Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special Issue "Multisensory integration": Research Report

Weak observer–level correlation and strong stimulus-level correlation between the McGurk effect and audiovisual speech-in-noise: A causal inference explanation

John F. Magnotti, Kristen B. Dzeda, Kira Wegner-Clemens, Johannes Rennig and Michael S. Beauchamp*

Department of Neurosurgery, Baylor College of Medicine, USA

ARTICLE INFO

Article history:
Received 3 May 2020
Reviewed Jul 13 2020
Revised 5 August 2020
Accepted 5 October 2020
Published online 17 October 2020

Keywords:
Audiovisual
Multisensory integration
Causal inference
Bayesian inference
Illusions

ABSTRACT

The McGurk effect is a widely used measure of multisensory integration during speech perception. Two observations have raised questions about the validity of the effect as a tool for understanding speech perception. First, there is high variability in perception of the McGurk effect across different stimuli and observers. Second, across observers there is low correlation between McGurk susceptibility and recognition of visual speech paired with auditory speech-in-noise, another common measure of multisensory integration. Using the framework of the causal inference of multisensory speech (CIMS) model, we explored the relationship between the McGurk effect, syllable perception, and sentence perception in seven experiments with a total of 296 different participants. Perceptual reports revealed a relationship between the efficacy of different McGurk stimuli created from the same talker and perception of the auditory component of the McGurk stimuli presented in isolation, both with and without added noise. The CIMS model explained this strong stimulus-level correlation using the principles of noisy sensory encoding followed by optimal cue combination within a common representational space across speech types. Because the McGurk effect (but not speech-in-noise) requires the resolution of conflicting cues between modalities, there is an additional source of individual variability that can explain the weak observer–level correlation between McGurk and noisy speech. Power calculations show that detecting this weak correlation requires studies with many more participants than those conducted to-date. Perception of the McGurk effect and other types of speech can be explained by a common theoretical framework that includes causal inference, suggesting that the McGurk effect is a valid and useful experimental tool.

* Corresponding author.
E-mail address: michael.beauchamp@pennmedicine.upenn.edu (M.S. Beauchamp).

https://doi.org/10.1016/j.cortex.2020.10.002

1. Introduction

Viewing the talker's face influences the perception of auditory speech, as exemplified by McGurk and MacDonald's discovery that pairing incongruent auditory and visual syllables can evoke the percept of a completely different syllable (McGurk & MacDonald, 1976). In the decades since its discovery, the McGurk effect has grown into one of the most popular experimental tools for assessing multisensory integration, with thousands of citations across both behavioral and neural sciences (Beauchamp, 2018).

Recently, doubts have arisen about the utility of the McGurk effect as a tool for understanding speech perception, including the suggestion that it should be "retired" (Rosenblum, 2019). In a large study, Basu Mallick and colleagues examined perception of 12 different McGurk stimuli by 165 participants tested in the laboratory (Basu Mallick et al., 2015). There was high variability, both across stimuli (rates ranging from 17% to 58%) and across participants (rates from 0% to 100%). This high variability has been used as an argument that the McGurk effect is unreliable and hence poorly suited for experimental study (Alsius et al., 2018).

A second critique of the McGurk effect arises from comparisons with other measures of speech perception. Viewing the talker's face benefits understanding of noisy auditory speech (Peelle & Sommers, 2015; Sumby & Pollack, 1954). Van Engen and colleagues (Van Engen et al., 2017) found high variability in visual enhancement of noisy speech and in rates of the McGurk effect, but low correlation between the two measures for any of the twelve types of noisy speech examined, with a maximum of 4% variance explained. Similarly, Brown and colleagues (Brown et al., 2018) found only a weak correlation between lipreading accuracy and McGurk susceptibility (3% of variance explained; 8% variance if lipreading responses were grouped by place-of-articulation).

We reasoned that modeling the processes underlying audiovisual speech perception might shed light on these observations. Two processes thought to underlie multisensory integration are *noisy sensory encoding* and *optimal cue combination*. Noisy sensory encoding assumes that observers to do not have veridical access to the physical properties of a stimulus, but only to a perceptual representation that is corrupted by sensory noise and that can vary on repeated presentations of identical stimuli (Deneve et al., 2001). Optimal cue combination assumes that when cues from different modalities are combined, they are weighted by the reliability of each modality, a process often referred to as Bayesian integration (Alais & Burr, 2004; Aller & Noppeney, 2019; Ernst & Banks, 2002; Magnotti et al., 2013) although there are alternative algorithms such as probability summation (Arnold et al., 2019). When there is a potential for the cues to arise from separate causes (nearly always the case in natural perception), optimal cue combination requires *causal inference*: judging whether the cues in the different modalities arise from the same physical cause. Causal inference is necessary because it is only beneficial to integrate cues generated by the same source; integrating cues from different sources leads to misestimation (French & DeAngelis, 2020; Kording et al., 2007; Shams & Beierholm, 2010). Humans are frequently confronted with multiple talkers, necessitating causal inference (Ma et al., 2009; Massaro, 1998; Noppeney & Lee, 2018; Vroomen, 2010) and individual differences in causal inference judgments have been used to characterize individual- and group-level differences in audiovisual speech perception (Baum et al., 2015; Gurler et al., 2015; Magnotti & Beauchamp, 2015; Magnotti et al., 2013; Stropahl et al., 2017).

The causal inference of multisensory speech (CIMS) model incorporates these processes into a principled framework that predicts perception of arbitrary combinations of auditory and visual speech (Magnotti & Beauchamp, 2017). The CIMS model has been used to explain a number of puzzling audiovisual speech phenomena, such as the increase in the McGurk effect observed with co-articulation (Magnotti, Smith, et al., 2018); the decrease in the McGurk effect observed with slow playback rates (Magnotti, Basu Mallick, & Beauchamp, 2018); and why the McGurk effect is produced by some incongruent syllables but not others (Magnotti & Beauchamp, 2017).

If the same model can account for weak observer–level correlation and strong stimulus-level correlation between the McGurk effect and speech-in-noise, it suggests that both types of speech are processed using common perceptual mechanisms, with the implication that the McGurk effect is a useful experimental tool. On the other hand, if models derived from the McGurk effect do not apply to other types of speech, it suggests that the McGurk effect has limited utility (Alsius et al., 2018; Rosenblum, 2019).

2. Methods

2.1. *Human Subject statement*

All experiments were approved by the Committee for the Protection of Human Subjects of Baylor College of Medicine.

2.2. *Data availability statement*

All data, code and materials, including experimental stimuli, are available at https://osf.io/C9EVY/

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedures was pre-registered prior to the research being conducted. No part of the study analyses was pre-registered prior to the research being conducted.

2.3. *Overview of the CIMS model*

For a full description, see (Magnotti & Beauchamp, 2017). Briefly, for each presentation of a given audiovisual stimulus, the model assumes each modality is encoded independently. For a single trial of a stimulus with auditory component S_A and visual component S_V, the model generates two vectors: the auditory representation $X_A \sim \mathcal{N}(S_A, \Sigma_A)$ and the visual representation $X_V \sim \mathcal{N}(S_V, \Sigma_V)$, where $\mathcal{N}(\mu, \Sigma)$ is a normal distribution with mean μ and variance Σ. Across many trials, the values of X_A and X_V will center around the exemplar locations S_A and S_V with variance equal to the modality-specific encoding variances (Σ_A and Σ_V).

The auditory and visual components of speech may emanate from a single talker ($C = 1$) or two separate talkers ($C = 2$). To form the $C = 1$ representation, the model assumes Bayesian inference (integration o cues according to their reliabilities). On each trial, the integrated representation is calculated as $X_{AV} = \Sigma_{AV}(\Sigma_A^{-1}X_A + \Sigma_V^{-1}X_V)$, where $\Sigma_{AV} = \left(\Sigma_A^{-1} + \Sigma_V^{-1}\right)^{-1}$. Across many trials, the distribution of the $C = 1$ representations will be the weighted average of the locations for S_A and S_V with the weighting controlled by the relative precision of the encoding matrices. For $C = 2$, the representation is the encoded representation of the auditory portion of the stimulus (here we assume observers always choose to report auditory modality when the cues are clearly from separate causes). Next, the log posterior ratio of $C = 1$ to $C = 2$ is calculated:

$$d = \log\frac{P(C = 1|X_A, X_v)}{P(C = 2|X_A, X_v)} = \log\frac{P(X_A, X_v|C = 1)}{P(X_A, X_v|C = 2)} + \log\frac{P(C = 1)}{P(C = 2)}$$

The prior probability of a common cause, $P(C = 1)$, is set to .50 (giving no prior bias toward one *vs* two causes), resulting in $P(C = 2) = .50$. $P(X_A, X_V|C = 1)$ is calculated for each syllable, S_i, individually. These probabilities are then combined, weighted by their respective prior probabilities to determine the overall conditional probability:

$$P\left(X_A, X_V|C = 1) = \sum_i P(X_A, X_V|S_i\right)\pi_S(S_i)$$

$P(X_A, X_V|C = 2)$ is calculated using a similar process for all possible incongruent syllable combinations. Their locations in representational space are determined as

$$\mu_{S_i,C=2} = \Sigma_{S_i,C=2}\left(\Sigma_{A'}^{-1}X_A + \Sigma_{V'}^{-1}X_V\right)\text{with } \Sigma_{S_i,C=2} = \left(\Sigma_{A'}^{-1} + \Sigma_{V'}^{-1}\right)^{-1}$$

X_A and X_V are the locations for the unisensory components. The matrices $\Sigma_{A'}$ and $\Sigma_{V'}$ are the original sensory noise matrices plus the variance of the syllable category that generated the exemplar. The probabilities are then calculated using the Gaussian density: $\mathcal{N}_P(X_{A,}X_V; \mu_{S_i,C=2}, \Sigma_{S_i,C=2})$ and weighted by their prior probabilities (assumed to be equal).

After the decision variable d is computed, it is converted to the probability of each causal structure: $P\Big(C = 1 \,|\, X_A, X_V) = \frac{1}{1+e^{-d}}$; $P(C = 2 \,|\, X_A, X_V\Big) = \frac{1}{1+e^{d}}$

Next, the $C = 1$ and $C = 2$ representations are combined, weighted by their likelihood:

$$X_{CIMS} = P(C = 1 \,|\, X_A, X_V)X_{AV} + P(C = 2 \,|\, X_A, X_V)X_A$$

This combination is done on a trial-by-trial basis, producing a non-linear combination of the original exemplars (X_A and X_V). To produce a categorical percept, the syllable that is most likely to have generated the integrated representation is determined: $P(X_A, X_V|S_i) = \mathcal{N}_P(X_{A,}X_V; \mu_{S_i,C=1}, \Sigma_{S_i,C=1})$, where $\mathcal{N}_P$ is the P-dimensional Gaussian density function, $\mu_{S_i,C=1}$ is the P-dimensional location of a particular syllable category and $\Sigma_{S_i,\ C=1} = \Sigma_i + \Sigma_{AV}$ is the sum of the category's variance-covariance matrix and the variance of X_{AV}. All syllables are assumed have equal prior probability, all locations within the representational space have equal prior probability, and the category variance-covariance matrices are assumed to be equal.

The CIMS model assumes that individuals vary in the precision with which they represent speech features, but that a given individual has a constant level of precision across different stimulus exemplars. Measurements of neural variability in responses with techniques such as single-trial fMRI, MEG and iEEG could prove or disprove this assertion. Individuals with less neural variability across trials would be expected to have more precise perceptual representations of speech features. Adding noise to the sensory stimulus is assumed to broaden the variance of the distribution of perceived locations in representational space. The sensory noise for a specific modality is assumed to be constant and characteristic of the specific observer.

2.4. *Overview of experimental procedures*

Participants viewed brief recordings of audiovisual speech and reported their percepts. Experiments 1–6 examined perception of syllables (both McGurk and congruent) using the online data collection service Amazon Mechanical Turk (Buhrmester et al., 2018). In a previous study, we found that online testing gives similar results as in-person testing (Magnotti, Smith, et al., 2018). Experiment 7 examined the McGurk effect and perception of noisy sentences, with data collected in-person at Baylor College of Medicine. A total of 262 different participants were tested online and 34 different participants were tested in person for a total of 296 different participants.

All data was analyzed using R (R Core Team, 2020). Variability was modeled using linear mixed effects models (LMEs) as implemented in the lme4 packager (Bates et al., 2015). LMEs provide a consistent approach for understanding the effect of both categorical and numeric independent variables (fixed effects) while taking into account other sources of variation (random effects such as participant effects or stimulus effects). To test the significance of the fixed effects, we report *t*-tests with Satterthwaite-approximated degrees of freedom, as implemented in the lmerTest package (Kuznetsova et al., 2017).

For the Mechanical Turk experiments, control stimuli were included in each experiment, consisting of congruent AV stimuli or noise-free A-only stimuli, to measure whether participants were able to correctly perform the task. Performance on control stimuli was high for every participant, indicating a willingness and ability to attend to the stimulus, correctly identify it, and select the appropriate response.

Sample sizes for the initial experiment ($n = 40$ per stimulus) were determined based on our previous use of similar sample sizes to observe stimulus differences in the McGurk effect. Because the stimuli chosen for the follow-up studies depended on the results of Experiment 1, they could not be pre-determined. Instead, we included replications of the primary effect (stimulus-level differences in auditory-only accuracy relate to stimulus-level McGurk responses) in subsequent experiments to bolster our confidence in the difference. Based on previous findings of no significant subject-level correlations between McGurk and speech-in-noise, we did not attempt a power analysis for experiment 7, and instead used a sample size comparable to previous studies.

For experiment 5, data were collected in two groups of $n = 25$. Six participants enrolled in both groups. To avoid unduly weighting the behavior of these six participants, their data from the second testing group was excluded from analysis, leaving $n = 19$ in the second group and a total of $n = 44$ participants.

2.5. *Online testing procedures for experiments 1 - 6*

Participants enrolled as workers in the Mechanical Turk service and requested to participate in our experiment, in return for compensation of $10 per hour. After accepting the assignment, they were directed to an informed consent statement, followed by completion of a demographic questionnaire. Before starting the experiment, participants were shown a demonstration video. Participants were instructed to resize their browser window and their computer audio as needed to make the demonstration video easily visible and audible. The demonstration video could be repeated as often as necessary by the participant. Then, the participants proceeded to the main experiment, which consisted of multiple trials. Within each trial, participants viewed a single stimulus and reported their percept using a forced-choice response, selecting among three possibilities, “ba” (the auditory component of the stimulus), “ga” (the visual component of the stimulus) or “da/tha” (the McGurk or fusion percept). In a previous study, we demonstrated similar results between this set of forced-choice responses and open-choice responding (Basu Mallick et al., 2015). Forced-choice has the advantages of reducing the time to analyze data and of reducing the experimenter degrees of freedom available when quantifying open-choice responses. No feedback was given to reduce demand characteristics.

2.6. *Experiment 1 stimulus creation*

The experimental stimulus set in Experiment 1 consisted of twenty different McGurk syllables (auditory “ba” paired with visual “ga”, AbaVga) where the “ba” was different in each syllable but the “ga” was identical. A female native speaker of American English was recorded voicing the syllable “ba” twenty times. The auditory component of each stimulus was imported into MATLAB and the volume of each clip was normalized by dividing the sound amplitude by the square root of the squared mean. The resulting waveform was scaled to prevent clipping, with each auditory track scaled to the same power. See Supplemental Table 1 for quantification of acoustic properties of the recordings.

The same talker was recorded saying a single “ga” using a Panasonic AG-HVX200AP video camera. The camera view showed the talker’s head and shoulders against a white background. The video obtained was imported into Adobe Premiere Pro CC 2015. Then, each of the twenty auditory “ba” recordings was dubbed onto the visual portion of the “ga” recording so that the auditory and visual components were synchronized. Each was exported to MOV format and Handbrake software was used to crop and convert the videos to 640 by 480 resolution in the MP4 format.

The control stimulus set in Experiment 1 consisted of audiovisual recordings of a different female native speaker of American English speaking three syllables for which the auditory and visual components were congruent, (AbaVba, AgaVga, AdaVda).

To avoid participant fatigue or adaptation, each participant was presented with five different McGurk stimuli (randomly selected from the entire battery of twenty) and the three congruent stimuli. Each McGurk stimulus was presented nine times and each congruent stimulus was presented three times, all randomly interleaved.

A total of 160 participants were recruited (57 female, 93 male, 10 did not specify). The 20 stimuli were tested in 4 batches of 5 stimuli each. The first 40 participants viewed the first five stimuli; the second 40 participants viewed the next five stimuli; and so on. Since each participant viewed one-quarter of the stimuli (5 out of 20), the final sample size was 40 participants for each of the twenty McGurk stimuli.

2.7. *Experiment 2*

From the twenty different McGurk stimuli presented in Experiment 1, we selected two stimuli at opposite ends of the distribution for further investigation, labeling them “S1” and “S2”. A total of 40 participants were recruited for experiment 2 (10 female, 25 male, 5 did not specify). Each participant was presented with 9 repetitions each of S1 and S2 and three repetitions of the congruent stimuli, randomly interleaved, and participants made forced choice responses.

2.8. *Experiment 3*

MATLAB was used to average the auditory components of S1 and S2 (after aligning auditory onsets) to create an intermediate stimulus labeled “S1.5” (the visual “ga” component of S1.5 was identical to S1 and S2).

The experimental stimulus set in Experiment 3 consisted of the auditory-only “ba” component of S1, S1.5 and S2. The control stimulus set consisted of the auditory-only component of the control stimuli in Experiment 1, auditory “ba”, “da” and “ga”. The visual component for all stimuli consisted of white text on a black square instructing participant to “Listen to the audio.” 40 participants (14 female, 23 male, 3 did not specify) were presented with nine repetitions of each of the experimental stimuli and three repetitions of each of the control stimuli, all randomly interleaved.

2.9. *Experiment 4*

The experimental stimuli were the audiovisual AbaVga syllables S1, S1.5 and S2. The control stimuli were the congruent syllables AbaVba, AgaVga and AdaVda. 40 participants (20 female, 19 male, 1 did not specify) were presented with nine repetitions of each McGurk stimulus and three repetitions of each congruent syllable, all randomly interleaved.

2.10. Experiment 5

Auditory noise was added to S1 and S2 by combining each stimulus with uniform white noise with signal to noise ratios (SNRs) of −30, −24, −18, −12, 0 dB and no noise. Final Cut Pro was used to create the noisy stimuli by combining the noise clip and the syllable clip and adjusting the decibel levels of each to the desired SNR. After noise was added, the volume of each clips was RMS normalized in MATLAB and the stimuli were imported into Final Cut Pro to add a blank visual screen for the visual component, followed by resizing in Handbrake to a 640:480 aspect ratio. 44 participants (21 female, 22 male, 1 did not specify). Each of the two stimuli (the "ba" extracted from S1 and S2) was presented six times at each of the six noise levels (72 total presentations). Control stimuli consisted of two examples of auditory "da" and two examples of auditory "ga" recorded by the same talker. The same six noise levels were added to each control stimulus and each was presented six times, ensuring an equal total number of experimental and control stimulus presentations (72 for each).

2.11. Experiment 6

The 44 participants from Experiment 5 were invited to return. 38 participants returned (18 female, 19 male, 1 did not specify) and rated the McGurk stimuli S1, S1.5 and S2 to allow for intra-participant comparison of noisy syllable and McGurk perception. Experimental stimuli consisted of 10 presentations each of the McGurk stimuli S1, S1.5 and S2. Control stimuli consisted of three presentations each of the congruent audiovisual syllables AbaVba, AgaVga and AdaVda.

2.12. Experiment 7 overview

34 native English speakers (18 female, 16 male) were tested in-person at the Core for Advanced MRI at Baylor College of Medicine. All tasks were presented using Matlab (Mathworks, Inc., Natick, MA, USA) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Visual speech was presented on a high-resolution screen (Display++ LCD Monitor, 32-in., 1,920 × 1,080, 120 Hz, Cambridge Research Systems, Rochester, UK). Auditory speech was presented through speakers on either side of the screen at a constant sound pressure level of 60 dB, a value similar to that of face-to-face conversation.

2.13. Experiment 7: McGurk and congruent syllables

First, participants viewed stimuli consisting of 2-s recordings of audiovisual syllables with no added noise. The syllables were recorded by five different talkers. Each participant was presented with a total of 270 trials: 20 repetitions × four talkers × three syllables (two congruent: AbaVba, AdaVda; one McGurk AbaVga) and 10 repetitions × one talker × three audiovisual syllables (all congruent: AbaVba, AdaVda, AgaVga), all randomly interleaved. Participants reported the identity of each syllable ("ba", "da", or "ga") with a button press.

2.14. Experiment 7: Noisy speech

Second, participants were presented with 3-s duration sentences recorded from a single male talker combined with auditory pink noise at a signal-to-noise ratio (SNR) of −16 dB, as used in a previous study (Van Engen et al., 2017). The sentences were presented either alone (noisy auditory-only, A) or paired with a video recording (noisy auditory + visual, AV). After the sentence ended, participants repeated the sentence aloud. Responses were scored for number of correct keywords (*e.g.,* "The **hot sun warmed** the **ground**," keywords in bold). Each participant was presented with 80 sentence trials, consisting of randomly interleaved presentations of 40 auditory-only and 40 audiovisual sentences. To prevent perceptual learning, individual sentences were never repeated.

3. Results

3.1. Section 1: Stimulus variability

If McGurk and other forms of speech are processed using common perceptual mechanisms, then variability across stimuli should create predictable changes. For instance, even for a single talker, there is substantial variability in repeated productions of the same speech token (Holmberg et al., 1994; Whalen et al., 2018). In the CIMS model, this variability can be modeled with a representational space that collapses all auditory and visual speech features onto a one-dimensional line with "ba", "da", and "ga" at neighboring locations. One production of the syllable "ba" (Stimulus 1) might lie near the prototypical "ba", while a second production (Stimulus 2) might lie further from the prototype (Fig. 1A).

An important feature of the CIMS model is noisy sensory encoding. While the physical properties of a given speech token place it at one location in representational space, sensory encoding is noisy—repeated presentations of the identical token produce a distribution of perceived locations whose mean is at the true location and whose variance is proportional to the amount of sensory noise. Over repeated presentations of Stimulus 1, its location far from the perceptual boundary means that despite sensory encoding noise, nearly all of the perceived locations are in the "ba" region of representational space (Fig. 1B). However, for repeated presentations of Stimulus 2, its location near a perceptual boundary means that sensory noise places some of the perceived locations in the "da" region of representation space (Fig. 1C).

The CIMS model assumes that auditory syllables and McGurk syllables are processed by the same perceptual mechanisms, resulting in predictable differences if Stimulus 1 and 2 are paired with an identical visual "ga" in a McGurk AbaVga stimulus. For Stimulus 1, optimal cue combination produces an integrated representation that lies predominantly in the "ba" region of representational space, resulting in primarily "ba" percepts (Fig. 1D). For Stimulus 2, the integrated representation lies primarily in the "da" region of

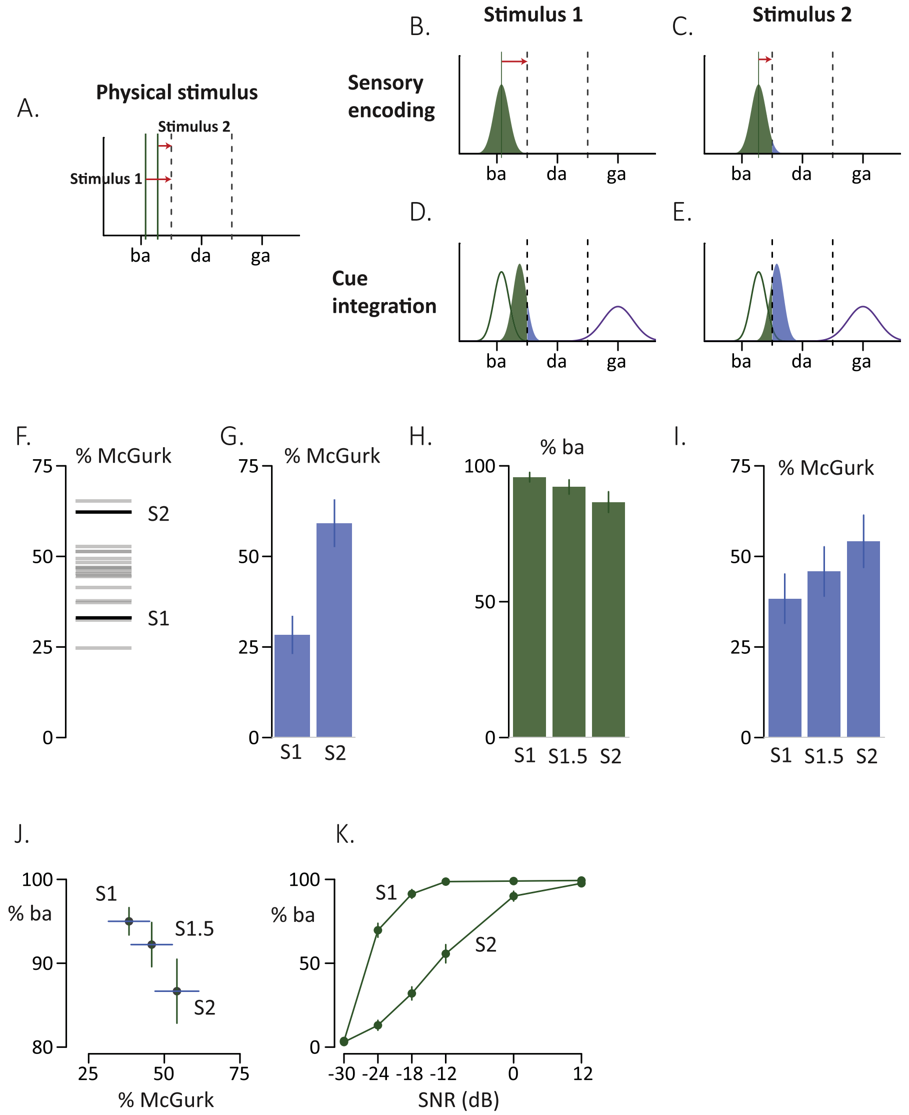

Figure 1 – The CIMS model applied to stimulus variability. (A) In the causal inference of multisensory speech perception (CIMS) model, the physical properties of auditory and visual speech can be collapsed onto a one-dimensional representational space with different regions of the space representing different tokens ("ba", "da" and "ga" shown). The physical properties of a token determine its location in representational space, as shown for two example/ba/tokens. The left token (Stimulus 1) is closer to the center of the "ba" " region of representational space and hence is a more prototypical "ba". The right token (Stimulus 2) is further from the center of the "ba" region of representational space, and hence is a less prototypical "ba" . Dashed lines indicate boundaries between different regions of representational space. (B) In the sensory encoding stage of the CIMS model, the physical properties of the stimulus are encoded with sensory noise, resulting in a distribution of encoded values for any given token. The mean of the distribution is determined by the physical properties of the stimulus and the variance of the distribution is determined by the sensory noise for that modality for that observer. For Stimulus 1, the stimulus is far from the perceptual boundary (distance indicated by red arrow), with the result that even after noisy encoding, most perceived locations are in the "ba" region of representational space, resulting in exclusively "ba" percepts (green shaded region). (C) For Stimulus 2, the stimulus is close to the perceptual boundary (distance indicated by red arrow) with the result that after noisy encoding, some perceived location s are in the "da" region of representational space (blue shaded region). (D) In the cue integration stage of the CIMS model, auditory cues (green lines) and visual cues (purple lines) are integrated using optimal cue combination, resulting in an average representation weighted by the reliability of each modality. For Stimulus 1, most perceived locations for the integrated representation are in the "ba" x region of representational space, resulting in mainly "ba" percepts (green shaded region). (E) For Stimulus 2, the location of the auditory component nearer the perceptual boundary means that many perceived locations for the integrated representation are in the "da" region of representational space, the McGurk fusion percept (blue shaded region). (F) In experiment 1, twenty different "ba" tokens recorded by the same talker were paired with a single "ga" from the same talker to create twenty different AbaVga McGurk stimuli. Each bar represents the % of McGurk fusion for a single stimulus. Two

representational space, corresponding to the McGurk fusion perception of “da” (Fig. 1E).

3.2. Experiments 1 and 2: finding strong and weak McGurk stimuli, with replication

To test these predictions, in Experiment 1 twenty productions of auditory “ba” were paired with the same visual “ga”. Across stimuli, there was substantial variability in the responses, ranging from 25% fusion responses to 65% fusion responses (Fig. 1F). Participants were also presented with unmanipulated control stimuli consisting of three different congruent audiovisual syllables (AbaVba, AdaVda, and AgaVga). Participants responded accurately to the control stimuli, demonstrating that they were engaged in the task and not responding randomly. Table 1 shows behavioral data from all experiments. Supplemental Table 1 shows measurements of acoustic properties for the McGurk stimuli.

From the twenty different McGurk stimuli, we selected two stimuli at opposite ends of the distribution for further investigation, labeling them "S1" and "S2" by analogy with the modeling above. To ensure the stimulus differences between S1 and S2 were reliable, in Experiment 2 we attempted to replicate the results of Experiment 1, presenting only S1 and S2. Similar rates of McGurk fusion responses were observed (Fig. 1G; S1: 33% original *vs* 28% replication; S2: 62% *vs* 59%). A linear mixed-effects analysis on fusion responses showed a significant effect of stimulus [t (80) = −4.3, $p = 10^{-5}$], but no effect of experiment [t (156) = −.4, p = .66] or stimulus-by-experiment interaction [t (80) = −.18, p = .86].

3.3. Experiments 3 and 4: comparing McGurk effect and clear auditory syllables

Next, we created a stimulus that was the average of S1 and S2 (labeled “S1.5″) in order to test two predictions of the CIMS model. First, CIMS predicts that if S1 and S2 lie at different locations in the representational space, then S1.5 should lie between them. Second, CIMS predicts that there should be a relationship between the perception of auditory “ba” syllables presented alone and in combination with visual “ga”.

To test the first prediction, in Experiment 3 we presented the three auditory-only “ba” components of S1, S1.5 and S2 to 40 participants. As expected, there was a decrease in the number of “ba” percepts from S1 to S1.5 to S2 (Fig. 1H; 96% to 92% to 87%).

To test the second prediction, in Experiment 4, we presented the three stimuli (S1, S2, and S1.5, each consisting of a different auditory "ba" paired with same visual "ga") to 40 participants. As expected, there was an increase in the McGurk fusion percepts from S1 to S1.5 to S2 (Fig. 1I). Replicating the results of Experiments 1 and 2, S2 produced more fusion responses than S1 (54% *vs* 38%; paired t-test, t = −2.4, p = .02). The S1.5 stimulus produced an intermediate level of fusion responses (46%).

To determine if there was a stimulus-level relationship between syllable perception and the McGurk effect, we plotted the rates of McGurk perception for S1, S1.5 and S2 against the perceptual accuracy for the auditory component of each stimulus (Fig. 1J). To test for the linearity of this relationship, we compared two different LMEs. The first LME assumed a linear relationship between stimuli (coded 0, .5 and 1) while the second LME allowed the stimuli to vary freely (categorical coding of stimuli). Comparing BIC between the two models yielded a better fit for the model assuming a linear relationship (BIC difference 9). For the winning model, the estimated slope across stimuli was 16 ± 6 [t (80) = 2.6, p = .01].

3.4. Experiments 5 and 6: comparing McGurk effect and noisy auditory syllables

The CIMS model predicts that adding noise to the sensory stimulus should broaden the distribution of perceived locations, which should differentially affect the different stimuli, with a bigger effect on the weaker S2 stimulus. To test this prediction, in Experiment 5 we added auditory noise to the auditory-only “ba” component of S1 and S2 and presented them to 40 participants (Fig. 1K). Neither stimulus was perceived as “ba” at the highest noise level (SNR of −30 dB). Decreasing the amount of added noise progressively increased the number of "ba" reports until both stimuli always evoked a “ba” percept at an SNR of 12 dB.

At intermediate noise levels, there were always fewer “ba” reports for Stimulus 2 than Stimulus 1. Fig. 1B and C provide a graphical depiction of the CIMS explanation for this observation. Adding noise broadens the distribution of perceived locations of a stimulus within the representational space. Stimulus 2 is a weaker stimulus as it lies close to the ba/da boundary. Broadening the distribution by adding noise means that the perceived location of S2 often falls outside the “ba” region of representational space, resulting in few “ba” reports. In contrast, Stimulus 1 is a stronger stimulus and lies far from the ba/da boundary. For the same noise level (distribution width), most presentations of S1 fall within the “ba” region of representational space, resulting in many “ba” reports.

An LME with fixed factors of noise (entered as SNR, with clear speech set to +12 dB), stimulus, and their interaction along with random effect of subject yielded significant main effects for stimulus [t (484) = 7.0, p = 10^{-11}], noise [t

stimuli were selected for further analysis, dark bars labelled "S1" and "S2", analogous to modeled Stimulus 1 and Stimulus 2. (G) In experiment 2, S1 and S2 were presented to a different set of participants. There was no significant difference between experiment 1 and the replication sample in experiment 2 (*p* = .66). (H) In experiment 3, a new stimulus (labeled "S1.5") was created by averaging S1 and S2 and the auditory-only "ba" component of S1, S1.5 and S2 was presented. Mean and SEM of % “ba” percepts for each stimulus are shown. (I) In experiment 4, the McGurk stimuli S1, S1.5 and S2 were presented, mean and SEM of % McGurk percepts for each stimulus are shown. (J) The rates of McGurk perception for S1, S1.5 and S2 were plotted against the perceptual accuracy for the auditory component of each stimulus, (I) versus (H). (K) In experiment 5, different levels of auditory noise were added to the auditory components of S1 and S2 and the rate of “ba” responses measured.

Table 1 – Summary of all experiments.

Exp	N	Congruent Performance (mean ± sem)			Responses to AbaVga (mean ± sem)			
		ba	da	ga	# Stim	ba	da	ga
1	128	97 ± 1	87 ± 2	98 ± 1	20	26 ± 3	46 ± 3	28 ± 2
2	40	97 ± 1	94 ± 3	98 ± 1	2	37 ± 5	44 ± 5	19 ± 4
3	40	96 ± 2	80 ± 6	98 ± 2				
4	40	99 ± 1	93 ± 3	97 ± 2	3	34 ± 5	46 ± 6	20 ± 4
5	44	98 ± 1	98 ± 1	88 ± 3				
6	38	96 ± 3	86 ± 5	100 ± 0	2	41 ± 7	41 ± 5	18 ± 4
7	34	97 ± 2	77 ± 6	97 ± 1	4	35 ± 7	56 ± 7	6 ± 3

Summary information details for each experiment (Exp) including the number of participants (N) and the mean and standard error of the mean (SEM) accuracy across participants for the control stimuli (Congruent Performance) and the response percentages to the McGurk AbaVga stimuli (first averaged across stimuli, then mean and SEM across participants).

(484) = 18.9, $p < 10^{-16}$] and their interaction [t (484) = −4.8, $p = 10^{-5}$]. Compared with clear speech, at −12 dB the number of “ba” reports decreased only slightly for S1 [99% *vs* 98%; paired *t* (43) = −.8, *p* = .42] but decreased greatly for S2 [98% *vs* 53%; *t* (43) = −8.0, $p = 10^{-10}$].

3.5. *Section 2: Participant variability*

There are two sources of participant variability in CIMS. The first source is individual differences in sensory encoding noise. Observer 1 might precisely encode speech, creating a narrow distribution of perceived locations in representational space, while Observer 2 might imprecisely encode speech, creating a broad distribution (Fig. 2A). When presented with an auditory “ba” near the perceptual boundary, the precise representation of Observer 1 will result in mainly “ba” percepts while the imprecise representation of Observer 2 will blur many of the estimates into the “da” regions of representational space (Fig. 2B).

The second source of participant variability in CIMS is individual differences in causal inference. Multisensory integration is only beneficial if the cues to be integrated were caused by the same real-world event (a single cause, C = 1); integrating cues generated by two different real-world events (C = 2) worsens perception. For presentation of clear or noisy syllables, all cues strongly indicate that C = 1, reducing the impact of participant variability in causal inference. In contrast, McGurk stimuli are created by dubbing incongruent auditory and visual syllables, creating a conflict between the temporal synchrony and spatial coincidence of the auditory and visual syllables (which suggest that C = 1) and the content disparity between the heard speech and the viewed mouth movement (which suggests that C = 2). For observers with a high tolerance for content disparity (leading them to infer that C = 1) optimal cue combination will more strongly weight the integrated representation of auditory and visual speech, usually resulting in the illusory McGurk percept. For observers with a low tolerance for content disparity (leading them to infer that C = 2) optimal cue combination will more strongly weight the reliable auditory-speech representation, usually resulting in a percept corresponding to the auditory token.

3.6. *Experiments 5 and 6: participant variability in McGurk effect and noisy syllable perception*

To test these ideas, we examined the performance for those participants who participated in both experiment 5 (where they were presented with noisy syllables) and experiment 6 (where they were presented with McGurk stimuli). To prevent floor or ceiling effects, the measure for noisy syllable perception was the accuracy of perception of stimulus S2 presented at the −12 dB noise level in experiment 5. There was substantial variability in perception of the McGurk effect, with rates ranging from 0% to 100% and in perception of noisy syllables, with accuracy ranging from 0% to 100%. However, across participants there was low correlation between the two values, *r* (36) = −.09, *p* = .60 (Fig. 2E). Participants were found in all quadrants of the plot, explained by CIMS as participants with low versus high encoding noise and low versus high tolerance for audiovisual disparity in their causal inference judgments.

Next, we considered all noise levels across participants by comparing two LMEs: the original LME fit to the noisy syllable data (dependent variable accuracy; fixed effects of stimulus, noise level and their interaction; random effect of subject) and a second LME that additionally contained subject-level McGurk perception (per stimulus). Comparing BIC between models, we found that knowing subject-level McGurk responses did not explain additional variance in noisy-syllable perception (BIC difference 11). Importantly, this absence of participant-level relationships between McGurk and noisy syllable perception occurred despite the presence of a stimulus-level relationship: S1 was more accurate than S2 (77% *vs* 48%, across all noise levels), but S2 produced more McGurk responses (61% *vs* 39%).

3.7. *Experiment 7: participant variability in McGurk effect and noisy sentence perception*

To assess whether there was an across–participant correlation for more complex forms of speech, in Experiment 7 we compared perception of the McGurk effect and perception of noisy sentences in 33 participants. There was substantial variability in the rate of perceiving the McGurk effect across

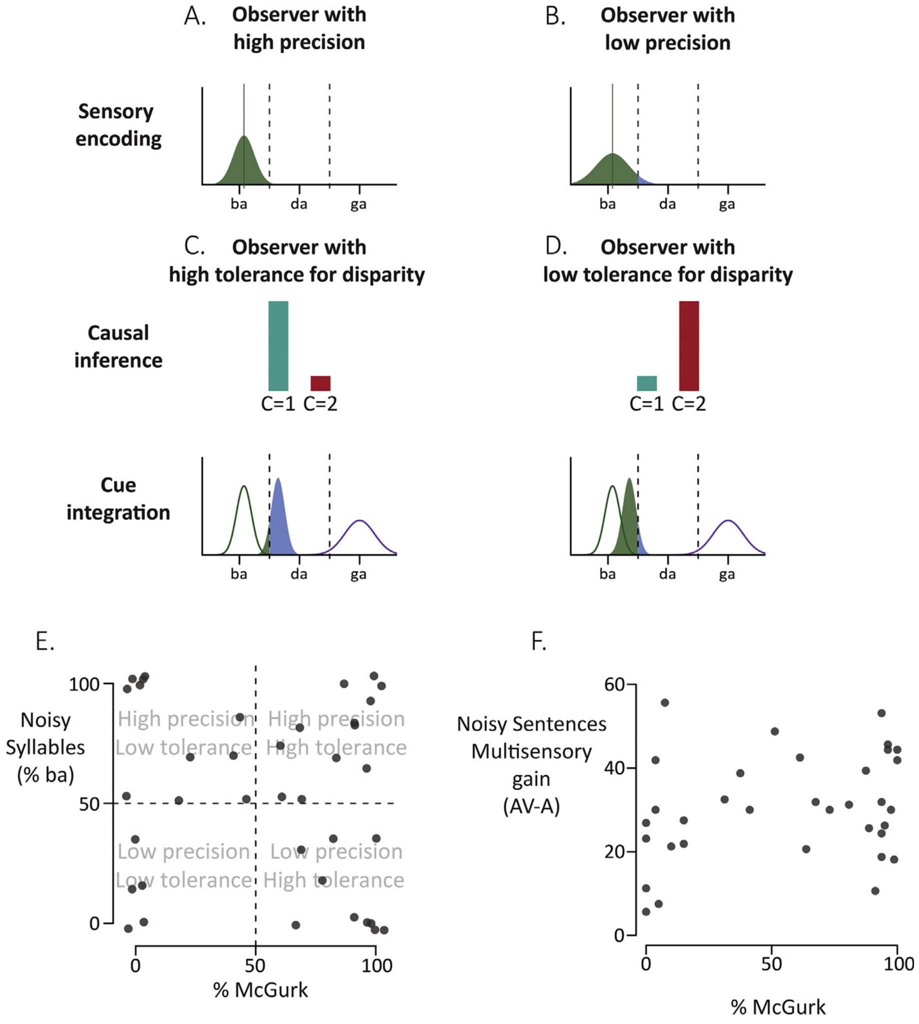

Figure 2 – The CIMS model applied to participant variability. (A) Variability in speech perception between participants can arise from individual differences in sensory encoding. For an observer with high precision, there will be a narrower distribution of perceived locations for a given stimulus (shown for Stimulus 1 from Fig. 1A). (B) For an observer with low precision, there will be a broader distribution of perceived locations for a given stimulus (shown for Stimulus 1 from Fig. 1A). (C) Even if encoding precision is identical, variability in speech perception between participants can also arise from individual differences in causal inference. For an observer with high tolerance for disparity presented with a McGurk stimulus consisting of auditory "ba" and visual "da", the observer infers that the auditory and visual cues arise from a single talker ($C = 1$, illustrated as relative heights of $C = 1$ and $C = 2$ bars). Optimal cue integration reflects this inference so that the integrated representation lies between the auditory and visual representations, with most percepts falling in the "da" region of representational space (blue shaded region, high rates of McGurk). (D) An observer with low tolerance for disparity infers that the auditory and visual speech in the McGurk stimulus arises from two different talkers ($C = 2$). Optimal cue integration reflects this inference so that the integrated representation is weighted by the auditory-only representation, with most percepts falling in the "ba" region of representational space (green shaded region, low rates of McGurk). (E) Rates of McGurk and accuracy of noisy auditory syllable presentation across participants, one symbol per participant, r (36) = −.09, p = .60. The noisy syllable measure is the accuracy of perception of stimulus S2 with −12 dB noise added from experiment 5 and the rate of McGurk perception is from experiment 6. Participants were distributed across quadrants with all combinations of low and high sensory encoding precision and low and high tolerance for audiovisual disparity. During plotting, the location of each symbol was randomly jittered by up to 2% in both the x and y directions to create visual separation between overlapping symbols; this can result in plotted values that lie outside the actual range of the data values (0%–100%). (F) Rates of McGurk and multisensory gain during noisy sentence perception across participants in experiment 7, one symbol per participant; r (31) = .261, p = .14. The audiovisual gain measure is the percentage of keywords understood during perception of noisy sentences with the talker visible minus the percentage of keywords understood during perception of auditory-only noisy sentences.

different participants, from 0% to 100% (Fig. 2F). For noisy sentences, participants correctly reported 14 out of 160 (9%) of keywords for auditory-only sentences but correctly reported 62 out of 160 (39%) of key words for audiovisual sentences, a 30% improvement [AV - A, paired t (33) = 14.0, p = 10^{-15}]. However, there was substantial variability in the visual benefit across participants, ranging from a 6% improvement in the number of key words recognized to a 56% improvement (Fig. 2F). For consistency with previous reports, in addition to multisensory gain (AV - A) we also calculated a visual enhancement index for each participant, VE = [(AV - A)/(1 - A)] (Van Engen et al., 2017). The mean visual enhancement was 34% (range 6%–63%).

Next, we compared the two axes of individual variability. There were low correlations between rates of McGurk effect and multisensory gain, r (31) = .261, p = .14 (Fig. 2F); between rates of McGurk effect and the visual enhancement index, r (31) = .256, p = .14; and between rates of McGurk effect and auditory-only performance, r (31) = −.06, p = .73. As in the syllable perception data, participants fell in all 4 quadrants of the plot, with high and low McGurk susceptibility and high and low multisensory gain.

4. Discussion

We used natural variation in the production of the syllable "ba" to understand the relationship between the McGurk effect and other speech perception tasks. Pairing twenty unique auditory "ba" syllables with a single visual "ga" produced stimuli that elicited reliably different levels of McGurk perception across large groups of subjects. For two stimuli that differed substantially in their McGurk strength, the auditory "ba" that was *less* effective at evoking the McGurk effect was recognized *more* accurately in both clear and noisy auditory-only perception tasks. The stimulus-level relationship between McGurk and auditory-only perception was well-described by the CIMS model. Across participants, the CIMS model also provided a straightforward explanation of why perception of noisy speech and McGurk stimuli are only weakly correlated. For speech-in-noise tasks, individual differences arise from variability in sensory encoding (how well one understands the individual speech tokens) while for the McGurk effect, individual differences arise from both sensory encoding and variability in how cue conflict between modalities is resolved.

4.1. *Stimulus differences in the McGurk effect*

Basu Mallick and colleagues (Basu Mallick et al., 2015) reported high variation in the McGurk effect across different stimuli (rates ranging from 17% to 58%) and participants (rates from 0% to 100%). While a careful study of many acoustic and visual properties of McGurk stimuli showed that taken together they accounted for about half of the variability in the frequency of the effect across stimuli and participants (Jiang & Bernstein, 2011), it is not clear if the factors contributing to variability in the McGurk effect are relevant for other forms of speech perception.

Speech production is known to be variable in both articulation and acoustics (Holmberg et al., 1994; Whalen et al., 2018), with voice onset time serving as a particularly useful measure (Abramson & Whalen, 2017). The stimuli in the Basu Mallick study were culled from different sources and contained different talkers, different views of the face and upper body, different native languages, and different video and audio quality. To better control these factors, in the present study we created a new corpus of McGurk stimuli that were all recorded from the same talker within a short time span and had the same visual component, making them closely matched for auditory properties and with identical visual properties. Nevertheless, the efficacy of each McGurk stimulus varied, and this variability was related to the perception of the auditory component of each stimulus in the direction predicted by the CIMS model. Auditory "ba" tokens modeled as being more prototypical were more likely to evoke a "ba" percept when presented alone, either with or without added noise, and were less likely to evoke a McGurk fusion percept when paired with a visual "ga".

4.2. *Relating the McGurk effect to other speech perception tasks*

Van Engen and colleagues (Van Engen et al., 2017) reported non-significant correlations between McGurk effect and visual enhancement of noisy speech, while Brown and colleagues (Brown et al., 2018) reported weak but significant correlations between McGurk effect and lipreading accuracy. Many other studies have also reported weak relationships between different measures of speech perception (Grant and Seitz, 1998, 2000; Sommers et al., 2005; Stevenson et al., 2012; Strand et al., 2018; Rennig et al., 2020; Stacey et al., 2020). If correlations between any two measures are weak, large sample sizes are necessary for accurate estimation. The present study found a non-significant relationship (p = .14) between McGurk perception and noisy speech perception, with an effect size of r = .26. To detect (p < .05) an effect of this size with 90% power would require a sample size of 151 participants, more than in most published studies of the McGurk effect (Magnotti & Beauchamp, 2018).

Perception of noisy speech and lipreading do not require causal inference, as all cues indicate that the speech arises from a single talker. In contrast, perception of the McGurk effect requires a causal inference judgment because of the conflicting cues from temporal synchrony (which suggests that $C = 1$) and content incongruity (which suggests that $C = 2$). Any model of cue conflicts requires additional machinery beyond sensory encoding, introducing additional individual variability. Evidence suggests that cross-subject variability in the tendency to bind audiovisual signals is found across a range of tasks, and that these differences are stable across time but are task-specific (Odegaard & Shams, 2016).

Findings that the McGurk effect shows different neural signatures than congruent audiovisual syllables (Erickson et al., 2014; Moris Fernandez et al., 2017; Sánchez-García et al., 2018) has been used as evidence that the McGurk effect is processed differently than other types of speech. An alternative explanation is that McGurk stimuli, but not congruent syllables, place greater demands on neural circuits underlying causal inference judgments.

Neurocomputational models of the McGurk effect show how the McGurk effect can be generated from biologically realistic neural responses. In the model of Cuppini and colleagues (Cuppini et al., 2017a, 2017b) distinct unisensory visual and auditory regions share reciprocal interconnections, as well as projecting downstream to a multisensory area that performs causal inference. Olasagasti and colleagues (Olasagasti et al., 2015) applied a hierarchical predictive coding framework to model sequential activation of causal units, incorporating the earlier arrival in the brain of visual speech information relative to auditory speech information.

These neurocomputational models are consistent with BOLD fMRI studies suggesting anatomical dissociations between brain areas responsible for sensory encoding and those responsible for causal inference judgments (Cuppini, Shams, et al., 2017; Rohe & Noppeney, 2015). Conversely, perception of noisy sentences calls on many cognitive processes in addition those required for syllable or McGurk perception (Davis & Johnsrude, 2007). Contextual information is thought to be modulated by top-down projections from frontal cortex, a different set of brain areas from the networks responsible for sensory encoding and causal inference judgments (Cope et al., 2017; Gau & Noppeney, 2016; Peelle & Sommers, 2015; Tuennerhoff & Noppeney, 2016). Given these neuroanatomical dissociations, it is unsurprising that individual differences in noisy speech perception are only weakly correlated with individual differences in McGurk perception.

5. Conclusions

One criticism of the McGurk effect is that it varies across stimuli and participants (Alsius et al., 2018; Rosenblum, 2019). This criticism is not compelling as it is equally true of other measures such as audiovisual speech-in-noise (Rennig et al., 2020; Van Engen et al., 2017). A second criticism of the McGurk effect, that it is only weakly correlated with other measures, is similarly non-specific: many measures of speech perception show weak pairwise correlations (Strand et al., 2018). CIMS and related models predict perception of both McGurk and other forms of speech with identical parameter sets, suggesting that both types of speech are processed using similar computations and that the McGurk effect can serve as a useful tool for interrogating everyday speech perception.

Selective publication of only the statistically significant results from underpowered studies is an important contributor to the replication crisis in science. A review of 119 published studies on the McGurk effect found an average group size of $n = 22$ participants (Magnotti & Beauchamp, 2018). Along with the current study, only a few studies have enrolled the large sample size ($n > 100$) necessary to accurately estimate group differences in the McGurk effect (Basu Mallick et al., 2015; Magnotti et al., 2015). In addition to variability across participants, there is also variability across different McGurk stimuli. This variability is consistent, such that a weak McGurk stimulus in one individual is also a weak stimulus in another (Basu Mallick et al., 2015; Magnotti & Beauchamp, 2015). Taken together, these two sources of variability underscore the importance of large n samples, tested using a variety of stimuli, when examining the McGurk effect or other phenomena with substantial variability (Magnotti & Beauchamp, 2018).

A more philosophical criticism of the McGurk effect is that the effect is somehow “unnatural” because stimuli are made by splicing together incongruent auditory and visual recordings. However, multiple talkers speaking at once is a common real-world situation that requires observers to process conflicting auditory and visual speech streams (French & DeAngelis, 2020; Kording et al., 2007). The McGurk effect, which requires resolution of conflicting auditory and visual cues, could therefore be a *better* model for understanding individual differences in real-world social interactions than simpler tasks not requiring conflict resolution. Of course, audiovisual speech perception is not a unitary phenomenon easily captured by a single behavioral measure (Soto-Faraco & Alsius, 2009). Rather than trying to classify different measures of speech perception as “good” or “bad”, we advocate creating an explicit model of the particular process of interest and using the model to guide selection of the appropriate stimulus and task.

Open Practices

The study in this article earned Open Materials and Open Data badges for transparent practices. Materials and data for the study are available at https://osf.io/C9EVY/.

Acknowledgments

This research was supported by NIH R01NS065395.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2020.10.002.

REFERENCES

Abramson, A. S., & Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics, 63*, 75–86.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257–262.

Aller, M., & Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. *Plos Biology, 17*, Article e3000210.

Alsius, A., Pare, M., & Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research, 31*, 111–144.

Arnold, D. H., Petrie, K., Murray, C., & Johnston, A. (2019). Suboptimal human multisensory cue combination. *Scientific Reports, 9*, 5155.

Basu Mallick, D. F., Magnotti, J. S., & Beauchamp, M. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review, 22*(5), 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, 1*, 1–48.

Baum, S. H., Stevenson, R. A., & Wallace, M. T. (2015). Behavioral, perceptual, and neural alterations in sensory and multisensory function in autism spectrum disorder. *Progress in Neurobiology, 134*, 140–160.

Beauchamp, M. S. (2018). Introduction to the special issue: Forty years of the McGurk effect. *Multisensory Research, 31*, 1–6. https://doi.org/10.1163/22134808-00002598

Brainard, D. H. (1997). The psychophysics Toolbox. *Spatial Vision, 10*, 433–436.

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*, Article e0207160.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of amazon's mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science, 13*, 149–154.

Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P. S., Wiggins, J., Dawson, C., Grube, M., Carlyon, R. P., Griffiths, T. D., Davis, M. H., & Rowe, J. B. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications, 8*, 2154.

Cuppini, C., Shams, L., Magosso, E., & Ursino, M. (2017). A biologically inspired neurocomputational model for audiovisual integration and causal inference. *The European Journal of Neuroscience, 46*, 2481–2498.

Cuppini, C., Ursino, M., Magosso, E., Ross, L. A., Foxe, J. J., & Molholm, S. (2017). A computational analysis of neural mechanisms underlying the maturation of multisensory speech integration in neurotypical children and those on the autism spectrum. *Frontiers in Human Neuroscience, 11*, 518.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research, 229*, 132–147.

Deneve, S., Latham, P. E., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience, 4*, 826–831.

Erickson, L. C., Zielinski, B. A., Zielinski, J. E., Liu, G., Turkeltaub, P. E., Leaver, A. M., & Rauschecker, J. P. (2014). Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Frontiers in Psychology, 5*, 534.

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

French, R. L., & DeAngelis, G. C. (2020). Multisensory neural processing: From cue integration to causal inference. *Current Opinion in Physiology, 16*, 8–13. https://doi.org/10.1016/j.cophys.2020.04.004

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *Neuroimage, 124*, 876–886.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104*, 2438–2450.

Grant, K. W., & Seitz, P. F. (2000). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *The Journal of the Acoustical Society of America, 107*, 1000–1011.

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception & Psychophysics, 77*(4), 1333–1341. https://doi.org/10.3758/s13414-014-0821-1

Holmberg, E. B., Hillman, R. E., Perkell, J. S., & Gress, C. (1994). Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in SPL across repeated recordings. *Journal of Speech ad Hearing Research, 37*, 484–495.

Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of experimental psychology Human perception and performance, 37*, 1193–1209.

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *Plos One, 2*, e943.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *lmerTest package: Tests in linear mixed effects models. 82* p. 26), 2017.

Magnotti, J. F., Basu Mallick, D., & Beauchamp, M. S. (2018). Reducing playback rate of audiovisual speech leads to a surprising decrease in the McGurk effect. *Multisensory Research, 31*, 19–38. https://doi.org/10.1163/22134808-00002586

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.

Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review, 22*, 701–709.

Magnotti, J. F., & Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *Plos Computational Biology, 13*, Article e1005229. https://doi.org/10.1371/journal.pcbi.1005229

Magnotti, J. F., & Beauchamp, M. S. (2018). Published estimates of group differences in multisensory integration are inflated. *Plos One, 13*, Article e0202908.

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in psychology, 4*, 798. https://doi.org/10.3389/fpsyg.2013.00798

Magnotti, J. F., Smith, K. B., Salinas, M., Mays, J., Zhu, L. L., & Beauchamp, M. S. (2018). A causal inference explanation for enhancement of multisensory integration by co-articulation. *Scientific Reports, 8*(1), 18032. https://doi.org/10.1038/s41598-018-36772-8

Massaro, D. W. (1998). *Perceiving talking faces : From speech perception to a behavioral principle*. Cambridge, Mass: MIT Press.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A bayesian explanation using high-dimensional feature space. *Plos One, 4*, Article e4638.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Moris Fernandez, L., Macaluso, E., & Soto-Faraco, S. (2017). Audiovisual integration as conflict resolution: The conflict of the McGurk illusion. *Human Brain Mapping, 38*, 5691–5705.

Noppeney, U., & Lee, H. L. (2018). Causal inference and temporal predictions in audiovisual perception of speech and music. *Annals of the New York Academy of Sciences, 1423*(1), 102–116. https://doi.org/10.1111/nyas.13615

Odegaard, B., & Shams, L. (2016). The brain's tendency to bind audiovisual signals is stable but not general. *Psychological Science, 27*, 583–591.

Olasagasti, I., Bouton, S., & Giraud, A. L. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 68*, 61–75.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169–181.

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

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2020). Face viewing behavior predicts multisensory gain during speech perception. *Psychonomic Bulletin & Review, 27*(1), 70–77. https://doi.org/10.3758/s13423-019-01665-y

Rohe, T., & Noppeney, U. (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception. *Plos Biology, 13*, Article e1002073.

Rosenblum, L. (2019). Audiovisual speech perception and the McGurk effect. In *Oxford Research Encyclopedia, Linguistics*. Oxford University Press.

Sánchez-García, C., Kandel, S., Savariaux, C., & Soto-Faraco, S. (2018). The time course of audio-visual phoneme identification: A high temporal resolution study. *Multisensory Research, 31*, 57–78.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences, 14*, 425–432.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*, 263–275.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of experimental psychology Human perception and performance, 35*, 580–587.

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, Perception & Psychophysics, 82*(7), 3544–3557. https://doi.org/10.3758/s13414-020-02042-x

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of experimental psychology Human perception and performance, 38*, 1517–1529.

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research: JSLHR, 61*, 1463–1486.

Stropahl, M., Schellhardt, S., & Debener, S. (2017). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: The Oldenburg Audio Visual Speech Stimuli (OLAVS). *Psychonomic Bulletin & Review, 24*, 863–872.

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

Tuennerhoff, J., & Noppeney, U. (2016). When sentences live up to your expectations. *Neuroimage, 124*, 641–653.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics, 79*, 396–403.

Vroomen, J. (2010). Causal inference in audiovisual speech. Comment on "Crossmodal influences on visual perception" by L. Shams. *Physics of Life Reviews, 7*, 289–290. discussion 295-288.

Whalen, D. H., Chen, W. R., Tiede, M. K., & Nam, H. (2018). Variability of articulator positions and formants across nine English vowels. *Journal of Phonetics, 68*, 1–14.