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, Minke J; Jürgens, Tim; Cornelissen, Frans W; Başkent, Deniz

Published in:
Vision Research

DOI:
10.1016/j.visres.2020.12.002

Publication date:
2021

Vision Research 180 (2021) 51–62

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

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

Minke J. de Boer [a,b,c,*], Tim Jürgens [d], Frans W. Cornelissen [a,b,1], Deniz Başkent [a,c,1]

[a] Research School of Behavioural and Cognitive Neuroscience (BCN), University of Groningen, Groningen, The Netherlands
[b] Laboratory of Experimental Ophthalmology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[c] Department of Otorhinolaryngology - Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[d] Institute of Acoustics, Technische Hochschule Lübeck, Lübeck, Germany

ARTICLE INFO

Keywords:
Emotion perception
Eye-tracking
Central scotoma
Age-related hearing loss
Audiovisual
Dynamic

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 on 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.

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 (Picou, 2016). The valence and arousal levels of sounds can affect mood, induce or reduce stress (Alvarsson et al., 2010; Husain et al., 2002), and the degree to which sounds attracts attention (Baumeister et al., 2001). 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 (Luo et al., 2018).

* Corresponding author at: Department of Otorhinolaryngology - Head and Neck Surgery, University Medical Center Groningen, P.O. Box 30.001, 9700 RB Groningen, The Netherlands. Internal postal code: BB21.
E-mail address: minke.de.boer@rug.nl (M.J. de Boer).
[1] These authors contributed equally.

https://doi.org/10.1016/j.visres.2020.12.002
Received 7 May 2020; Received in revised form 6 November 2020; Accepted 6 December 2020
Available online 24 December 2020

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 (see, e.g., Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004). 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 (Colijn et al., 2017) 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 recognition of most facial expressions requires detecting small, detailed movements (Ekman & Friesen, 1977). 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 (Barnes et al., 2011), 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 (Boucart et al., 2008). Additionally, Johnson et al. (2017) 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 (Raphael et al., 1980). To properly perceive these cues, usable hearing in the low frequency range, up to 750 Hz, is necessary (Ling, 1976). 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 (Moore, 1996). Fully in line with this, studies show that both adults and children with hearing loss perform worse in auditory emotion recognition (Most & Aviner, 2009; Rigo & Lieberman, 1989). Additionally, Most and Aviner (2009) 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 (Nagels et al., 2020). 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 (Goy et al., 2016).

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. (2005) 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 (Taylor et al., 2018). 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 (Roth et al., 2011).

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 (Hayhoe & Ballard, 2005; de Boer et al., 2020; Vo et al., 2012). 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 (Hooge & Erkelens, 1998). Additionally, studies have shown that AMD patients typically develop a preferred retinal locus (PRL, Cummings et al., 1985; Schuchard, 1994), 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 (Fletcher & Schuchard, 1997; Sunness et al., 1996). 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 (Collignon et al., 2008; Jessen et al., 2012), 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 (Hooge & Erkelens, 1998), 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 (Bertera, 1988; Cornelissen et al., 2005). Furthermore, Cornelissen and colleagues (2005) 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.

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 (Bänziger, Mortillaro, & Scherer, 2012; de Boer et al., 2020). 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.

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.

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, Bach, 1996, 2007). 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.

2.3. Stimuli

The stimuli used in the experiment were taken from the Geneva Multimodal Emotion Portrayals (GEMEP) core set (for a detailed description, see: Bänziger et al., 2012), 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 (Russell, 1980). The reason for using many emotions was to avoid any ceiling effects that are often found in emotion research (e.g., Hunter et al., 2010; Kokinous et al., 2015; Moraitou et al., 2013), 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 (de Boer et al., 2020) 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.

Table 1
The selected emotion categories used in the experiment. The emotions are distributed over the quadrants of the valence-arousal scale (Russell, 1980).

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

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 Fig. 1b-c. The simulated scotoma extended roughly 17° horizontally and 11.5° visual angle vertically (731 × 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. 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 Fig. 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.

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, Williges, Oetting, Hohmann, & Jürgens, 2017) was used, which was inspired by the HI simulation of Nejime and Moore (1997). 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 (Moore, 1998). 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, 2001). 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 (Bennett & Hohmann, 2012).

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 (Brand & Hohmann, 2002) of an average HI listener was achieved. This procedure was done based on average categorical loudness data (Oetting, Hohmann, Appell, Kollmeier, & Eqert, 2016). 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, Vlaming, and Dahlquist (2010). 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

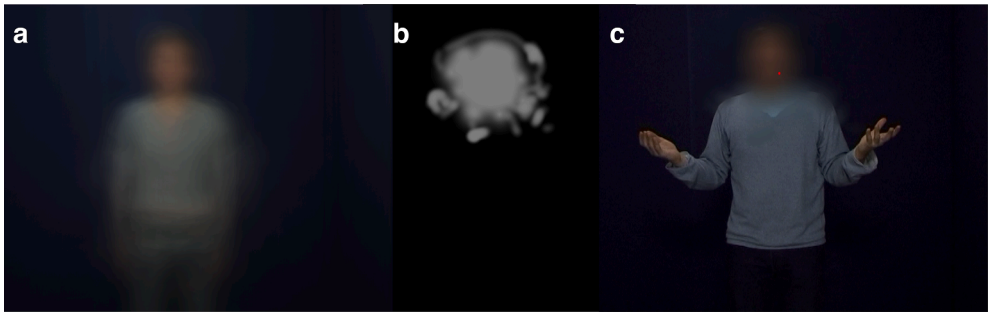

Fig. 1. a) Still image created by averaging together all frames of all videos. This image preceded stimulus presentation in all conditions, except in the A and dA conditions. b) Shape of the scotoma mask, drawn approximately to scale. The scotoma was gaze-contingent and the center of the scotoma was 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 × 1080 pixels (43 × 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, Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) and Eyelink Toolbox (Cornelissen et al., 2002) 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 greater than 1° of drift in more than one consecutive drift check.

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 Fig. 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.

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 × 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 h to complete. The experiment was separated over two test sessions performed on separate days to avoid fatigue.

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	

2.8. Analyses of behavioral data

Accuracy scores for each condition and emotion were first converted to unbiased hit-rates (Wagner, 1993) 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.

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 lead 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 point 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 Euclidian distance between the fixation and each AOI was calculated. To test whether the Euclidian 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). Euclidian distances were averaged per trial. In the model, the averaged Euclidian 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. Results

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 Fig. 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 is 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 (de Boer et al., 2020). 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 =

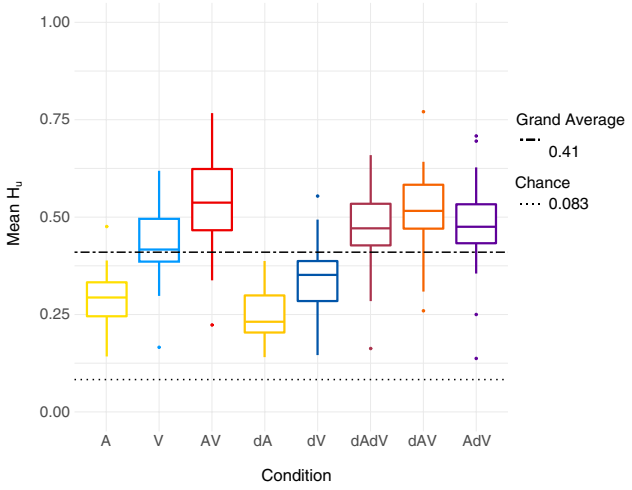

Fig. 2. Task performance for each condition, shown as unbiased hit-rates. Averaged across emotions and blocks. Each box shows the data between the first and third quartiles. The horizontal solid line in each box denotes the median. The whiskers extend to the lowest/highest value still within 1.5 * inter-quartile 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.

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).

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.

3.2. Saccadic amplitude differences

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

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 lead to somewhat smaller saccadic amplitudes.

3.3. Fixation duration differences

Fig. 4 shows fixation duration, averaged over all stimuli and participants, for each condition and the two test sessions. As in Fig. 3, Fig. 4 only shows fixation durations for fixations made during the first 1000 ms 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.

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

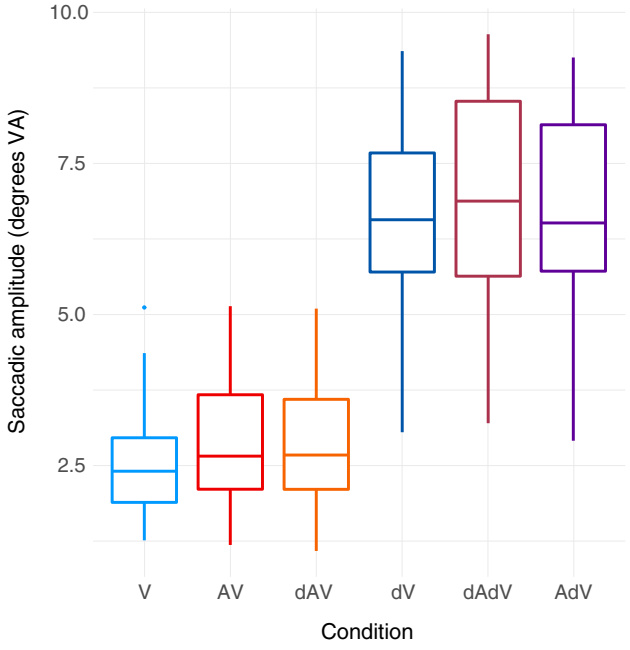

Fig. 3. Saccadic amplitude in degrees of visual angle for correct responses in each condition, averaged over stimuli and participants. The horizontal solid line in each box denotes the median. Colors for each condition correspond to the same colors in Fig. 2.

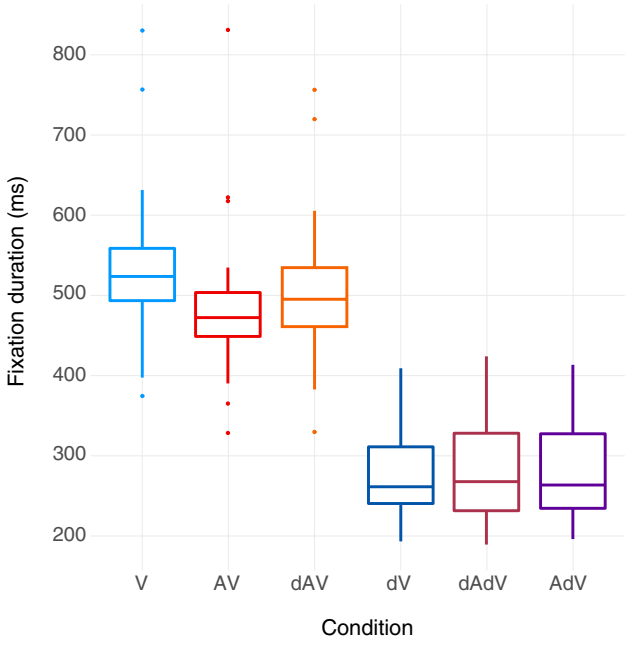

Fig. 4. Fixation duration in ms for correct responses in each condition, averaged over stimuli and participants. The horizontal solid line in each box denotes the median. Colors for each condition correspond to the same colors in Fig. 2.

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 lead to a decrease in fixation durations.

3.4. Fixation distance differences between conditions

Fixation heatmaps for the first 1000 ms of gaze data for audio-only conditions, conditions with intact video, and conditions with degraded video are shown in Fig. 5. The heatmaps are overlaid on a 1000 ms 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 Fig. 6. Differently colored bars indicate the different conditions, the x-axis shows the different AOIs. As before, only fixation data for the first 1000 ms 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.

Fig. 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).

To summarize, participants moved their fixations further from the actor's face and closer to the left hand when 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.

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 (Chandrasekaran et al., 2009; Peelle & Sommers, 2015). Because of this order, visual cues provide information about the onset of the acoustic signal, but also about the amplitude envelope of the speech (Chandrasekaran et al., 2009). Therefore, in speech, early visual make auditory cues more predictable, yet auditory cues cannot increase the predictability of visual cues. This natural asynchrony between visual and auditory cues could one of the reasons for the fact that intact vision can compensate for a degradation

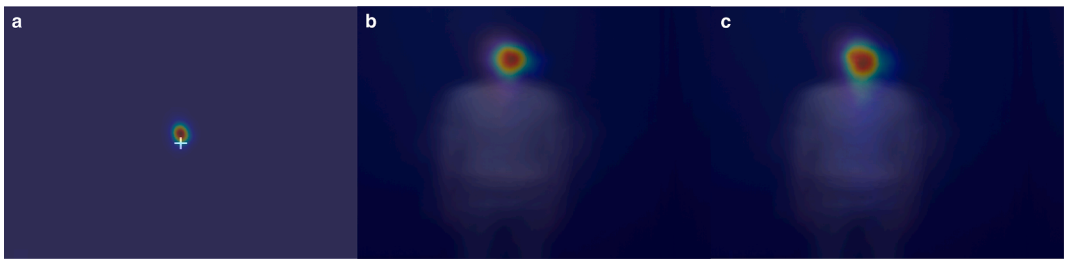

Fig. 5. Fixation heatmaps overlaid on a 1000 ms window averaged video image. a) Fixation heatmap for the audio-only conditions (A, dA). b) Fixation heatmap for conditions with intact video (V, AV, dAV). c) Fixation heatmap for conditions with degraded video (dV, dAdV, AdV). Heatmaps for individual conditions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

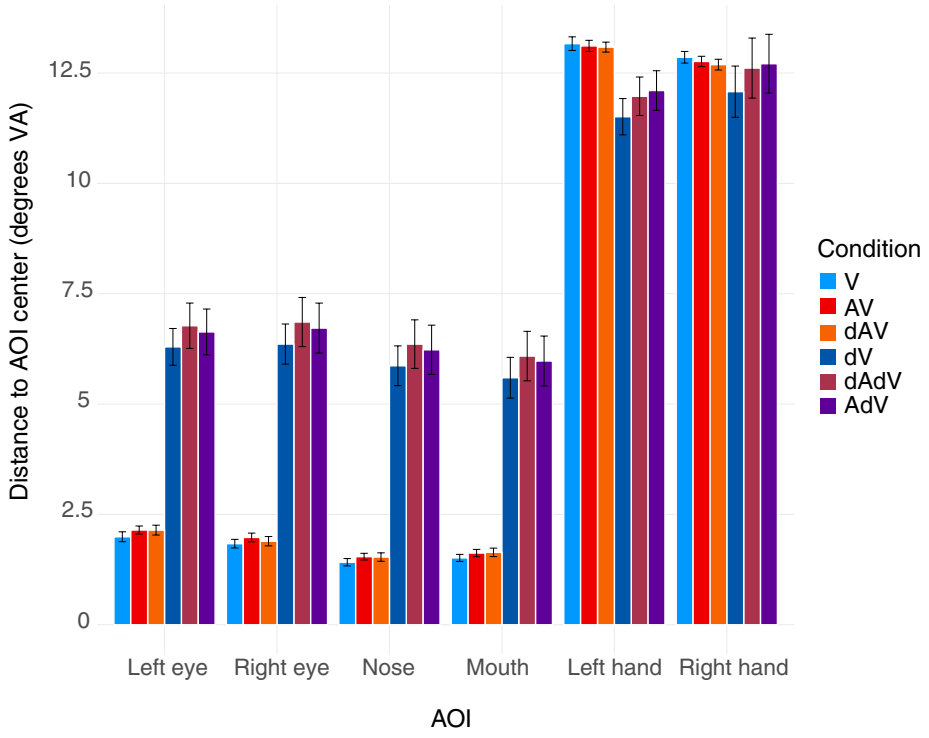

Fig. 6. Euclidian fixation distance to AOI center in degrees of visual angle for each condition, averaged over stimuli and participants. Error bars denote the SEM. Colors for each condition correspond to the same colors in Fig. 2.

in auditory information, while auditory information cannot fully do so for a degradation in visual information.

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 (Collignon et al., 2008; Jessen et al., 2012), 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.

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 × 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 (see Cheung & Legge, 2005 for a review) and in control observers with simulated scotoma's (Varsori et al., 2004; Walsh & Liu, 2014), 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 (de Boer et al., 2020), 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.

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 (Hooge & Erkelens, 1998) and with findings of longer fixation durations with simulated scotoma's for visual search tasks (Bertera, 1988; Cornelissen et al., 2005). 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 (2012) 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. (1997) 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 (2005) 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, Deubel, Schneider, and Findlay (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 (Bertera, 1988; Cornelissen et al., 2002), 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.

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.

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 (see, e.g., Collignon et al., 2008) 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.

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.

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.

CRediT authorship contribution statement

M.J. de Boer: Conceptualization, Methodology, Software, Investigation, Formal analysis, Project administration, Visualization, Resources, Writing - original draft, Data curation. **T. Jürgens:** Software, Resources, Writing - original draft. **F.W. Cornelissen:** Conceptualization, Methodology, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition. **D. Başkent:** Conceptualization, Methodology, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.visres.2020.12.002.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029.

Alvarsson, J. J., Wiens, S., & Nilsson, M. E. (2010). Stress recovery during exposure to nature sound and environmental noise. *International Journal of Environmental Research and Public Health, 7*(3), 1036–1046. https://doi.org/10.3390/ijerph7031036.

Bach, Michael (1996). The Freiburg Visual Acuity Test???Automatic measurement of visual acuity. *Optometry and Vision Science, 73*(1), 49–53. https://doi.org/10.1097/00006324-199601000-00008.

Bach, M. (2007). The Freiburg Visual Acuity Test-Variability unchanged by post-hoc re-analysis. *Graefes Archive for Clinical and Experimental Ophthalmology, 245*(7), 965–971. https://doi.org/10.1007/s00417-006-0474-4.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion, 12*(5), 1161–1179. https://doi.org/10.1037/a0025827.

Barnes, C. S., De l'Aune, W., & Schuchard, R. A. (2011). A test of face discrimination ability in aging and vision loss. *Optometry and Vision Science, 88*(2), 188–199. https://doi.org/10.1097/OPX.0b013e318205a17c.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323.

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

Bertera, J. H. (1988). The effect of simulated scotomas on visual search in normal subjects. *Investigative Ophthalmology & Visual Science, 29*(3), 470–475. https://iovs.arvojournals.org/article.aspx?articleid=2160114.

Bisgaard, N., Vlaming, M. S., & Dahlquist, M. (2010). Standard Audiograms for the IEC 60118-15 Measurement Procedure. *Trends in Amplification, 14*(2), 113–120. https://doi.org/10.1177/1084713810379609.

Boucart, Muriel, Dinon, Jean.-François., Despretz, Pascal, Desmettre, Thomas, Hladiuk, Katrine, & Oliva, Aude (2008). Recognition of facial emotion in low vision: A flexible usage of facial features. *Visual Neuroscience, 25*(4), 603–609. https://doi.org/10.1017/S0952523808080656.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357.

Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America, 112*(4), 1597–1604. https://doi.org/10.1121/1.1502902.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. PLoS Computational Biology, 5(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436.

Cheung, Sing.-Hang., & Legge, Gordon. E. (2005). Functional and cortical adaptations to central vision loss. *Visual Neuroscience, 22*(2), 187–201. https://doi.org/10.1017/S0952523805222071.

Colijn, J. M., Buitendijk, G. H. S., Prokofyeva, E., Alves, D., Cachulo, M. L., Khawaja, A. P., Cougnard-Gregoire, A., Merle, B. M. J., Korb, C., Erke, M. G., Bron, A., Anastasopoulos, E., Meester-Smoor, M. A., Segato, T., Piermarocchi, S., de Jong, P. T. V. M., Vingerling, J. R., Topouzis, F., Creuzot-Garcher, C., … Zwiener, I. (2017). Prevalence of Age-Related Macular Degeneration in Europe. Ophthalmology, 124 (12), 1753–1763. https://doi.org/10.1016/j.ophtha.2017.05.035.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research, 1242*, 126–135. https://doi.org/10.1016/j.brainres.2008.04.023.

Cornelissen, F. W., Bruin, K. J., & Kooijman, A. C. (2005). The Influence of Artificial Scotomas on Eye Movements during Visual Search. *Optometry and Vision Science, 82* (1), 27–35. https://doi.org/10.1097/01.OPX.0000150250.14720.C5.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, 34*(4), 613–617.

Cummings, Roger. W., Whittaker, Stephen. G., Watson, Gale. R., & Budd, James. M. (1985). Scanning Characters and Reading with a Central Scotoma. *Optometry and Vision Science, 62*(12), 833–843. https://doi.org/10.1097/00006324-198512000-00004.

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

Ekman, P., & Friesen, W. V. (1977). Facial action coding system.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433. https://doi.org/10.1038/415429a.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002.
Fletcher, D. C., & Schuchard, R. A. (1997). Preferred retinal loci relationship to macular scotomas in a low-vision population. *Ophthalmology, 104*(4), 632–638. https://doi.org/10.1016/S0161-6420(97)30260-7.
Ghitza, O. (2001). 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*(3), 1628–1640. https://doi.org/10.1121/1.1396325.
Goy, H., Pichora-Fuller, M. K., Singh, G., & Russo, F. A. (2016). Perception of emotional speech by listeners with hearing aids. Canadian Acoustics, 44(3). //jcaa.caa-aca.ca/index.php/jcaa/article/view/2962.
Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009.
Henderson, J. M., Mcclure, K. K., Pierce, S., & Schrock, G. (1997). Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Perception & Psychophysics, 59*(3), 323–346. https://doi.org/10.3758/BF03211901.
Hooge, I. T. C., & Erkelens, C. J. (1998). Adjustment of fixation duration in visual search. *Vision Research, 38*(9), 1295–IN4. https://doi.org/10.1016/S0042-6989(97)00287-3.
Hunter, E. M., Phillips, L. H., & MacPherson, S. E. (2010). Effects of age on cross-modal emotion perception. *Psychology and Aging, 25*(4), 779–787. https://doi.org/10.1037/a0020528.
Husain, G., Thompson, W. F., & Schellenberg, E. G. (2002). Effects of Musical Tempo and Mode on Arousal, Mood, and Spatial Abilities. Music Perception, 20(2), 151–171. https://doi.org/10.1525/mp.2002.20.2.151.
Jessen, S., Obleser, J., & Kotz, S. A. (2012). How bodies and voices interact in early emotion perception. PLoS One, 7(4), e36070.
Johnson, A. P., Woods-Fry, H., & Wittich, W. (2017). Effects of magnification on emotion perception in patients with age-related macular degeneration. *Investigative Ophthalmology & Visual Science, 58*(5), 2520–2526. https://doi.org/10.1167/iovs.16-21349.
Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.
Kokinous, J., Kotz, S. A., Tavano, A., & Schroger, E. (2015). The role of emotion in dynamic audiovisual integration of faces and voices. Social Cognitive and Affective Neuroscience, 10(5), 713–720. https://doi.org/10.1093/scan/nsu105.
Ling, D. (1976). *Speech and the Hearing-Impaired Child: Theory and Practice*. Alexander Graham Bell Association for the Deaf.
Luo, X., Kern, A., & Pulling, K. R. (2018). Vocal emotion recognition performance predicts the quality of life in adult cochlear implant users. *The Journal of the Acoustical Society of America, 144*(5), EL429–EL435. https://doi.org/10.1121/1.5079575.
McIlreavy, L., Fiser, J., & Bex, P. J. (2012). Impact of simulated central scotomas on visual search in natural scenes. *Optometry and Vision Science, 89*(9), 1385–1394. https://doi.org/10.1097/OPX.0b013e318267a914.
Moore, B. C. J. (1996). Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear and Hearing, 17*(2), 133–161. https://doi.org/10.1097/00003446-199604000-00007.
Moraitou, D., Papantoniou, G., Gkinopoulos, T., & Nigritinou, M. (2013). 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*(3), 139–147. https://doi.org/10.1111/psyg.12016.
Most, T., & Aviner, C. (2009). Auditory, visual, and auditory-visual perception of emotions by individuals with cochlear implants, hearing aids, and normal hearing. *Journal of Deaf Studies and Deaf Education, 14*(4), 449–464. https://doi.org/10.1093/deafed/enp007.
Moore, B. C. J. (1998). *Cochlear Hearing Loss*. Wiley.
Nagels, L., Gaudrain, E., Vickers, D., Matos Lopes, M., Hendriks, P., & Başkent, D. (2020). Development of vocal emotion recognition in school-age children: The EmoHI test for hearing-impaired populations. PeerJ, 8, e8773. https://doi.org/10.7717/peerj.8773.
Nejime, Y., & Moore, B. C. J. (1997). 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*(1), 603–615. https://doi.org/10.1121/1.419733.
Oetting, D., Hohmann, V., Appell, J. E., Kollmeier, B., & Eqert, S. D. (2016). Spectral and binaural loudness summation for hearing-impaired listeners. *Hearing Research, 335*, 179–192. https://doi.org/10.1016/j.heares.2016.03.010.
Orbelo, D. M., Grim, M. A., Talbott, R. E., & Ross, E. D. (2005). 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*(1), 25–32. https://doi.org/10.1177/0891988704272214.
Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. https://doi.org/10.1163/156856897X00366.
Picou, E. M. (2016). How hearing loss and age affect emotional responses to nonspeech sounds. *Journal of Speech, Language, and Hearing Research, 59*(5), 1233–1246. https://doi.org/10.1044/2016_JSLHR-H-15-0231.
Raphael, L. J., Borden, G. J., & Harris, K. S. (1980). *Speech science primer: Physiology, acoustics, and perception of speech*. Lippincott Williams & Wilkins.
Rigo, T. G., & Lieberman, D. A. (1989). Nonverbal sensitivity of normal-hearing and hearing-impaired older adults. *Ear and Hearing, 10*(3), 184–189. https://doi.org/10.1097/00003446-198906000-00008.
Roth, T. N., Hanebuth, D., & Probst, R. (2011). Prevalence of age-related hearing loss in Europe: A review. *European Archives of Oto-Rhino-Laryngology, 268*(8), 1101–1107. https://doi.org/10.1007/s00405-011-1597-8.
Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714.
Schuchard, R. (1994). Preferred retinal locus: A review with application in low vision rehabilitation. *Low Vision and Vision Rehabil, 7*, 243–256.
Siebe, T., Williges, B., Oetting, D., Hohmann, V., & Jürgens, T. (2017). Evaluation einer modularen Auralisation von sensorineuraler Schwerhörigkeit. *Annual meeting of the German Society for Audiology, Aalen*.
Sunness, J. S., Applegate, C. A., Haselwood, D., & Rubin, G. S. (1996). fixation patterns and reading rates in eyes with central scotomas from advanced atrophic age-related macular degeneration and stargardt disease. *Ophthalmology, 103*(9), 1458–1466. https://doi.org/10.1016/S0161-6420(96)30483-1.
Taylor, D. J., Edwards, L. A., Binns, A. M., & Crabb, D. P. (2018). Seeing it differently: Self-reported description of vision loss in dry age-related macular degeneration. *Ophthalmic and Physiological Optics, 38*(1), 98–105. https://doi.org/10.1111/opo.12419.
Varsori, M., Perez-Fornos, A., Safran, A. B., & Whatham, A. R. (2004). Development of a viewing strategy during adaptation to an artificial central scotoma. *Vision Research, 44*(23), 2691–2705. https://doi.org/10.1016/j.visres.2004.05.027.
Võ, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. Journal of Vision, 12(13), 1–14. https://doi.org/10.1167/12.13.3.
Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3–28. https://doi.org/10.1007/BF00987006.
Walsh, D. V., & Liu, L. (2014). Adaptation to a simulated central scotoma during visual search training. *Vision Research, 96*, 75–86. https://doi.org/10.1016/j.visres.2014.01.005.
Walker, R., Deubel, H., Schneider, W. X., & Findlay, J. M. (1997). Effect of Remote Distractors on Saccade Programming: Evidence for an Extended Fixation Zone. *Journal of Neurophysiology, 78*(2), 1108–1119. https://doi.org/10.1152/jn.1997.78.2.1108.