Journal of Autism and Developmental Disorders (2022) 52:28–37
https://doi.org/10.1007/s10803-021-04916-x

Audiovisual Speech Perception in Children with Autism Spectrum Disorders: Evidence from Visual Phonemic Restoration

2044

Julia Irwin[1,3] · Trey Avery[3,4,5] · Daniel Kleinman[3] · Nicole Landi[1,2,3]

Accepted: 8 February 2021 / Published online: 25 February 2021

Abstract

Children with autism spectrum disorders have been reported to be less influenced by a speaker's face during speech perception than those with typically development. To more closely examine these reported differences, a novel visual phonemic restoration paradigm was used to assess neural signatures (event-related potentials [ERPs]) of audiovisual processing in typically developing children and in children with autism spectrum disorder. Video of a speaker saying the syllable /ba/ was paired with (1) a synthesized /ba/ or (2) a synthesized syllable derived from /ba/ in which auditory cues for the consonant were substantially weakened, thereby sounding more like /a/. The auditory stimuli are easily discriminable; however, in the context of a visual /ba/, the auditory /a/ is typically perceived as /ba/, producing a visual phonemic restoration. Only children with ASD showed a large /ba/-/a/ discrimination response in the presence of a speaker producing /ba/, suggesting reduced influence of visual speech.

Keywords Audiovisual · Speech · Autism · Phonemic restoration

Introduction

In face-to-face interactions, listeners can both hear and see what a speaker is saying. Visual information about speech has been shown to influence what listeners hear, increasing identification of the speech signal both in the context of background noise (Sumby and Pollack 1954) and in clear listening conditions, where mismatched auditory and visual speech results in a new percept (known as the McGurk effect; McGurk and MacDonald 1976), suggesting that face-to-face communication involves integration across visual and heard speech. This influence of visible speech on what is heard has been demonstrated in infancy (e.g., Rosenblum et al. 1997; Yeung and Werker 2013), and typical speech and language development is thought to take place in this audiovisual (or AV) context, fostering native language acquisition (e.g., Lewkowicz and Hansen-Tift 2012; Meltzoff and Kuhl 1984). Moreover, the presence of visual articulatory information facilitates perception of heard speech, speeding up cortical processing of the auditory signal (van Wassenhove et al. 2005). Critically, the speaking face does not simply provide redundant information to the speech sound, but includes functional cues that supplement the auditory information. Examples of this include that the visible speech signal has been shown to change perception even when the auditory portion of the signal can be clearly heard (McGurk and MacDonald 1976) and that blind listeners show a difference in vowel space as compared to sighted listeners, suggesting that access to the visual signal influences the development of speech perception and organization (Ménard et al. 2013).

✉ Julia Irwin
julia.irwin@yale.edu

Trey Avery
trey.avery@philips.com

Daniel Kleinman
daniel.kleinman@yale.edu

Nicole Landi
nicole.landi@yale.edu

[1] Southern Connecticut State University, New Haven, CT, USA

[2] University of Connecticut, Storrs, CT, USA

[3] Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA

[4] Electrical Geodesics Inc, 500 E 4th Ave #200, Eugene, OR 97401, USA

[5] Philips Healthcare, Amsterdam, Netherlands

Autism Spectrum Disorders (ASD) are characterized by difficulties in social communication as well as restricted and repetitive behaviors (American Psychiatric Association 2013). Children with ASD have been reported to have atypical processing of audiovisual (AV) speech in comparison to their typically developing (TD) peers. Specifically, children with ASD are less influenced (relative to TD children) by the speaker's face in (a) audiovisual speech-in-noise tasks, where the listener must integrate the face and voice to identify what was said; and (b) mismatched audiovisual speech tasks, where the face and voice are different signals, that, if integrated, will lead to a visually influenced percept (Feldman et al. 2018; Foxe et al. 2015; Iarocci et al. 2010; Irwin et al. 2011; Mongillo et al. 2008; Smith and Bennetto, 2007; Smith et al. 2017; Woynaroski et al. 2013). These findings are typically interpreted as evidence of weaker integration of the face and voice in ASD, and sometimes as evidence of broader integration deficits in processing in children with ASD. Given that weakened integration could significantly impair a listener's ability to recover or disambiguate a speaker's message in noisy listening environments, such differences in processing could account for some of the observed language comprehension difficulties observed in children with ASD. Notably, Shic et al. (2014) reported that 6-month-old infants who go on to receive a diagnosis of autism spectrum disorder show a pronounced reduction in gaze to the eyes, nose and mouth of a face when speech is being produced (e.g., in comparison to a smiling face), suggesting that the root of difficulties in integration of auditory and visible speech associated with ASD may begin early in development.

In addition to these behavioral and eye tracking findings, a few studies have utilized event-related potentials (ERPs), which provide an objective neural index of perception, to explore AV speech processing in individuals with ASD (Russo et al. 2010). ERP designs, like eye-tracking studies, can be particularly useful for studying speech perception in children or lower-functioning individuals with ASD because they do not require an overt behavioral response. For example, Magnée et al. (2008) found that adults with ASD do not show typical congruency-associated N2 effects (e.g., to a face and voice with matching vs. mismatching emotional valence), suggesting that adults with ASD are less sensitive to mismatching AV speech stimuli. Megnin et al. (2012) reported reduced P2 amplitude for words in the context of a speaking face in adolescents with ASD. Taken together, these studies indicate that individuals with ASD have difficulty using visual speech information during perception of a speaking face. This does not appear to be due entirely to less looking to the face of a speaker: Irwin et al. (2011) controlled for this by examining AV speech perception only when children were fixated on the face of the speaker, and still reported significant differences between children with ASD and TD controls in visual influence on heard speech.

While existing studies are suggestive of atypical response to AV speech in children with ASD, common paradigms used to study AV speech may not be ideal for use in this population because they make substantial processing demands beyond perception of audio and visual speech. Specifically, studies that use speech-in-noise and/or mismatched (or McGurk-type) AV tasks may require additional attentional or executive resources, which are noted areas of weakness for many children with ASD (Irwin and DiBlasi 2017; Stevenson et al. 2017). The McGurk Effect creates a percept that differs from either the visual or auditory signal alone because of conflict between the two modalities; however, these percepts are independently identified as poorer exemplars of a category than matched A + V speech (Brancazio 2004). Because labeling of ambiguous stimuli places demands on executive functions, this task may be particularly challenging for those with ASD (Eigsti and Shapiro 2003). Additionally, studying AV speech perception using paradigms that utilize auditory noise is problematic because noise can be particularly aversive for individuals with ASD in the perception of speech (Alcántara et al. 2004). Both speech-in-noise and mismatched AV stimuli could confer additional variability for children with ASD, making interpretation of group differences more difficult.

In order to examine visual influence on heard speech in children with autism spectrum disorder, we have developed a measure that can serve as an alternative to assessing audiovisual speech processing without involving either noise or auditory and visual category conflict (also see Jerger et al. 2014 for a related approach). This paradigm, which we describe in detail in Irwin et al. (2018), uses restoration of weakened auditory tokens with visual stimuli. Two types of stimuli are presented to the listener: clear exemplars of an auditory consonant–vowel syllable (in this case, /ba/), and syllables in which the auditory cues for the consonant are substantially weakened, creating a stimulus which is more /a/-like, from this point on referred to as */a/*. The auditory stimuli are created by synthesizing speech based on a natural production of the syllable and systematically flattening the formant transitions to create the */a/*. Video of the speaker's face does not change (always producing /ba/), but the auditory stimuli (/ba/ or */a/*) vary. Thus, when the */a/* auditory stimulus is dubbed over the visual /ba/, a visual influence will effectively "restore" the weakened auditory cues so that the stimulus is perceived as a /ba/, akin to a visual phonemic restoration effect (Kashino 2006; Samuel 1981; Warren 1970).

To provide a sensitive measure of AV speech processing in typically developing children and those with ASD, we recorded electroencephalography (EEG) during this visual phonemic restoration paradigm. Specifically, we utilize a

Table 1 Sociodemographics, nonverbal IQ and social functioning scores by group

	ASD	TD	Difference between groups
N	18	34	
Sex	13 M (72%), 5 F (28%)	16 M (47%), 18 F (53%)	$\chi^2(1)=3.08$, $p=.082$
Age (years)[a]	7.6–14.9, $M=10.7$, $SD=2.4$	5.3–13.2, $M=9.7$, $SD=2.1$	$t(48)=-1.41$, $p=.164$
WASI PRI Composite[a]	58–132, $M=94.33$, $SD=19.8$	85–148, $M=105.16$, $SD=13.6$	$t(48)=2.28$, $p=.027$*
ADOS-2 Comparison Score[b]	6–14, $M=9$, $SD=2.1$	N/A	
ADI-R[c] A	9–27, $M=20.5$, $SD=4.86$	N/A	
ADI-R[c] B Verbal	7–24, $M=16.93$, $SD=5.80$	N/A	
ADI-R[c] C	2–11, $M=6.00$, $SD=2.64$	N/A	
ADI-R[c] D	1–5, $M=3.45$, $SD=1.63$	N/A	

[a] Test statistics have df = 48 because two TD participants did not have recorded ages or WASI scores

[b] ADOS-2 $n=16$

[c] ADI-R $n=14$

modified passive listening paradigm to elicit auditory ERP responses to /a/ and /ba/ paired with a face producing /ba/. In this paradigm, if the visual /ba/ causes the auditory /a/ to be perceived as /ba/ (phonemic restoration), then ERP responses to both tokens should be similar. However, if audio and visual speech are not integrated (no phonemic restoration) we would instead expect a differential, mismatch-like response (MMR),[1] resulting from differential perception of the two tokens. Here we hypothesize that children with ASD, who are suspected to exhibit deficits in AV integration, will be less likely to use visual speech to effectively restore the /a/ sound to a /ba/ percept. That is, we predicted that children with ASD would exhibit a larger difference between the waveforms to the two tokens (MMR) when we combined the auditory /a/ with the visual /ba/, relative to typically developing (TD) controls.

Method

Participants

Seventy-one monolingual American English-speaking children were recruited from the greater *New Haven* community through flyers, local walks and events and listservs of organizations that serve individuals with ASD. Participants were excluded from participating if they had uncorrected hearing or vision problems. Participants were excluded from analysis if they could not be classified either as children with ASD or as TD children ($n=10$), technical issues led to an incomplete data file (ASD: $n=3$; TD: $n=3$), or there were not enough usable trials in each condition to permit analysis (ASD: $n=1$; TD: $n=2$; criteria described below). The final data set included 52 participants: 18 children with ASD and 34 TD children. Sociodemographic information for each sample is reported in Table 1. All participants were monolingual American English-speaking children, were right-handed, and had normal or corrected-to-normal vision and hearing by parent report. Children in the TD group had no history of developmental delays per parent report. For characterization purposes, all participants in the ASD group had a clinical diagnosis of ASD. In addition, 16 participants completed the Autism Diagnostic Observation Schedule—second edition [ADOS-2; Lord et al. 2012] and met criteria for an autism spectrum disorder, and parents of 14 children completed the Autism Diagnostic Interview-Revised [ADI-R; Rutter et al. 2003]. Two participants' records for the ADOS-2 and four for the ADI-R are unavailable. Performance IQ subtests from the Wechsler Abbreviated Scale of Intelligence (Wechsler 2011) were used to account for individual differences in some analyses reported below. All data were collected according to the ethical guidelines laid out by the *Yale University* Institutional Review Board. Written consent was obtained from participants' caregivers and written assent was obtained from the child participants.

Audiovisual Stimuli and Experimental Paradigm

Participants completed three short audiovisual (AV) experiments designed to examine the neural basis of AV speech integration in our participants. The primary experiment of interest (the AV speech experiment) utilized the novel phonemic restoration procedure introduced above and will be reported here.

To produce the stimuli for the AV speech experiment, the /ba/ and /a/ synthesized auditory stimuli were dubbed

[1] We call this a mismatch-like or mismatch response (MMR) because it is elicited here in response to two different tokens in an equiprobable rather than an oddball presentation, and thus should not be confused with a mismatch negativity (MMN) response.

onto video of the speaker producing /ba/, with the acoustic onsets synchronized with the visible articulation time locked to a single video frame. The stimuli for the AV speech were created by videotaping and recording an adult male speaker of English producing the syllable /ba/. Using Praat, we extracted acoustic parameters for the token, including formant trajectories, amplitude contour, voicing and pitch contour (Boersma 2001). Critically, the token had rising formant transitions for F1, F2, and to a lesser extent F3, characteristic of /ba/. To create our /ba/ stimulus, we synthesized a new token of /ba/ based on these values. To create our /a/ stimulus we then modified the synthesis parameters by changing the onset values for F1 and F2 to reduce the extent of the transitions and lengthened the transition durations for F1, F2 and F3, and then synthesized a new stimulus. Video duration was 1567 ms, with speech onset at 829 ms and offset at 1225 ms for both syllables. Stimuli are publicly available at https://doi.org/10.17605/OSF.IO/EHVG8.

Instructions and a practice trial were presented prior to the start of the EEG session. The AV experiment contained 200 fully randomized presentations of equiprobable /ba/ and /a/ tokens (100 trials each). On each trial, the video was presented and followed by a 600 ms ITI. Total EEG session time (including two other experiments not reported here) was approximately 45 min, depending on the amount of EEG net rehydration between experiments.

EEG Data Collection

EEG data was collected with a Philips Neuro (formerly Electrical Geodesics Inc) EEG System using 128 Ag/AgCl electrodes embedded in soft sponges woven into a geodesic array. The EEG sensor nets were soaked for up to ten minutes prior to use in a warm potassium chloride solution. Impedance for all electrodes was kept below 40 kΩ throughout the experimental run (impedances were re-checked between experiments— approximately every 15 min). Online recordings at each electrode used the vertex electrode as the reference and were later referenced to the average reference. EEG was continuously recorded using Net Station 4.5 on a Mac Pro. Stimuli were presented using E-Prime version 2.0.8.90 (Psychology Software Tools, Inc., Sharpsburg, PA, USA) on a Dell computer running Windows XP. Audio stimuli were presented from an audio speaker centered 85 cm above the participant at 65 decibels. Visual stimuli were presented on a Dell 17-inch flat panel monitors 60 cm from the participant.

ERP Data Preprocessing

EEG preprocessing was performed using EEGLAB v2019.1 (Delorme and Makeig 2004) and ERPLAB v7.0 (Lopez-Calderon and Luck 2014). To maximize the amount of data in analyses, two preprocessing stages (notch filtering and ICA-based removal of ocular movements) were included for individual participants only when necessary and when they increased the number of usable trials; all other stages were applied to data from all participants. (Potential differences between groups in data preprocessing are considered below and accounted for via statistical analyses.)

For participants with excessive line noise, notch filters were applied at every multiple of 60 Hz up to the Nyquist frequency of the data (60, 120, 180, 240 Hz; order = 180). For all participants, the PREP pipeline (Bigdely-Shamlo et al. 2015) was used to remove line noise and to identify and replace bad EEG channels via spherical spline interpolation. Independent Components Analysis (ICA; Makeig et al. 1996) and the ICLabel algorithm (Pion-Tonachini et al. 2019) were used to identify ocular movements in the EEG data (blinks and lateral eye movements). Data were subsequently band-pass filtered from 0.3 to 30 Hz (Butterworth filter, 12 dB/oct roll-off) and re-referenced to the average of all channels. The EEG data were segmented into epochs time-locked to the onset of speech, including 100 ms before and 800 ms after speech onset, and baseline-corrected to the pre-stimulus window. Horizontal eye movements (HEOG) were measured as the difference between channels 125 and 128, which are located to the right and left of the eyes; and vertical eye movements (VEOG) were measured as four differences between pairs of channels above and below the eyes (8 and 126; 14 and 126; 21 and 127; 25 and 127). An epoch was rejected for containing a lateral eye movement if HEOG exceeded $|\mu V| \geq 55$, and rejected for containing a blink if VEOG exceeded $|\mu V| \geq 150$, within sliding 80-ms windows. For participants with reasonable ICA decompositions for whom IC removal improved data retention, these artifact rejection criteria were applied to the window [− 100 100] ms, then ocular movements were removed, and then the criteria were applied to the entire window. (As the visual stimulus during speech onset affected the perceived sound, this ensured that trials were always rejected if participants made eye movements during the critical window.) For participants for whom IC component removal did not improve data retention, these artifact rejection criteria were applied once to the entire window (and components were not removed). Finally, for all participants, each EEG channel was marked as bad—and interpolated from nearby channels—in every epoch in which it varied by more than 200 µV (max–min). Channels that were marked as bad in at least 33% of epochs were marked bad (and interpolated) in all epochs. Epochs that contained at least 20 bad channels were discarded from analysis.

Data from each participant were preprocessed in four ways: both with and without notch filtering, and both with and without removal of ocular movements (these options were crossed). Optional preprocessing stages were applied only when they improved the data quality and the number of

Table 2 Summary statistics for EEG preprocessing and data retention by group

	ASD	TD	Difference between groups
Notch filter applied	$n=11$ (61%)	$n=26$ (76%)	$\chi^2(1)=1.35$, $p=.245$
Ocular movements removed	$n=16$ (89%)	$n=29$ (85%)	$\chi^2(1)=0.13$, $p=.718$
Channels interpolated	$M=6.3$, $SD=3.2$	$M=8.9$, $SD=5.4$	$t(50)=1.85$, $p=.071$
Usable /ba/ trials	$M=79$, $SD=15$	$M=82$, $SD=19$	$t(50)=0.43$, $p=.670$
Usable /a/ trials	$M=79$, $SD=15$	$M=82$, $SD=18$	$t(50)=0.45$, $p=.653$

usable trials. After selecting the best preprocessing pipeline for each participant, participants with fewer than 20 usable trials in any condition, and participants with more than 20% of channels interpolated, were excluded from analysis (ASD: $n=1$; TD: $n=2$). Information about how often preprocessing stages were applied, how many channels were interpolated, and how many usable trials remained in each condition are shown in Table 2 separately for each group. A 2×2 repeated measures ANOVA with group and speech stimulus as within-subject factors revealed that the number of usable trials (the dependent variable) did not differ between groups or speech stimuli (with no interaction), all $F<0.20$, all $p>0.65$.

ERP Data Analysis

Data were analyzed using the LIMO EEG toolbox (v2.0; Pernet et al. 2011), which implements a two-step method of cluster-based permutation testing (Maris and Oostenveld 2007) that permits multiple comparisons across channels and latencies while keeping the family wise error rate in check. In the first step, single-trial data from each participant was used to calculate the effect of speech stimulus (/ba/ vs. /a/) at all EEG channels and all latencies between 200 and 375 ms, a time window identical to one that we used in a previous AV ERP speech study to measure audiovisual mismatch effects (Irwin et al. 2018). Second, the resulting beta coefficients for the effect of speech stimulus, and a between-subjects variable indicating group status (ASD vs. TD), were submitted to cluster-based permutation testing. This test uses a bootstrap procedure, randomly sampling participants with replacement 1000 times, to determine the null distribution of the spatiotemporal clustering statistics (the strength of an effect across adjacent channels and/or samples) for the interaction between group (for randomly determined groups) and speech stimulus (using participants' actual data). This null distribution is then used to determine the statistical significance of the observed clustering statistics. A significant result would indicate that the relationship between group and speech stimulus is significant, but the resulting cluster(s) would be descriptive in nature, as the test does not indicate that the relationship is significant at specific time points and/or channels (Maris and Oostenveld 2007; Sassenhagen and Draschkow 2019).

Results

Main Analysis

The cluster-based permutation test revealed a significant interaction between group status (ASD vs. TD) and the effect of speech stimulus (/ba/ vs. /a/), $p=.043$. Within the latency range analyzed (200–375 ms), this relationship was strongest from 242 to 375 ms over a cluster of left temporal channels, where /a/ minus /ba/ was more positive for participants with ASD. This channel cluster is shown in Fig. 1, where the 14 channels that contributed to the cluster for at least 25 ms within the time window are highlighted. For each group and speech stimulus, scalp maps are shown in Fig. 2 and mean waveforms for the cluster are shown in Fig. 3 (plotted using the *ggplot2* software package; Wickham 2016).

To determine whether (as predicted) this interaction was driven by participants with ASD showing larger effects of the speech stimulus than TD participants, single-trial means (average µV between 242 and 375 ms) for each cluster (as shown in Fig. 1) were output for analysis in R (v. 3.6.1; R Core Team 2019) using linear mixed-effects modeling (lme4 v. 1.1–21; Bates et al. 2015b; see also Baayen et al. 2008), and denominator degrees of freedom were estimated via the Satterthwaite approximation (lmerTest v. 3.1–0; Kuznetsova et al. 2017). The model contained fixed effects of group (TD = − 0.5, ASD = + 0.5) and speech stimulus (/ba/ = − 0.5, /a/ = + 0.5), as well as their interaction. The random effects structure was initially maximal (a random intercept for participants, a random slope for speech stimulus, and random correlations). As the random slope for speech stimulus accounted for no additional variance beyond the random intercept, it was removed to facilitate convergence (Bates et al. 2015a). A priori contrasts (emmeans v. 1.4; Lenth 2016) were conducted to examine the effect of speech stimulus separately for each group.

Given that the time window and channel cluster were selected for single-trial analysis based on the results of

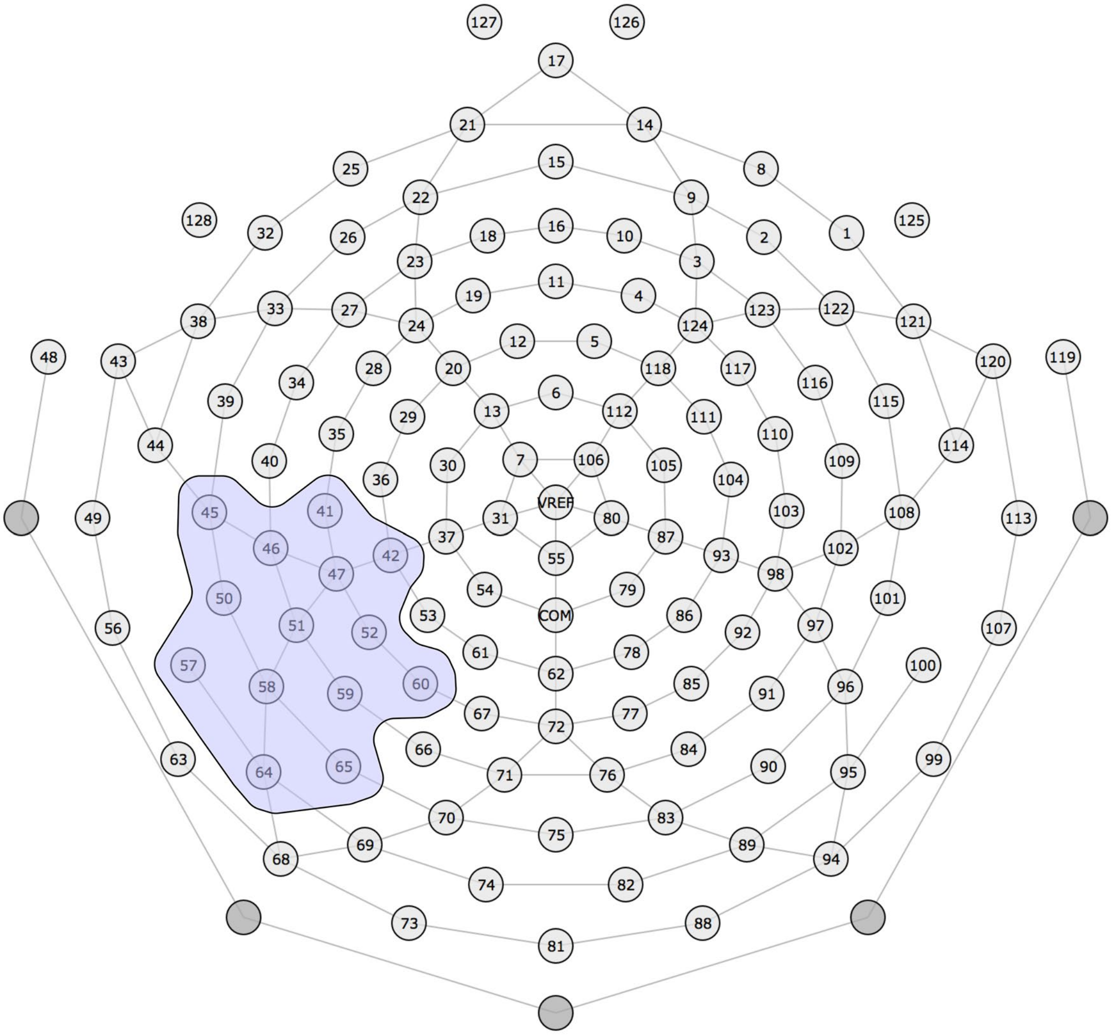

Fig. 1 The electrode montage. Electrodes that contributed to a cluster (as determined by the cluster-based permutation test) for at least 25 ms between 200 ms and 375 ms post-stimulus are highlighted in blue

the cluster test, drawing new statistical inferences from the interaction between group and speech stimulus in the mixed-effects model would be a form of ‘double-dipping’. Accordingly, although we note that that interaction was significant, $B = 1.03$, $SE = 0.37$, $t(8347) = 2.77$, $p = .006$, this indicates only that the results of the cluster test can be fairly represented by the time window and channel cluster selected for analysis. More importantly, contrasts revealed that the effect of speech stimulus was significant for participants with ASD, $B = 0.75$, $SE = 0.30$, $t(8346) = 2.47$, $p = .013$, but did not reach significance for TD participants, $B = -0.28$, $SE = 0.22$, $t(8348) = -1.30$, $p = .194$. In other words, the interaction observed in the cluster-based permutation test arose because participants with ASD showed statistically different ERPs to /ba/ vs. /a/ stimuli, but TD participants did not.

Analysis Controlling for Individual Differences

Due to differences between samples in demographic characteristics, we ran a second analysis to control for additional participant-level variables. Specifically, we repeated the cluster-based permutation test described above with five additional between-subjects variables representing sex, age, composite WASI performance IQ score, whether or not the participant’s data was notch-filtered, and whether or not the participant’s data had ocular components removed. All five variables were entered into the regression model at the same time as the variable of interest (group). (Data from two TD participants were discarded for this analysis because their ages and WASI scores were not recorded.) After controlling for these other variables, the interaction between group and speech stimulus was marginally significant, $p = .064$.

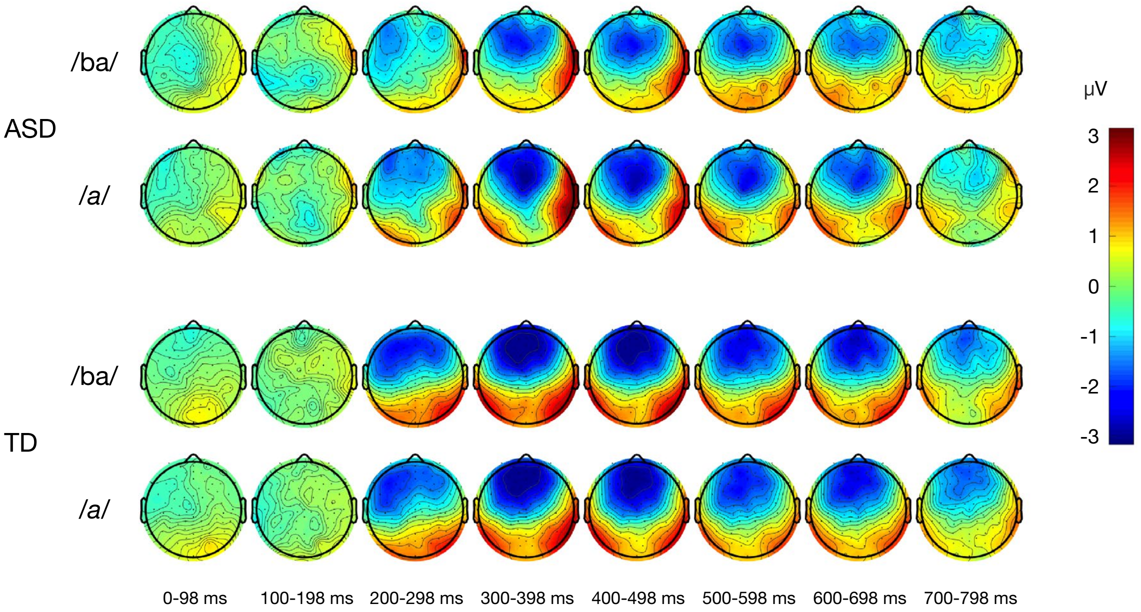

Fig. 2 Scalp maps depicting mean μV within every 100 ms window for each group (ASD: top panel; TD: bottom panel) and stimulus type (/ba/: top rows:/*a*/: bottom rows). The interaction between group and stimulus type revealed by the cluster-based permutation test is driven by the appearance, in participants with ASD but not in TD participants, of a greater left temporal positivity to /*a*/ than /ba/ in the scalp maps covering the time window from 200 ms to 400 ms

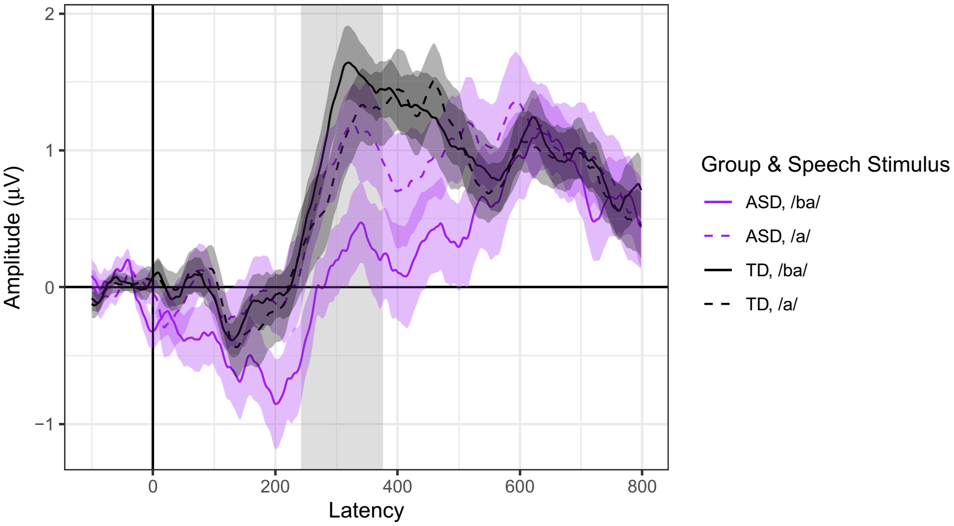

Fig. 3 Grand mean ERP waveforms for the cluster identified by the cluster-based permutation test, shown separately for each group and stimulus type. The time window used in subsequent single-trial analyses (242–375 ms) is highlighted in grey. Error ribbons = 1 SEM

Analysis with a Broader Time Window

To evaluate the extent to which our results were contingent on our selection of time window (200–375 ms), we ran two additional analyses that were identical to those reported above except that they considered the entire post-stimulus window (0–800 ms). In a cluster-based permutation test, statistical significance is evaluated by comparing an observed cluster mass against a null distribution of cluster masses, and considering a longer time window means that the largest cluster observed by chance will be (on average) larger. This in turn can increase the observed cluster *p*-value, which means that a cluster that reaches the significance threshold at a shorter time window may fail to do so at a longer one. Consistent with this possibility, the interaction between group and speech stimulus was not significant either when group status was the only between-subjects variable in the analysis, $p = .196$, or when the other covariates were considered simultaneously, $p = .242$.

Analysis of Time Course Effects

Finally, we considered whether the size of the effect changed over the course of the task (e.g., due to participants' attention flagging over time) by examining interactions between trial number and the effect of interest. Of note, trial retention did not vary much between the first and second halves of the experiment, as both groups had an average of between 39 and 43 usable trials (out of 50 total trials) in every combination of speech stimulus and experiment half (trials 1–100 vs. trials 101–200). Single-trial analyses were conducted as described above, except with one additional predictor (trial number, centered) entered as a fixed effect and allowed to interact with all other fixed and random effects. The main effect of trial number and all of its interactions were not significant, all $|t| < 1.29$, all $p > .20$. Importantly, the interactions between group and speech stimulus remained significant, $t(8323) = 2.78$, $p = .005$, and the effect of speech stimulus remained significant for participants with ASD, $t(8319) = 2.46$, $p = .014$, but not for TD participants, $t(8330) = -1.35$, $p = .177$. Thus, there is no evidence that the interaction between group and the effect of speech stimulus is attributable to changes in attention over time.

Discussion

Children with autism spectrum disorder have been reported to be less influenced by visible speech even when controlling for gaze to the face of a speaker. However, the underlying mechanisms associated with differences in audiovisual speech perception are still largely unknown. We employed a novel visual phonemic restoration task in a passive ERP paradigm to assess neural signatures of audiovisual processing in typically developing children and in children with ASD. The two speech stimuli, an auditory consonant–vowel syllable /ba/ and a syllable /*a*/ in which the auditory cues for the consonant was substantially weakened, were paired with video of a speaker producing /ba/. In this paradigm, the video of the speaker producing /ba/ can visually restore perception of the consonant when paired with the auditory /*a*/, leading to a perception of /ba/. If this phonemic restoration occurs, it should lead to an attenuated ERP discrimination response (or mismatch response, MMR), providing a measure of integration of the visual and auditory speech signals. Indeed, we found that TD children had an attenuated MMR response relative to children with ASD, suggesting reduced audiovisual integration in the latter group.

It is possible that previous reports of reduced AV integration in children with ASD are a function of the characteristics of the stimuli used. The current findings extend previous work by using a paradigm with no auditory noise or mismatched AV stimuli in order to measure AV integration in children with ASD without placing additional demands on higher-level executive systems. This approach was designed to provide a more sensitive test of visual influence on heard speech than has been used previously.

Our results are consistent with a number of theoretical approaches to perceptual processing in autism spectrum disorder. Reduced global processing, sometimes described as "weak central coherence" (Booth and Happé 2018; Frith and Happé 1994), suggests that integration of details into a meaningful whole is more difficult and requires sustained effort for individuals with ASD. Similarly, Markram and Markram (2010) characterized sensory perception in ASD as an intense focus on low-level stimuli, thought to impair integration. In combination with poorly modulated gaze to others' faces, a focus on local details on the face when a speaker talks, much of which is not linguistically informative, could reduce exposure of children with ASD to critical visible articulatory information (Irwin and Brancazio 2014). Less visual influence on heard speech could also be a function of weak temporal binding (Stevenson et al. 2014) where the visual and auditory speech signals are not perceived as emanating from the same source. These accounts both may explain differences in children with ASD and suggest that visual speech may not be as informative to a child listener with ASD as to their TD peers. Developmentally, this could lead to cascading negative effects in social communication, already primary difficulties for children with ASD.

The current work must be viewed in light of several limitations. First, as we did not have an auditory-only control condition for exploring the /ba/ vs. /*a*/ contrast, we cannot establish whether our participants with ASD had atypical patterns of speech perception in the auditory domain. (Note, for example, that Chen et al. 2019, reported differences in auditory processing of speech in young children with ASD using the auditory brain stem response.) Second, given poorly modulated gaze to the faces of others in children with ASD, it is possible that a greater discrimination response may be a function of reduced gaze to the speaker's face (Irwin and Brancazio 2014), which was not measured in the current study. Further, Dunn et al. (2008) used auditory tones to elicit mismatch negativities (MMN) in children with ASD and reported a typical MMN response when children attended to the stimuli. This suggests that future work should include an active task, to explore whether attention may lead to increased integration in perception of AV speech. Third, we note that the interaction between group and condition was only significant in the analysis on our preselected epoch, indicating that the effects are sensitive to our choice of time window. As this may simply be an issue of power, we are currently exploring this effect with a larger sample of children. Finally, as we tested only one phonemic contrast here, future studies should replicate these findings with additional contrasts and with longer utterances, such as words.

Conclusion

These findings provide preliminary evidence of impaired integration of auditory with visual speech signals in ASD in a passive AV speech processing paradigm. Children with ASD—but not TD children—discriminated between /ba/ and /a/ auditory tokens in the presence of video of the speaker producing /ba/ in this passive condition, indicating reduced influence of visual speech in children with ASD. In this passive condition, children with ASD were less visually influenced than their TD peers. Our findings are consistent with previous work that has reported differences in AV integration and visual influence in individuals with ASD (e.g., Iarocci et al. 2010; Irwin et al. 2011; Smith and Bennetto 2007). Reduced visual influence on heard speech puts children with ASD at increased risk for communication and language deficits, which could lead to cascading negative effects on social communication (Stevenson et al. 2018). Given the ubiquity and utility of visual articulation in day-to-day communication and its role in language acquisition, a deeper understanding of the contribution of the factors that underlie AV speech processing differences in ASD may also be informative for the development of interventions that focus on speech.

Acknowledgments Supported by NIH grants DC013864 and DC011342. We recognize and thank Larry Brancazio for the development of the phonemic restoration paradigm.

References

Alcántara, J. I., Weisblatt, E. J., Moore, B. C., & Bolton, P. F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *Journal of Child Psychology and Psychiatry, 45*(6), 1107–1114.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 2008,* 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Available from arXiv:1506.04967v2 (stat.ME).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. https://doi.org/10.18637/jss.v067.i01.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics, 9,* 16. https://doi.org/10.3389/fninf.2015.00016.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345.

Booth, R. D., & Happé, F. G. (2018). Evidence of reduced global processing in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 48*(4), 1397–1408.

Brancazio, L. (2004). Lexical Influences in Audiovisual Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance, 30*(3), 445–463.

Chen, J., Liang, C., Wei, Z., Cui, Z., Kong, X., Dong, C. J., et al. (2019). Atypical longitudinal development of speech-evoked auditory brainstem response in preschool children with autism spectrum disorders. *Autism Research, 12*(7), 1022–1031.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134,* 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

Dunn, M. A., Gomes, H., & Gravel, J. (2008). Mismatch negativity in children with autism and typical development. *Journal of Autism and Developmental Disorders, 38*(1), 52–71.

Eigsti, I. M., & Shapiro, T. (2003). A systems neuroscience approach to autism: Biological, cognitive, and clinical perspectives. *Mental Retardation and Developmental Disabilities Research Reviews, 9*(3), 206–216.

Feldman, J. I., Dunham, K., Cassidy, M., Wallace, M. T., Liu, Y., et al. (2018). Audiovisual multisensory integration in individuals with autism spectrum disorder: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews, 95,* 220–234.

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H. P., Russo, N. N., Blanco, D., et al. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cerebral Cortex, 25,* 298–312.

Frith, U., & Happé, F. (1994). Autism: Beyond "theory of mind." *Cognition, 50*(1–3), 115–132.

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., & Chua, R. (2010). Visual influences on speech perception in children with autism. *Autism, 14*(4), 305–320.

Irwin, J., Avery, T., Brancazio, L., Turcios, J., Ryherd, K., & Landi, N. (2018). Electrophysiological indices of audiovisual speech perception: Beyond the McGurk effect and speech in noise. *Multisensory Research*, *31*(1–2), 39–56. https://doi.org/10.1163/22134808-00002580.

Irwin, J. R., & Brancazio, L. (2014). Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. *Frontiers, Language Sciences*. https://doi.org/10.3389/fpsyg.2014.00397.

Irwin, J., & DiBlasi, L. (2017). Audiovisual speech perception: A new approach and implications for clinical populations. *Language and Linguistics Compass*. https://doi.org/10.1111/lnc3.12237.

Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Development, 82*(5), 1397–1403.

Jerger, S., Damian, M. F., Tye-Murray, N., & Abdi, H. (2014). Children use visual speech to compensate for non-intact auditory speech. *Journal of Experimental Child Psychology, 126,* 295–312.

Kashino, M. (2006). Phonemic restoration: The brain creates missing speech sounds. *Acoustic Science Technology, 27*(6), 318–321.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82,* 1–26. https://doi.org/10.18637/jss.v082.i13.

Lenth, R. V. (2016). Least-squares means: The (R) Package (lsmeans). *Journal of Statistical Software, 69*(1), 1–33. https://doi.org/10.18637/jss.v069.i01.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America, 109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8,* 213. https://doi.org/10.3389/fnhum.2014.00213.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule* (2nd ed.). Torrance, CA: Western Psychological Services.

Magnée, M. J., De Gelder, B., Van Engeland, H., & Kemner, C. (2008). Audiovisual speech integration in pervasive developmental disorder: Evidence from event-related potentials. *Journal of Child Psychology Psychiatry, 49*(9), 995–1000.

Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems, 8,* 145–151.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024.

Markram, K., & Markram, H. (2010). The intense world theory—A unifying theory of the neurobiology of autism. *Frontiers in Human Neuroscience, 4,* 224.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 246,* 746–748.

Megnin, O., Flitton, A., Jones, C. R. G., de Haan, M., Baldeweg, T., & Charman, T. (2012). Audiovisual speech integration in autism spectrum disorders: ERP evidence for atypicalities in lexical-semantic processing. *Autism Research, 5*(1), 39–48. https://doi.org/10.1002/aur.231.

Meltzoff, A. N., & Kuhl, P. K. (1984). Infants' representation of events: Studies in imitation, cross-modal perception, and categorization. *Infant Behavior and Development, 7,* 243. https://doi.org/10.1016/S0163-6383(84)80305-7.

Ménard, L., Toupin, C., Baum, S. R., Drouin, S., Aubin, J., & Tiede, M. (2013). Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America, 134*(4), 2975–2987. https://doi.org/10.1121/1.4818740.

Mongillo, E. A., Irwin, J. R., Whalen, D. H., Klaiman, C., Carter, A. S., & Schultz, R. T. (2008). Audiovisual processing in children with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders, 38,* 1349–1358.

Pernet, C. R., Chauveau, N., Gaspar, C., & Rousselet, G. A. (2011). LIMO EEG: a toolbox for hierarchical LInear MOdeling of ElectroEncephaloGraphic data. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2011/831409.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). The ICLabel dataset of electroencephalographic (EEG) independent component (IC) features. *Data in Brief, 25,* 104101. https://doi.org/10.1016/j.dib.2019.104101.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/. Accessed 18 February 2021.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception and Psychophysics, 59*(3), 347–357. https://doi.org/10.3758/bf03211902.

Russo, N., Foxe, J. J., Gomes, H., Brown, A., Altschuler, T., Gomes, H., et al. (2010). Multisensory Processing in Children with Autism: High-density electrical mapping of auditory-somatosensory integration. *Autism Research, 3,* 1–15.

Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism diagnostic interview-revised*. Los Angeles, CA: Western Psychological Services.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*(4), 474–494.

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology, 56*(6), e13335. https://doi.org/10.1111/psyp.13335.

Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry, 75*(3), 231–237.

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology Psychiatry, 48*(8), 813–821.

Smith, E., Zhang, S., & Bennetto, L. (2017). Temporal synchrony and audiovisual integration of speech and object stimuli in autism. *Research in Autism Spectrum Disorders, 39,* 11–19.

Stevenson, R. A., Baum, S. H., Segers, M., Ferber, S., Barense, M. D., & Wallace, M. T. (2017). Multisensory speech perception in autism spectrum disorder: From phoneme to whole-word perception. *Autism Research, 10*(7), 1280–1290.

Stevenson, R. A., Segers, M., Ncube, B. L., Black, K. R., Bebko, J. M., Ferber, S., et al. (2018). The cascading influence of multisensory processing on speech perception in autism. *Autism, 22*(5), 609–624.

Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., et al. (2014). Multisensory temporal integration in autism spectrum disorders. *Journal of Neuroscience, 34*(3), 691–697.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212–215.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*(3917), 392–395.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II)*. San Antonio, TX: NCS Pearson.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org. Accessed 18 February 2021.

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2013). Multisensory speech perception in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 43*(12), 2891–2902.

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24*(5), 603–612.

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