shown in Figure 1, and individual visual acuity, contrast sensitivity, and PTA's are displayed in Supplementary Table A.1.

Besides hearing and vision, cognitive functioning of healthy older participants was screened for using the Montreal Cognitive Assessment (MoCA). All included participants scored at or above the cut-off for normal cognitive functioning (26 points). Additional exclusion criteria were neurological or psychiatric disorders, dyslexia, and the use of medication that could influence normal brain functioning.

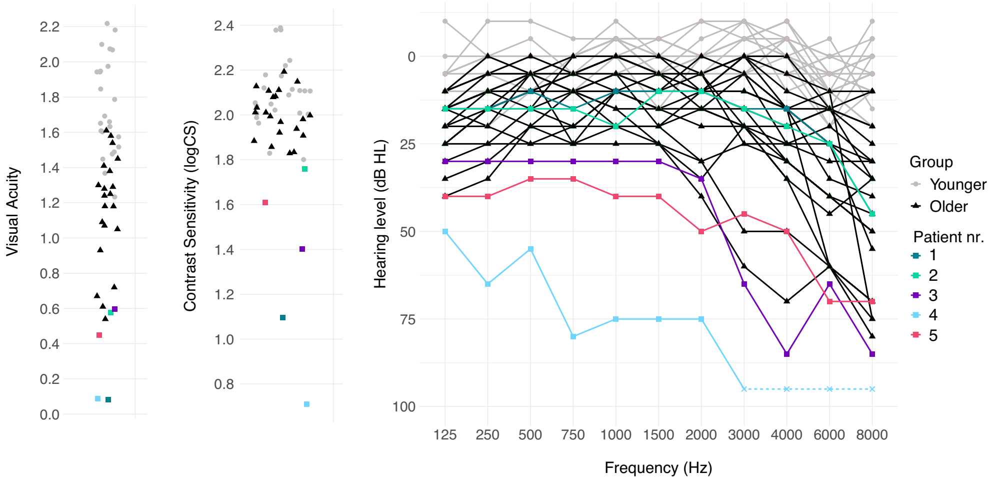


Figure 1. Individual levels of visual acuity (left) and contrast sensitivity (middle, in logCS), measured binocularly, for younger, older, and patient participants. Left: individual hearing thresholds in dB HL for the better ear for younger, older, and patient participants. Note: one patient participant (4) did not respond when the frequencies ≥ 3000 Hz were presented at 90 dB HL, at which point testing stopped to not further damage hearing. The thresholds in the figure were set at 95 dB HL to indicate this, the actual hearing thresholds for those audiometric test frequencies are unknown.

4.2.3 Stimuli

Audiovisual emotional expressions taken from the Geneva Multimodal Emotion Portrayals (GEMEP) core set[36] were used as stimuli during the experiment. A short demo showing only the face of the actor can be found at the Geneva Emotion Recognition Test (GERT) demo at: https://www.unige.ch/cisa/emotional-competence/home/exploring-your-ec/. The GEMEP core set consists of 145 audiovisual video recordings (mean duration: 2.5 s, range: 1-7 s) of emotional expressions portrayed by ten professional French-speaking Swiss actors (five females) of different ages (mean: 37.1 years, range: 25-57 years). The lexical content of the expressions was one of two pseudo-speech sentences with no semantic content, but resembling the phonetic sounds in western languages ("nekal ibam soud molen!" and "koun se mina lod belam?"). Out of the 17 emotions in GEMEP, 12 were selected for the main experiment, such that they would be equally distributed over the quadrants of the valence-arousal scale. See Table 1 for the 12 emotions and how they are distributed over the valence-arousal scale[70]. Portrayals from two actors that were found to be less clearly recognizable in previous work[114] were used during practice trials to familiarize participants with the stimulus materials and the task. Thus, a total of 96 unique stimuli were used in the main experiment and a total of 24 unique stimuli in the practice trials.

Table 1. The selected emotion categories. The emotions are evenly distributed over the quadrants of the valence-arousal scale[70].

		Valence	
		Negative	**Positive**
Arousal	**High**	Amusement Joy Pride	Fear Despair Anger
	Low	Pleasure Relief Interest	Irritation Anxiety Sadness

4.2.4 Visual stimulus degradation

A gaze-contingent relative scotoma was produced using custom MATLAB scripts. A semi-circular, yet irregular, shape that was centered on gaze position, was used to mimic the estimated vision loss in an individual with progressed binocular age-related macular degeneration (AMD), see Figure 2b-c. The shape of the scotoma was not based on an actual scotoma, but based on the fact that the macula spans a roughly circular region in central vision. However, as the vision loss of an individual with AMD will hardly ever be perfectly circular, an irregular shape was used. The scotoma was shown in one of four different orientations in each trial: original (as in Figure 2b), horizontally flipped, vertically flipped, and both horizontally and vertically flipped. Orientation was randomized between trials. The scotoma's size was roughly 17° x 11.5° visual angle (VA; 731 x 497 pixels) and had soft edges. Most AMD individuals do not perceive a hole in the location of their visual field defect, but distortions or blur[112]. Because of this, we decided to blur rather than remove the region of the video that the scotoma covered. A Gaussian low-pass filter (using the MATLAB function *fspecial* and *imfilter*), was used to create a blurred version of the video. The filter had a cut-off frequency (at full width at half maximum, FWHM) of 0.15 cycles/deg. Then, this filtered version was overlaid on the original – unfiltered – video, and the alpha-layer of the scotoma image (see Figure 2b) served to indicate what region of the video should be hazy and how strongly.

Participants were informed that the scotoma was gaze-contingent and that they could use compensatory eye-movements in order to peripherally view regions in the video they found relevant. Participants were informed that looking away from the actor could help them in still seeing the expressed emotion on the video, but were informed neither on the direction nor on the size of the eye-movements they should make in order to do so.

4.2.5 Auditory stimulus degradation

Degradation of the audio signal was done using customized MATLAB scripts aimed at approximating three characteristics of sensorineural hearing impairment: increased absolute thresholds, loudness recruitment, and the effects of broader auditory filters on narrowband envelopes in the auditory system. The processing used here was inspired by the hearing impairment (HI) simulation of Nejime and Moore[125]. The audio manipulation consisted of two sequential modules: one for envelope processing, and one for loudness perception. The envelope module was designed to produce perceptual effects of broader auditory filters (i.e., impaired frequency resolution), while the loudness module simulated raised audiometric thresholds and loudness recruitment.

The envelope-processing module created narrowband envelopes as they are assumed to be present in the impaired auditory system via broader auditory filtering, while the fine structure should be preserved as in normal hearing. Therefore, the input audio signal was processed with a Gammatone filter bank with bandwidths of two equivalent rectangular bandwidths (ERBs), representing impaired auditory filters, at one ERB distance across center frequencies between 80 Hz and 10 kHz. The filter bandwidth of two ERB was selected as representative for moderate sensorineural HI[126]. Within each frequency band the envelope was extracted using the Hilbert transform, which each served as the target HI envelope. Hilbert envelopes from broader filters were then multiplied onto Hilbert fine structure signals in each frequency band. Normal narrowband envelopes can be partially recovered from a NH fine structure signal by NH listeners[127]. To minimize this unwanted recovery of envelopes, thus to provide "degraded envelopes" inside the normal auditory system of the participants in this study, an iterative procedure was used whereby the output of the multiplication procedure was passed through the NH filter bank again and the fine structure extracted using the Hilbert transform was multiplied again by the target impaired envelopes. Ten such iterations were used in the present study, resulting in a high average correlation coefficient of 0.83 with the desired HI envelopes after modeled NH auditory processing using speech as a signal[128].

After the envelope processing module, the loudness module sets the sound level in each frequency band such that the NH participants listening to this simulation had a similar loudness perception as an (average) HI listener. For this manipulation, the output signal of the envelope-processing module was fast Fourier transformed (FFT-ed) into six octave-spaced channels with frequencies between 250 Hz and 8 kHz. The sound level in each channel was extracted from the output signal and the categorical loudness ratings as used in the procedure of Brand & Hohman[129] were calculated based on average HI categorical loudness data[130], which served as target loudness. The sound levels were then attenuated in an expansive fashion such that (average) NH listeners' loudness perception of the sound level matches the target HI loudness. Finally, the spectral signal was transformed back into the time domain using the inverse FFT. The loudness module thus also set the simulated audiometric thresholds. For the present study the degradations were implemented by taking the thresholds from a moderate hearing impairment, similar to the standard audiogram N3 as defined in Bisgaard et al.[131]. The thresholds were 40, 40, 45, 54, 62, and 70 dB HL at audiometric frequencies of 250, 500, 1000, 2000, 4000, and 8000 Hz, respectively.

After these two modules, the sound level of the final output signal was root-mean-square (RMS) equalized to the intact audio, to ensure any effects found were not a side-effect of an overall decrease in presentation level.

4.2.6 Experimental set-up

The experiment was performed in a dark and quiet room, with the monitor providing the only illumination. Participants sat in front of the monitor at a viewing distance of 70 cm with their head placed in a chin- and forehead rest to minimize head movements. Stimuli were presented full-screen on a 24.5-inch monitor with a resolution of 1920 x 1080 pixels (43 x 24.8 degrees). The average screen luminance was 38 cd/m^2, measured from the approximate head location of the participant. An Apple MacBook Pro (mid 2015 model) was connected to the monitor and controlled the stimulus presentation. The audio was produced by the internal soundcard of this computer and presented binaurally through Sennheiser HD 600 over-ear headphones (Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany). The sound level was calibrated to be at a comfortable and audible level, at a long-term RMS average of 65 dB SPL. Participants used an external mouse for responding. Stimulus display and response recording was controlled using the Psychophysics Toolbox (Version 3)[71–73] and Eyelink Toolbox[74] extensions of MATLAB (Version R2015b; The Mathworks, Inc., Natick, MA, USA).

Participant's eye movements were measured with an Eyelink 1000 Plus eye-tracker (SR Research Ltd., Ottawa, Ontario, Canada), running software version 4.51. Monocular gaze data was acquired at a sampling frequency of 1000 Hz. The eye-tracker was located just below the monitor. A calibration procedure preceded the experiment using the built-in 9-point calibration routine. Calibration accuracy was verified with the validation procedure in which the same nine points were displayed again. The experiment would start if the calibration accuracy was sufficient (i.e., average error of less than 0.5° and a maximum error of less than 1°). Drift was checked for after every fourth trial and after each break. The calibration procedure was repeated if the participant moved during breaks and whenever there was more than 1° of drift in more than one consecutive drift check.

Table 2. Experimental conditions used in the experiment. Both modalities were either shown as they are (intact), degraded, or absent.

		Video		
		Intact	**Degraded**	**Absent**
Audio	**Intact**	AV	AdV	A
	Degraded	dAV	dAdV	dA
	Absent	V	dV	

4.2.7 Procedure

During the experiment, participants were asked to identify the emotions expressed in the GEMEP core set videos. The videos were presented in eight different stimulus presentation conditions, listed in Table 2. Participants were asked to respond as accurately as possible in a forced-choice discrimination paradigm. Participants were further requested to blink as little as possible during the trial and maintain careful attention to the stimuli.

Each trial was preceded by either a full-screen image of the averaged frames of all videos (for all conditions with video, see Figure 2a) or a fixation cross (for A and dA; conditions without video), which was presented for a random duration between 600 and 1600 ms. For conditions with video, this averaged image was displayed instead of a fixation cross to allow participants to already orient their gaze, which could be especially beneficial in the conditions where a scotoma was present. Then, the stimulus was presented, for 1 to 7 seconds, depending on the specific video. For A and dA, the fixation cross remained on screen. After stimulus presentation, a response screen appeared. On this screen, all twelve emotions were presented with a label, grouped in a circular fashion by valence and arousal. The participant could, in a forced-choice response format, click with the mouse pointer on the emotion label that corresponded to the identified emotion. All twelve emotions were always presented on the response screen, and grouped in a circular fashion by valence and arousal. The response screen remained visible until a response was made. The participant's response (the emotion label) was recorded as well as whether the response was correct or not.

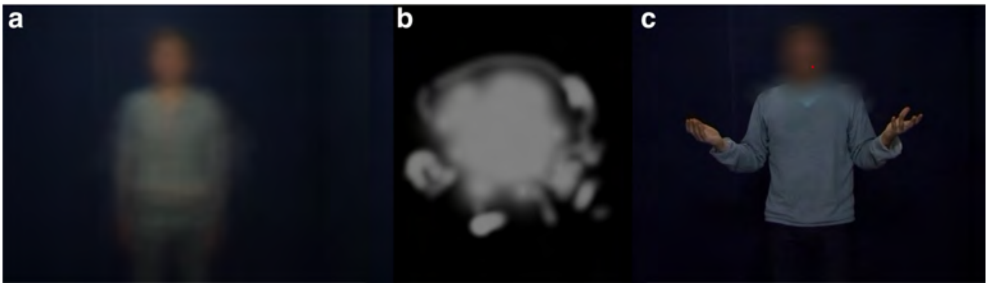


Figure 2. a) Still image created by averaging together all frames of all videos. This image preceded stimulus presentation in all conditions with video. b) Shape and approximate size of the scotoma mask. The scotoma was gaze-contingent with its center positioned on the point of gaze. Four different orientations were used during the experiment (randomly intermixed): as shown in this figure, left-right flipped, up-down flipped, and left-right and up-down flipped. c) Scotoma overlaid on a still image of one video. The red dot indicates the point of gaze, this dot was not visible to participants.

Each participant was presented with all 96 videos (twelve emotions x eight actors) in all eight conditions, each individual video was thus presented eight times. The experiment was separated into six experimental blocks and in each block, all eight conditions were presented in sub-blocks containing one sixth of the stimuli (i.e., 16 trials per sub-block, 128 trials per experimental block). The order of conditions between experimental blocks was counterbalanced using balanced Latin Squares within and across participants. For young participants, the stimulus order for each condition was fully randomized. For older participants, the stimulus order was pseudo-randomized: they saw the videos from a set of four pseudo-randomly chosen actors (two male, two female) in the first session, and the videos from the remaining four actors in the second session. Stimulus order within each set of four actors was randomized. The reason for this change was that we had expected many older participants would drop out of the study after one session due to the length of the experiment. With this change, at least we would have balanced data after one session (i.e., all emotions presented equally often in all conditions). In the end, none of the older participants dropped out for this reason.

The experiment was preceded by 64 practice trials (eight practice trials for each condition) to acquaint the participants with the stimulus material and the task. For the practice trials, all conditions were presented in the following fixed order: AV, V, A, AdV, dAV, dV, dA, dAdV. Stimulus order within each practice block was randomized. During the practice block, participants received minimal feedback after each trial on their given response (i.e., correct/incorrect). No feedback was provided during the experiment.

Overall, the experiment consisted of 832 trials, including the 64 practice trials, and took about 2.5 hours to complete. The experiment was separated over two test sessions performed on separate days to avoid fatigue. Participants were able to take a self-paced break every 32 trials and were encouraged to take breaks in order to maintain concentration and prevent fatigue. The experiment continued upon a mouse-click from the participants and the eye-tracker was recalibrated if the participant moved during the break.

4.2.8 Data analyses

The data analysis was performed in two stages. The first analysis stage focused on intact conditions (A, V, and AV). The second stage focused on the effects of audio and video degradation (dA, dV, AdV, dAV, and dAdV). All data (that is, accuracy scores, fixation durations, saccadic amplitudes, and fixation proportions) were analyzed in R (version 3.6.0; R Foundation for Statistical Computing, Vienna, Austria — https://cran.r-project.org) with linear regression models (using *lmer* from the *lme4* package, version 1.1-21). Since our main interest was in the effect of age, only the main effects of *age group* and interactions with *age group* were followed-up by post-hoc tests. Other variables (e.g., *condition*, *emotion*) were added if they improved the model. For both stages, the best model was found by comparing Akaike Information Criterion (AIC) values for the different models. The criterion for picking a more complex model was an AIC decrease of at least two[164]. Significance of main effects and interactions of the final models were assessed with an Analysis of Deviance table (type III Wald chisquare test) with the *Anova* function from the *car* package (version 3.0-3). Significant effects were followed up by post-hoc tests to test how age groups differ. Post-hoc tests were performed using *lsmeans* from the *emmeans* package (version 1.4.1). Note that many of our analyses were exploratory, meaning that we did not have clear hypotheses a priori for these analyses (especially concerning the effect of different conditions for both age groups). In those cases, the corrections for multiple comparisons were generally not strict, and some of the findings may not survive more stringent corrections.

4.2.8.1 Analyses of behavioral data

Accuracy scores for each condition and emotion were first converted to unbiased hit-rates[77,114,138] to account for any response biases. The unbiased hit-rate (H_u) is different from the regular hit-rate in that it also considers false alarms. It can be calculated by squaring the number of correct responses for a category, and dividing that by the number of occurrences of that category times the number of times this particular response was used. In our study, an example of the H_u for the emotion Joy would be the $\text{Joy}_{\text{correct}}^2/(\text{Joy}_{\text{occurrence}}*\text{Joy}_{\text{responded}})$. Because of this, if a participant often responds to Joy correctly (i.e., $\text{Joy}_{\text{correct}}$ is high), but this is due to a bias towards responding Joy (i.e., $\text{Joy}_{\text{responded}}$ is high than $\text{Joy}_{\text{occurrence}}$), the unbiased hit-rate will be lower than the regular hit-rate to account for this bias. The unbiased hit-rates were arcsine transformed[165] to create a normal distribution. Then, a linear regression analysis was performed with the arcsine transformed H_u as the dependent variable.

In both stages, the base model included the *condition* (with three/eight levels), *age group* (young/old), and their interaction as fixed effects. Then, *participant* was included as a random intercept and *emotion* was included in steps (i.e., first as random intercept, then as main effect, then in interaction with age group and/or condition), making the model more complex with each step. Additionally, we tested whether the inclusion of random slopes for *condition* and/or *emotion* improved the model. As mentioned, the AIC was used to test whether the model improved with the added complexity and in addition, if the more complex model did not converge, the addition was excluded. Post-hoc tests were performed using Bonferroni correction for multiple comparisons in the first stage, and with the False Discovery Rate (FDR) correction for multiple comparisons in the second stage.

4.2.8.2 Analyses of eye-tracking data

For the eye-tracking data, the built-in data-parsing algorithm of the Eyelink eye-tracker was used to extract fixations from the raw eye-tracking data. Only data from conditions in which the video was present (all except A and dA) were analyzed, as in the conditions without video participants would have mostly been fixating on the fixation cross throughout the trial. All analyses were restricted to eye movements made during stimulus presentation, and only those made within 1000 ms after stimulus onset. No gaze data after 1000 ms were considered to limit data analysis to the duration of the shortest movie, which lasted 1000 ms. In addition, this aimed to discard any data that no longer was task-related, i.e. after a participant decided on a response, which is increasingly likely to occur at a longer interval after stimulus onset. Trials with single blinks longer than 300 ms during the first 1000 ms of stimulus presentation were discarded. Additionally, only trials with a correct response were included, as our main interest was in gaze behavior prior to correct recognition. Focusing on correct responses allowed examining whether changes in gaze behavior due to information degradation and availability of audio were adaptive and lead to good performance.

Mixed linear regressions were performed to test for the effects of *age group*, *condition*, and *emotion* on fixation durations and saccadic amplitudes. For fixation proportions, *AOI* was included as an additional fixed effect. Random intercepts were included for *participant* and *movie* and random slopes for *condition* were included if they improved the model.

Fixation durations and saccadic amplitudes were extracted from the parsed data file. Saccades with amplitudes larger than the diagonal of the monitor, which was 49.6 degrees, were filtered out, removing less than 1% of saccades. An exploratory mixed linear regression was performed for both fixation duration and saccadic amplitude.

Additionally, we performed an Area-of-Interest (AOI) based analysis on fixations for those conditions in which the video was present. For fixation proportions on AOIs, the eyes (left and right), nose, mouth, and hands (left and right) of the actors were chosen as AOIs. Because the stimuli are dynamic, the AOIs were dynamic as well. Coordinates of the AOI positions for each stimulus and each frame were extracted using Adobe® After Effects® (Version 15.1.1; Adobe Inc., San Jose, CA, USA). The coordinates for the face AOIs were obtained by applying the 'Face Tracking (Detailed Features)' method of Adobe® After Effects®, which automatically tracks many face features. Face track points at each frame were visually inspected and manually edited whenever the tracking software failed to track them correctly. For the hand AOIs, the 'Track Motion' method of Adobe® After Effects® was used. A single tracker point per hand was used to track position. The tracker point was placed roughly in the center of the hand. Again, tracking was inspected visually and manually edited where needed (for more details on face and hand tracking, see de Boer et al.[114]). Coordinates of all obtained face and hand track points for each stimulus were stored in text-files and used to create point AOIs. For the eyes we used the coordinates of the 'Left/Right Eyebrow Outer' for the *x*-position of the lateral corner, 'Left/Right Eyebrow Inner' for the *x*-position of the medial corner, 'Left/Right Eyebrow Middle' for the top, and the middle between the *y*-positions of 'Left Pupil' and 'Nose tip' for the bottom, indicating the eye–nose border. The individual AOIs for the left and right eye were later merged for analysis. For the nose we used the eye–nose border as the top, the nose–mouth border (middle between the *y*-positions of 'Right Nostril' and 'Mouth Top'), the *x*-position of 'Right Nostril' and the *x*-position of 'Left Nostril' for the lateral corners. For the mouth AOI: the *x*-position of 'Mouth Right' and the *x*-position of 'Mouth Left' for the lateral corners, the nose–mouth border for the top, and the *y*-position of 'Mouth Bottom' for the bottom. Each AOI was expanded by 10 pixels on each side (20 pixels across the horizontal and vertical axes), except at the eye–nose and nose–mouth borders. Overlap between AOIs was avoided. The actual size of each AOI varied across actors and frames e.g. due to some actors being closer to the camera. Note that left and right are in reference to the actor, not the observer. Thus, the left eye and hand are generally on the right side of the screen and vice versa for the right eye and hand.

Fixation proportions on the AOIs were defined as follows: for all of the N fixation time-points, the fixation proportion is the proportion of N that is located on a given AOI. These proportions were then averaged over each trial, resulting in a mean fixation proportion on each AOI for each trial. These means were finally arcsine-transformed. A mixed linear regression was performed on the arcsine-transformed mean proportions.

4.2.9 Data from patient participants

We collected data from five individuals (two males, mean age = 69, SD = 4.44, range: 66-77) with some form of macular degeneration and, for three cases, also some hearing loss. All patient participants were screened in the same way the healthy younger and older participants were screened. Unlike in the healthy older participants, the MoCA was not administered in patient participants because their vision and hearing loss may negatively affect the outcome and lead to the spurious conclusion that their cognitive functioning is poorer. In addition, standard automated perimetry (HFA Central 10-2 protocol) was obtained and all filled in the Dutch versions of the Speech and Spatial Qualities (SSQ 5.6, home version) and the Visual Functioning Questionnaire (VFQ-25/NL, home version) to assess how they experience their hearing and vision impairments. HFA results are included in Supplementary Figure A.1, questionnaire outcomes are summarized in Supplementary Tables A.2 and A.3.

For patient participants, the general set-up was the same as for healthy participants; each patient participant was presented with all 96 stimuli in sub-blocks of 16 trials. However, only the A, V, and AV conditions were used, which in principle should correspond to the dA, dV, and dAdV conditions because of the patient's vision and hearing impairments. The experiment was thus also preceded by only 24 practice trials (eight practice trials for each condition) in which the conditions were shown in the following order: AV, V, A. In total, the experiment for the patient participants consisted of 312 trials, including the 24 practice trials and took about 1.5 hours to complete. The experiment was completed in one session. We also collected eye-tracking data from the patient participants, but calibrating the eye-tracker properly proved impossible due to their central visual field defect. Therefore, the eye-tracking data from the patient participants was too noisy to properly analyze and we only describe patient participants' emotion recognition accuracy results.

4.3 Results

4.3.1 Age effects on accuracy for intact conditions

Emotion recognition performance is shown in unbiased hit-rates in Figure 3. Please note that while analyses were performed on the arcsine transformed $H_{u,}$ Figure 3 plots non-transformed H_u for interpretability. Figure 3 shows that, overall, the performance (quantified as unbiased hit rates) of older participants was lower than that of the younger ones. It also appears that, for both age groups, performance was lowest in A, intermediate in V, and best in AV. The best regression model (i.e., the most complex model with the lowest AIC value) to test this included *condition* (only A, V, and AV) in interaction with *age group*, *condition* in interaction with *emotion*, and *emotion* in interaction with *age group*. *Participant* was included as random intercept, with a random slope for *condition*. Thus, the formula for the final model was: H_{u_asin} ~ condition*age + condition*emotion + emotion*age + (condition|participant). All main effects were significant (all $p < 0.001$). Additionally, the interactions between *condition* and *emotion* ($Chi^2_{22} = 117.9$, $p < 0.001$) and between *age group* and *emotion* ($Chi^2_{11} = 37.4$, $p < 0.001$) were significant. The interaction between *age group* and *condition* was not significant

(*Chi*$^2_2$ = 5.3, *p* = 0.07).

Follow-up post-hoc tests on the main effect of *condition* confirmed that performance was lowest for A, intermediate at V, and best for AV (all *p* < 0.001). The significant main effect of *age group* confirmed that older participants performed poorer than younger participants (difference estimate = 0.17, *t* = 4.64, *p* < 0.001). Older participants performed significantly poorer than young participants for all emotions, except for the emotions Joy (difference estimate = 0.08, *t* = 1.52, *p* = 0.13) and Anxiety (difference estimate = 0.09, *t* = 1.91, *p* = 0.06), even though the latter differences were in the same direction as for the other emotions.

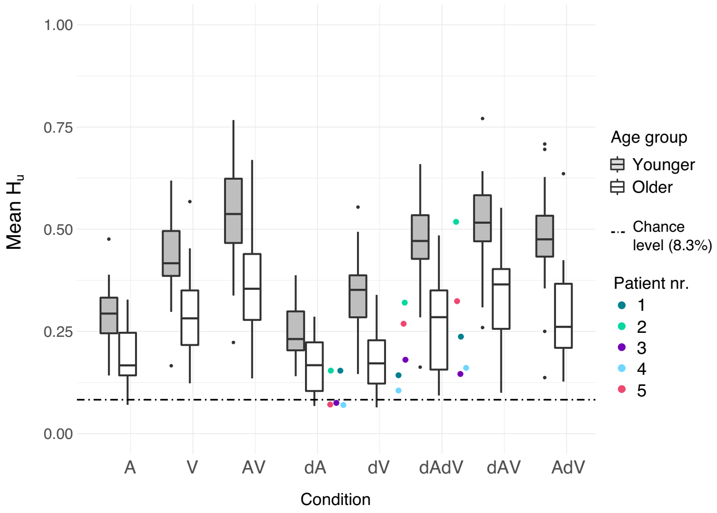


*Figure 3. Task performance for each condition and age group, shown as unbiased hit-rates. Performance is averaged across emotions and blocks. Each box shows the data between the first and third quartiles. The horizontal solid line in each box denotes the median. The whiskers extend to the lowest/highest value still within 1.5 * inter-quartile range (IQR), data outside the 1.5 * IQR are plotted as dots. Performance for young participants is shown in light grey boxes, performance for older participants is shown in white boxes. Performance for individual patient participant is shown in the colored dots. Note that these participants did not receive degraded stimuli, but their hearing and visual acuity tests indicate that their perception is degraded. Thus, for patient participants, dA corresponds to stimuli presented in A, likewise for dV and dAdV. The dashed line indicates chance level performance.*

4.3.2 Age effects on accuracy for degraded conditions

To investigate the effects of degradations, a regression model with *condition (all conditions)* in interaction with *age group*, *condition* in interaction with *emotion*, and *emotion* in interaction with *age group* was performed. *Participant* was included as a random intercept, but without a random slope for *condition*, as this led to a singular fit. Thus, the formula for the final model was: H_{u_asin} ~ condition*age + condition*emotion + emotion*age + (1|participant) All main effects were significant (all *p* < 0.001). Additionally, there were significant interactions between *condition* and *emotion* (*Chi*$^2_{77}$ = 393.7, *p* < 0.001), between *age group* and *emotion* (*Chi*$^2_{11}$ = 108.5, *p* < 0.001), and between *age group* and *condition* (*Chi*$^2_7$ = 42.9, *p* < 0.001).

Follow-up post-hoc tests showed that older participants had lower accuracy for all conditions (all $p < 0.002$) and all emotions (all $p < 0.009$), including positive ones. The significant interaction between *age group* and *emotion* indicates that the differences between younger and older participants were not the same for all emotions. Additionally, while the patterns across conditions appeared very similar for both age groups, there were subtle differences, see Table 3. For instance, degrading video seemed to reduce performance more in older than in younger participants. Note that Table 3 only lists sensible comparisons, e.g., A is compared to dA, but not to dV.

Additionally, Figure 3 shows that the five patient participants that were included had a similar emotion recognition accuracy as the included older healthy participants had in the degraded A, V, and AV conditions. These preliminary data support the idea age-related sensory changes can affect audiovisual emotion recognition, and our degradations captured some of these effects in individuals with no sensory impairments.

Table 3. Contrasts for the age group by condition interaction for recognition accuracy. The table shows the model estimate differences with the FDR adjusted p-values in parentheses. Significant differences are indicated by bold typeface.

Comparison	Age group	
	Younger	Older
A – dA	**0.06 (0.001)**	**0.05 (0.008)**
V – dV	**0.09 (<0.001)**	**0.16 (<0.001)**
AV – dAdV	**0.08 (<0.001)**	**0.12 (<0.001)**
AV – dAV	0.02 (0.337)	0.03 (0.109)
AV – AdV	**0.06 (<0.001)**	**0.09 (<0.001)**
dAdV – dAV	**-0.07 (<0.001)**	**-0.09 (<0.001)**
dAdV – AdV	-0.02 (0.180)	-0.03 (0.109)

4.3.3 Effects of auditory and visual functioning on emotion recognition accuracy

Overall, older participants had poorer hearing and vision than the younger participants, even though the older participants perceived themselves as having normal hearing and vision. This was tested by a two-sample t-test (function *t.test* from the R *stats* package, version 4.0.3), equal variances not assumed. The differences between younger and older participants were significant for all screening outcomes: PTA ($t_{26.1} = -6.86$, $p < 0.001$, $\text{mean}_{\text{younger}} = 0.89$, $\text{mean}_{\text{older}} = 14.46$), visual acuity ($t_{39.9} = 6.67$, $p < 0.001$, $\text{mean}_{\text{younger}} = 1.75$, $\text{mean}_{\text{older}} = 1.16$), and contrast sensitivity ($t_{41.2} = 2.55$, $p = 0.015$, $\text{mean}_{\text{younger}} = 2.10$, $\text{mean}_{\text{older}} = 2.0$). Because of these differences, an additional model was constructed that included *PTA*, visual acuity (*VA*), and contrast sensitivity (*CS*): $H_{\text{u\_asin}}$ ~ condition*age + condition*emotion + emotion*age + PTA + VA + CS + (1|participant). However, the effects of *PTA* ($Chi^2_1 = 0.46$, $p = 0.50$), *VA* ($Chi^2_1 = 0.06$, $p = 0.81$), and *CS* ($Chi^2_1 = 1.16$, $p = 0.28$) were not significant while the effect of *age group* ($Chi^2_1 = 5.48$, $p = 0.02$) was still significant, indicating that the poorer hearing and vision of the older participants seemed not to be the reason for their lower emotion recognition accuracy.

4.3.4 Age effects on fixation duration for intact conditions

Figure 4 shows that, on average, older participants tended to have shorter fixation durations than younger participants. In addition, there seems to be a small effect of condition.

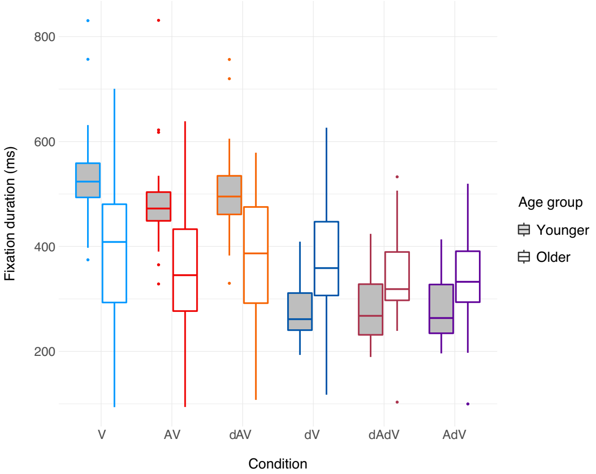


*Figure 4. Fixation durations in ms for all conditions and age groups. As for Figure 3, fixation durations are averaged across emotions and blocks. Each box shows the data between the first and third quartiles. The horizontal solid line in each box denotes the median. The whiskers extend to the lowest/highest value still within 1.5 * IQR, data outside the 1.5 * IQR are plotted as dots. Performance for young participants is shown in light grey boxes, performance for older participants is shown in white boxes.*

The regression models confirmed this. The best model included *condition* and *age group* as main effects only, a random intercept for *participant*, with a random slope for *condition*, and a random intercept for *movie*. The formula for the final model was: duration ~ condition + age + (condition|participant) + (1|movie). The main effects of *condition* (Chi^2_1 = 27.6, p < 0.001) and of *age group* (Chi^2_1 = 12.3, p < 0.001) were significant. A follow-up of these main effects showed that fixations were of longer duration in the V compared to the AV condition (difference estimate = 53.2, t = 5.25, p < 0.001). Additionally, older participants made fixations of shorter duration than younger participants (difference estimate = 126, t = 3.43, p = 0.001).

4.3.5 Age effects on fixation duration for degraded conditions

From Figure 4, it can be seen that younger participants adapt their gaze to the degraded video by making fixations with a shorter duration. Older participants do not seem to show the same adaptation, or they do so to a smaller degree. The best model to test this included *age group* and *condition* as main effects as well as their interaction. Random intercepts were included for *participant* and *movie*, but without any random slopes as these led to a singular fit. Thus, the formula for the final model was: duration ~ condition*age + (1|participant) + (1|movie). Both the main effect of *age group* (Chi^2_1 = 18.0, p < 0.001) and of *condition* (Chi^2_5 = 2243.3, p < 0.001) were significant, as well as the interaction between *condition* and *age group* (Chi^2_5 = 599.1, p < 0.001).

Post-hoc tests of the interaction between *condition* and *age group* showed that, in general, participants decreased fixation duration in conditions with degraded video. However, the differences were much smaller for older participants than for younger participants. For younger participants, the decrease in mean fixation durations with degraded video compared to intact video was significant (all $p < 0.001$) and on average 225 ms, while for older participants the average decrease was significant in most cases ($p < 0.013$), except for the comparisons between AV and dV ($p = 0.507$) and dAV and dV ($p = 0.340$), but was only 11 ms. There even appeared to be a small increase in fixation duration when comparing AV and dV in older participants, although this difference was not significant (difference estimate = -7.6, p = 0.507). For both groups, fixation durations were longest in the V condition and fixation duration did not differ between AV and dAV.

4.3.6 Age effects on saccadic amplitude for intact conditions

Figure 5 shows that older adults generally made saccades with a smaller amplitude than younger adults. The regression models confirmed this. The best model included *condition* and *age group* as main effects only, a random intercept for *participant*, with a random slope for *condition*, and a random intercept for *movie*. The formula for the final model was: amplitude ~ condition + age + (condition|participant) + (1|movie).

The main effects of *condition* (Chi^2 (1) = 13.0, $p < 0.001$) and of *age group* (Chi^2 (1) = 15.9, $p < 0.001$) were significant. A follow-up of these main effects showed that saccades were larger in the AV compared to the V condition (difference estimate = 0.24, t = 3.60, $p < 0.001$). In addition, older participants made smaller saccades than younger participants (difference estimate = 1.01, $t = 3.91$, $p < 0.001$).

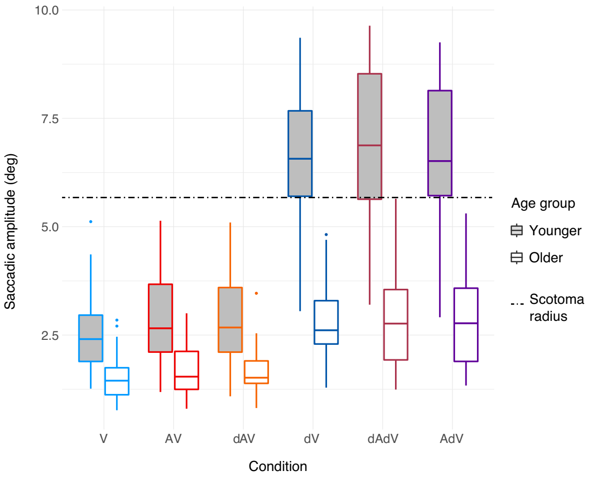


*Figure 5. Saccadic amplitudes in degree of visual angle for all conditions and age groups. Amplitudes are averaged across emotions and blocks. Each box shows the data between the first and third quartiles. The horizontal solid line in each box denotes the median. The whiskers extend to the lowest/highest value still within 1.5 * IQR, data outside the 1.5 * IQR are plotted as dots. Performance for young participants is shown in light grey boxes, performance for older participants is shown in white boxes. The dashed line indicates the minimal radius of the scotoma.*

4.3.7 Age effects on saccadic amplitudes for degraded conditions

From Figure 5, a similar result to what was observed for fixation duration, is seen for saccadic amplitudes. Participants adapt their gaze to degraded video by making larger saccades in those conditions, but older participants seem to make smaller adjustments than younger ones.

The regression model confirmed this. The best model included *condition* and *age group* as main effects as well as their interaction. Random intercepts were included for *participant* and *movie*, but without any random slopes as these led to a singular fit. The formula for the final model was: amplitude ~ condition*age + (1|participant) + (1|movie). Both the main effects of *condition* (*Chi*^{2} (5) = 4713.4, $p < 0.001$) and *age group* (*Chi*^{2} (1) = 12.6, $p < 0.001$), as well as the interaction (*Chi*^{2} (5) = 889.2, $p < 0.001$) were significant.

The follow-up post-hoc comparisons had results similar to those for fixation duration. All participants adapted their gaze to degraded video by making larger saccades, although the differences were smaller for older participants. For younger participants, the increase in saccadic amplitudes for degraded video conditions was on average 3.70 deg (from 2.83 deg in intact video conditions to 6.54 deg in degraded video conditions), while for older participants the increase was only 1.20 deg (from 1.58 deg in intact video conditions to 2.78 deg in degraded video conditions). The increases in saccadic amplitudes were significant for all comparisons between degraded and intact video conditions and for both age groups (all $p <$ 0.001) Additionally, only younger participants made significantly smaller saccades in V compared to the AV and dAV conditions (AV – V = 0.29, $p = 0.005$; dAV – V = 0.24, $p = 0.015$). For older participants there was a trend in the same direction (AV – V = 0.16, $p = 0.225$; dAV – V = 0.17, $p = 0.225$).

4.3.8 Age effects on fixation proportions for intact conditions

Figure 6 shows that all participants fixate more on the face than on the hands of the actors. Additionally, it appears that younger participants distribute their fixations more or less equally across the face AOIs, but that older participants focus mostly on the mouth.

The final model included *age group* in interaction with *AOI*, and *AOI* in interaction with *emotion*. Random intercepts were added for *participant* and *movie*, but no random slopes were added as these led to a singular fit. *Condition* did not have a significant effect on fixation proportions, both as a main effect and in interaction with any of the other variables (all $p >$ 0.33) and was therefore taken out of the final model. Thus, the formula for the final model was: proportion ~ AOI*age + AOI*emotion + (1|participant) + (1|movie). All main effects were significant (all $p < 0.001$), as well as the interaction between *age group* and *AOI* (*Chi*^{2} (3) = 263.3, $p < 0.001$) and between *AOI* and *emotion* (*Chi*^{2} (33) = 122.4, $p < 0.001$).

A follow-up of the interaction between *age group* and *AOI*, using an FDR-corrected post-hoc test, showed that older participants fixated more often on the mouth than younger participants (difference estimate = 0.08, $t = 4.33$, $p < 0.001$), but less often on the eyes (difference estimate = 0.16, $t = 8.97$, $p < 0.001$).

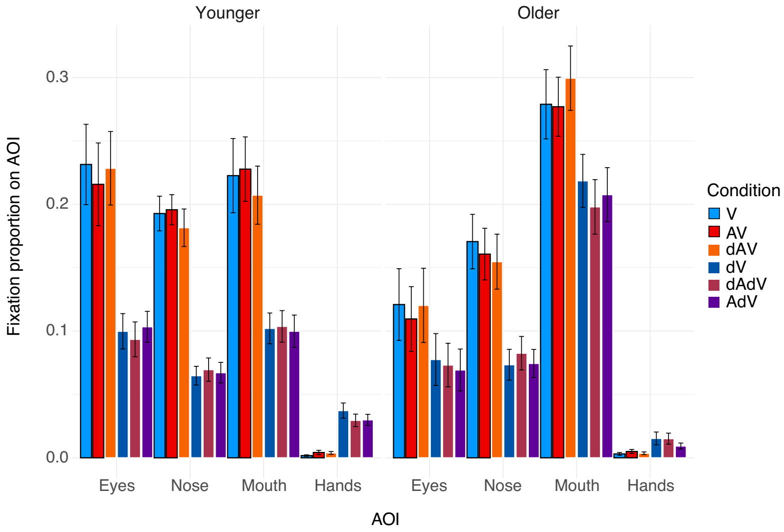


Figure 6. Mean fixation proportions on the face and hand AOIs (Areas of Interest) for all conditions and both age groups, and averaged over emotions and blocks. Error bars denote the standard error of the mean (SEM). Intact conditions are indicated by a black outline.

4.3.9 Age effects on fixation proportions for degraded conditions

Fixation proportions for all conditions and both age groups are shown in Figure 6. For both age groups, participants fixated less on the face AOIs in conditions with degraded video. Additionally, the bias for older participants to fixate more on the mouth was also present for the dAV condition, perhaps even stronger, and remained present under degraded video.

The best regression model included main effects of *AOI*, *age group*, *condition*, and *emotion*, as well as interactions between *AOI*, *age group*, and *condition*, and between *AOI* and *emotion*. Thus, the final model formula was: proportion ~ AOI*age*condition + AOI*emotion + (1|participant) + (1|movie). All main effects were significant (all $p < 0.012$). Additionally, there were significant interactions between *age group* and *AOI* (*Chi*^{2} (3) = 10.0, $p = 0.018$), between *AOI* and *condition* (*Chi*^{2} (15) = 1023.8, $p < 0.001$), *AOI* and *emotion* (*Chi*^{2} (33) = 59.1, $p = 0.003$), and between *age group*, *AOI*, and *condition* (*Chi*^{2} (15) = 130.6, $p < 0.001$). Because our main interest was in age effects, only the interactions between *age group* and *AOI*, and between *age group*, *AOI*, and *condition* were followed-up with post-hoc tests.

The *age group*-by-*AOI* interaction showed that, overall, young participants fixated significantly more often on the face AOI than the hands (all $p < 0.001$), with no differences between the fixation proportions on the face AOI (all $p > 0.266$). Conversely, while older participants also fixated more on the face than on the hands (all $p < 0.001$), they additionally fixated more on the mouth than on both the nose (difference estimate = 0.19, $t = 5.40$, $p < 0.001$) and the eyes (difference estimate = 0.25, $t = 4.22$, $p < 0.001$). All comparisons for the *age group*-by-*AOI*-by-*condition* interaction are shown in Supplementary Table A.4. In general, all participants fixate less on the face AOIs in degraded video conditions, and young participants additionally fixate more on the hands in those conditions. The differences were generally smaller for older than for younger participants. Lastly, young participants fixated less on the mouth for the dAV condition compared to AV (with a similar trend for dAV compared to V), but older participants fixated more on the mouth for the dAV condition compared to both AV and V.

In summary, results from the first analysis stage showed that older participants had lower accuracy scores than younger participants, but older participants were as capable of integrating auditory and visual information as younger participants were. There was no evidence for a 'positivity effect' for older participants, as their accuracy was lower for all emotions. Additionally, older participants made smaller saccades and fixations of shorter durations than younger participants. Lastly, older participants fixated mostly on the mouth of the actor, while younger ones distributed their fixations roughly equally over the actors' face.

From the second analysis stage, we found that, for both age groups, audio degradation did not reduce performance if the degraded audio was accompanied by intact video. Moreover, presenting degraded audio and degraded video simultaneously did not reduce performance more than only degrading the video and leaving the audio intact. Lastly, older participants did not adapt their gaze behavior as much as young participants.

4.4 Discussion

Our main finding is that older participants were as good as younger participants at integrating audio and video during the recognition of emotions presented using the AV stimulus materials. Likewise, both groups were equally good at compensating for degraded audio. However, in contrast to these comparable relative effects, older participants were systematically poorer at recognizing emotions than younger adults. Their recognition accuracy was lower in all conditions and for nearly all emotions compared to that of the of younger participants. This age effect could not be explained by a difference in visual and auditory functioning. Both age groups had a higher accuracy in the video-only than in the audio-only conditions, and accuracy was highest during AV presentation. Notably, the differences in performance between these conditions were similar for both age groups. Additionally, degrading the video always reduced recognition accuracy, regardless of whether the degraded video was presented in isolation or together with audio, while degraded audio only reduced accuracy when it was presented in isolation. This suggests that participants rely more strongly on the visual than on the auditory information when judging emotions with these stimulus materials.

In addition to these differences in recognition accuracy, we found that older participants had a strong fixation bias towards the mouth of the actor, while young participants distributed their fixations more evenly across the face. When presented with the video degradations, younger participants made much larger saccades, presumably in an attempt to move the scotoma away from the face and view the face with their peripheral vision. While older participants did so too, their increase in saccadic amplitude was much smaller. Consequently, their saccades were not large enough to move the scotoma away from the face. Our results thus confirm that emotion recognition deteriorates with age and we additionally show that age also affects gaze behavior.

Lastly, even though we have not formally analyzed the data from the patient participants due to the small sample size, their data still provide some useful preliminary insights. In general, the patient participants performed similarly as the older healthy participant group did in the degraded conditions, with both groups being of similar age. This similarity is an indication that our stimulus degradations captured at least some of the consequences of actual hearing and vision loss on emotion recognition. However, individual differences in performance between patient participants were very large, and were presumably at least partly related to their vision and hearing loss. For example, patient participant 2 had relatively good visual acuity (0.58) and contrast sensitivity (1.76 logCS), relatively little visual field loss, and only some hearing loss in higher frequencies (and normal PTA: 16.3 dB HL). In all conditions, this patient participant had the highest accuracy. In contrast, patient participant 4 had both poor visual acuity (0.09) and contrast sensitivity (0.71 logCS), had much more visual field loss, and was completely deaf in one ear and had severe hearing loss in the other ear (PTA: 68.3 dB HL), and this patient participant had very low accuracy in all conditions. Perceived auditory and visual functioning, measured with the SSQ and VFQ-25 respectively, were loosely correlated with the results from the screening. Although other factors, such as age, education level, and how long they have had impaired vision and hearing likely also contribute to differences in emotion recognition accuracy across patient participants, it appears that differences in visual field loss, visual acuity, contrast sensitivity, and hearing levels at least partially explain the individual performance differences.

4.4.1 Older and younger adults integrate audiovisual information for emotion recognition similarly

For all intact conditions (A, V, AV) and all emotions, older participants showed lower emotion recognition accuracy than young participants. This is in line with other findings[148,149]. Additionally, we found that the addition of another modality did not change the accuracy difference between older and younger participants that was observed for unimodal modalities. Rather, when only considering the intact conditions, there was no significant age group by condition interaction, indicating that the difference in accuracy remained roughly the same across A, V, and AV conditions. Therefore, unlike what has been previously reported[122,151], we find that older participants are as good as younger participants at integrating auditory and visual information, but not better. Wieck and Kunzmann[151] already proposed that divergent findings could be due to differences in the quality of the emotion expression. They hypothesized that older adults only benefit from additional information (in other modalities) if that additional information clearly points towards the same emotion. In our experiment, due to the large number of different emotions included, the emotional cues in each modality may have been subtler and more complex than in previous studies, such that integrating auditory and visual cues does not necessarily resolve all ambiguity. The chance of that happening is much smaller when there are fewer emotions being portrayed; Wieck and Kunzmann[151] only presented two emotions (anger and sadness), and Hunter et al.[122] presented four (fear, sadness, disgust, and anger). In our study, in contrast, twelve emotions (of which six were negative) were used, and some were closely related (e.g., anger and irritation). We consider our approach a more ecologically valid approximation of real life, in which people do not always display their emotions very consistently and clearly, do not limit themselves to core emotions only but instead display a wide range of emotions. Therefore, we claim that our results are a relatively good representation of emotion recognition abilities in daily life, and the earlier studies may not have been sufficiently sensitive as a result of using too few emotion categories.

It is worth noting that the difference in accuracy between younger and older participants is not (fully) driven by poorer vision and hearing in the older group, as shown by our analysis in section 4.3.3 (*Effects of auditory and visual functioning on emotion recognition accuracy*).

4.4.2 The ability to compensate for sensory degradation remains stable with age

For both age groups, we found that our signal degradations decreased recognition accuracy. When presented in isolation (i.e., unimodal degraded stimulus presentation), degraded audio/video (dA, dV) led to lower accuracy than for unimodal intact audio/video stimulus presentation (A, V). Besides this, older participants showed roughly the same pattern across degraded conditions as younger participants did: degraded video combined with intact (AdV) or degraded audio (dAdV) led to a similar decrease in accuracy compared to AV. Only degrading audio (dAV), however, did not lead to a decrease in accuracy compared to AV. Therefore, it seems that, at least for the task and materials used here, participants could fully compensate for the degraded audio by relying more on the visual information. In contrast, relying more on intact auditory information to compensate for degraded video was not possible. Moreover, these effects were the same for both the younger and older participants. This similarity suggests that, although emotion recognition ability may decline with age, the ability to compensate for sensory degradation seems to remain stable with advance age.

4.4.3 No evidence for a positivity effect, but an overall emotion recognition reduction with age

We found that older adults' recognition accuracy was poorer compared to young participants' accuracy for both positive and negative emotions. There was therefore no evidence for a positivity effect in our data, contradicting some previous findings[123,153–155]. Again, this discrepancy with literature could be related to the large number of emotions that were used in the current study. The task of discriminating between many different emotions, and additionally integrating auditory and visual information, which were sometimes degraded, likely lead to a high cognitive load. There is evidence that high cognitive load reduces or completely diminishes the positivity effect[166,167]. Additionally, previous findings of a positivity effect may have been related to the fact that these studies used little positive emotions. All these studies[123,153–155] only used the six basic emotions (happiness, surprise, sadness, fear, anger, disgust). Only two emotions of the six basic emotions are positive, and only happiness is very clearly positive, while surprise is a bit more ambiguous. Therefore, the reason that these studies find that recognition of positive emotions is preserved with age, may be solely due to the fact that it is easier to correctly guess the positive emotions if there are only two positive emotions in the stimulus set.

4.4.4 Older adults tend to fixate more on the mouth, while younger adults distribute fixations evenly across the face

For intact conditions, older participants had a strong tendency to fixate on the mouth of the actor, which is in line with previous findings[158,159]. This bias towards fixating on the mouth was traded off by a decrease in fixations on the eyes. Younger participants, however, distributed their fixations more evenly over the actor's face. Both age groups hardly ever fixated on the hands of the actor. The bias of older adults to fixate on the mouth (or at least, bottom half of the face) more has been indicated to be related to their preserved ability for recognizing positive emotions[159], as a prototypical expression of happiness is most clearly recognizable by the smiling mouth[37,90]. However, here we showed that while older adults generally have this bias, the accuracy difference between younger and older adults still remains for positive emotions. Therefore, it remains to be examined why this bias exists in older adults. Contrary to our hypothesis, the fixation bias towards the mouth remained in multimodal conditions, but this is line with the finding that the age effect for performance also remained in multimodal conditions. In addition to the difference in fixation proportions, older participants on average had shorter fixation durations and additionally made smaller saccades.

4.4.5 Reduced gaze adaptation in older adults

All participants adapted their gaze to degraded video presentation (dV, AdV, dAdV), but did not adapt their gaze in response to degraded audio (dAV), for which there were no significant differences with AV. For both age groups the gaze adaptations to degraded video were apparent as a decrease in fixation durations, an increase in saccadic amplitudes, and a decrease in fixation proportions on all face AOIs. However, these changes were much smaller for older participants than they were for younger participants. For example, younger participants increased saccadic amplitudes from on average 2.5 degrees of visual angle in intact video conditions (V, AV, dAV) to about 6 degrees for degraded video conditions (dV, AdV, dAdV). In contrast, older participants showed saccadic amplitudes of on average 1.5 degrees for intact video conditions, and increased to on average 2.5 degrees for degraded video conditions. Since the scotoma extended 17 by 11.5 degrees, making saccades of 6 degrees, as the young participants generally did, would be sufficient to move the scotoma away from the face.

These results suggest that there was a limitation in older adults' vision, eye movements, or cognitive processing that makes it impossible or less optimal to make the large gaze adaptations that younger adults do, although it is uncertain what exactly. One possible explanation is that older adults consistently make hypometric saccades, and because of this never 'reach' the target with their gaze. However, several studies on the effects of age on saccade dynamics do not show an effect on saccadic amplitude or accuracy[168–170], making this an unlikely explanation for our findings. A potentially straightforward explanation is that within the relatively short time span of fixation, older observers are not capable of attending to items that are far away from their point of gaze. Indeed, it has been shown that when given the same amount of time to inspect a display, older adults have a narrower spatial spread of attention compared to younger adults[171] and a smaller useful field of view (i.e., the visual area in which useful information can be acquired within a brief timespan)[172,173]. Therefore, we propose that within the typical duration of their fixations, the older participants in our study were incapable of attending to the face if it was far out in their visual periphery and therefore optimized their performance by fixating closer to it.

Note that we only analyzed trials with correct recognition, as we assumed that this would inform on whether the adapted gaze behavior would lead to good performance. However, an extra analysis (not included here) showed that there was no difference in gaze behavior for incorrect versus correct recognition for both age groups. Based on this, it can be concluded that observers settle on a gaze adaptation strategy (consciously or unconsciously) that optimizes performance as much as the restrictions of that participant's visual and cognitive systems allow.

4.4.6 Limitations and future directions

Our findings, especially those related to the fact that visual information seems more important than auditory information, may be strongly dependent on the specific materials used here. The video stimuli had very rich visual cues, including both facial expressions and body language, and possibly less clear auditory cues, which only included prosodic but not semantic information. Therefore, future studies should test the assumption that vision can compensate for degraded audition (be it simulated or real) by using different audiovisual emotion materials, for example by including sentences with meaningful semantic content.

In addition, we cannot rule out that our results were not driven by differences in other factors that have been indicated to impact emotion recognition processes, such as education level[148], cultural differences, and cognitive functioning[174]. While the present study confirmed with the MoCA that none of our older participants showed signs of cognitive impairment, we did not directly assess cognitive functioning in both groups. Likewise, we did not assess participants' education level and as most of the younger participants were university students, it is possible that there was a difference in education level between the younger and older participant groups.

Lastly, the fact that we analyzed eye-movements over a relatively short time period of 1000 ms, may have affected what differences we observed between age groups. For example, it is possible that older adults needed more time during the trial to start exhibiting adapted gaze behavior and a short temporal analysis window may not have captured this properly. However, as mentioned in the methods section (see Data analyses – Analyses of eye-tracking data), the time period was chosen to fit the length of the shortest video and to ensure that only task-related gaze data was included. It may be worthwhile to study this by using emotion stimuli that morph from neutral to emotional over different time spans and study whether morph duration affects age differences in gaze.

4.4.7 Conclusions

Altogether, the present data show that audiovisual integration for emotion recognition remains intact with age, even though ageing seems to lead to a general decrease in emotion recognition abilities. Additionally, we have shown that both younger and older adults adapt their perceptual strategies in response to degraded visual information, although older adults make smaller adaptations than younger adults. These smaller adaptations may be related to the smaller useful field of view in older adults. Therefore, rehabilitation programs aimed at expanding the useful field of view[175] and teaching adapted viewing behavior to visually impaired individuals may improve their emotion recognition. However, before implementing this, further studies into the mechanisms and benefits of gaze adaptation are necessary.

4.5 Acknowledgements

We gratefully acknowledge Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer for permission for using the GEMEP core set and for publishing sample images from the core set. Additionally, we would like to thank Merel de la Rie for help with data collection.

4.6 Funding

The first author was supported by a BCN-BRAIN grant from The Graduate School of Medical Sciences (GSMS), University of Groningen, the Netherlands. This project was supported by the following foundation: the Landelijke Stichting voor Blinden en Slechtzienden, contributing through UitZicht (Grant number: Uitzicht 2014 – 28 – Pilot). The funding organizations had no role in the design or conduct of this study.

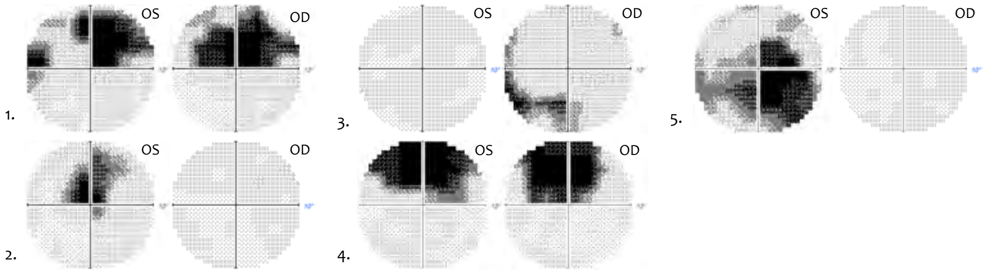
4.7 Data availability

The datasets generated for this study can be found in the DataverseNL repository via https://doi.org/10.34894/NCS9IG. All data is publicly available.

Supplementary Material

Supplementary Table A.1. demographics for all participants (younger, older, and patients). Included information is age at time of participation, visual acuity (VA), contrast sensitivity (CS), and pure-tone average (PTA; based on four frequencies) of the best ear. Non-normal values for vision and/or hearing are indicated by bold typeface.

Participant	Younger				Older				Patient			
	Age	VA	CS	PTA	Age	VA	CS	PTA	Age	VA	CS	PTA
1	21	1.95	2.38	0	68	**0.61**	2.01	7.5	68	**0.08**	**1.1**	11.3
2	21	1.61	2.17	2.5	69	0.93	1.83	6.3	68	**0.58**	**1.76**	16.3
3	23	2.10	2.12	-1.3	63	**0.72**	2.03	**21.3**	66	**0.60**	**1.4**	**45**
4	23	2.07	2.04	-3.8	71	1.3	2.01	**28.8**	67	**0.09**	**0.71**	**68.3**
5	26	1.85	2.05	5	66	1.54	2.13	12.5	77	**0.45**	**1.61**	**43.8**
6	21	1.79	2.24	5	69	1.07	2.08	10				
7	25	1.66	2.01	-1.3	66	1.28	1.99	15				
8	19	2.07	2.05	-3.8	70	1.38	1.98	35				
9	20	2.22	2.02	7.5	62	1.29	2.08	**21.3**				
10	24	1.52	2.11	0	64	1.18	1.83	8.8				
11	27	1.57	2.39	7.5	61	**0.54**	2.11	8.8				
12	20	1.94	2.11	1.3	72	**0.67**	1.97	**25**				
13	29	1.65	2.39	0	71	1.24	1.92	3.8				
14	24	1.45	2.11	5	65	1.09	1.88	18.8				
15	19	1.63	1.83	0	64	1.45	2.15	7.5				
16	24	2.18	2.22	-2.5	61	1.61	2.11	20				
17	19	1.98	2.01	-6.3	65	1.25	1.94	7.5				
18	22	1.69	2.38	0	66	1.58	1.91	20				
19	26	1.48	2.09	3.8	66	1.18	2.19	8.8				
20	23	1.23	2.11	2.5	69	1.05	2.00	7.5				
21	28	1.49	1.99	2.5	63	1.41	1.86	13.8				
22	21	1.37	2.00	-2.5								
23	22	1.60	1.96	-1.3								
24	26	1.94	1.8	1.3								
Mean	23	1.75	2.10	0.89	66	1.16	2.00	14.6	69	0.36	1.32	36.9
SD	2.91	0.27	0.15	3.54	3.36	0.31	0.11	8.39	4.44	0.26	0.42	23.4



Supplementary Figure A.1. HFA data (Central 10-2 protocol) for all five patient participants. Data from the left eye is displayed on the left, data from the right eye is displayed on the right.

Supplementary Table A.2. Scores for the Speech and Spatial Qualities Questionnaire (SSQ) for individual patient participants. Scores in bold typeface are mean scores for each main scale (Speech, Spatial, Qualities) and the sub-scales defined by Gatehouse & Akeroyd[187]. All scores are on a scale of 0-10, with higher scores indicating better perceived hearing.

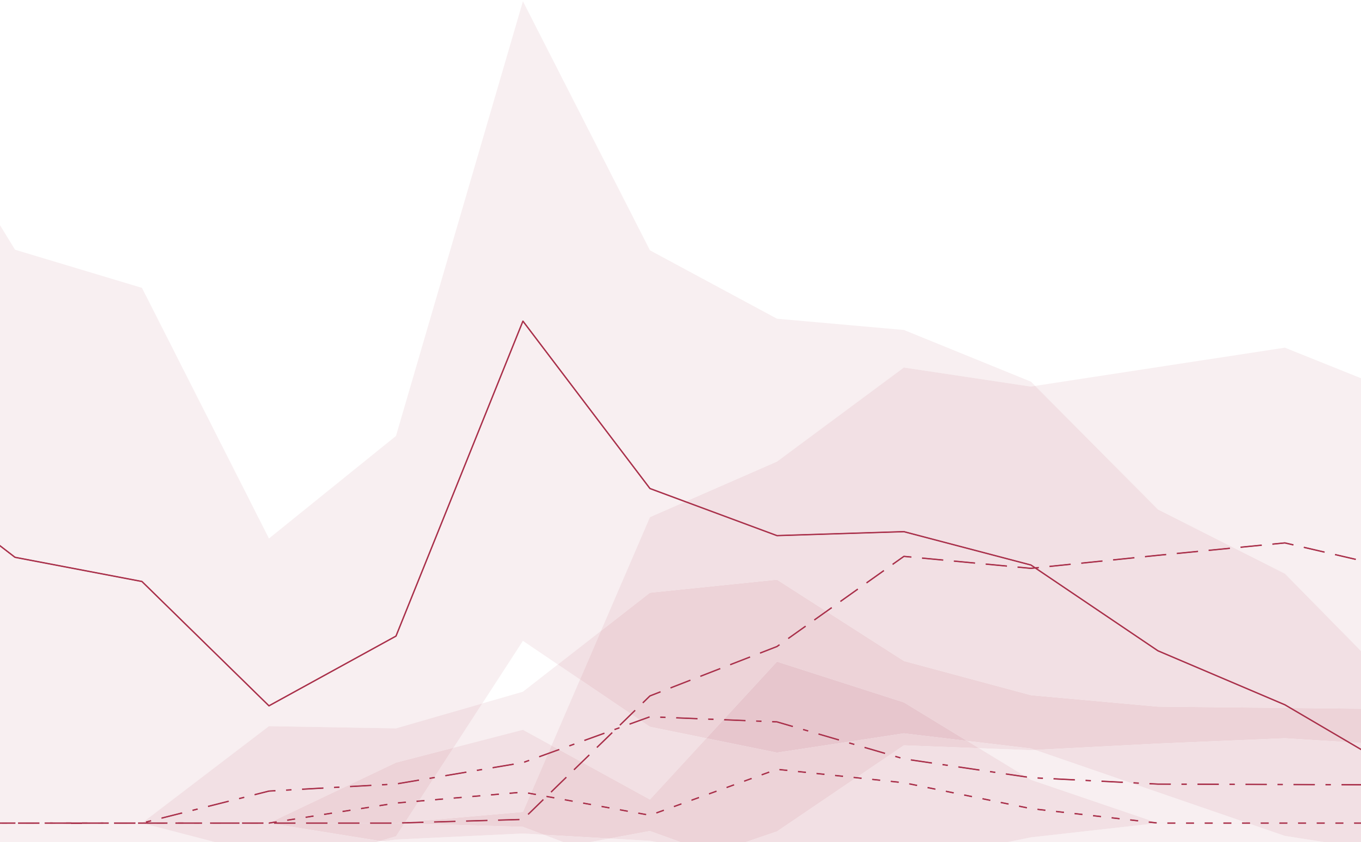
	Score				
Patient participant	**1**	**2**	**3**	**4**	**5**
(Sub-)scale					
Speech	**8.86**	**4.57**	**5.07**	**4.71**	**2.29**
Speech in quiet	10	7	8	7	6
Speech in noise	8.5	5	3.75	3.75	1.75
Speech in speech contexts	8.5	3.25	4.75	4.5	0.75
Multiple speech-stream processing and switching	8.67	2.67	4.33	3.67	0.67
Spatial	**8.29**	**6.88**	**6.82**	**4.06**	**2.69**
Localization	8.17	6.33	8	3.5	3
Distance and movement	8.33	6.67	5.89	4.22	2.44
Qualities	**9.35**	**8.17**	**6.89**	**7**	**4.67**
Sound quality and naturalness	9.4	9.8	7.8	7.8	5.2
Identification of sound and objects	9.8	8.6	8.2	7.8	6.8
Segregation of sounds	10	9.33	8.67	7	2
Listening effort	8.67	5	3.33	4.33	4.33

Supplementary Table A.3. Scores for the Visual Functioning Questionnaire 25-item (VFQ-25) for individual patient participants. The additional questions were used as well, so the scores are based on the 39-item VFQ. Scores are recoded according to the manual, mean scores for each scale are shown. The composite score is the mean of scores for all scales except the general health scale. All scores are on a scale of 0-100, with higher scores indicating better perceived vision.

	Score				
Patient participant **Scale**	**1**	**2**	**3**	**4**	**5**
General health	82.5	65	77.5	65	25
General vision	30	55	65	30	60
Ocular pain	100	50	87.5	50	37.5
Near activities	25	62.5	70.83	20.83	65
Distance activities	45.83	50	79.17	29.17	50
Driving	0	41.67	83.33	0	66.67
Color vision	25	100	100	50	75
Peripheral vision	50	75	100	75	75
Vision specific subscales					
Social functioning	75	75	100	41.67	91.67
Mental health	90	45	40	50	65
Role difficulties	68.75	31.25	62.5	25	56.25
Dependency	93.75	56.25	93.75	56.25	75
Composite score	**54.84**	**58.33**	**80.19**	**38.90**	**65.19**

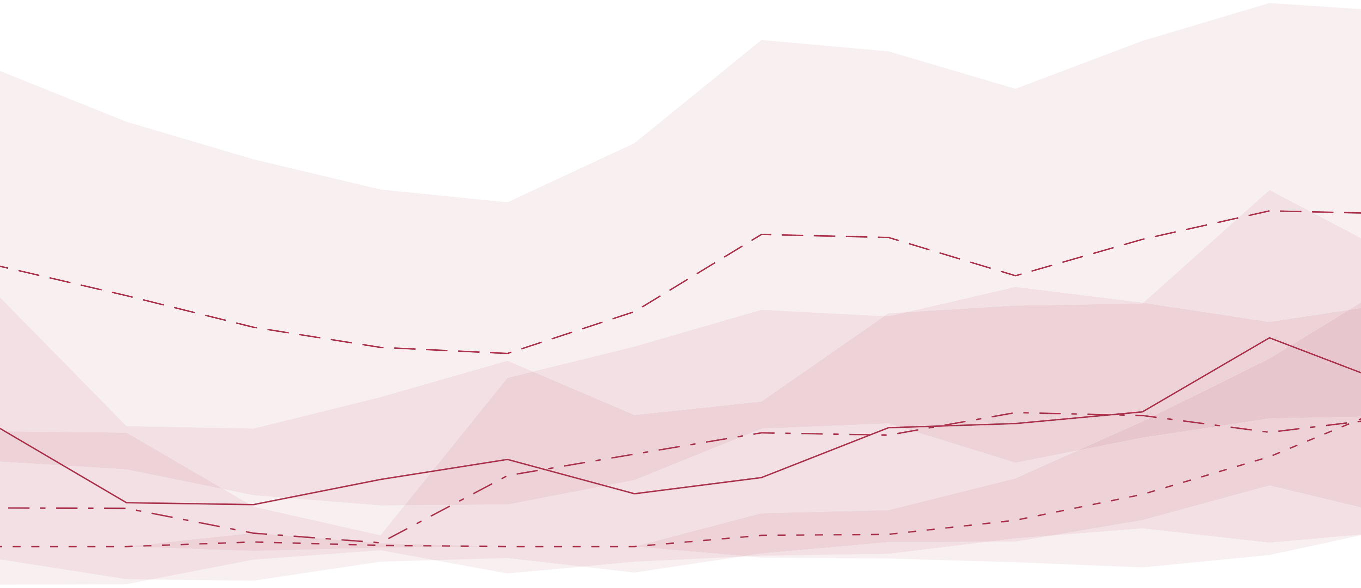
Supplementary Table A.4. Contrasts for the age group by AOI by condition interaction for fixation proportions. It shows the model estimate differences with the FDR adjusted p-values in parentheses. Significant differences are indicated by bold typeface.

	Eyes		**Nose**		**Mouth**		**Hands**	
	Young	Old	Young	Old	Young	Old	Young	Old
V – AV	0.01 (0.312)	0.02 (0.269)	0.00 (0.973)	0.01 (0.319)	-0.00 (0.697)	-0.01 (0.626)	-0.00 (0.859)	-0.00 (0.969)
V – dAV	0.003 (0.761)	0.02 (0.269)	0.02 (0.070)	0.02 (0.263)	0.02 (0.069)	**-0.04** **(0.003)**	-0.00 (0.856)	-0.00 (0.993)
V – dV	**0.18** **(<0.001)**	**0.08** **(<0.001)**	**0.19** **(<0.001)**	**0.16** **(<0.001)**	**0.15** **(<0.001)**	**0.07** **(<0.001)**	**-0.08** **(<0.001)**	-0.03 (0.231)
V – dAdV	**0.18** **(<0.001)**	**0.08** **(<0.001)**	**0.19** **(<0.001)**	**0.14** **(<0.001)**	**0.15** **(<0.001)**	**0.10** **(<0.001)**	**-0.07** **(<0.001)**	-0.02 (0.231)
V – AdV	**0.17** **(<0.001)**	**0.08** **(<0.001)**	**0.19** **(<0.001)**	**0.15** **(<0.001)**	**0.16** **(<0.001)**	**0.08** **(<0.001)**	**-0.07** **(<0.001)**	-0.01 (0.518)
AV – dAV	-0.01 (0.411)	0.00 (0.964)	0.02 (0.070)	0.00 (0.822)	**0.03** **(0.019)**	**-0.03** **(0.012)**	-0.00 (0.933)	0.00 (0.969)
AV – dV	**0.16** **(<0.001)**	**0.07** **(<0.001)**	**0.19** **(<0.001)**	**0.15** **(<0.001)**	**0.16** **(<0.001)**	**0.08** **(<0.001)**	**-0.07** **(<0.001)**	-0.02 (0.231)
AV – dAdV	**0.17** **(<0.001)**	**0.06** **(<0.001)**	**0.19** **(<0.001)**	**0.12** **(<0.001)**	**0.15** **(<0.001)**	**0.11** **(<0.001)**	**-0.06** **(<0.001)**	-0.02 (0.231)
AV – AdV	**0.15** **(<0.001)**	**0.06** **(<0.001)**	**0.19** **(<0.001)**	**0.13** **(<0.001)**	**0.16** **(<0.001)**	**0.09** **(<0.001)**	**-0.07** **(<0.001)**	-0.01 (0.592)
dAV – dV	**0.17** **(<0.001)**	**0.07** **(<0.001)**	**0.17** **(<0.001)**	**0.15** **(<0.001)**	**0.13** **(<0.001)**	**0.12** **(<0.001)**	**-0.07** **(<0.001)**	-0.03 (0.231)
dAV – dAdV	**0.18** **(<0.001)**	**0.06** **(<0.001)**	**0.16** **(<0.001)**	**0.12** **(<0.001)**	**0.13** **(<0.001)**	**0.14** **(<0.001)**	**-0.06** **(<0.001)**	-0.02 (0.231)
dAV – AdV	**0.16** **(<0.001)**	**0.06** **(<0.001)**	**0.16** **(<0.001)**	**0.13** **(<0.001)**	**0.14** **(<0.001)**	**0.12** **(<0.001)**	**-0.06** **(<0.001)**	-0.01 (0.518)
dV – dAdV	0.01 (0.581)	-0.001 (0.785)	-0.01 (0.636)	-0.03 (0.11)	-0.00 (0.916)	0.02 (0.181)	0.01 (0.508)	0.00 (0.969)
dV – AdV	-0.01 (0.411)	-0.001 (0.785)	-0.01 (0.636)	-0.02 (0.319)	0.01 (0.697)	0.01 (0.626)	0.01 (0.613)	0.01 (0.621)
dAdV – AdV	-0.02 (0.162)	-0.00 (0.964)	-0.00 (0.973)	0.01 (0.490)	0.01 (0.671)	-0.01 (0.333)	-0.00 (0.888)	0.01 (0.622)



Chapter 5

General Discussion



Preface

Emotion recognition is a dynamic and multimodal process and crucial for successful communication. Despite its importance, while it is known that age and sensory impairments negatively affect emotion recognition, it is not well understood how age and sensory impairments affect the integration of auditory and visual emotion cues. Therefore, in this thesis, I strived to first create a comprehensive understanding of audiovisual integration in emotion recognition from communication and subsequently investigate the impact of age and sensory impairments on this. The main reason for a lack of understanding of the audiovisual integration process for emotion recognition, in general and more specifically with age and sensory impairments, is that emotion recognition has typically been investigated in a single modality. Most commonly studied are static facial expressions, which may not be representative of face-to-face emotion expressions and thus limit the generalizability of previous findings. Therefore, stimuli that are very close to what we encounter in daily life were used for the work presented in this thesis. The emotional expressions in the used stimulus set were judged to be fairly authentic and believable in a validation study[36]. In addition, the stimuli allow for a controlled study, as all participants see exactly the same expressions, allowing to generalize findings over participants. Moreover, a good number of different emotions is included in the set, of which some are closely related, presenting more subtle distinctions between emotions. This allows us to generalize over more emotions that can be encountered in real life and also to study how well related emotions can be differentiated. Because the actors speak a nonsensical sentence, this stimulus set does not contain any semantic information and the focus is on the effects of prosody and facial and bodily expressions. Using this stimulus set, three experimental studies were carried out to answer the research questions stated in the introduction. In this chapter, I will discuss the findings of the studies and how they answer these research questions.

5.1 Integration of intact visual and auditory emotion cues

The first research question was: in normal vision and hearing, with rich, dynamic emotion cues, how do auditory and visual information contribute to audiovisual integration for emotion recognition? For all studies, emotion recognition with only the video (V) presented was better than when only the audio (A) was presented, and, averaged over emotions, best with audiovisual (AV) presentation. However, the data presented in the Chapter 2 showed that, for most emotions, emotion recognition did not improve significantly when audio was added to the video. It thus appears that the information in the video was already so rich, or the information in the audio too ambiguous or not rich enough, that the addition of audio did not supplement the information provided by the video. However, as the difference in recognition accuracy between V and AV stimulus presentation was significant when averaged over emotions, it is possible that the data in Chapter 2 were underpowered to show the difference for individual emotions. Regardless, even if the data were underpowered, the differences between V and AV were small, as otherwise statistical power would not have been an issue. Notwithstanding, emotion recognition was always better with V than with A, supporting the idea that visual information may be more important/reliable than auditory information, which is in line with earlier works[55,63–65].

As I used more natural stimuli than many previous studies, it is possible that also in real life there could be a greater reliance on visual cues than auditory cues. However, no other stimuli were used in this thesis, so absolute accuracy levels may be strongly reliant on the specific stimulus materials used here. Relatedly, Paulmann and Pell[62], who also used dynamic audiovisual stimuli, found that emotion recognition from facial expressions was better than recognition from prosody, but not better than recognition from semantic content. Therefore, the fact that in this thesis observers seemed to rely more on the visual than on the auditory information, may be related to the fact that the auditory information came from prosody only without related semantic content. It is even possible that our participants actively ignored the audio when the video was available, as they could not understand the sentences the actors were speaking. As a result, they also would not have properly processed the prosodic cues, possibly explaining the low contribution of audio. One way to test whether this may have happened, is to use audiovisual stimuli with meaningful semantic content, but in a language the participants do not understand (and is preferably unrelated to the language of the study population), and stimuli without meaningful semantic content in the native language of the participants. With such a design, the semantic content is never helpful for emotion recognition, but the observers may ignore the audio in one case (unknown language), and use it in the other (native language).

In addition to a lack of meaningful semantic content, the video actually contained two distinct cues of emotion (the face and the body), while the audio only contained a single emotion cue (prosody). If emotional expressions from the face and body can indeed be viewed as two distinct cues, it is not surprising that the video was perceived as more reliable. In this case, audiovisual stimulus presentation would have contained three emotion cues (face, body, and prosody), V contained two cues (face and body), while A contained only one cue (prosody). Future studies could test whether this played a role by comparing uni- and multimodal stimulus presentations with two distinct emotion cues (face + body vs. face + prosody, vs. body + prosody) or by adding a distinct cue to the audio, such as meaningful semantic content, such that the number of distinct emotion cues in the video and audio are more comparable. Adding meaningful semantic content would also provide stronger evidence on whether, in real life, the visual cues are also deemed more reliable, or alternatively, whether the semantic content is so reliable, that the balance between visual and auditory cues shifts towards the audio.

5.2 Compensation for degraded audio is possible by relying more on the video, but compensating for degraded video is not possible

As for the second question (how do (simulated) vision and hearing loss affect emotion recognition and audiovisual integration for emotion recognition?), the data from Chapter 3 showed that the stimulus degradations affected emotion recognition when presented in isolation (i.e., A/V only presentation), but compensation was possible in some cases. Importantly, combining visual and auditory degradation did not exacerbate the isolated effects. In fact, emotion recognition accuracy with AV presentation when both modalities were degraded (i.e., comparable with face-to-face conversation for someone who is both vision and hearing impaired) was higher than recognition accuracy for degraded A/V presentation and quite similar to intact AV presentation. Compensation for degraded audio was possible if the video was available, and compensation was as good with intact video as it was with degraded video. Thus, degraded audio did not lead to a decline in emotion recognition, because the visual cues could compensate for the decreased reliability of the auditory cues. Alternatively, it is possible that the audio still contained some relevant information, such as on- and offset cues. Conversely, full compensation for the simulated macular degeneration was not possible, although the addition of any audio (intact or degraded) to any video (intact or degraded) greatly facilitated performance. It is possible that relying more on intact audio does not allow for compensation for the degraded video because of the greater importance of visual cues compared to auditory cues for emotion recognition (at least with these stimuli). In addition, in audiovisual speech, visual cues generally precede auditory cues by several hundred milliseconds allowing visual cues to provide information about the onset of the acoustic signal and the speech envelope[28,29]. Thus, in speech, and possibly also in emotion communication, early visual cues can be used to predict auditory information, but auditory cues cannot increase the predictability of visual cues.

Again, in real life conversations where there is meaningful semantic content, it is quite possible that the relative contribution of auditory and visual information is different and that, because of this, mechanisms for compensating for sensory impairments are also different. For example, it could be that the auditory information is more important than visual information in conversations with meaningful semantic content and that compensation for a hearing loss is not fully possible then.

Preliminary data from patients with macular degeneration and hearing loss showed similar emotion recognition performance as the older observers who were tested with degraded stimuli, thus the simulated degradations seemed to capture some of the important consequences of actual sensory impairments on emotion recognition (Chapter 4). As this was data from only five patients, and the severity of their sensory impairment was variable, it would be beneficial to include a much larger group in future studies, preferably with better selection of sensory impairments (i.e., similar type and degree of impairment).

5.3 Perceptual strategies are adapted flexibly to information reliability

The third research question was: do healthy observers adapt their perceptual strategies to the availability and reliability of visual and auditory information? We found that observers indeed flexibly adapt their perceptual strategies. It has long been known that eye movements are tightly linked to the task at hand and observers rarely fixate irrelevant objects[45]. Adaptations in perceptual strategies thus indicates that different objects were relevant in the different conditions. Chapter 2 showed that participants viewed the mouth less when audio was present and more when audio was lacking, possibly to compensate for the missing auditory cues. This finding indicates that gaze is not only guided by the visual information, which remained the same in AV and V stimulus presentation, but also by the presence or absence of auditory information. Thus, when auditory cues are present, observers use them and adapt their perceptual strategies to them, even if this adaptation does not result in improved task performance. This gaze shift towards the mouth in video only stimulus presentation was not replicated in the following chapters, possibly because observers in those experiments felt adapting their perceptual strategies to the simulations was more important, and did not focus as much on optimizing their gaze for the presence or absence of the audio.

Chapters 3 and 4 showed that the observers made slightly smaller saccades in V than in AV, suggesting that their eye-movements were somewhat more precise when there was no audio. In addition, observers changed their perceptual strategies with degraded video by making larger saccades and fixations at a larger distance from the face. This viewing behavior indicates that they were trying to move the scotoma away from the face and compensate for the video degradation as much as possible. It has been proposed that people with a central vision loss develop a preferred retinal locus (PRL), also called a pseudo-fovea, that replaces their degraded fovea[116,117]. Our observers showed viewing behavior consistent with them using a PRL, as the observers' fixation distance to the actors' face was similar to the scotoma radius. The increase in saccadic amplitude is likely related to this PRL behavior; the saccades made in conditions with degraded video were large enough to move the scotoma fully away from the actors' face. It should be noted here that it is unlikely that the participants in our studies actually developed a pseudo-fovea, as it likely takes much longer to develop than the few hours that our participants spend on the task. For example, in one study, normally sighted participants were trained on a letter recognition task using peripheral vision, and received training of two hours per day for four days. The participants showed increased letter recognition accuracy and improved reading speeds at the trained location[176], indicating that the participants developed a pseudo-fovea.

5.4 Ageing leads to a general decline in emotion recognition, but does not reduce integration or compensation abilities

Lastly, I examined how integration and compensation abilities and adaptations of perceptual strategies change with age. Chapter 4 showed that age affected both general emotion recognition and perceptual strategies. Overall, we found that while ageing leads to a general decline in emotion recognition[148,149], it does not have an effect on audiovisual integration abilities. Contrary to the findings from Chapter 4, it has been proposed that audiovisual integration is enhanced in ageing, both in general, for example for stimulus detection or localization, but also for speech recognition[177,178] and in particular for emotion perception[122,151]. This enhancement is suggested to serve as compensation for a decline in the senses[177,179]. The data presented in Chapter 4 show that older observers benefitted as much from audiovisual integration as younger observers, but not more. In addition, the older observers included in my study were as good at compensating for degraded audio as the younger observers were and were also not able to compensate for degraded video. Therefore, if I can speculate on what this finding may mean in real-life applications, older adults that experience degraded hearing, but still have intact vision, may likewise show good emotion recognition because of this compensation.

Despite the intact ability to compensate for degraded audio, older observers showed less adaptation of perceptual strategies than younger observers. As gaze adaptation may be a possible strategy to compensate for degraded vision, it is noteworthy that the older observers seemed incapable of adapting their gaze as efficiently as young adults. At the moment, it is uncertain what limits older observers in these adaptations, and further research is needed to uncover this. A possible explanation is that older observers are not capable of attending away from their point of gaze, evidenced by a narrower spatial spread of attention[171] and smaller useful field of view (UFOV)[172,173] in older compared to younger adults. If an incapacity to attend away from fixation is indeed the cause for smaller gaze adaptation, then training to expand the useful field of view may increase the ability to compensate for degraded vision. UFOV training has been shown to increase neuropsychological measures of attention[175], and thus might be a promising direction for rehabilitation of visually impaired individuals targeted at improving emotion recognition.

5.5 Future directions

Future studies could focus on uncovering why for some emotions the addition of audio to the video did not improve recognition performance, even though the performance was not at the ceiling level (Chapter 2). One possibility is that the limitation for recognition performance is not related to perceptual processes, but instead related to decision processes. For example, it is possible that an observer can perfectly perceive a certain facial movement of the actor (say, raised eyebrows), but this movement may map to several emotion categories. As one emotion contains many facial- and body movements and voice fluctuations, and with each of these cues mapping to multiple emotion categories, the actual decision process may be quite noisy and ultimately the limiting factor for recognition performance. So even if one were able to perfectly perceive all these cues, they still may not always recognize the expressed emotion correctly. The experiments carried out for the purpose of this thesis were not designed to specifically test whether perception or decision processes are the main limiting factor, so this should be studied in future experiments. One manner of doing this could be to systematically remove or distort certain cues (for example by covering the eye region) and test the effect this has on emotion recognition accuracy. It is probable that these future experiments will indicate that more cues is not always better. Another possibility is that even when recognition accuracy does not improve, emotion processing is more efficient with multimodal cues. The findings on adapted perceptual strategies already point that this may be the case. For example, the fact that observers looked less at the mouth with audiovisual stimulus presentation (Chapter 2) could be an indication that they did not need to look at the mouth because they received the same emotional cues from the audio and that processing of the emotional cues was less effortful than when only the video was presented.

Further, if one wishes to study emotion recognition in a setting with an even higher ecological validity, studying it during face-to-face communication is possible nowadays due to the fast developments in eye-tracking research. It is possible to have a participant converse with either a trained actor or a second participant and simultaneously record the eye-movements (and ideally face movements) of both conversation partners. When using an eye-tracker such as the Pupil Invisible (Pupil Labs GmbH, Germany), a head-worn binocular eye-tracker that closely resembles a pair of regular glasses, the participants would not be hindered in their emotion production and recognition. Such a set-up would also allow examining whether certain proposed fixation biases are maintained during an emotional conversation, such as the idea that the eye region is most important[180–182]. Alternatively, fixations could be located on what are proposed as the most informative regions for the expressed emotion[90]. If professional actors are used, the research environment is still relatively controlled (as they would be more capable of producing similar expressions for multiple participants) and there is an additional bonus of being able to study both emotion perception and production simultaneously. However, if one still wishes to test a large range of emotions, this would require the researcher to make their participants angry, sad, scared, etcetera, which is not desirable from an ethical point of view. In addition, the experiment would either take much longer, or the researcher could only test a few emotions per participant. This is because if a participant is e.g. angry, it will take some time before they calm down enough to properly perceive and produce other emotions, especially positive ones.

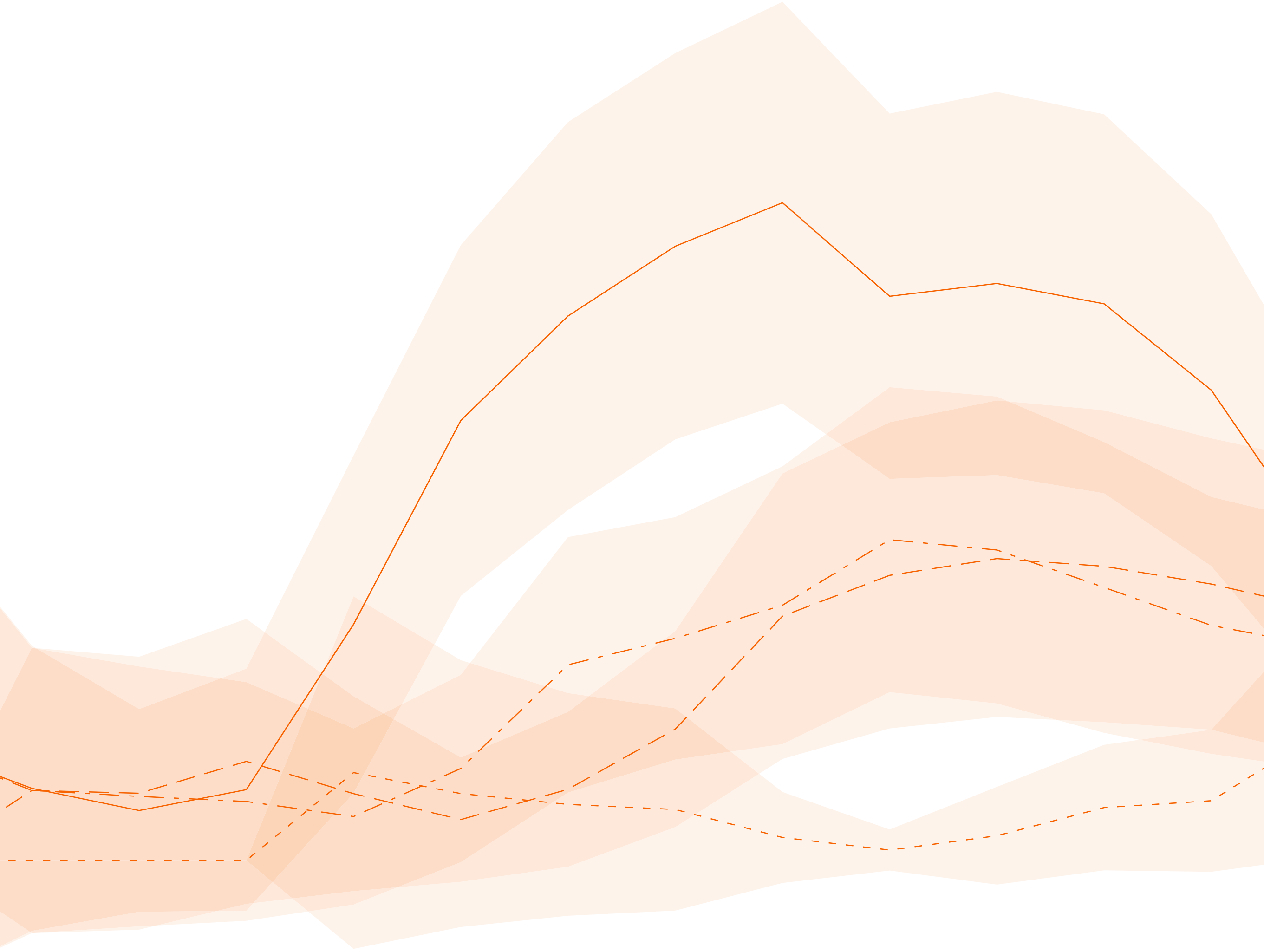
5.6 Clinical implications

Overall, the results from Chapters 3 and 4 suggest that hearing impaired individuals need not have difficulties with emotion recognition during face-to-face conversations, compared to individuals without sensory impairments, as they may be able to compensate for their degraded hearing by relying more on the visual cues. Conversely, based on these studies, visually impaired individuals may experience problems with emotion recognition, because even if they rely more on the auditory cues, they may still be unable to compensate for the degraded visual cues. Lastly, individuals who are both vision and hearing impaired are not expected to show reduced emotion recognition compared to individuals with solely a visual impairment and normal hearing based on the results of this thesis. However, as mentioned above, compensation for degraded vision might be possible to a greater degree during actual interactions when there is meaningful semantic content. In addition, in real sensory impairments, there will likely be some long-term adaptation to the impairment (such as cortical reorganization)[34,35], while the findings from this thesis only relate to acute effects. The long-term adaptation could additionally change the abilities for compensation that were found in Chapters 3 and 4. Therefore, it is important to extensively study audiovisual integration for emotion recognition in individuals with sensory impairments before any rehabilitation strategies are incorporated in the care for sensory impaired individuals. If there is a greater understanding of the underlying mechanisms of long-term adaptation, it may even be possible to exploit these mechanisms in rehabilitation to enhance emotion recognition.

Should future studies indeed indicate that individuals with impaired vision cannot compensate for their impairment by relying more on the audio, it is necessary to focus on improving the perception of visual emotion cues in order to improve their emotion recognition. Visual training is being applied in many visual impairments, mainly focusing on improving mobility and activities of daily living in hemianopia[183,184] and glaucoma[185], and mainly focused on improving reading in macular degeneration[186]. For macular degeneration, the main strategy is teaching patients to use a peripheral region of their visual field to replace their damaged fovea (i.e., a pseudo-fovea)[116,117], whereas visual training for hemianopia and glaucoma is focused on teaching patients to routinely make saccades into their blind visual field. As I have shown in this thesis that observers adapt their perceptual strategies to the available information, teaching these adapted perceptual strategies to individuals with sensory impairments may improve their emotion recognition. For example, we found that the mouth was fixated more when the audio was absent (Chapter 2), and hearing-impaired individuals may therefore benefit from looking at the mouth more often. Similarly, training patients to form a pseudo-fovea in a suitable location may improve their emotion recognition. The observers from Chapter 3 shifted their gaze downwards in conditions with degraded video, which would correspond to a pseudo-fovea in the upper visual field. This allowed the observers to view the face with their peripheral vision. At the same time, they could still see large body and hand movements that, although covered by the simulated scotoma, would still be visible due to the relative nature of the scotoma. Thus, for patients with a relative central scotoma, forming a pseudo-fovea in the upper visual field may benefit their emotion recognition. However, more research is necessary to determine the optimal perceptual strategy for emotion recognition with a central scotoma.

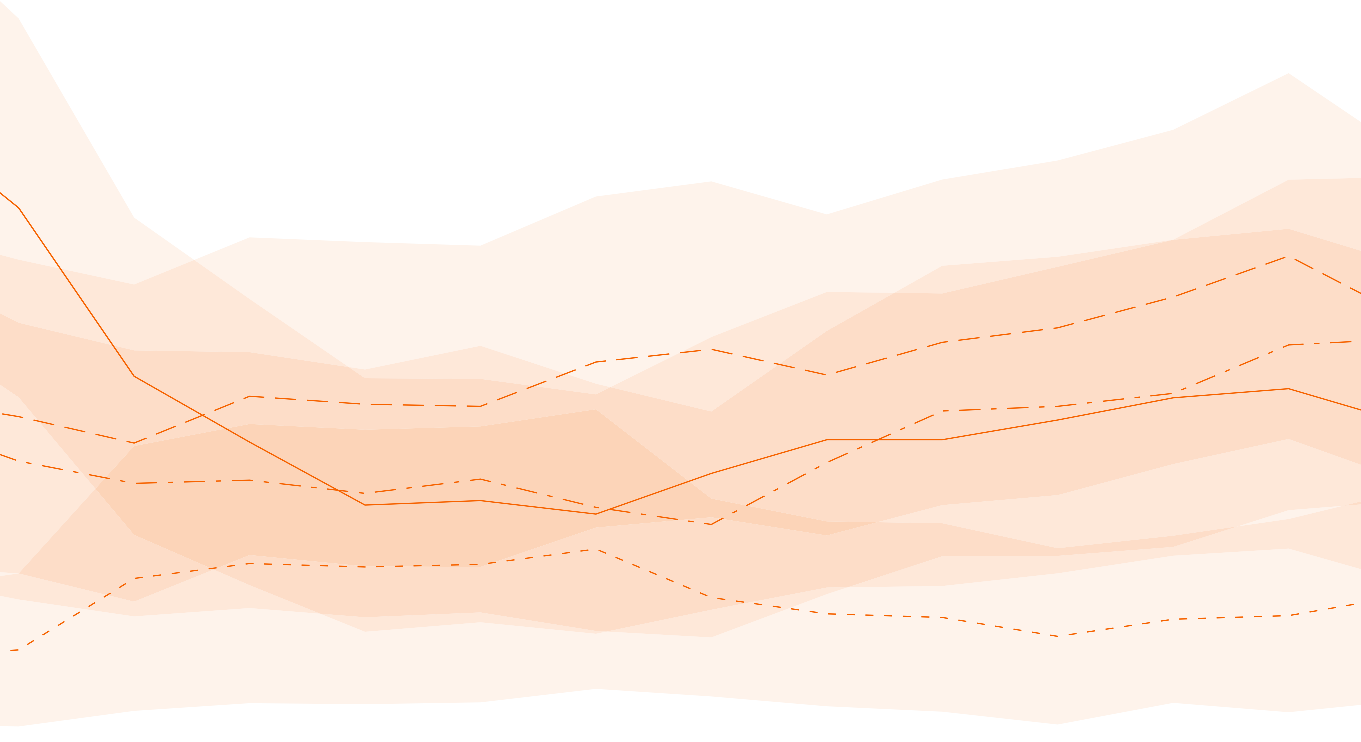
5.7 Conclusions

In my thesis, I examined the role of sensory impairments and age on audiovisual emotion recognition. Overall, the studies suggest that visual information is relied on more than auditory information for emotion recognition. Possibly, due to a greater reliability and predictability of visual cues in the used stimuli for the expressed emotion compared to auditory cues, compensating for degraded audio was possible by relying more on the video, but compensating for degraded video by relying more on the audio was not possible. Importantly, while emotion recognition declines with age, audiovisual integration and the ability to compensate for a degraded modality remain intact. Additionally, this thesis demonstrates that additional measurements besides recognition accuracy (here, eye movements) are crucial for understanding integration and compensation mechanisms. For example, the results from Chapter 2 showed that, for many emotions, recognition accuracy was the same for V and AV stimulus presentation, but perceptual strategies always differed between the different modalities. Thus, even if emotion recognition accuracy does not improve in multimodal situations, it is still possible that the recognition process is more efficient or less effortful with multimodal information. In conclusion, this thesis shows that when visual cues are rich, emotion recognition is affected by impaired vision, but not by impaired hearing, that observers flexibly adapt their perceptual strategies to the available emotion cues, and that ageing does not necessarily reduce the audiovisual integration capability for emotion recognition.



Appendices

- References
- Summary
- Nederlandse samenvatting
- Acknowledgements
- About the author



References

1. Plutchik, R. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist* **89**, 344–350 (2001).
2. Darwin, C. *The Expression of the Emotions in Man and Animals*. (Oxford University Press, 1872).
3. Ekman, P. & Friesen, W. V. Constants across cultures in the face and emotion. *Journal of personality and social psychology* **17**, 124 (1971).
4. Ekman, P. & Friesen, W. V. Facial action coding system. (1977).
5. Elfenbein, H. A. & Ambady, N. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin* **128**, 203–235 (2002).
6. Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G. & Caldara, R. Cultural Confusions Show that Facial Expressions Are Not Universal. *Current Biology* **19**, 1543–1548 (2009).
7. Russell, J. A. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin* **115**, 102–141 (1994).
8. Matsumoto, D. Cultural Influences on the Perception of Emotion. *Journal of Cross-Cultural Psychology* **20**, 92–105 (1989).
9. Stein, B. E. & Stanford, T. R. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience* **9**, 255–266 (2008).
10. Alais, D. & Burr, D. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology* **14**, 257–262 (2004).
11. Jacobs, R. A. Optimal integration of texture and motion cues to depth. *Vision Research* **39**, 3621–3629 (1999).
12. Diederich, A. & Colonius, H. Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time. *Perception & Psychophysics* **66**, 1388–1404 (2004).
13. Mahoney, J. R., Li, P. C. C., Oh-Park, M., Verghese, J. & Holtzer, R. Multisensory integration across the senses in young and old adults. *Brain Research* **1426**, 43–53 (2011).
14. Meredith, M. A. & Stein, B. E. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research* **365**, 350–354 (1986).
15. Meredith, M. A. & Stein, B. E. Spatial determinants of multisensory integration in cat superior colliculus neurons. *Journal of Neurophysiology* **75**, 1843–1857 (1996).
16. Meredith, M. A., Nemitz, J. W. & Stein, B. E. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J. Neurosci.* **7**, 3215–3229 (1987).
17. Recanzone, G. H. Auditory Influences on Visual Temporal Rate Perception. *Journal of Neurophysiology* **89**, 1078–1093 (2003).
18. Ernst, M. O. & Bülthoff, H. H. Merging the senses into a robust percept. *Trends in Cognitive Sciences* **8**, 162–169 (2004).
19. Howard, I. P. & Templeton, W. B. *Human spatial orientation.* 533 (John Wiley & Sons, 1966).
20. Stawicki, M., Majdak, P. & Başkent, D. Ventriloquist Illusion Produced With Virtual Acoustic Spatial Cues and Asynchronous Audiovisual Stimuli in Both Young and Older Individuals.

Multisensory Research **32**, 745–770 (2019).

21. Thurlow, W. R. & Jack, C. E. Certain Determinants of the "Ventriloquism Effect". *Percept Mot Skills* **36**, 1171–1184 (1973).
22. Shams, L., Kamitani, Y. & Shimojo, S. Visual illusion induced by sound. *Cognitive Brain Research* **14**, 147–152 (2002).
23. Grant, K. W. & Seitz, P.-F. The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America* **108**, 1197–1208 (2000).
24. Sumby, W. H. & Pollack, I. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America* **26**, 212–215 (1954).
25. Summerfield, Q. Lipreading and Audio-Visual Speech Perception. *Philosophical Transactions: Biological Sciences* **335**, 71–78 (1992).
26. Baskent, D. & Bazo, D. Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment. *Ear and hearing* **32**, 582–592 (2011).
27. Middelweerd, M. J. & Plomp, R. The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America* **82**, 2145–2147 (1987).
28. Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A. A. The Natural Statistics of Audiovisual Speech. *PLoS Computational Biology* **5**, e1000436 (2009).
29. Peelle, J. E. & Sommers, M. S. Prediction and constraint in audiovisual speech perception. *Cortex* **68**, 169–181 (2015).
30. McGurk, H. & MacDonald, J. Hearing lips and seeing voices. *Nature* **264**, 746 (1976).
31. Huey, E. B. *The Psychology and Pedagogy of Reading.* Pp. 469 (Macmillan, 1908).
32. Yarbus, A. L. *Eye Movements and Vision*. (Plenum Press, 1967). doi:10.1007/978-1-4899-5379-7.
33. Merabet, L. B. & Pascual-Leone, A. Neural reorganization following sensory loss: the opportunity of change. *Nature Reviews Neuroscience* **11**, 44–52 (2010).
34. Singh, A. K., Phillips, F., Merabet, L. B. & Sinha, P. Why Does the Cortex Reorganize after Sensory Loss? *Trends in Cognitive Sciences* **22**, 569–582 (2018).
35. Voss, P., Collignon, O., Lassonde, M. & Lepore, F. Adaptation to sensory loss. *WIREs Cognitive Science* **1**, 308–328 (2010).
36. Bänziger, T., Mortillaro, M. & Scherer, K. R. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion* **12**, 1161–1179 (2012).
37. Bassili, J. N. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology* **37**, 2049 (1979).
38. de Gelder, B., Teunisse, J.-P. & Benson, P. J. Categorical perception of facial expressions: Categories and their internal structure. *Cognition & Emotion* **11**, 1–23 (1997).
39. de Gelder, B. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**, 3475–3484 (2009).
40. Jessen, S. & Kotz, S. A. On the role of crossmodal prediction in audiovisual emotion percep-

tion. *Frontiers in human neuroscience* **7**, 369 (2013).

41. Banse, R. & Scherer, K. R. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology* **70**, 614 (1996).

42. Juslin, P. N. & Laukka, P. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin* **129**, 770 (2003).

43. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).

44. Itti, L. & Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* **40**, 1489–1506 (2000).

45. Hayhoe, M. & Ballard, D. Eye movements in natural behavior. *Trends in Cognitive Sciences* **9**, 188–194 (2005).

46. Võ, M. L.-H., Smith, T. J., Mital, P. K. & Henderson, J. M. Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision* **12**, 1–14 (2012).

47. Lansing, C. R. & McConkie, G. W. Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics* **65**, 536–552 (2003).

48. de Gelder, B. & Vroomen, J. The perception of emotions by ear and by eye. *Cognition & Emotion* **14**, 289–311 (2000).

49. Massaro, D. W. & Egan, P. B. Perceiving affect from the voice and the face. *Psychonomic bulletin & review* **3**, 215–221 (1996).

50. Paulmann, S., Titone, D. & Pell, M. D. How emotional prosody guides your way: Evidence from eye movements. *Speech Communication* **54**, 92–107 (2012).

51. Rigoulot, S. & Pell, M. D. Seeing Emotion with Your Ears: Emotional Prosody Implicitly Guides Visual Attention to Faces. *PLOS ONE* **7**, e30740 (2012).

52. Etzi, R., Ferrise, F., Bordegoni, M., Zampini, M. & Gallace, A. The Effect of Visual and Auditory Information on the Perception of Pleasantness and Roughness of Virtual Surfaces. *Multisensory Research* **31**, 501–522 (2018).

53. Samermit, P., Saal, J. & Davidenko, N. Cross-Sensory Stimuli Modulate Reactions to Aversive Sounds. *Multisensory Research* **32**, 197–213 (2019).

54. Taffou, M., Guerchouche, R., Drettakis, G. & Viaud-Delmon, I. Auditory–Visual Aversive Stimuli Modulate the Conscious Experience of Fear. *Multisensory Research* **26**, 347–370 (2013).

55. Collignon, O. *et al.* Audio-visual integration of emotion expression. *Brain research* **1242**, 126–135 (2008).

56. Kokinous, J., Kotz, S. A., Tavano, A. & Schroger, E. The role of emotion in dynamic audiovisual integration of faces and voices. *Social cognitive and affective neuroscience* **10**, 713–720 (2015).

57. Buchan, J. N., Paré, M. & Munhall, K. G. The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research* **1242**, 162–171 (2008).

58. Smith, M. L., Cottrell, G. W., Gosselin, F. & Schyns, P. G. Transmitting and Decoding Facial Expressions. *Psychol Sci* **16**, 184–189 (2005).

59. Groner, R., Walder, F. & Groner, M. Looking at faces: Local and global aspects of scanpaths.

in vol. 22 523–533 (Elsevier, 1984).

60. Walker-Smith, G. J., Gale, A. G. & Findlay, J. M. Eye Movement Strategies Involved in Face Perception. *Perception* **6**, 313–326 (1977).

61. Dael, N., Mortillaro, M. & Scherer, K. R. Emotion expression in body action and posture. *Emotion* **12**, 1085 (2012).

62. Paulmann, S. & Pell, M. D. Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion* **35**, 192–201 (2011).

63. Bänziger, T., Grandjean, D. & Scherer, K. R. Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). *Emotion* **9**, 691 (2009).

64. Jessen, S., Obleser, J. & Kotz, S. A. How bodies and voices interact in early emotion perception. *PLoS One* **7**, e36070 (2012).

65. Wallbott, H. G. & Scherer, K. R. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology* **51**, 690–699 (1986).

66. Skuk, V. G. & Schweinberger, S. R. Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices. *PloS one* **8**, e81691 (2013).

67. Takagi, S., Hiramatsu, S., Tabei, K. & Tanaka, A. Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality. *Frontiers in integrative neuroscience* **9**, 1 (2015).

68. Bach, M. The Freiburg Visual Acuity Test-variability unchanged by post-hoc re-analysis. *Graefe's Archive for Clinical and Experimental Ophthalmology* **245**, 965–971 (2006).

69. Bach, M. The Freiburg Visual Acuity Test-automatic measurement of visual acuity. *Optometry & Vision Science* **73**, 49–53 (1996).

70. Russell, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology* **39**, 1161–1178 (1980).

71. Brainard, D. H. The psychophysics toolbox. *Spatial vision* **10**, 433–436 (1997).

72. Kleiner, M. *et al.* What's new in Psychtoolbox-3. *Perception* **36**, 1 (2007).

73. Pelli, D. G. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision* **10**, 437–442 (1997).

74. Cornelissen, F. W., Peters, E. M. & Palmer, J. The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods* **34**, 613–617 (2002).

75. Crosse, M. J., Liberto, G. M. D. & Lalor, E. C. Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **36**, 9888–9895 (2016).

76. Stevenson, R. A. *et al.* Identifying and quantifying multisensory integration: a tutorial review. *Brain topography* **27**, 707–730 (2014).

77. Wagner, H. L. On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior* **17**, 3–28 (1993).

78. Calvert, G. A. Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cereb Cortex* **11**, 1110–1123 (2001).

79. Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G. & Fendrich, R. Visual-auditory interactions in sensorimotor processing: saccades versus manual responses. *J Exp Psychol Hum Percept*

Perform **20**, 131–153 (1994).

80. Stein, B. E. & Meredith, M. A. *The merging of the senses.* (1993).
81. Meredith, M. A. & Stein, B. E. Interactions among converging sensory inputs in the superior colliculus. *Science* **221**, 389–391 (1983).
82. Angelaki, D. E., Gu, Y. & DeAngelis, G. C. Multisensory integration. *Curr Opin Neurobiol* **19**, 452–458 (2009).
83. Stanford, T. R. & Stein, B. E. Superadditivity in multisensory integration: putting the computation in context. *NeuroReport* **18**, 787 (2007).
84. Baron-Cohen, S., Wheelwright, S. & Jolliffe, T. Is There a 'Language of the Eyes'? Evidence from Normal Adults, and Adults with Autism or Asperger Syndrome. *Visual Cognition* **4**, 311–331 (1997).
85. Schyns, P. G., Petro, L. S. & Smith, M. L. Dynamics of Visual Information Integration in the Brain for Categorizing Facial Expressions. *Current Biology* **17**, 1580–1585 (2007).
86. Bal, E. *et al.* Emotion Recognition in Children with Autism Spectrum Disorders: Relations to Eye Gaze and Autonomic State. *Journal of Autism and Developmental Disorders* **40**, 358–370 (2010).
87. Lischke, A. *et al.* Intranasal oxytocin enhances emotion recognition from dynamic facial expressions and leaves eye-gaze unaffected. *Psychoneuroendocrinology* **37**, 475–481 (2012).
88. Eisenbarth, H. & Alpers, G. W. Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion* **11**, 860–865 (2011).
89. Blais, C., Fiset, D., Roy, C., Régimbald, C. S. & Gosselin, F. Eye fixation patterns for categorizing static and dynamic facial expressions. *Emotion* **17**, 1107 (2017).
90. Calder, A. J., Young, A. W., Keane, J. & Dean, M. Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance* **26**, 527–551 (2000).
91. Hsiao, J. H. & Cottrell, G. Two Fixations Suffice in Face Recognition. *Psychol Sci* **19**, 998–1006 (2008).
92. Peterson, M. F. & Eckstein, M. P. Looking just below the eyes is optimal across face recognition tasks. *PNAS* **109**, E3314–E3323 (2012).
93. Posner, M. I. Orienting of attention. *Quarterly Journal of Experimental Psychology* **32**, 3–25 (1980).
94. Thompson, B., Hansen, B. C., Hess, R. F. & Troje, N. F. Peripheral vision: Good for biological motion, bad for signal noise segregation? *Journal of Vision* **7**, 12–12 (2007).
95. Picou, E. M. How Hearing Loss and Age Affect Emotional Responses to Nonspeech Sounds. *J Speech Lang Hear Res* **59**, 1233–1246 (2016).
96. Alvarsson, J. J., Wiens, S. & Nilsson, M. E. Stress Recovery during Exposure to Nature Sound and Environmental Noise. *International Journal of Environmental Research and Public Health* **7**, 1036–1046 (2010).
97. Husain, G., Thompson, W. F. & Schellenberg, E. G. Effects of Musical Tempo and Mode on Arousal, Mood, and Spatial Abilities. *Music Perception* **20**, 151–171 (2002).
98. Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. D. Bad is Stronger than Good. *Review of General Psychology* **5**, 323–370 (2001).

99. Luo, X., Kern, A. & Pulling, K. R. Vocal emotion recognition performance predicts the quality of life in adult cochlear implant users. *The Journal of the Acoustical Society of America* **144**, EL429–EL435 (2018).

100. Colijn, J. M. *et al.* Prevalence of Age-Related Macular Degeneration in Europe. *Ophthalmology* **124**, 1753–1763 (2017).

101. Barnes, C. S., De l'Aune, W. & Schuchard, R. A. A Test of Face Discrimination Ability in Aging and Vision Loss. *Optometry and Vision Science* **88**, 188 (2011).

102. Boucart, M. *et al.* Recognition of facial emotion in low vision: A flexible usage of facial features. *Visual Neuroscience* **25**, 603–609 (2008).

103. Johnson, A. P., Woods-Fry, H. & Wittich, W. Effects of Magnification on Emotion Perception in Patients With Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **58**, 2520–2526 (2017).

104. Raphael, L. J., Borden, G. J. & Harris, K. S. *Speech science primer: Physiology, acoustics, and perception of speech*. (Lippincott Williams & Wilkins, 1980).

105. Ling, D. *Speech and the Hearing-Impaired Child: Theory and Practice*. (Alexander Graham Bell Association for the Deaf, 1976).

106. Moore, B. C. J. Perceptual Consequences of Cochlear Hearing Loss and their Implications for the Design of Hearing Aids. *Ear and Hearing* **17**, (1996).

107. Most, T. & Aviner, C. Auditory, Visual, and Auditory–Visual Perception of Emotions by Individuals With Cochlear Implants, Hearing Aids, and Normal Hearing. *J Deaf Stud Deaf Educ* **14**, 449–464 (2009).

108. Rigo, T. G. & Lieberman, D. A. Nonverbal Sensitivity of Normal-Hearing and Hearing-Impaired Older Adults: *Ear and Hearing* **10**, 184–189 (1989).

109. Nagels, L. *et al.* Development of vocal emotion recognition in school-age children: The EmoHI test for hearing-impaired populations. *PeerJ* **8**, e8773 (2020).

110. Goy, H., Pichora-Fuller, M. K., Singh, G. & Russo, F. A. Perception of emotional speech by listeners with hearing aids. *Canadian Acoustics* **44**, (2016).

111. Orbelo, D. M., Grim, M. A., Talbott, R. E. & Ross, E. D. Impaired Comprehension of Affective Prosody in Elderly Subjects Is Not Predicted by Age-Related Hearing Loss or Age-Related Cognitive Decline. *Journal of Geriatric Psychiatry and Neurology* **18**, 25–32 (2005).

112. Taylor, D. J., Edwards, L. A., Binns, A. M. & Crabb, D. P. Seeing it differently: self-reported description of vision loss in dry age-related macular degeneration. *Ophthalmic and Physiological Optics* **38**, 98–105 (2018).

113. Roth, T. N., Hanebuth, D. & Probst, R. Prevalence of age-related hearing loss in Europe: a review. *European Archives of Oto-Rhino-Laryngology* **268**, 1101–1107 (2011).

114. de Boer, M. J., Başkent, D. & Cornelissen, F. W. Eyes on Emotion: Dynamic Gaze Allocation During Emotion Perception From Speech-Like Stimuli. *Multisensory Research* 1–31 (2020) doi:10.1163/22134808-bja10029.

115. Hooge, I. Th. C. & Erkelens, C. J. Adjustment of fixation duration in visual search. *Vision Research* **38**, 1295-IN4 (1998).

116. Cummings, R. W., Whittaker, S. G., Watson, G. R. & Budd, J. M. Scanning Characters and Reading with a Central Scotoma: *Optometry and Vision Science* **62**, 833–843 (1985).

117. Schuchard, R. Preferred retinal locus: a review with application in low vision rehabilitation. *Low Vision and Vision Rehabil* **7**, 243–256 (1994).

118. Fletcher, D. C. & Schuchard, R. A. Preferred Retinal Loci Relationship to Macular Scotomas in a Low-vision Population. *Ophthalmology* **104**, 632–638 (1997).

119. Sunness, J. S., Applegate, C. A., Haselwood, D. & Rubin, G. S. Fixation Patterns and Reading Rates in Eyes with Central Scotomas from Advanced Atrophic Age-related Macular Degeneration and Stargardt Disease. *Ophthalmology* **103**, 1458–1466 (1996).

120. Bertera, J. H. The effect of simulated scotomas on visual search in normal subjects. *Invest. Ophthalmol. Vis. Sci.* **29**, 470–475 (1988).

121. Cornelissen, F. W., Bruin, K. J. & Kooijman, A. C. The Influence of Artificial Scotomas on Eye Movements during Visual Search. *Optometry and Vision Science* **82**, 27–35 (2005).

122. Hunter, E. M., Phillips, L. H. & MacPherson, S. E. Effects of age on cross-modal emotion perception. *Psychology and Aging* **25**, 779–787 (2010).

123. Moraitou, D., Papantoniou, G., Gkinopoulos, T. & Nigritinou, M. Older adults' decoding of emotions: age-related differences in interpreting dynamic emotional displays and the well-preserved ability to recognize happiness: Emotion decoding in ageing. *Psychogeriatrics* **13**, 139–147 (2013).

124. Siebe, T., Williges, B., Oetting, D., Hohmann, V. & Jürgens, T. Evaluation einer modularen Auralisation von sensorineuraler Schwerhörigkeit. in 1–4 (2017).

125. Nejime, Y. & Moore, B. C. J. Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America* **102**, 603–615 (1997).

126. Moore, B. C. J. *Cochlear Hearing Loss*. (Wiley, 1998).

127. Ghitza, O. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *The Journal of the Acoustical Society of America* **110**, 1628–1640 (2001).

128. Bennett, R. M. C. & Hohmann, V. Simulation of reduced frequency selectivity found with cochlear hearing loss using a model based procedure. *Annual Meeting of the German Society of Audiology* (2012).

129. Brand, T. & Hohmann, V. An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America* **112**, 1597–1604 (2002).

130. Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B. & Ewert, S. D. Spectral and binaural loudness summation for hearing-impaired listeners. *Hearing Research* **335**, 179–192 (2016).

131. Bisgaard, N., Vlaming, M. S. M. G. & Dahlquist, M. Standard Audiograms for the IEC 60118-15 Measurement Procedure. *Trends in Amplification* **14**, 113–120 (2010).

132. Cheung, S.-H. & Legge, G. E. Functional and cortical adaptations to central vision loss. *Visual Neuroscience* **22**, 187–201 (2005).

133. Varsori, M., Perez-Fornos, A., Safran, A. B. & Whatham, A. R. Development of a viewing strategy during adaptation to an artificial central scotoma. *Vision Research* **44**, 2691–2705 (2004).

134. Walsh, D. V. & Liu, L. Adaptation to a simulated central scotoma during visual search training. *Vision Research* **96**, 75–86 (2014).

135. McIlreavy, L., Fiser, J. & Bex, P. J. Impact of Simulated Central Scotomas on Visual Search in Natural Scenes: *Optometry and Vision Science* **89**, 1385–1394 (2012).

136. Henderson, J. M., Mcclure, K. K., Pierce, S. & Schrock, G. Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Perception & Psychophysics* **59**, 323–346 (1997).

137. Fischer, M. E. *et al.* Multiple Sensory Impairment and Quality of Life. *Ophthalmic Epidemiol* **16**, 346–353 (2009).

138. de Boer, M. J., Jürgens, T., Cornelissen, F. W. & Başkent, D. Degraded visual and auditory input individually impair audiovisual emotion recognition from speech-like stimuli, but no evidence for an exacerbated effect from combined degradation. *Vision Research* **180**, 51–62 (2021).

139. Roets-Merken, L. M., Zuidema, S. U., Vernooij-Dassen, M. J. F. J. & Kempen, G. I. J. M. Screening for hearing, visual and dual sensory impairment in older adults using behavioural cues: A validation study. *International Journal of Nursing Studies* **51**, 1434–1440 (2014).

140. Gates, G. A. & Mills, J. H. Presbycusis. *The Lancet* **366**, 1111–1120 (2005).

141. Wong, W. L. *et al.* Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet Global Health* **2**, e106–e116 (2014).

142. Steinmetz, J. D. *et al.* Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *The Lancet Global Health* **9**, e144–e160 (2021).

143. Christensen, J. A., Sis, J., Kulkarni, A. M. & Chatterjee, M. Effects of age and hearing loss on the recognition of emotions in speech. *Ear Hear* **40**, 1069–1083 (2019).

144. Guthrie, D. M., Declercq, A., Finne-Soveri, H., Fries, B. E. & Hirdes, J. P. The Health and Well-Being of Older Adults with Dual Sensory Impairment (DSI) in Four Countries. *PLOS ONE* **11**, e0155073 (2016).

145. Roets-Merken, L. M., Zuidema, S. U., Vernooij-Dassen, M. J. & Kempen, G. I. Screening for hearing, visual and dual sensory impairment in older adults using behavioural cues: A validation study. *International journal of nursing studies* **51**, 1434–1440 (2014).

146. Saunders, G. H. & Echt, K. V. An Overview of Dual Sensory Impairment in Older Adults: Perspectives for Rehabilitation. *Trends in Amplification* **11**, 243–258 (2007).

147. Schneider, J. M. *et al.* Dual Sensory Impairment in Older Age. *J Aging Health* **23**, 1309–1324 (2011).

148. Gonçalves, A. R. *et al.* Effects of age on the identification of emotions in facial expressions: a meta-analysis. *PeerJ* **6**, e5278 (2018).

149. Ruffman, T., Henry, J. D., Livingstone, V. & Phillips, L. H. A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews* **32**, 863–881 (2008).

150. Olderbak, S., Wilhelm, O., Hildebrandt, A. & Quoidbach, J. Sex differences in facial emotion perception ability across the lifespan. *Cognition and Emotion* **33**, 579–588 (2019).

151. Wieck, C. & Kunzmann, U. Age differences in emotion recognition: A question of modality?

Psychology and Aging **32**, 401–411 (2017).

152. Lambrecht, L., Kreifelts, B. & Wildgruber, D. Age-related decrease in recognition of emotional facial and prosodic expressions. *Emotion* **12**, 529–539 (2012).
153. Calder, A. J. *et al.* Facial expression recognition across the adult life span. *Neuropsychologia* **41**, 195–202 (2003).
154. Orgeta, V. & Phillips, L. H. Effects of Age and Emotional Intensity on the Recognition of Facial Emotion. *Experimental Aging Research* **34**, 63–79 (2007).
155. West, J. T. *et al.* Age Effects on Emotion Recognition in Facial Displays: From 20 to 89 Years of Age. *Experimental Aging Research* **38**, 146–168 (2012).
156. Carstensen, L. L. & DeLiema, M. The positivity effect: a negativity bias in youth fades with age. *Current Opinion in Behavioral Sciences* **19**, 7–12 (2018).
157. Mather, M. & Carstensen, L. L. Aging and motivated cognition: the positivity effect in attention and memory. *Trends in Cognitive Sciences* **9**, 496–502 (2005).
158. Sullivan, S., Ruffman, T. & Hutton, S. B. Age Differences in Emotion Recognition Skills and the Visual Scanning of Emotion Faces. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **62**, P53–P60 (2007).
159. Wong, B., Cronin-Golomb, A. & Neargarder, S. Patterns of Visual Scanning as Predictors of Emotion Identification in Normal Aging. *Neuropsychology* **19**, 739–749 (2005).
160. Blais, C., Fiset, D., Roy, C., Saumure Régimbald, C. & Gosselin, F. Eye fixation patterns for categorizing static and dynamic facial expressions. *Emotion* **17**, 1107–1119 (2017).
161. Khosdelazad, S. *et al.* Comparing static and dynamic emotion recognition tests: Performance of healthy participants. *PLOS ONE* **15**, e0241297 (2020).
162. Deary, I. J. *et al.* Age-associated cognitive decline. *British Medical Bulletin* **92**, 135–152 (2009).
163. Martini, A. European Working Group on Genetics of Hearing Impairment Infoletter 2, European Commission Directorate. *Biomedical and health research programme (HEAR)* (1996).
164. Wieling, M., Montemagni, S., Nerbonne, J. & Baayen, R. H. Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language* **90**, 669–692 (2014).
165. Sokal, R. R. & Rohlf, F. J. *Biometry: the principles and practice of statistics in biological research*. (W. H. Freeman, 1995).
166. Knight, M. *et al.* Aging and goal-directed emotional attention: Distraction reverses emotional biases. *Emotion* **7**, 705–714 (2007).
167. Noh, S. R. & Isaacowitz, D. M. The effects of varying contextual demands on age-related positive gaze preferences. *Psychology and Aging* **30**, 356–368 (2015).
168. Mack, D. J. *et al.* The effect of age and gender on anti-saccade performance: Results from a large cohort of healthy aging individuals. *European Journal of Neuroscience* **52**, 4165–4184 (2020).
169. Pratt, J., Dodd, M. & Welsh, T. Growing Older Does Not Always Mean Moving Slower: Examining Aging and the Saccadic Motor System. *Journal of Motor Behavior* **38**, 373–382 (2006).
170. Warabi, T., Kase, M. & Kato, T. Effect of aging on the accuracy of visually guided saccadic eye movement. *Annals of Neurology* **16**, 449–454 (1984).
171. Lawrence, R. K., Edwards, M. & Goodhew, S. C. Changes in the spatial spread of attention

with ageing. *Acta Psychologica* **188**, 188–199 (2018).

172. Coeckelbergh, T. R. M., Cornelissen, F. W., Brouwer, W. H. & Kooijman, A. C. Age-Related Changes in the Functional Visual Field: Further Evidence for an Inverse Age × Eccentricity Effect. *The Journals of Gerontology: Series B* **59**, P11–P18 (2004).
173. Sekuler, A. B., Bennett, P. J. & Mamelak, M. Effects of Aging on the Useful Field of View. *Experimental Aging Research* **26**, 103–120 (2000).
174. Phillips, L. H., Channon, S., Tunstall, M., Hedenstrom, A. & Lyons, K. The role of working memory in decoding emotions. *Emotion* **8**, 184–191 (2008).
175. Edwards, J. D., Fausto, B. A., Tetlow, A. M., Corona, R. T. & Valdés, E. G. Systematic review and meta-analyses of useful field of view cognitive training. *Neuroscience & Biobehavioral Reviews* **84**, 72–91 (2018).
176. Chung, S. T. L., Legge, G. E. & Cheung, S. Letter-recognition and reading speed in peripheral vision benefit from perceptual learning. *Vision Research* **44**, 695–709 (2004).
177. de Dieuleveult, A. L., Siemonsma, P. C., van Erp, J. B. F. & Brouwer, A.-M. Effects of Aging in Multisensory Integration: A Systematic Review. *Front. Aging Neurosci.* **9**, (2017).
178. Freiherr, J., Lundström, J. N., Habel, U. & Reetz, K. Multisensory integration mechanisms during aging. *Front. Hum. Neurosci.* **7**, (2013).
179. de Boer-Schellekens, L. & Vroomen, J. Multisensory integration compensates loss of sensitivity of visual temporal order in the elderly. *Exp Brain Res* **232**, 253–262 (2014).
180. Birmingham, E., Bischof, W. F. & Kingstone, A. Saliency does not account for fixations to eyes within social scenes. *Vision Research* **49**, 2992–3000 (2009).
181. Henderson, J. M., Williams, C. C. & Falk, R. J. Eye movements are functional during face learning. *Memory & Cognition* **33**, 98–106 (2005).
182. Hessels, R. S., Cornelissen, T. H. W., Hooge, I. T. C. & Kemner, C. Gaze behavior to faces during dyadic interaction. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* **71**, 226–242 (2017).
183. Pambakian, A., Currie, J. & Kennard, C. Rehabilitation Strategies for Patients With Homonymous Visual Field Defects. *Journal of Neuro-Ophthalmology* **25**, 136–142 (2005).
184. Pambakian, A. L. M., Mannan, S. K., Hodgson, T. L. & Kennard, C. Saccadic visual search training: a treatment for patients with homonymous hemianopia. *Journal of Neurology, Neurosurgery & Psychiatry* **75**, 1443–1448 (2004).
185. Sabel, B. A. & Gudlin, J. Vision Restoration Training for Glaucoma: A Randomized Clinical Trial. *JAMA Ophthalmology* **132**, 381–389 (2014).
186. Maniglia, M., Cottereau, B. R., Soler, V. & Trotter, Y. Rehabilitation Approaches in Macular Degeneration Patients. *Front. Syst. Neurosci.* **10**, (2016).
187. Gatehouse, S. & Akeroyd, M. Two-eared listening in dynamic situations: Audición con dos oídos en situaciones dinámicas. *International Journal of Audiology* **45**, 120–124 (2006).

Summary

Human communication involves emotional expressions, which are often carried by multisensory signals. Emotion recognition, therefore, requires audiovisual integration of these signals. It is already known that advanced age and sensory impairments can negatively impact emotion recognition. Even though sensory impairments are common in older adults, much is still unknown about the interplay between age and sensory impairments for emotion recognition. When a sensory impairment affects only a single modality (e.g., vision or hearing), compensation for the impaired modality by relying more on the intact modality is likely possible. However, in the case of an impairment in both modalities, compensation is less likely. This may be especially so in older age, where age-related cognitive changes may alter compensation mechanisms as well. Moreover, more people are reaching an older age, which is accompanied by an increase in the prevalence of sensory impairments. Therefore, it is urgent to understand the mechanisms of audiovisual integration with sensory impairments, in particular in older individuals. To investigate this properly, however, it is necessary to first create a basic understanding of audiovisual integration for emotion recognition. Therefore, the primary aim of the research presented in my thesis was to comprehensively investigate how auditory and visual emotional information are combined and how audiovisual interactions change with simulated sensory impairments. A secondary aim was to understand how age affects these outcomes, as integration mechanisms employed by young observers may be different in older observers.

To systematically address these aims, I examined how well observers recognize dynamic audiovisual emotions, presented via videos, and how emotion recognition, assessed via identification accuracy, and perceptual strategies, assessed via eye-tracking, vary under changing availability and reliability of the visual and auditory information. Reliability of the modalities was altered by simulating a central visual field defect (as occurring in macular degeneration) in the video, and simulating a high-frequency hearing loss (as in age-related hearing loss) in the audio.

The main results are that older observers are as effective at audiovisual integration as younger observers, although they show a general decline in overall emotion recognition accuracy. In addition, both younger and older observers can compensate for degraded auditory information when the video is intact. For either group, compensation for degraded visual information was not possible with our stimuli. The gaze data showed that younger and older observers both flexibly adapt their gaze, but the degree of adaptation seemed to differ. When audio is removed, more fixations are on the mouth of the actor, and less on the eyes. When video is degraded, larger saccades are made to move the simulated defect away from the actor's face. Older observers, however, showed less gaze adaptation, possibly indicating they were not capable of efficient adaptation.

Concluding, the research presented in my thesis shows that audiovisual integration and compensation abilities remain intact with age, despite a general decline in emotion recogni-

tion accuracy. Compensation for degraded audio is possible by relying more on the video, but not vice versa. Older observers adapt their perceptual strategies in a different, and perhaps less efficient, manner than younger observers. Importantly, I demonstrate that it is crucial to use additional measurements besides recognition accuracy (here, eye-tracking) in order to understand audiovisual integration and compensation mechanisms for emotion recognition. Additional measurements such as eye-tracking allow for examining whether the reliance on visual and auditory information alters with age and different reliabilities of the modalities, even when there is no change in emotion recognition accuracy.

Nederlandse samenvatting

Een van de belangrijkste vaardigheden voor sociale wezens als de mens is goed kunnen communiceren. Communicatie is noodzakelijk voor het overbrengen van ideeën, maar ook om relaties met anderen op te bouwen, en hiermee ook voor het algemeen welzijn. Hoewel het verstaan van andermans woorden noodzakelijk is voor succesvolle communicatie, is het essentieel om ook de emotie achter die woorden te herkennen om de intenties van deze partner echt te begrijpen. Emoties worden geuit via verschillende modaliteiten, maar voornamelijk via visuele (gezichtsuitdrukkingen) en auditieve (stem) signalen, en voor het herkennen van emoties is integratie van deze signalen dan ook noodzakelijk. Uit eerder onderzoek is al bekend dat zowel een hoge leeftijd als zintuiglijke beperkingen een negatieve invloed hebben op emotieherkenning. Hoewel zintuiglijke beperkingen vaak voorkomen in de oudere bevolking, is nog veel onbekend over de wisselwerking tussen leeftijd en zintuiglijke beperkingen voor emotieherkenning. Wanneer een zintuiglijke beperking slechts in een enkele modaliteit voorkomt (zoals bij slechtziendheid of gehoorverlies), is compensatie voor de beperking waarschijnlijk mogelijk door meer gebruik te maken van de intacte modaliteit. Wanneer er echter een beperking is in zowel het zicht als gehoor, zal compensatie waarschijnlijk minder goed mogelijk zijn. Dit zal nog sterker het geval zijn bij een hoge leeftijd, waar leeftijd gerelateerde veranderingen in cognitief functioneren tevens invloed hebben op de onderliggende mechanismen voor compensatie. Nu mensen steeds ouder worden, en de prevalentie van zintuiglijke beperkingen ook zal stijgen, is het belangrijk dat de kennis over audiovisuele integratie voor emotieherkenning bij een beperking in de zintuigen wordt uitgebreid, in het bijzonder bij ouderen. Echter, om dit goed te onderzoeken, is het noodzakelijk om eerst basiskennis op te bouwen over audiovisuele integratie bij emotieherkenning. Daarom was het primaire doel van dit proefschrift om uitgebreid te onderzoeken hoe visuele en auditieve emotionele informatie gecombineerd worden en hoe audiovisuele interacties veranderen wanneer sprake is van (gesimuleerde) zintuiglijke beperkingen. Een secundair doel was om te begrijpen hoe veroudering deze interacties beïnvloedt, aangezien strategieën voor het integreren van visuele en auditieve signalen mogelijk veranderen naarmate men ouder wordt.

Om deze doelen systematisch aan te pakken, heb ik onderzocht hoe goed mensen dynamische en audiovisuele emoties, gepresenteerd via video's, herkennen. Daarnaast heb ik onderzocht hoe emotieherkenning, geëvalueerd via nauwkeurigheid, en kijkstrategieën, geëvalueerd via oogbewegingen, variëren bij veranderende beschikbaarheid en betrouwbaarheid van de visuele en auditieve informatie. Betrouwbaarheid van de visuele en auditieve informatie was aangepast door het simuleren van een centraal gezichtsvelddefect (als voorkomend bij maculadegeneratie) in de video, en het simuleren van hoogfrequent gehoorverlies (als voorkomend bij ouderdomsslechthorendheid) in de audio.

De voornaamste resultaten zijn dat zowel jongeren als ouderen emoties beter herkennen wanneer zowel visuele als auditieve informatie beschikbaar is en emotieherkenning is beter met alleen visuele dan alleen auditieve informatie. Ook zijn ouderen net zo goed in het integreren van visuele en auditieve emotionele informatie als jongeren. Echter is er wel een algehele achteruitgang in emotieherkenning bij ouderen. Daar komt bij dat zowel jongere als oudere waarnemers volledig kunnen compenseren voor gesimuleerd gehoorverlies wanneer de video intact is. Compensatie voor gesimuleerde maculadegeneratie was echter voor beide groepen niet mogelijk met de gebruikte stimuli. Oogbewegingsdata toont aan dat zowel jongeren als ouderen flexibel hun kijkstrategieën aanpassen aan de beschikbaarheid en betrouwbaarheid van de informatie, hoewel de mate van aanpassing verschilt. Wanneer alleen de video beschikbaar was zonder de bijbehorende audio, werd meer naar de mond, en minder naar de ogen van de acteur gekeken. Bij gesimuleerde maculadegeneratie werden grotere saccades (snelle oogbewegingen) gemaakt en verder weg van het gezicht van de acteur gefixeerd om de gezichtsuitdrukkingen toch met het perifere zicht te kunnen zien. Ouderen maakten kleinere aanpassingen in hun kijkstrategieën, waarschijnlijk omdat ze niet in staat waren om hun aandacht ver van het punt van fixatie te richten. Daarnaast hadden ouderen een algemene bias om meer naar de mond van de acteur te kijken, terwijl jongeren hun fixaties gelijk verdeelden tussen de mond en de ogen.

Concluderend heeft het onderzoek in dit proefschrift aangetoond dat audiovisuele integratie intact blijft met veroudering, hoewel veroudering leidt tot een algehele achteruitgang van emotieherkenning. Compensatie voor minder betrouwbare audio is mogelijk door de visuele signalen beter te gebruiken, maar het omgekeerde is niet mogelijk. Ouderen passen hun kijkstrategieën op een andere, en wellicht minder efficiënte, manier aan dan jongeren. Belangrijker nog is dat ik in dit proefschrift aantoon dat het cruciaal is om meerdere uitkomstmaten te gebruiken naast de nauwkeurigheid van emotieherkenning (hier door het gebruik van oogbewegingsmetingen) om audiovisuele integratie en compensatiemechanismen voor emotieherkenning te begrijpen. Aanvullende uitkomstmaten zoals het meten van oogbewegingen maken het mogelijk om te onderzoeken of de weging van visuele en auditieve informatie verandert met leeftijd en betrouwbaarheid van de informatie, ook wanneer er geen verandering is in de nauwkeurigheid van emotieherkenning.

Acknowledgements

Doing a PhD and especially writing a thesis is very much a solitary journey, as finishing the project is mainly your own responsibility. Despite its individualistic nature, this thesis would not be here without the help, support, and advice from many people. Therefore, some thanks to them are in order.

First, I would like to express my appreciation for my wonderful supervisors Frans Cornelissen and Deniz Başkent, who have shared the load of supervising me. Frans, during my major project you (and Hinke and Barbara) showed me that science is hard, but that this difficulty is probably half the fun. Your encouragement and suggestions helped made it possible that I could pursue this PhD project, which has been both fun and difficult. I very much appreciate you showing me that psychophysics is amazing and that eye-tracking is so much more than just filming someone's eye while they do a task. Deniz, thank you for all your enthusiasm and kind words throughout my PhD. I am also grateful for you being a role model for succeeding in academia as a woman and at the same time pointing out that this is still not an unremarkable achievement. I sometimes feel that you and Frans are complete opposites, which has made decision making and writing challenging sometimes. However, also because of this, I had the opportunity to experience two very different supervisors, and it challenged me to decide what I thought would be the optimal solution instead of relying mostly on a single source.

Thank you both for all your support, feedback, and discussions throughout these (nearly) five years and teaching me, more than anything, that doing science means perseverance and being both independent and collaborative.

I am thankful for all the technical help and ideas regarding analyses and statistics that I received from Jan-Bernard Marsman, Remco Renken, and Paolo Toffanin. I have learned a lot from all of you, and my manuscripts have certainly improved by incorporating your suggestions. Paolo, special thanks to you for providing so many interesting, unexpected, and nonsensical topics for discussion during lunch and coffee breaks.

Tim Jürgens, thank you for a very pleasant collaboration regarding the hearing impairment simulations. It was amazing that we could use your simulations and thank you for providing me with the opportunity to come to Lübeck to discuss our work in person and see your labs.

I would like to thank the reading committee members Prof. dr. Marleen Janssen, Prof. dr. Stefan van der Stigchel, and Prof. Raymond van Ee for taking the time to review this thesis.

The ENT department has been a lovely place to work in. I am grateful for all my fellow PhDs at ENT, you made work more enjoyable and it was always comforting to hear we all run into

problems with our research. Also, thank you for the good talks during lunch, coffee, and research meetings. Sina, I thoroughly enjoyed sharing an office with you and thank you for your efforts in making sure we did not get a third office mate.
Of course, even though I did not have an office there, many thanks to everyone from the Visual Neuroscience Group. Thank you for all your input during lab meetings and talks before and after. Birte and Rijul, thank you in particular with your help with analyses of eye-tracking data. Alessandro, I cannot thank you enough for explaining how the Eyelink should be operated and especially for all the Matlab scripts you made that I could use for my experiments and analyses. I don't think I would have been able to do the studies I did without those scripts, at least it would have taken me much more time.

I gratefully acknowledge the assistance from all the more senior and support staff from ENT, Opthalmology, and the CNC. Sonja, Pim, Emile, Anita, Etienne, Thomas, Laura, Terrin, Wiebe, and Diek: thank you all for your comments, suggestions, and praises during the research meetings. In addition, I thank you for showing and teaching me all the different branches of audiology related research. Ria and Jennifer, thank you for helping with everything regarding finances. Kim, thank you for helping me setup and execute my patient measurements. And Hedwig, thank you for helping me find keys whenever they were lost and never forgetting to ask how my project is going.

I also wish to thank Frank Zaal for introducing me to scientific research and making me enthusiastic about pursuing a career in academia because of your fine supervision during my Bachelor thesis.

Merel, bedankt dat je mij geholpen hebt met het verzamelen van data en dat ik je mocht begeleiden bij het schrijven van je scriptie. Het was prettig om met je samen te werken en je aanwezigheid maakte dat de dataverzameling net even wat minder saai was.

Famke en Inge, bedankt dat jullie mijn paranimfen wilden zijn! Het voelt goed dat twee mensen die ik al zo lang ken mij helpen met de laatste loodjes en mij bijstaan op één van de spannendste momenten in mijn leven.

The research presented in thesis would not have been possible without the generous contribution of all the participants that agreed to take part in my studies.

While this thesis would not be here without all the technical and practical help and the supervision I received from all of my colleagues, possibly equally important for its completion is the mental support I received from my friends and family.

Luka, Marije, Famke en Yvonne, ik ben heel blij dat ik jullie heb leren kennen tijdens het introkamp van BW, ik had toen nooit kunnen voorspellen wat een goede vrienden we zouden worden. Ik kijk met een glimlach terug op onze studietijd, inclusief gezamelijk stressen over tentamens, de stapavondjes, en het samen sporten en eten. De stapavonden zitten er inmiddels niet echt meer in, maar de gezelligheid is er zeker niet minder op geworden. Meiden, bedankt voor alle steun en gezelligheid, fijn dat ik af en toe lekker tegen jullie aan kon zeuren. Helaas wonen we niet meer met z'n allen in Groningen, en is samen afspreken wat lastiger geworden, maar ik heb er alle vertrouwen in dat we nog lang vrienden zullen blijven.

Inge, als geen ander snapte jij de struggles van het promoveren en hierdoor konden we ons verhaal bij elkaar kwijt en elkaar helpen relativeren. Natuurlijk hebben we altijd ook genoeg andere gespreksstof, zeker aangezien we elkaar al zo lang kennen. Ik vond het super leuk dat je ook in Groningen woonde, zodat we regelmatig samen konden gaan hardlopen, eten, en bijkletsen na werk. Juliette, we zien elkaar zelden en spreken elkaar eigenlijk ook niet zo vaak, maar toch is het elke keer als ik je wel zie net alsof we elkaar dagelijks zien. Ik heb het gevoel dat ik met alles wel bij je kan komen en je bent altijd erg goed in dingen relativeren, waardoor het altijd fijn is als ik je weer eens gesproken hebt. Het lijkt me wel een goed streven om elkaar proberen vaker te zien dan dat ik je vader zie. Jaap en Jacco, hoewel ik jullie nauwelijks spreek, mogen jullie toch niet ontbreken in dit rijtje. Wanneer we elkaar wel spreken is het altijd gezellig en ik ben blij dat we toch enigszins contact hebben gehouden na school (ook al is dat wellicht grotendeels te danken aan Inge en Juliette).

Ik wil ook al mijn trainingsmaatjes van Survivalrun Groningen bedanken. Jullie zorgden er regelmatig voor dat ik me zowel fysiek als mentaal even kon uitleven. Hard trainen wanneer ik eigenlijk gaar en moe was, leidde ertoe dat ik altijd met een voldaan gevoel, en zware benen en armen, weer naar huis fietste. Ik mis jullie gezelligheid en de trainingen nu ik niet meer in Groningen woon, wie weet komen we elkaar nog eens tegen op een run.

Ook Femke, Alle-Jan, Carin, Arsèn, Mark, Hilde, Auke, Djoeke, en Kees wil ik hier graag bedanken. Ik ben dan weliswaar als aanhangsel in de groep gekomen, toch gaven jullie mij al snel het gevoel dat ik er echt bij hoorde. Bedankt voor alle mooie avonden en weekends (wanneer gaan we weer naar Disney?), de pubquizzen het afgelopen jaar, de goede gesprekken, en de gezelligheid.

Ook ben ik veel dank verschuldigd aan mijn familie. Bedankt voor jullie steun en pogingen te begrijpen wat ik doe. Mam, bedankt voor al je hulp bij praktische zaken en dat ik altijd bij je terecht kan en thuis kan komen. Papa en Ennadien, bedankt dat jullie me altijd ophaalden in Amersfoort als ik weer eens langskwam, voor de gezelligheid, en het lekkere eten. Peter en Hilly, bedankt voor jullie interesse en betrokkenheid in wat ik doe, en mij het gevoel te geven dat ik onderdeel ben van de familie Veldman.

Lieve Menno, ik kan met recht zeggen dat je mijn steun en toeverlaat bent. Je staat altijd voor me klaar en helpt me te relativeren wanneer ik dat zelf niet kan. Ook dwing je me (met zachte hand) om uitdagingen aan te gaan, iets wat ik niet altijd waardeer, maar wat er uiteindelijk mede voor gezorgd heeft dat ik het aandurfde een PhD te gaan doen en met toch enige zekerheid mijn verdediging inga. Mijn dank hiervoor is groot en ik hoop dan ook dat je me zal blijven uitdagen. Ik kan niet zeggen of je net zo voor mij klaar stond als ik voor jou, maar ik kan je wel verzekeren dat ik heel veel aan je steun gehad heb de afgelopen jaren.

About the author

Minke de Boer was born on July 30 1992 in Utrecht, the Netherlands. In 2011, she started her bachelor program in Human Movement Sciences at the University of Groningen. During her bachelor she developed an interest in perceptual processes and explored this interest in a research internship investigating the control of hand movements during a lateral interception task. After obtaining her Bachelor's degree in 2014, she enrolled in the research master program in Behavioural and Cognitive Neurosciences at the University of Groningen to pursue her passion for research. During her master program, she further specialized in studying perceptual processes during her research projects. For the first project, she studied how TMS stimulation of the early visual cortex influences illusory perception, while for her second project she investigated with fMRI how the lateral occipital complex of the human brain processes different spatial frequencies in images of objects. Minke obtained her Master's degree in 2016 and, shortly after, a scholarship from the research school of Behavioural and Cognitive Neurosciences for her PhD project in a collaborative project between the Visual Neuroscience Group and the department of Otorhinolaryngology of the University Medical Center Groningen. In her PhD research she used psychophysics and eye-tracking to study how visual and auditory information are integrated during emotion recognition, and how (simulated) sensory impairments affect this integration process, resulting in several journal publications and conference contributions.
Currently, Minke is involved as a postdoc in the Promise project at the Visual Neuroscience Group where she is optimizing a new method of visual field testing that uses eye-tracking to map the visual field.

List of publications

Eyes on Emotion: Dynamic Gaze Allocation During Emotion Perception From Speech-Like Stimuli.
de Boer, M.J., Başkent, D. & Cornelissen, F.W. *Multisensory Research* 1–31 (2020)

Degraded visual and auditory input individually impair audiovisual emotion recognition from speech-like stimuli, but no evidence for an exacerbated effect from combined degradation.
de Boer, M.J., Jürgens, T., Cornelissen, F.W. & Başkent, D. *Vision Research* **180**, 51–62 (2021)

Auditory and visual integration for emotion recognition and compensation for degraded signals are preserved with age.
de Boer, M.J., Jürgens, T., Başkent, D & Cornelissen, F.W. *Trends in Hearing* (in press)

Conference contributions

Audiovisual integration in emotion recognition and compensation for sensory loss are preserved with age
de Boer, M.J.*, Jürgens, T., Cornelissen, F.W., Başkent, D.
Oral presentation at the Audiological Research Cores in Europe (ARCHES) meeting, December 2019, online

Perceptual strategies are only loosely coupled to perceived emotions
de Boer, M.J.*, Başkent, D., Cornelissen, F.W.
Poster presentation at the European Conference on Visual Perception, August 2019, Leuven, Belgium

Audio-visual interactions shape perceptual strategies in emotion recognition for communication
de Boer, M.J.*, Başkent, D., Cornelissen, F.W.
Oral presentation at the Perception Day conference, December 2018, Nijmegen, the Netherlands

Audio-visual interactions in emotion perception for communication
de Boer, M.J.*, Başkent, D., Cornelissen, F.W.
Poster presentation at the European Conference on Visual Perception, August 2018, Trieste, Italy

Audio-visual interactions in emotion perception for communication
de Boer, M.J.*, Başkent, D., Cornelissen, F.W.
Poster presentation at the Doctoral Symposium of the Eye-Tracking Research and Application conference, June 2018, Warsaw, Poland

Object selective areas in the lateral occipital complex preferentially process high spatial frequencies
Halbertsma, H.N.*, **de Boer, M.J.**, Cornelissen, F.W., Nordhjem, B.
Poster presentation at the European Conference on Visual Perception, August 2016, Barcelona, Spain

* Presenting author