The McGurk effect in the time of pandemic: Age-dependent adaptation to an environmental loss of visual speech cues

Kateřina Chládková^{1,2} · Václav Jonáš Podlipský³ · Natalia Nudga⁴ · Šárka Šimáčková³

Accepted: 19 November 2020 / Published online: 14 January 2021
© The Psychonomic Society, Inc. 2021

Abstract

Seeing a person's mouth move for [ga] while hearing [ba] often results in the perception of "da." Such audiovisual integration of speech cues, known as the McGurk effect, is stable within but variable across individuals. When the visual or auditory cues are degraded, due to signal distortion or the perceiver's sensory impairment, reliance on cues via the impoverished modality decreases. This study tested whether cue-reliance adjustments due to exposure to reduced cue availability are persistent and transfer to subsequent perception of speech with all cues fully available. A McGurk experiment was administered at the beginning and after a month of mandatory face-mask wearing (enforced in Czechia during the 2020 pandemic). Responses to audio-visually incongruent stimuli were analyzed from 292 persons (ages 16–55), representing a cross-sectional sample, and 41 students (ages 19–27), representing a longitudinal sample. The extent to which the participants relied exclusively on visual cues was affected by testing time in interaction with age. After a month of reduced access to lipreading, reliance on visual cues (present at test) somewhat lowered for younger and increased for older persons. This implies that adults adapt their speech perception faculty to an altered environmental availability of multimodal cues, and that younger adults do so more efficiently. This finding demonstrates that besides sensory impairment or signal noise, which reduce cue availability and thus affect audio-visual cue reliance, having experienced a change in environmental conditions can modulate the perceiver's (otherwise relatively stable) general bias towards different modalities during speech communication.

Keywords Audiovisual integration · Speech perception · Cue reweighting · Cognitive adaptation

Without necessarily realizing it, most adults lipread when communicating with others using speech. This is because spoken language interaction often involves seeing your interlocutor's face as well as hearing their voice, and from early infancy humans are hardwired to look at people's faces, and the massive input they receive does not go wasted (Burnham & Dodd, 2004; ter Schure, Junge, & Boersma, 2016). Utilizing

characteristics of speech units perceivable via different modalities makes speech perception more robust and resistant to various types of noise and cue degradation (Sumbly & Pollack, 1954). In this study, we tested whether an environmental reduction in the availability of visual speech cues results in perceptual adaptation affecting even speech communication events that do provide full access to visual cues.

The McGurk effect is a well-known speech-perception illusion that exemplifies the role of both auditory and visual cues during speech comprehension. McGurk and McDonald (1976) demonstrated that upon viewing clips of a speaking woman whose articulators moved for a [ga] syllable, but with the soundtrack containing a time-aligned [ba] syllable, participants reported hearing a "da" syllable. The robust visual cue of no lip closure during the syllable-initial consonant overrode the auditory cues indicating a labial constriction, resulting in a percept of a consonantal category least incompatible with the mismatching cues from the two modalities (i.e., one having an intermediate place of articulation). The McGurk effect has since been replicated in a number of studies, using different

✉ Kateřina Chládková
katerina.chladkova@ff.cuni.cz

¹ Institute of Psychology, Czech Academy of Sciences, Prague, Czech Republic

² Institute of Czech Language and Theory of Communication, Faculty of Arts, Charles University, Nám. Jana Palacha 2, 116 38 Praha, Czech Republic

³ Department of English and American Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic

⁴ Institute of Phonetics, Faculty of Arts, Charles University, Praha, Czech Republic

speech sounds, and across labs and language communities (see Alsius, Paré, & Munhall, 2018). The effect is surprisingly persistent, even for informed subjects, and rather stable within individuals over time (Basu Mallick, Magnotti, & Beauchamp, 2015), though quite variable across individuals (Magnotti et al., 2015). What determines the inter-individual variability remains, for the most part, unknown (Brown et al., 2018).

Although cognitive traits such as attentional control, working memory, or processing speed do not seem to predict the extent to which an individual will exhibit the McGurk effect (Brown et al., 2018), age or even cultural background might. Visual reliance has been shown to increase with age. Hirst, Stacey, Cragg, Stacey, and Allen (2018) found that 3-to-9-year-old children are less likely to display the McGurk effect than are older children and adults, attributing such age dependence to a general developmental shift from auditory towards visual sensory dominance in speech perception (for a similar developmental change, see Maidment, Kang, Stewart, & Amitay, 2015). Sekiyama, Soshi, and Sakamoto (2014) reported that the age-related change in the magnitude of the audio-visual integration extends across the life span, with older adults (60+ years old) relying on visual speech cues more than younger adults (19–21 years old); however, Tye-Murray, Sommers, and Spehar (2007b) found the reverse. Being raised in a particular culture and with a particular native language may also affect the extent to which listeners use visual speech cues. For instance, Japanese perceivers seem to rely on visual cues less than American English perceivers, perhaps because of the cultural avoidance of looking directly at the interlocutor's face (Sekiyama & Tohkura, 1993), although not all studies report such between-language differences (Magnotti et al., 2015).

Besides the somewhat inconclusive findings about the effects of age and cultural or language background, there is ample evidence that the weighting of visual versus auditory cues is robustly affected by another factor—namely, their sensory availability. If one type of speech cues is degraded or no longer accessible as a result of a person's sensory impairment, the individual can recalibrate their cross-modal cue reliance. A person who loses vision may be able to compensate for the loss by relying more on the auditory information (Erber, 1979; Wan, Wood, Reutens, & Wilson, 2010), and someone who suffers a hearing impairment may improve their speech perception by lip-reading more (Tye-Murray, Sommers, & Spehar, 2007a). The sensory availability of individual cues can also depend on the characteristics of a particular communication event, for instance a talker's face may be blocked from view or the auditory signal may be masked by noise. Humans use perceptual cues to speech sounds in effective ways

and can promptly compensate for the immediate loss of cues from one modality by relying more on information received via the other modality. As is predictable, increasing auditory noise promotes visual cue reliance (Hirst et al., 2018; Sekiyama et al., 2014), whereas added visual noise (such as blurring the talker's face) leads to a decrease in visual cue reliance (Bejjanki, Clayards, Knill, & Aslin, 2011; Moro & Steeves, 2018).

In sum, the degree of visual versus auditory cue reliance and of audio-visual integration can change on a long-term basis as a result of an individual's sensory impairment (or perhaps even cultural constraints), as well as temporarily as a result of altered audibility or visibility of the stimulus at a given moment. A question that remains is whether—even after access to all speech cues is regained—speech perception continues to be affected by the prior experience with altered availability of cues from the individual modalities. Aftereffects in multimodal speech perception are evidenced by studies on perceptual recalibration showing that (fully accessible) visual cues enable perceivers to disambiguate a speech sound intermediate between two categories and update the auditory content of those categories, which influences subsequent (auditory) speech-sound identification (Bertelson, Vroomen, & de Gelder, 2003; Ullas, Formisano, Eisner, & Cutler, 2020). Shifts in cue weighting brought about by a person's recent experience of persistently lower availability, and hence lower informativeness, of cues via a particular modality could also lead to perceptual adaptation—a decreased reliance on cues from that modality, even if they in fact are currently available in the speech signal.

There is one study suggesting that, apart from immediate sensory uncertainty, language users' visual versus auditory reliance is likely modulated by prior experience with the environmental variability of the speech categories. Bejjanki et al. (2011) argue that individuals' speech categorization is best predicted by a model that accounts not only for the trial-by-trial sensory reliability of the cues but also for accumulated knowledge about the variability of the particular cue in the environment. As the authors note, assessing how environmental within-category variability of visual versus auditory cues contributes to speech categorization is challenging because it is virtually impossible to estimate the environmental variability to which each individual has previously been exposed.

A community-wide wearing of face masks lends itself as a rare opportunity to assess whether an environmental change in visual versus auditory cue distributions affects speech categorization. Regularly interacting with each other wearing face masks effectively removes the visual cues from most communication events, which means that the visual correlates of speech categories become less relevant compared with the auditory information. The probability distribution of visual

cues thus gets more dampened (i.e., less informative) than that of the auditory cues. If language users take into account the (change in) encountered environmental variability, their reliance on visual speech cues after having been exposed to visually impoverished (face-masked) speech should become attenuated. Crucially, this environmentally conditioned, reduction of visual-cue use may represent a general setting of the speech recognition system reflected in perception even when both auditory and visual cues are subsequently available.

Millions of speakers have currently experienced exactly this environmental change: As a result of the wearing of face masks during the 2020 pandemic, compulsory in many parts of the world, the visual speech cues suddenly became much less commonly available. We tested whether this environmental change modulates the way in which language users categorize speech sounds. We predicted that the novel but persistent experience of communicating with reduced visual speech cues would result in a decreased perceptual reliance on visual cues and/or increased reliance on auditory cues, affecting speech perception even during communication without face masks. That is, after a period of visual speech cue deprivation, listeners may identify an auditory[ba] + visual[ga] stimulus less often as *ga* and/or more often as *ba*. Potentially, the experience of reduced availability of visual cues might also make perceivers less likely to integrate information from the two modalities (i.e. give *da* responses). As previous studies have indicated that age and gender may affect the degree of visual reliance in speech, we explored whether these two factors further modulate any environmentally conditioned changes in speech categorization.

Method

Participants

The participants were native speakers of Czech, ages 16–55 years, residing in the Czech Republic at the time of the experiment. We recruited participants for a between-subjects design as well as participants for a within-subjects design.

In the between-subjects design, we tested 161 participants at Session 1, and 202 different participants at Session 2. An additional 29 participants took part, but were excluded because they fell outside the target age range (defined as ± 2 standard deviations around the mean age, $n = 17$), were children ages 15 or younger ($n = 4$), or because, at Session 2, they indicated prior participation at Session 1 without being members of our longitudinal target group ($n = 8$). Figure 1 shows the participants' age distribution per session. These participants were recruited via social media, from extended networks of followers of various academic and research institutes and science-oriented groups. They can thus be thought of as representing a general, probably above-average educated, on-line-active, Czech population.

For the within-subjects design, we recruited 41 undergraduate students from the Department of English and American Studies at Palacky University Olomouc, who received course credit for participating in both sessions. They were between 19 and 27 years old.

Stimuli

The stimuli were videos of eight speakers saying /ba/, /da/, /ga/ in American English, from the study by Basu Mallick

Fig. 1 Ages of the participants in the between-subjects pool, per session. Colored bars show the per-age counts of all the cross-sectional individuals who completed the experiment at each session, and in case of Session 2, did not participate in Session 1. Black vertical lines mark the age range

included in the experiment (i.e., 16–55 years), which represents approximately two standard deviations from the mean in each session while also excluding children ages 15 or younger. The number of participants within the defined age range was 161 in Session 1 and 202 in Session 2

et al. (2015, downloaded from <http://openwetware.org/wiki/Beauchamp:Stimuli>). We used the audio-visual incongruent [ba]_{Audio}/[ga]_{Video} stimuli from three female and three male speakers, and the three congruent [ba]_{AV}, [da]_{AV}, [ga]_{AV} stimuli from a fourth female and a fourth male speaker. In our complete stimulus set, each of the six speakers' incongruent [ba]_{Audio}/[ga]_{Video} stimuli was repeated 6 times, thus yielding 36 “McGurk” trials, and each of the other two speakers' three congruent stimuli were repeated twice, thus yielding 12 “control” trials. The 48 resulting trials were fully randomized for each experimental run. The stimulus files had a duration of approximately 2.5 s, starting with a 1-s long initial silence, continuing with a 0.35-s long syllable, and finally ending in another, approx. 1.2-s portion of silence.

The American English stimuli were adopted for the sake of time. When it was essential to capture the initial perceptual stage and perform the first round of testing immediately after the face-mask mandate, creating and piloting Czech-language McGurk stimulus set would have prolonged experiment preparation. Fortunately, the difference between Czech and English stops is in the implementation of the voicing within the three place contrasts /p/–/b/, /t/–/d/, /k/–/g/, specifically longer voice onset time (VOT) in English than Czech (Skarnitzl, 2011), while any differences in the perception of place of articulation, relevant for the McGurk effect, are negligible. Based on informal piloting, it is in fact unlikely that the congruent stimuli, containing VOT values within the Czech VOT range, sounded recognizably foreign to Czech perceivers. Even if they occasionally did, nonnative stimuli do not preclude the effects of visual information on speech categorization. The McGurk effect has been repeatedly demonstrated in studies with various cross-language designs (e.g., Burnham & Dodd, 2018; Gelder, Bertelson, Vroomen, & Chen, 1995; Hardison 1999; Kuhl, Tsuzaki, Tohkura, & Meltzoff, 1994; Sekiyama & Tohkura, 1993). Crucially, identical stimuli were used at both times of measurement, allowing us to assess shifts in perception reliably.

Procedure

The Czech speakers' perception was assessed at two time points. Session 1 was administered between March 21 and 23, 2020, shortly after the government of the Czech Republic mandated compulsory mouth and nose coverage at all places except people's homes for all individuals, including children older than 2 years (effective as of March 19, 2020). Session 2 was administered one month later, between April 18 and April 30. Within the intervening period, the regulation was generally being strictly followed, people wore face masks at public places, in parks, at work; TV reporters, public figures, and politicians always appeared and made speeches in the media with face masks on. The population thus not only wore face masks themselves, but was consistently exposed to others' speaking with their mouth and nose covered.

The two testing sessions were identical. The experiment was an online-administered multiple forced-choice task (set up and run on an online platform; Goeke, Finger, Diekamp, Standvoss, & König, 2017). At the start of each trial, a still of a video appeared on the screen and was played once the participant clicked the play button, without the possibility to replay the video. The participant then indicated which syllable the talker said by selecting one of the six buttons marked “ba”, “da”, “ga”, “pa”, “ta”, “ka”. Note that we included both the voiced and the voiceless initial consonant series because the English-spoken voiced stop consonant categories /b, d, g/ (i.e. the initial consonants of the syllables comprising our stimulus material) are mostly realized as voiceless during the consonantal constriction and may thus sound as the voiceless /p, t, k/ to Czech listeners. There was a total of 48 trials, and the entire experimental session took about 10 minutes to complete.

At the beginning of each experimental session, participants received written on-screen instructions on the procedure itself and on their task being to indicate what a speaker in the video said, without revealing the manipulated nature of the stimuli or the purpose of the experiment. Before the experiment proper, they also received two practice trials with audio-visually congruent stimuli. At the end of the session, participants indicated whether they had been wearing headphones during the task. Additionally, at the end of session 2, participants indicated whether they had already taken part in session 1. The software automatically collected information about the type of device that the experiment was run on.

Data coding and statistical model

Responses to the 12 congruent, control stimuli were used as a criterion for participant inclusion. Due to the lack of experimenter control on a participant's progress through the experiment, and particularly on their attention and engagement throughout the online-administered task, we analyzed data only from those participants who correctly identified all 12 congruent syllables (i.e., responded *ba* or *pa* to all [ba]_{AV} stimuli, *da* or *ta* to all [da]_{AV} stimuli, and *ga* or *ka* to all [ga]_{AV} stimuli; this excluded 33 participants from session 1 and 38 from session 2 in the cross-sectional design, and seven participants' session 1 and nine participants' session 2 data in the longitudinal design).

The remaining number of participants was 128 in session 1 (91 women, 37 men) and 164 in session 2 (127 women, 37 men) for the cross-sectional analysis, and 41 in the longitudinal analysis (30 women, 11 men), out of which 25 provided includable data in both sessions. The participants' responses to the 36 incongruent McGurk stimuli were used to assess the size of the McGurk effect (i.e., extent to which participants integrated information across the two modalities) as well as the participants' relative reliance on exclusively the auditory versus exclusively the visual modality. Only responses that

were given after the syllable was played were analyzed (i.e., excluding trials at which reaction time was smaller than 200 ms relative to syllable offset; this excluded fewer than 1% of trials). We analyzed the following binomial (0, 1) measures: integrated, auditory-only, and visual-only perception. Integrated perception received a score of 1 whenever a McGurk [ba]_{Audio}/[ga]_{Video} stimulus elicited a *da* or *ta* response, and 0 otherwise. Visual-only was coded 1 whenever a [ba]_{Audio}/[ga]_{Video} stimulus elicited a *ga* or *ka* response, and auditory-only was coded 1 whenever a [ba]_{Audio}/[ga]_{Video} stimulus elicited a *ba* or *pa* response.

The integrated, the visual-only, and the auditory-only scores were each analyzed with a generalized linear mixed model (package *lme4* in R; Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2019; using *glmer()* with a binomial logit link function). The cross-sectional models contained Session (coded −1 vs. +1, for sessions 1 and 2 respectively), participant Gender (coded as −1 female vs. +1 male) and participant Age (continuous, mean-centered) as fixed factors, including the main effects, as well as two-way interactions of Session and Gender, and Session and Age. Main effects were also modelled for the categorical variables Headphones (−1 without headphones, +1 with headphones) and Device (−1 computer [i.e., laptop or desktop], +1 mobile [i.e., phone or tablet]).¹ The random structure modelled intercepts per participant and per item (i.e., one of the six stimulus speakers).

The longitudinal models contained Session (−1 for session 1 vs. +1 for session 2), participant Gender (−1 female vs. +1 male), and their interaction. These models included per-participant and per-item random intercepts, as well as per-participant slopes for the (now within-subjects) factor Session. Age was not included as a fixed factor since the longitudinal group was rather homogeneous in age (range: 19–27 years), and neither were the Headphones and Device factors, as all the students did the experiment with headphones and on a computer, as instructed.

Results

Figure 2 plots individual raw data; Tables 1 and 2 show the summaries of the three cross-sectional and the three longitudinal models, respectively.

The cross-sectional analysis of the integrated percept revealed a main effect of Gender, showing that the integrated percept (i.e., the McGurk effect) occurs more often for women (mean = 60%, CI [40%, 78%]) than for men (mean = 34%, CI

[17%, 58%]). There was a main effect of Device, meaning that the McGurk effect was much more likely in participants who completed the task using a computer (mean = 68%, 95% CI [49%, 83%]) than in those using a mobile device (mean = 27%, CI [12%, 50%]). The analysis of auditory-only responses yielded a main effect of Gender, indicating that men gave more auditory-only responses than women did (men = 53%, CI [29%, 76%]; women = 24%, [12%, 43%]). A main effect of Headphones showed that auditory-only responses were more frequent with headphones than without them (with headphones = 49%, CI [26%, 72%], without = 27%, CI [14%, 47%]). The analysis of visual-only responses yielded a significant effect of Headphones, with visual-only responses being slightly more frequent without headphones than with them (without headphones = 3%, CI [2%, 5%]; with = 1%, CI [0.5%, 2%]). Finally, and perhaps most importantly for the present research question, the model for visual-only responses detected a significant interaction of Age and Session, suggesting that the visual reliance changed between sessions and did so depending on participant age.

Table 3 and Fig. 3a elucidate the two-way interaction of Age and Session by comparing the estimated means and confidence intervals (*ggeffects*; Lüdtke, 2018). The younger participants in session 1 were likely to give visual-only responses more often than were the younger participants in session 2, and for the older participants it was the other way round. Note that although significant, this effect is rather small (i.e., a difference of about 3% across the two sessions). For comparison with the other modality, we inspected the same hypothetical interaction for the auditory-only, as well as for the integrated responses. These are plotted in Fig. 3b and c, respectively. Speculatively, the significant but rather small visual-only decrease in the younger group seems to be compensated by a relatively large average—but highly variable and thus nonsignificant—auditory decrease. Also, the integrated percept seems to attenuate numerically between sessions 1 and 2 somewhat, but this attenuation is not meaningful because of the great between-participants variance.

As for the longitudinal models, the integrated percept was modulated by Gender, showing that women have a larger proportion of McGurk responses than men do (women mean = 79%, CI [49%, 94%]; men mean = 28%, CI [5%, 75%]). For auditory-only perception, there was a main effect of Gender whereby women gave fewer auditory-only responses than men (women = 5%, CI [1%, 26%]; men = 65%, CI [11%, 97%]), and an interaction of Gender and Session. The estimated means for this two-way interaction are plotted in Fig. 4b. It is seen that women have generally lower auditory-only reliance than men and this gender difference seems a bit more pronounced in session 2 than in session 1. Note that the number of men ($n = 11$) in the longitudinal pool was rather small, and thus this interaction might as well be coincidental. For visual-only perception, there was a main effect of Session,

¹ A generalized mixed-model analysis did not detect any significant main or interaction effects of age, gender, and session on the type of device used or on headphone use. It thus appears that the use of computers versus mobile devices, and of headphones versus device speakers, were distributed comparably across ages, gender, and sessions in the pool of participants.

Fig. 2 Raw data: proportion visual-only, integrated, and auditory-only perception per participant and per session in the cross-sectional design (top two graphs) and in the longitudinal design (bottom two graphs)

Table 1 Fixed-effects summary for the cross-sectional part of the experiment

Predictors	Integrated perception			Auditory-only perception			Visual-only perception		
	Odds ratio	95% CI	<i>p</i>	Odds ratio	95% CI	<i>p</i>	Odds ratio	CI	<i>p</i>
Intercept	0.9	[0.38, 2.09]	.801	0.6	[0.25, 1.43]	.253	0.02	[0.01, 0.03]	<.001
Session (–ses.1 +ses.2)	1.08	[0.78, 1.49]	.654	0.9	[0.62, 1.33]	.609	1.04	[0.77, 1.39]	.807
Gender (–female +male)	0.59	[0.42, 0.81]	.001	1.9	[1.28, 2.80]	.001	0.96	[0.71, 1.29]	.768
Age (centered)	0.99	[0.96, 1.02]	.467	1.01	[0.98, 1.05]	.438	0.99	[0.97, 1.02]	.645
Device (–comp +mobile)	0.41	[0.28, 0.58]	<.001	2.66	[1.74, 4.06]	<.001	1.34	[0.98, 1.82]	.063
Headphones (–no +yes)	0.73	[0.53, 1.01]	.059	1.61	[1.10, 2.36]	.015	0.61	[0.45, 0.83]	.001
Session : Gender	1.02	[0.74, 1.41]	.915	1.1	[0.75, 1.61]	.643	0.99	[0.74, 1.33]	.96
Session : Age	1	[0.97, 1.03]	.913	0.99	[0.95, 1.02]	.449	1.04	[1.01, 1.07]	.009

Note. Effects significant at $\alpha = 0.05$ are in boldface

Table 2 Fixed-effects summary for the longitudinal part of the experiment

Predictors	Integrated perception			Auditory-only perception			Visual-only perception		
	Odds ratio	CI	<i>p</i>	Odds ratio	CI	<i>p</i>	Odds ratio	CI	<i>p</i>
Intercept	1.2	[0.31, 4.75]	.791	0.32	[0.05, 1.87]	.204	0.01	[0.00, 0.04]	<.001
Session (–ses.1 +ses.2)	1	[0.71, 1.42]	.984	0.97	[0.62, 1.52]	.909	0.35	[0.16, 0.75]	.007
Gender (–female +male)	0.32	[0.11, 0.94]	.038	5.82	[1.26, 26.89]	.024	1	[0.38, 2.61]	.993
Session : Gender	0.73	[0.52, 1.04]	.082	1.79	[1.11, 2.88]	.017	0.96	[0.62, 1.48]	.838

Note. Effects significant at alpha = 0.05 are in boldface

showing that exclusively visual-based responses were more frequent in session 1 than in session 2; see Fig. 4a and Table 3 for the means and confidence intervals.

Discussion

This study tested whether the experience of reduced access to visual speech cues changes the way adults attend to the visual modality during speech perception. The fully naturalistic “training” involved a month-long mandatory wearing of face masks anywhere outside people’s homes, practiced by the entire Czech community during the 2020 pandemic. The pre-test and posttest was a forced-choice McGurk experiment administered online with a cross-sectional sample representing a general educated population (ages 16–55 years), as well as with a longitudinal sample of university students.

We measured participants’ integration of auditory and visual cues as well as the degree to which they relied on each domain separately. While no reliable effects of testing session were detected for audio-visual integration, the time of testing affected reliance on visual cues depending on age: after communicating with face-masks for about a month, younger individuals decreased their visual reliance, even though the talkers’ mouths were fully visible in the online task. In

contrast, older persons increased their visual-cue reliance. The between-session difference for the younger adults from the cross-sectional sample was corroborated by the results of the longitudinal sample. Though rather small in size (see Table 3), the effect is most probably a genuine finding for the following reasons. First, although we used stimuli from a longitudinal study that had reported little change in individuals’ audio-visual speech perception (Basu Mallick et al., 2015), we detected a difference between sessions ascribable to the environmental change. Second, in terms of direction and magnitude, the pattern of change was the same for the young adults from the cross-sectional and the longitudinal sample. We therefore conclude that after the partial loss of visual cues lasting a month, young adults indeed adapt their speech-perception faculty by lowering their reliance on vision.

This pattern of adaptation did not occur in older adults, whose visual reliance instead somewhat increased. It is possible that the difference between the younger and older participants was not in their adaptation strategy but in the rate of progress. The adaptability to a changed environment decreases with age (as evidenced, e.g., by more efficient auditory adaptation in early than late vision loss; Wan et al., 2010). The older adults, with greater entrenchment of visual cues (Sekiyama et al., 2014), could thus have been slower in adapting to the environmental change though following the

Table 3 Estimated means and 95% confidence intervals of visual-only perception (in %) for various participant ages throughout the cross-sectional sample and for the longitudinal sample

Experiment	Age (years)	Session 1		Session 2	
		Mean	CI	Mean	CI
Cross-sectional	16	3.4%	[1.4%, 7.8%]	1.2%	[0.5%, 2.7%]
	24	2.4%	[1.2%, 4.7%]	1.5%	[0.7%, 2.9%]
	32	1.7%	[0.9%, 3.2%]	1.9%	[1.0%, 3.4%]
	42	1.1%	[0.5%, 2.4%]	2.5%	[1.2%, 4.9%]
	55	0.7%	[0.2%, 2.1%]	3.6%	[1.3%, 9.3%]
Longitudinal	19–27	2.5%	[0.9%, 7.0%]	0.3%	[0%, 2.2%]

Note. Age in the cross-sectional sample was modelled as a continuous variable; the five ages listed in the table are arbitrarily selected points to illustrate the trajectory of the age effect

Fig. 3 Model-estimated (a) visual-only, (b) integrated, and (c) auditory-only response scores per age and session in the cross-sectional experiment. Middle curves are means and ribbons show 95% confidence

intervals. Note the different y-axis scales across the three graphs. Recall that a significant effect of Age $\times$ Session was detected for the visual-only responses, depicted in graph (a)

same trajectory. Alternatively, the age effect could be due to differing compensatory strategies. First, age-related hearing deterioration (Sommers et al., 2011) discourages auditory-cue reliance in older perceivers. Second, age (and linguistic maturation) affects people's attention to specific parts of the interlocutor's face: children (and nonnative speakers) focus more on the talker's mouth, while adults (and native speakers)

focus more on the eyes (Birulés, Bosch, Pons, & Lewkowicz, 2020; Morin-Lessard, Poulin-Dubois, Segalowitz, & Byers-Heinlein, 2019), which also provide some visual speech cues (Jordan & Thomas, 2011). Therefore, the younger perceivers (with better hearing) stand a greater chance of improving the extraction of speech cues from the auditory signal, while the older perceivers (eye-lookers) need to boost visual-cue

Fig. 4 Model-estimated (a) visual-only perception and (b) integrated, and (c) auditory-only perception per gender and session in the longitudinal experiment. Points are means, and whiskers show 95% confidence intervals. Recall that a significant effect of Session was detected for the visual-

only responses, depicted in graph (a), and an interaction effect of Session $\times$ Gender was detected for the auditory-only responses, depicted in graph (c)

extraction even from faces with covered mouths. The two proposed explanations are testable in future research. Lab-based studies could examine whether the naturalistic training effect that we detected is replicable in controlled settings, and eye-tracking could help determine whether any age-dependent adaptations are due to differential attention to specific parts of a talker's face.

Next, we found that Czech women and men differ in how they utilize the multimodal information in speech. Previous studies with different populations provide inconclusive findings on gender differences in visual cue reliance (Alm & Behne, 2015; Aloufy, Lapidot, & Myslobodsky, 1996; Irwin, Whaler, & Fowler, 2006; Tye-Murray et al., 2007b). With a large sample and appropriate modelling of individual variation across subjects and stimulus items, we detected a robust gender effect: Men rely exclusively on auditory information more than women do (53% vs. 24%), who in turn more often show audio-visual integration (60% vs. 34%). Therefore, the previously debated gender difference may not only lie in the ability to lipread (as suggested by previous studies), but possibly also in the degree to which the two genders preferentially attend to the auditory signal.

Additionally, our longitudinal comparisons indicated a gender-specific exposure-induced change: Women seemed to reduce, and men increase, their auditory-cue reliance, though the effect was not very convincing because of high variance, especially for men, and because it was not detected in the cross-sectional sample. We do not offer any interpretation of this particular result which can be addressed in future studies of exposure-induced cue reweighting in audio-visual perception of speech.

This study advances our understanding of cross-modal perceptual learning in speech in adults and the modulating factors. Prior work on perceptual adaptation documented visually guided recalibration of the auditory content of speech-sound categories (e.g., Bertelson et al., 2003). Another line of research (e.g., Moro & Steeves, 2018) showed that reduced access to specific cues, due to momentary noise in stimuli or long-term physiological changes on the part of the perceiver (e.g., Tye-Murray et al., 2007a), modifies the likelihood of the McGurk effect for the stimuli at hand. However, it has been unclear whether the experience of cue deprivation causes perceptual adaptation (cue reweighting) with aftereffects on the perception of speech with regained access to all cues. Our study addressed this question and showed that altered environmental experience can bring about persistent perceptual adjustment affecting a person's degree of reliance on a specific modality. Effects of the environmental reliability of cues across different modalities are attested for scenarios where people learn novel multimodal categories (Bankieris, Bejjanki, & Aslin, 2017; Jacobs, 2002). The present study demonstrates that an environmentally altered cue reliability modulates cross-modal perceptual processing, even for representations as well established as linguistic categories.

Bejjanki et al. (2011) proposed that the encountered environmental variability of multimodal speech cues can affect speech categorization. The authors noted though that testing this proposal is an intriguing task since individuals' prior experience is difficult to control. Here, we managed to overcome that limitation by exploiting a community-wide change in spoken language interaction and demonstrated that environmental variability indeed affects language users' reliance on a particular kind of cue. Interestingly, this finding has

implications for the widely observed between-subjects variability in the McGurk effect, virtually unexplainable by individuals' cognitive traits (Brown et al., 2018). The present results suggest it could be each individual's environmental experience that determines the degree to which they attend to the auditory and visual modality, and/or integrate them. This possibility is definitely worth pursuing in future research.

Admittedly, whether or not our findings are generalizable to any language community worldwide is uncertain. The adaptation effect found here might be due to the unprecedented community-wide reduction of access to visual speech cues. Perhaps in societies where the wearing of face masks is common during any epidemic, people do not experience a perceptual cue reweighting every time. Or perhaps differing frequency of face-mask use is exactly the reason why some studies found differential degrees of audio-visual integration for participants from different cultures (Sekiyama & Tohkura, 1993). The effect we found here may thus not be universal in its size, direction, or nature. In any case, we conclude that in human adults, an abrupt change in real-life environment can lead to adjustments of mature cognitive capacities (speech recognition). At least in young persons these adjustments seem to compensate efficiently for the environmental changes. How cross-modal learning develops across the lifespan is worth investigating further.

Author note This work was supported by Charles University research grant Primus/17/HUM/19 and by Czech Science Foundation grant 18-01799S.

Open practices statement The data and materials for all experiments are available at <https://osf.io/ud8gf/>. The experiment was not preregistered.

Data availability The data sets generated and analyzed during the current study are available in the OSF repository, at <https://osf.io/ud8gf/>.

References

- Alm, M., & Behne, D. (2015). Do gender differences in audio-visual benefit and visual influence in audio-visual speech perception emerge with age? *Frontiers in Psychology*, 16(1014). <https://doi.org/10.3389/fpsyg.2015.01014>
- Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research*, 31, 111–144. <https://doi.org/10.1163/22134808-00002565>
- Aloufy, S., Lapidot, M., & Myslobodsky, M. (1996). Differences in susceptibility to the “blending illusion” among native Hebrew and English speakers. *Brain and Language*, 53, 51–57. <https://doi.org/10.1006/brln.1996.0036>
- Bankieris, K. R., Bejjanki, V. R., & Aslin, R. N. (2017). Sensory cue-combination in the context of newly learned categories. *Scientific Reports* 7, 10890. <https://doi.org/10.1038/s41598-017-11341-7>
- Basu Mallick, D., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, 22, 1299–1307. <https://doi.org/10.3758/s13423-015-0817-4>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLOS ONE*, 6(5), e19812. <https://doi.org/10.1371/journal.pone.0019812>
- Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, 14, 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x
- Birulés, J., Bosch, S., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language, Cognition and Neuroscience*. <https://doi.org/10.1080/23273798.2020.1762905>
- Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. F. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PLOS ONE*, 13(11), e0207160. <https://doi.org/10.1371/journal.pone.0207160>
- Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, 45, 204–220. <https://doi.org/10.1002/dev.20032>
- Burnham, D., & Dodd, B. (2018). Language-general auditory-visual speech perception: Thai-English and Japanese-English McGurk effects. *Multisensory Research*, 31(1/2), 79–110.
- Erber, N. P. (1979). Auditory-visual perception of speech with reduced optical clarity. *Journal of Speech, Language, and Hearing Research*, 22(2), 212–223.
- Gelder, B. D., Bertelson, P., Vroomen, J., & Chen, H. C. (1995, September 18–21). *Inter-language differences in the McGurk effect for Dutch and Cantonese listeners*. Paper presented at the Fourth European Conference on Speech Communication and Technology, Madrid, Spain.
- Goeke, C., Finger, H., Diekamp, D., Standvoss, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. *International Conference on Computational Social Science IC²S²*. Retrieved from www.labvanced.com
- Hardison, D. M. (1999). Bimodal speech perception by native and non-native speakers of English: Factors influencing the McGurk effect. *Language Learning*, 49, 213–283. <https://doi.org/10.1111/0023-8333.49.s1.7>
- Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., & Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports*, 8, 1–12. <https://doi.org/10.1038/s41598-018-30798-8>
- Irwin, J. R., Whaler, D. H., & Fowler, C. A. (2006). A sex difference in visual influence on heard speech. *Perception & Psychophysics*, 68, 582–592. <https://doi.org/10.3758/BF03208760>
- Jordan, T. R., & Thomas, S. M. (2011). When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audio-visual speech perception. *Attention, Perception, & Psychophysics* 73, 2270. <https://doi.org/10.3758/s13414-011-0152-4>
- Kuhl, P. K., Tsuzaki, M., Tohkura, Y. I., & Meltzoff, A. N. (1994). *Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces*. Paper presented at the Third International Conference on Spoken Language Processing, Yokohama, Japan.
- Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3, 772. <https://doi.org/10.21105/joss.00772>
- Magnotti, J. F., Basu Mallick, D., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect

- in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, 233, 2581–2586. <https://doi.org/10.1007/s00221-015-4324-7>
- Maidment, D. W., Kang, H. J., Stewart, H. J., & Amitay, S. (2015). Audiovisual integration in children listening to spectrally degraded speech. *Journal of Speech, Language, and Hearing Research*, 58(1), 61–68. https://doi.org/10.1044/2014_JSLHR-S-14-0044
- McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748. <https://doi.org/10.1038/264746a0>
- Morin-Lessard, E., Poulin-Dubois, D., Segalowitz, N., & Byers-Heinlein, K. (2019). Selective attention to the mouth of talking faces in monolinguals and bilinguals aged 5 months to 5 years. *Developmental Psychology*, 55(8), 1640–1655. <https://doi.org/10.1037/dev0000750>
- Moro, S. S., & Steeves, J. K. (2018). Audiovisual plasticity following early abnormal visual experience: Reduced McGurk effect in people with one eye. *Neuroscience Letters*, 672, 103–107. <https://doi.org/10.1016/j.neulet.2018.02.031>
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology*, 5, 323. <https://doi.org/10.3389/fpsyg.2014.00323>
- Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 427–444. [https://doi.org/10.1016/S0095-4470\(19\)30229-3](https://doi.org/10.1016/S0095-4470(19)30229-3)
- Skarnitzl, R. (2011). *Znělostní kontrast nejen v češtině* [The voicing contrast in Czech and other languages]. Prague, Czech Republic: Epoque.
- Sommers, M. S., Hale, S., Myerson, J., Rose, N., Tye-Murray, N., & Spehar, B. (2011). Listening comprehension across the adult lifespan. *Ear and Hearing*, 32(6), 775–781. <https://doi.org/10.1097/AUD.0b013e3182234cf6>
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212–215. <https://doi.org/10.1121/1.1907309>
- ter Schure, S., Junge, C., & Boersma, P. (2016). Discriminating non-native vowels on the basis of multimodal, auditory or visual information: Effects on infants' looking patterns and discrimination. *Frontiers in Psychology*, 7, 525. <https://doi.org/10.3389/fpsyg.2016.00525>
- Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007a). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, 28, 656–668. <https://doi.org/10.1097/AUD.0b013e31812f7185>
- Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007b). The effects of age and gender on lipreading abilities. *Journal of the American Academy of Audiology*, 18, 883–892. <https://doi.org/10.3766/jaaa.18.10.7>
- Ullas, S., Formisano, E., Eisner, F., & Cutler, A. (2020). Audiovisual and lexical cues do not additively enhance perceptual adaptation. *Psychonomic Bulletin & Review*, 27, 707–715. <https://doi.org/10.3758/s13423-020-01728-5>
- Wan, C. Y., Wood, A. G., Reutens, D. C., & Wilson, S. J. (2010). Early but not late-blindness leads to enhanced auditory perception. *Neuropsychologia*, 48, 344–348. <https://doi.org/10.1016/j.neuropsychologia.2009.08.016>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.