BRILL *Multisensory Research (2017)* DOI:10.1163/22134808-00002586 brill.com/msr

Reducing Playback Rate of Audiovisual Speech Leads to a Surprising Decrease in the McGurk Effect

John F. Magnotti [1,*,**], **Debshila Basu Mallick** [2,**] **and Michael S. Beauchamp** [1]

[1] Department of Neurosurgery and Core for Advanced MRI, Baylor College of Medicine, Houston, TX, USA

[2] Department of Psychology, Rice University, Houston, TX, USA

Received 10 November 2016; accepted 3 June 2017

Abstract

We report the unexpected finding that slowing video playback decreases perception of the McGurk effect. This reduction is counter-intuitive because the illusion depends on visual speech influencing the perception of auditory speech, and slowing speech should increase the amount of visual information available to observers. We recorded perceptual data from 110 subjects viewing audiovisual syllables (either McGurk or congruent control stimuli) played back at one of three rates: the rate used by the talker during recording (the natural rate), a slow rate (50% of natural), or a fast rate (200% of natural). We replicated previous studies showing dramatic variability in McGurk susceptibility at the natural rate, ranging from 0–100% across subjects and from 26–76% across the eight McGurk stimuli tested. Relative to the natural rate, slowed playback reduced the frequency of McGurk responses by 11% (79% of subjects showed a reduction) and reduced congruent accuracy by 3% (25% of subjects showed a reduction). Fast playback rate had little effect on McGurk responses or congruent accuracy. To determine whether our results are consistent with Bayesian integration, we constructed a Bayes-optimal model that incorporated two assumptions: individuals combine auditory and visual information according to their reliability, and changing playback rate affects sensory reliability. The model reproduced both our findings of large individual differences and the playback rate effect. This work illustrates that surprises remain in the McGurk effect and that Bayesian integration provides a useful framework for understanding audiovisual speech perception.

Keywords

McGurk effect, playback rate, audiovisual speech, multisensory integration, individual differences

* To whom correspondence should be addressed. E-mail: magnotti@bcm.edu

** These authors contributed equally to the manuscript

 DOI:10.1163/22134808-00002586

1. Introduction

The McGurk effect is a striking demonstration of visual influence on auditory speech perception in which pairing incongruent auditory and visual syllables create a different, fusion percept (e.g., auditory 'ba' + visual 'ga' = 'da'; McGurk and MacDonald, 1976). Studies have demonstrated large inter-participant (Basu Mallick *et al*., 2015; Strand *et al*., 2014), and inter-stimulus differences in perception of this effect (Basu Mallick *et al*., 2015; Jiang and Bernstein, 2011; MacDonald and McGurk, 1978; Magnotti and Beauchamp, 2015). Some subjects almost never perceive the effect and some almost always do, regardless of the stimulus being viewed; similarly, there are consistent differences across stimuli, with some stimuli more effective at evoking the illusion.

Talkers vary widely in their natural speech rate, and these rate differences could contribute to efficacy differences across McGurk stimuli. To test this idea, the effect of talker speed could be examined with several possible experimental methods. McGurk stimuli created by talkers with naturally different talking speeds could be tested. However, this approach would confound talker speed with other inter-talker differences. Another method would be to record multiple stimuli from the same talker instructed to talk at different rates (Fixmer and Hawkins, 1998; Munhall *et al*., 1996). However, this approach confounds rate changes with other changes that occur when talkers consciously slow their speech, such as the exaggerated auditory and visual speech features produced during infant directed speech (Golinkoff *et al*., 2015). A third method, used in the present study, is to computationally manipulate playback rate. By slowing or speeding the playback rate, talker speed can be changed arbitrarily without changes in any other auditory or visual speech feature. Modern software tools allow for increasing and decreasing the playback rate of auditory stimuli while maintaining acoustic similarity with the natural rate.

To provide a theoretical basis for our behavioral observations, we turned to the framework of Bayesian inference, the most influential model in the field of multisensory integration (Ernst and Banks, 2002; Seilheimer *et al*., 2014). Bayesian models of multisensory integration have proven useful in understanding audiovisual speech perception under conditions in which the auditory and visual speech are either offset in time (i.e., temporal disparity; Magnotti *et al*., 2013) or contain incongruent speech features (i.e., content disparity; Magnotti and Beauchamp, 2017). Adopting a similar approach, we examined how well perception at different playback rates could be fit by a Bayesian model that assumed changes in sensory noise across playback rate.

2. Method

2.1. Participant Recruitment and Testing Environment

Participants ($n = 110$) were recruited, consented, tested, and compensated using the Amazon Mechanical Turk (MTurk) service, following a protocol approved by the Rice University Institutional Review Board. Amazon MTurk is a web-based platform that provides services for participant recruitment, testing, and payment. Registered users of Amazon MTurk browse a list of available tasks and view a description of the work and compensation before selecting a particular task to complete. All participants in our study received $3.00; median time of completion was 15 minutes. To validate this approach, in a previous study we compared McGurk perception tested in the laboratory and using Amazon MTurk and found similar results (Basu Mallick *et al*., 2015).

Prior to the experiment, participants answered demographic questions relating to gender: female ($n = 46$), male ($n = 60$); ages within a range: 18–25 ($n = 25$), 26–35 ($n = 56$), 36–45 ($n = 19$); proficiency with the English language: yes ($n = 107$), no ($n = 0$); English is first language: yes ($n = 98$), no ($n = 9$). Of 110 participants, three did not answer any questions; one additional participant did not answer the gender question.

2.2. Stimuli

The stimulus set consisted of eight different McGurk syllables (auditory 'ba' + visual 'ga' recorded from eight different talkers) and three congruent audiovisual syllables (audiovisual 'ba', 'da', and, 'ga') recorded from a separate talker. The McGurk stimuli were the same as the ones used in experiment 2 of Basu Mallick *et al*. (2015). These stimuli were originally created by dubbing an auditory 'ba' over the video of a congruent visual 'ga' for each of the eight talkers.

The stimuli were sped up and slowed down in real-time by modifying the JavaScript *playbackRate* value. The syllables were played back at three different rates: natural, slow and fast. The slow rate was set to 50% natural speed by adjusting the video playback rate (JavaScript *playbackRate* attribute set to 0.5); the fast rate was set to 200% natural speed (*playbackRate* = 2.0). Stimuli at each playback rate may be viewed at: mcgurkplaybackrate.herokuapp.com.

Changing the playback rate of the stimuli also modifies the asynchrony between the mouth movements and the voice. To ensure the playback rate manipulation was not merely enhancing talker-specific differences in natural mouth movement/voice asynchrony, we calculated a measure of asynchrony for each talker in the natural rate condition. We measured asynchrony using a time-lag correlation method: the time lag that produced the largest Pearson correlation between visual change in the movie and the auditory volume envelope was used as the mouth/voice asynchrony.

2.3. Procedure

After completing the demographic questionnaire, participants viewed a congruent syllable video and were asked to adjust the size and volume so that they could both see and hear the talker. The participants were giving the instructions: "You will see videos and hear audio clips of a person saying syllables. Please watch the screen at all times. After each video, press a button to indicate what the person said. If you are not sure, take your best guess."

To reduce experimental time, only four of the eight McGurk syllables were viewed in a single group of participants. In the first group ($n = 60$), subjects viewed 10 repetitions of each of four McGurk syllables at each of the three playback rates, plus two repetitions of each of the three congruent syllables at each playback rate, for a total of 138 trials. In the second group ($n = 58$), subjects viewed 10 repetitions at each playback rate of the other four McGurk syllables and two repetitions of the three congruent syllables at each playback rate (138 total trials). Each set of stimuli were randomized across talkers and playback rates. The same randomized set of stimuli was presented to all participants. Eight subjects were members of both groups (total unique $n = 110$). Two participants completed the experiment twice and we excluded their second set of results (leading to $n = 58$ for group 2).

2.4. Responses and Scoring

Following the presentation of each audiovisual syllable, subjects clicked a button to indicate their percept. A forced-choice response format with three choices was used ('ba', 'da', 'ga'). For McGurk (auditory 'ba' + visual 'ga') syllables, 'da' responses were scored as McGurk responses, 'ba' as auditory, and 'ga' as visual. On congruent trials, responses were scored as either correct or incorrect.

2.5. Subject Grouping Based on Unisensory Response Type

After plotting individual subject response data for the McGurk stimuli, we noticed that two groups of participants could be discerned — those that made predominantly *visual* responses to McGurk stimuli and those that made predominantly *auditory* responses. We did not expect such a grouping, as we and others find visual responses to McGurk stimuli to occur infrequently (Strand *et al.*, 2014). To understand if this subject-level difference could explain some of the variability in our results, we classified each subject into VIS and AUD groups, based on whether they had more visual or auditory responses to McGurk stimuli.

2.6. Data Analysis

We used a linear mixed effects (LME) analysis (R package lme4; Bates *et al.*, 2015) to assess changes in McGurk perception across playback rate. Playback

rate (categorical levels: natural, slow, fast) was treated as a fixed factor; subject and subject-by-stimulus interactions were treated as random effects. To assess how the playback rate effect varied across stimuli, we also included stimulus and stimulus-by-playback rate interactions as fixed effects (including stimulus as a fixed effect in addition to its interaction with playback rate makes the interaction coefficients easier to interpret, but does not affect the overall fit of the model). The dependent measure in the LME was the average McGurk response for each stimulus at each playback rate for each subject. The LME analysis provides an effect estimate (in units of % McGurk response), its standard error, and an estimated p-value for each categorical level of the fixed factors and their interactions. The estimated p-values were obtained using the R Package lmerTest (Kuznetsova *et al*., 2016). Because all variables are categorical, we chose the baseline condition to correspond to stimulus 1 at the natural playback rate.

To assess the grouping of subjects into AUD and VIS, we created a new LME model that added group (AUD *vs*. VIS) and group-by-playback rate interactions as fixed effects. We also considered adding group-by-stimulus interactions, but such a model did not increase the model fit sufficiently to offset the increase in degrees of freedom, as judged by Bayesian Information Criterion (BIC), which considers model fit and number of parameters (BIC increased by 98 for the model with more parameters). Because the original model already included subject-by-stimulus interactions, group-by-stimulus interactions were not necessary.

To assess the effect of playback rate on congruent syllable perception, we calculated a single accuracy score per playback rate for each subject by combining responses to the three congruent syllables ‘ba’, ‘da’, and ‘ga’. We used an LME analysis with accuracy as dependent measure, playback rate as a fixed effect, and subject as a random effect. The baseline condition corresponded to estimated accuracy in the natural playback rate condition.

2.7. Bayes-Optimal Inference Model: Construction

We created a one-dimensional Bayes-optimal model of speech perception to test if the large individual differences in perception across playback rates were consistent with subjects following an optimal integration rule (Bejjanki *et al*., 2011; Ma *et al*., 2009). In our model, multisensory speech syllables are represented as a point along a single axis. This axis represents a low-dimensional projection of the high-dimensional audiovisual word space (Ma *et al*., 2009). We placed the syllables along this axis such that ‘da’ is located halfway between ‘ba’ and ‘ga’. The actual values used (we picked 0, 5, and 10) are not important except to set the relative scale. The relative distance is critical for obtaining accurate fits. Placing ‘da’ intermediate to ‘ba’ and ‘ga’ accords

with both their visual feature information (e.g., place of articulation) and auditory feature information (e.g., the second formant transition) and has been used by other models of the McGurk effect (Magnotti and Beauchamp, 2017; Olasagasti *et al*., 2015). The location of the syllables is fixed across playback rates, estimating changes in prototype location caused by changing playback rate is not possible because of the low number of conditions tested.

On each trial, the model assumes subjects encode the auditory and visual speech cues with unbiased noise and that the sensory noise follows a Gaussian distribution (Andersen, 2015; Bejjanki *et al*., 2011; Ma *et al*., 2009) with zero mean, and variance fitted per subject and condition: $X_{\mathrm{A}} \sim N(\mu_{\mathrm{ba}}, \sigma_{\mathrm{A}}^2)$; $X_{\mathrm{V}} \sim N(\mu_{\mathrm{ga}}, \sigma_{\mathrm{V}}^2)$. After encoding the separate cues, the unisensory representations are integrated according to Bayes' rule to produce a multisensory representation: $x_{\mathrm{AV}} = \sigma_{\mathrm{AV}}^2(\frac{x_{\mathrm{A}}}{\sigma_{\mathrm{A}}^2} + \frac{x_{\mathrm{V}}}{\sigma_{\mathrm{V}}^2})$; $\sigma_{\mathrm{A}V}^2 = (\frac{1}{\sigma_{\mathrm{A}}^2} + \frac{1}{\sigma_{\mathrm{V}}^2})^{-1}$. This final representation is a linear combination of each unisensory cue, with the weights given to each cue determined by their relative variance. The multisensory representation is then classified using a linear decision rule to determine the syllable most likely to have generated the encoded multisensory representation. Because the linear combination of Gaussians results in another Gaussian: $X_{\mathrm{AV}} \sim N(\mu_{\mathrm{AV}}, \sigma_{\mathrm{AV}}^2)$; $\mu_{\mathrm{AV}} = \sigma_{\mathrm{AV}}^2(\frac{\mu_{\mathrm{ba}}}{\sigma_{\mathrm{A}}^2} + \frac{\mu_{\mathrm{ga}}}{\sigma_{\mathrm{V}}^2})$, predictions from the model (proportion of 'ba', 'da', and 'ga' responses for each playback rate) can be calculated analytically using the area under the Gaussian curve, rather than via simulation.

In the current model, sensory noise represents the trial-to-trial variability in the neural response to a fixed stimulus that leads to perceptual variability and contributes to response variability. Sensory noise is thus distinct from changes in a response criterion or a guessing strategy, but rather controls the relative weighting of the auditory and visual cues in the calculation of the location of the multisensory representation.

2.8. Bayes-Optimal Inference Model: Fitting

The Bayes-optimal model was instantiated in R (R Core Team, 2016). Best-fitting variance parameters were found for each subject data using an exhaustive search method. Source code for model fitting is available at the authors' public website: openwetware.org/wiki/Beauchamp:DataSharing.

To set a scale for the representational space, we fixed syllable locations at three specific points, such that 'da' was midway between 'ba' and 'ga' (we used the values 0, 5, and 10). The same representational space was used for each subject and each playback rate; we did not optimize the locations to the current data.

For our one-dimensional model, we needed two variance parameters per playback rate: one for the auditory modality and one for the visual modality.

Because the final multisensory representation is determined only by the relative sensory noise in each modality and we did not independently manipulate auditory or visual sensory noise, we fixed the auditory variance at 0.5 and fit a single parameter that adjusted the visual variance, reducing the number of free parameters to 1 per subject per condition. In total, we fit three free parameters per subject to explain responses in the three categories ('ba', 'da', and 'ga') across the three playback rates (natural, slow, and fast).

Because the parameter search space for the visual sensory noise was small (a parameter range from 0.02 to 12.2 was adequate for the range of possible behavioral data), we used an exhaustive search method rather than a stochastic optimization technique. In our method, we pre-calculated the predicted responses rates for a range of sensory noise ratios (determined by ex, where x took on 100 000 values evenly spaced from -4 to 2.5, inclusive) and then matched subject response rates to these pre-calculated values. This approach allows us to find an optimal parameter value without multiple restarts or other techniques designed to avoid fitting to local optima.

2.9. Bayes-Optimal Inference Model: Validation

The goal of the current model is to understand whether changes in McGurk perception caused by changes to playback rate are consistent with an optimal-integration model. The standard method for determining if behavior is Bayes-optimal is to compare multisensory behavior with predictions obtained by fitting to unisensory data (Ernst and Banks, 2002). Because of differences across McGurk stimuli, this validation method requires obtaining auditory-only and visual-only recognition performance for each playback rate for each stimulus for each subject: a prohibitive amount of trials for our on-line data collection method that is optimized for many subjects rather than many trials. We consider total fit error as measure of model suitability, noting that without unisensory data other suboptimal integration models will explain the data equally well. In summary, a small model error suggests that optimal integration is a plausible explanation, but does not rule out other models.

3. Results

Changing the playback rate had a pronounced effect on frequency of the McGurk effect (Fig. 1A). At the natural playback rate, McGurk responses averaged 53% $\pm$ 3% (standard error of the mean, SEM) which decreased to 42% $\pm$ 2% with slow playback, but remained unchanged for fast playback (53% $\pm$ 3%). We used a linear mixed effects (LME) model to estimate how playback rate affected McGurk response, with fixed factors of playback rate, stimulus, and their interactions; subject and subject-by-stimulus interactions

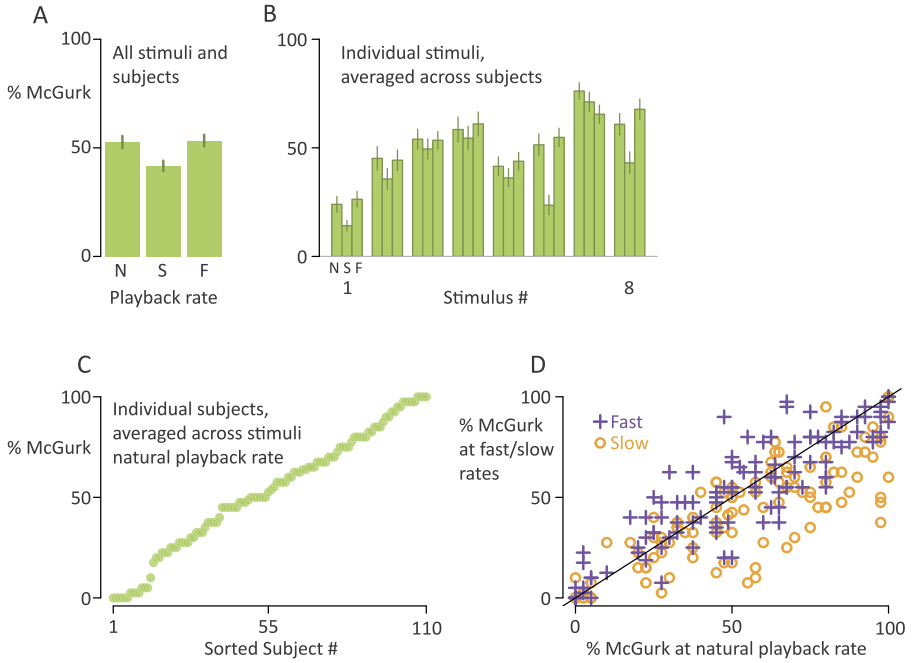

Figure 1. Effect of playback rate on perception of McGurk stimuli. (A) Grand mean McGurk response percentages at each playback rate (N: Natural, S: Slow, F: Fast) averaged across all presentations of all stimuli across all subjects. Error bars are standard error of the mean (SEM) across subjects. (B) Mean McGurk response percentages at each playback rate for each of the eight stimuli. (C) Mean response percentages (averaged across stimuli) for individual subjects at the natural playback rate. (D) Relationship between mean McGurk response percentages in the natural playback rate (x-axis) and the fast (y-axis, purple $+$) or the slow (y-axis, orange open circles) playback rates. The black solid line indicates equal McGurk perception between the natural playback rate and the fast/slow rate.

were random factors. The estimated effect of slow playback was $-10\% \pm 3\%$, $p = 0.004$; compared with $2\% \pm 3\%$, $p = 0.49$ for fast playback.

For congruent syllables, there was little effect of playback rate, with high accuracy across rates (natural, mean $= 95\% \pm 1\%$ standard error of the mean, SEM; slow $= 92\% \pm 1\%$; fast $= 97\% \pm 1\%$). An LME model on congruent stimuli estimated the change in accuracy for slow playback at $-3\% \pm 1\%$, $p = 0.006$; no change for fast playback (estimate $= 1\% \pm 1$, $p = 0.23$). Control analyses (described in detail below) suggest that the change in congruent accuracy was not strongly related to the change in McGurk responses.

Consistent with previous studies, there was substantial variability in McGurk responses at both the stimulus level and the subject level. Even at the natural playback rate, individual stimuli varied in their ability to evoke the McGurk effect, ranging from 24% for the weakest stimulus to 76% for the

strongest stimulus. Looking across stimuli, the LME estimated that slowed playback reduced McGurk frequency all eight stimuli (considering main effects of stimulus and playback rate, and stimulus by playback rate interactions, collapsing across subjects; estimate range from -4% to -28%), consistent with the aggregated behavioral data (Fig. 1B). Although the overall estimated effect for fast playback was small (and positive), the model did estimate a significant stimulus-by-playback rate interaction for stimulus 7 in the fast playback rate (estimate $= -13\%$, $p = 0.006$). For individual subjects, the frequency of the McGurk effect, averaged across stimuli, ranged from 0% in the least susceptible individual to 100% in the most susceptible individual (Fig. 1C). Slow playback rate evoked less or equal McGurk responses in 79% (87/110) of subjects (Fig. 1D).

A related question is the percept reported by subjects when they do not report the McGurk effect. Across rates, each individual subject tended to make either auditory or visual responses when they did not the report the illusion, rather than a mixture (Fig. 2A). Guided by this observation, we split participants into a group of subjects with mainly visual responses on trials on which they did not perceive the illusion (VIS group; $n = 58$) and another group with mainly auditory responses (AUD group; $n = 52$). This division accounted for 94% of all responses (i.e., on only 6% of McGurk trials did a VIS subject report an auditory percept or vice versa).

To assess the suitability of this visually prominent distinction, we expanded our LME model with additional fixed factors of group (AUD *vs*. VIS) and group-by-playback rate interactions. The model yielded a main effect of group (effect estimate $= 27\% \pm 5\%$, $p = 10^{-7}$), as McGurk susceptibility was higher for the VIS subjects than the AUD subjects (61% *vs*. 37%), in part reflecting the broader range of susceptibilities in AUD subjects (0% to 99% for AUD *vs*. 19% to 100% for VIS; averaged across stimuli and playback rates; Fig. 2B). The model did not show an overall effect of slow (estimate $= -4\% \pm 3\%$, $p = 0.24$) or fast playback rate on McGurk responses (estimate $= 1\% \pm 3\%$, $p = 0.67$), but a more complex picture emerged when considering group-by-playback interactions and stimulus-by-playback interactions. The model showed a significant interaction with slow playback rate for the VIS group (estimate $= -12\% \pm 2\%$, $p = 10^{-6}$) but no interaction with fast playback (estimate $= 2\% \pm 2\%$, $p = 0.46$), indicating the effect of slow playback was stronger in the VIS group than AUD group and that the effect of fast playback was similar. The model also showed significant stimulus-by-playback rate interactions for stimulus 6 in the slow condition (estimate $= -17\% \pm 5\%$, $p = 0.0002$) and for stimulus 7 in the fast condition (estimate $= -13\% \pm 5\%$, $p = 0.006$). Although complex, the results from the LME analysis suggest the reduction in McGurk responses was confined to the slow playback

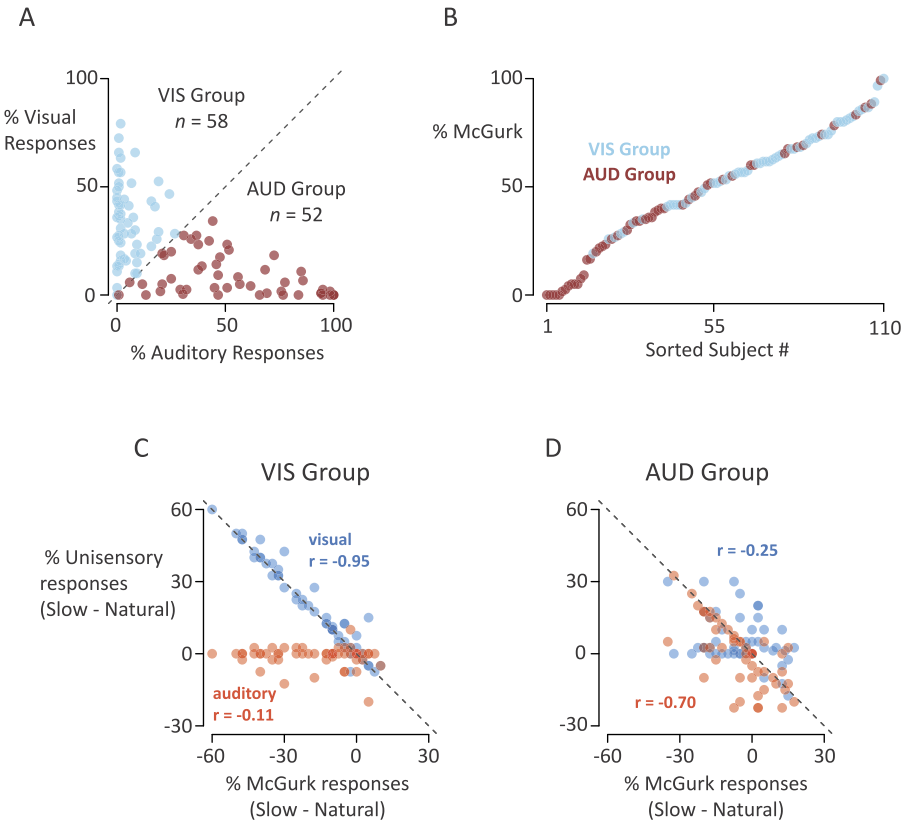

Figure 2. (A) Mean auditory *vs*. visual response percentages across all playback rates. Each point represents a single subject. Most points lie close to an axis, indicating a preference for either auditory or visual responses, rather than along the line of equality (dashed line). Subjects are classified into the VIS Group if they have equal or more visual than auditory responses (above the gray line; $n = 58$) or into the AUD Group if they have less visual responses than auditory responses (below the gray line; $n = 52$). (B) Mean McGurk response percentages for individual subjects, averaged across stimuli and playback rates. VIS Group is plotted as blue circles. AUD Group is plotted as brown circles. (C) Relationship between McGurk and unisensory responding. Slowing the playback rate resulted in fewer McGurk responses, shown as negative values along the x-axis, and a corresponding increase in unisensory responses, shown as positive values along the y-axis. For each subject, two symbols are plotted: blue symbols show the visual response percentages and red symbols show the auditory response percentages. For VIS subjects, visual responses increased strongly in tandem with McGurk response decreases (Spearman correlation $r = -0.95$) while auditory responses did not ($r = -0.11$). (D) For AUD subjects, auditory responses increased as McGurk responses decreased ($r = -0.70$) while visual responses did not ($r = -0.25$).

rate (except for stimulus 7), and was larger for subjects in the VIS group, perhaps caused by a floor effect in the AUD group, which had substantially lower overall McGurk responses.

Because the LME analysis focused solely on the change in McGurk responses, we conducted a complementary analysis to show that the change in McGurk responses corresponded to an increase in a preferred responses modality. We sorted subjects in each group by how much the slow playback rate influenced their McGurk perception, and compared this value with their change in auditory and visual responses at slow playback rate *vs*. natural rate (Fig. 2C and 2D). Subjects in the VIS group who experienced less McGurk percepts at slow rates increased their report of visual percepts (Spearman rank correlation, $r = -0.95$) more than their auditory percepts ($r = -0.11$; comparison of correlations using Fisher r-to-z transformation: $z = 8.8$, $p = 10^{-18}$). Conversely, in AUD subjects, decreases in McGurk frequency at slow playback rates corresponded more strongly to increases in the frequency of auditory percepts ($r = -0.70$) than visual percepts ($r = -0.25$; $z = 3.0$, $p = 0.002$).

3.1. Control Analyses

3.1.1. Effect of Non-Perceivers

In the primary analysis, we retained subjects that perceived 0% McGurk in the natural playback rate ($n = 13$) because our initial prediction was the slowed playback rate would *increase* rates of the McGurk Effect. We ran an additional *post hoc* LME analysis without these subjects; the main effects were unchanged (reduction of McGurk responses for slow playback, estimate = 11%, $p = 0.003$; no change for fast playback, estimate = 3%, $p = 0.49$), and similar stimulus-by-playback rate interactions were observed.

3.1.2. Effect of Congruent Trial Performance

We found that slowed playback caused a numerical reduction in congruent accuracy for 25% of subjects. To assess if this reduction in accuracy was related to the change in McGurk effect, we first correlated the change in congruent accuracy and the change in McGurk perception, which yielded no significant correlation, $r = -0.01$, $p = 0.99$. Second, we re-ran the LME analysis only on the subjects that had at least 83% accuracy (corresponding to getting only one trial incorrect) for the congruent stimuli in each condition ($n = 87$). This analysis produced the same qualitative result, with a negative effect of slowed playback (estimate = -9.4%, $p = 0.01$), no effect of fast playback (estimate = 0.4%, $p = 0.91$), and stimulus-by-playback rate interactions showing varying effect estimate magnitudes across stimuli.

We also considered congruent trial performance separately for AUD and VIS subjects. Overall accuracy was similar across playback rates for

AUD subjects (mean and standard error for natural rate: 93% ± 2%; slow: 96% ± 1%; fast: 97% ± 1%) and VIS subjects (natural: 91% ± 2%; slow: 95% ± 1%; fast: 97% ± 1%). An LME on congruent accuracy with subject as random factor and fixed factors of playback rate (natural rate was baseline), subject group (AUD group was baseline), and their interactions (natural rate for AUD group was baseline) yielded no significant effects for the VIS group parameter (estimate = −3% ± 2%, $p = 0.16$) or any VIS group by playback rate interactions (VIS-slow: estimate = 1% ± 2%, $p = 0.63$; VIS-fast: estimate = 3% ± 2%, $p = 0.27$). No large effects for congruent accuracy were found for slow playback (estimate = 3% ± 2%, $p = 0.12$) or fast playback (estimate = 3% ± 2%, $p = 0.06$) relative to natural playback.

3.1.3. Effect of Audiovisual Asynchrony

Although we experimentally manipulated the playback rate of the stimuli, this manipulation also affects the onset asynchrony between the mouth movements and the speech sounds. To assess if the playback rate effect was driven largely by a change in asynchrony, we correlated the measured asynchrony in each stimulus (which varied across stimuli due to talker idiosyncrasies) with the change in McGurk response (from natural to slow). We did not find a strong linear relationship between asynchrony and the effect of slowed playback (Pearson correlation, $r = -0.01$, $p = 0.99$).

3.1.4. Effect of Voice Onset Time

We also considered natural variation across stimuli in voice onset time (VOT). We found no correlation between VOT and overall McGurk effect ($r = -0.26$, $p = 0.54$) nor between VOT and the change in McGurk responses from natural to slowed playback ($r = -0.12$, $p = 0.78$), suggesting that VOT is not a major contributor to overall McGurk perception or the playback rate differences in the McGurk effect for these stimuli.

3.2. Bayes-Optimal Inference Model

A striking feature of our data was the presence of large individual differences observed at multiple levels: how often each subject perceived the illusion, what subjects perceived when they did not perceive the illusion, and how much playback rate changed their perception. Current thinking suggests that Bayes-optimal integration is a common feature of multisensory integration, but can these individual differences be explained using Bayesian inference?

To answer this question, we constructed a Bayes-optimal integration model and fit it to our data. The model assumes that subjects vary in how precisely they encode auditory and visual speech at each playback rate. Figure 3 illustrates the application of the model to one subject from the VIS Group and one subject from the AUD Group. The model uses three steps to explain how

McGurk stimuli are perceived. First, the individual modalities are encoded with unbiased Gaussian sensory noise. For the VIS Group Subject (Fig. 3A), the visual modality has *lower* sensory noise than the auditory modality; for the AUD Group Subject (Fig. 3D), the visual modality has *higher* sensory noise than the auditory modality. Second, the encoded unisensory representations are integrated according to their relative sensory noise (higher weight given to the cue with less sensory noise) to produce a multisensory representation. For the VIS Group Subject, the multisensory representation is closer to the unisensory visual representation; for the AUD Group Subject, the multisensory representation is closer to the unisensory auditory representation. Finally, the multisensory representation is categorized as either 'auditory', 'McGurk' or 'visual' based on its location in the representational space. Across many trials, the multisensory representations will have a Gaussian distribution, located closer to the more reliable modality (the modality with less sensory noise) and with a variance lower than the variance of either modality alone. We can use the location and variance of the multisensory representations to directly calculate the probability of a given response option across many trials — the model's prediction of that subject's mean response proportions across trials. For the VIS Group subject, the multisensory representations at the natural playback rate are located on the border of the visual and McGurk regions, producing a mixture of visual and McGurk responses; for the AUD Group subject, the multisensory representations are located on the border of the auditory and McGurk regions, producing a mixture of auditory and McGurk responses.

The model assumes that changing the playback rate changes the relative precision of the visual and auditory encoding with only a single parameter fit to the ratio (for simplicity, Fig. 3 figure shows the model with a fixed auditory precision and variable visual precision). For the VIS Group Subject, the slow playback rate led to reduced visual sensory noise, leading to multisensory representations closer to the unisensory visual representation, less frequent McGurk percepts, and more frequent visual percepts (Fig. 3B, 3C). For the AUD Group Subject, the slow playback rate led to increased visual sensory noise (equivalent to decreased auditory sensory noise because the two are modeled as a ratio), multisensory representations that were closer to the unisensory auditory representation, less frequent McGurk percepts, and more frequent auditory percepts (Fig. 3E, 3F). Fast playback rates lead to only minor changes in sensory noise and little difference in predicted percepts for both subjects.

Across all subjects, the model performed well at describing behavior, with mean prediction error (average root mean squared error across conditions and subjects) of 4.0% $\pm$ 5.2% (standard deviation across subjects). The best-fitting parameters for all subjects produced a two-response solution: 'ba' and 'da' for

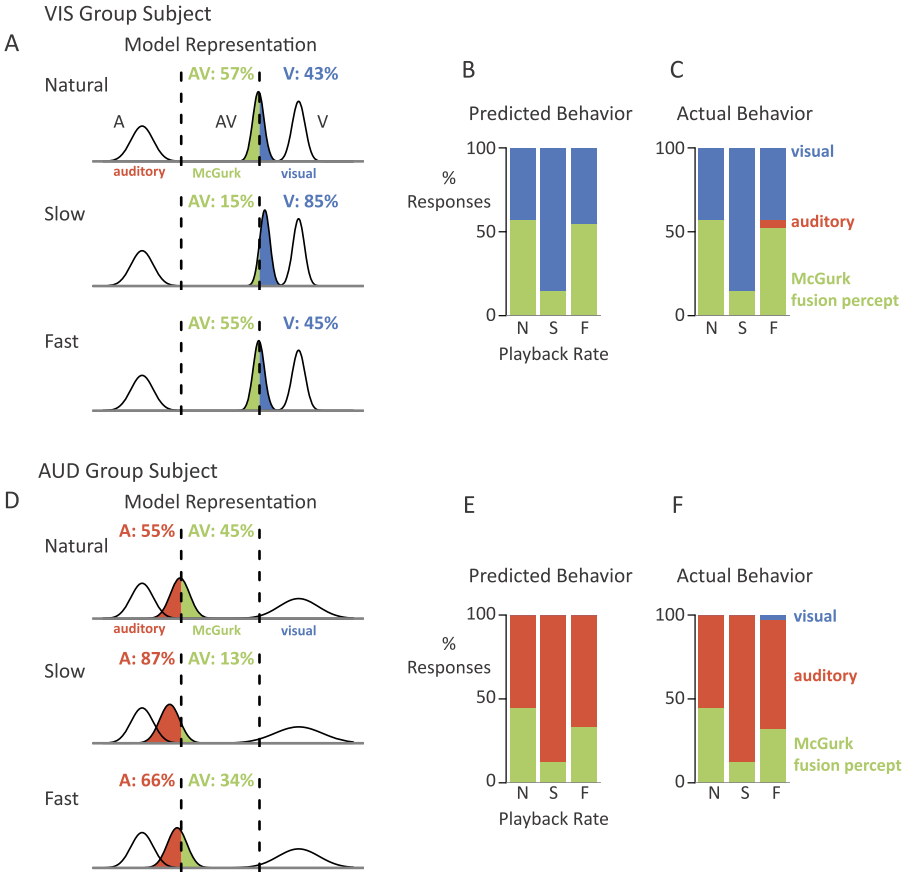

Figure 3. Bayes-optimal model of playback rate. (A) Model representation across playback rates. The model assumes the three response categories (auditory, McGurk, and visual) can be represented as points along a single axis. When presented with a McGurk stimulus, subjects encode the auditory and visual cues with noise (Gaussians labelled A and V, centered over auditory and visual, respectively). The amount of sensory noise determines the variance of the Gaussian. For a subject from the VIS Group, the sensory noise is lower (narrower Gaussian) for the visual modality than for the auditory modality. After integrating the unisensory cues, the distribution of multisensory representations (Gaussian labeled AV) is closer to the visual syllable than to the auditory syllable because the cues are weighted by their relative variance. The area under the AV Gaussian within each response region (separated by the vertical dashed lines) determines the predicted response proportions (percentages labelled with same color as response option). In the slow playback rate, the sensory noise for the visual modality is reduced, pulling the AV representation even more into the visual region than in the natural playback rate. In the fast playback rate, the sensory noise for the visual modality is similar to the natural playback rate, leading to a similar location for the AV representations and similar predicted percepts. (B) Model predictions of response percentages for each playback rate. For the natural playback rate, the VIS Group Subject is predicted to have a mixture of McGurk (green) and visual responses (blue). For the

AUD subjects or 'da' and 'ga' for VIS subjects. The prediction error reflects the small number of trials in which subjects chose a third response ('ga' for AUD subjects or 'ba' for VIS subjects).

4. Discussion

Our results replicate previous studies showing dramatic individual differences in the McGurk effect: some subjects never perceived the illusion while others always did, and different McGurk stimuli showed large variations in their ability to elicit the illusion (Basu Mallick *et al*., 2015). In the present study, we found an unexpected relationship between playback rate and McGurk perception. Slow playback rates reduced the frequency of the McGurk effect while leaving the perception of congruent syllables largely unaffected. 79% of subjects showed a reduction in McGurk responses and these subjects could be classified into two groups. On trials in which they did not perceive the McGurk effect, VIS subjects reported primarily visual percepts and AUD reported primarily auditory percepts.

To provide an explanatory framework for these results, we applied a Bayes-optimal integration model of speech perception. The model assumes that sensory noise varies across playback rates, and that the sensory noise in each modality determines the weight of that modality in the final percept. Fitting the model allowed us to estimate the relative sensory noise in each condition for each subject. For VIS subjects, slow playback rate decreased the relative sensory noise in the *visual* modality, pulling the integrated representation into the visual region of perceptual space, resulting in more visual responses and fewer McGurk percepts. Conversely, for AUD subjects, slow playback rate decreased the relative sensory noise in the *auditory* modality, pulling the integrated representations into the auditory region of perceptual space, resulting in more auditory responses and fewer McGurk percepts. The McGurk effect demonstrates the influence of visual mouth movements on speech perception. Therefore, slow playback rates might be expected to make speechreading

slow playback rate, there is still a mixture, but visual responses have increased and McGurk responses decreased. For the fast playback rate, the predictions are largely unchanged from the natural condition. (C) Actual subject responses for each playback rate. The subject behavior closely matches the model predictions in (B). (D) For the AUD Group Subject, the sensory noise for the auditory modality is lower than for the visual modality. This difference causes the AV representations to be pulled closer to the auditory syllable than to the visual syllable. In the slow playback rate, the sensory noise for the auditory modality is reduced, leading to more auditory responses. In the fast playback rate, the sensory noise is shows a small change from the natural playback rate, leading to small changes in response levels. (E) The model predicts a mixture of auditory and McGurk responses across playback rates, with the fewest McGurk responses for the slow playback rate. (F) The subject's response pattern matches the model predictions (E).

(decoding of visual speech information) more accurate, leading to increased frequency of the effect. Our model suggests that speechreading was in fact more accurate for subjects in the VIS group, but that if the brain applies Bayesian inference, this sensory noise decrease leads to more visual responses rather than more McGurk responses. Bayes-optimal integration has been repeatedly shown to be a key feature of many multisensory behaviors (Alais and Burr, 2004; Angelaki *et al*., 2009; Ernst and Banks, 2002; Ma *et al*., 2006; Seilheimer *et al*., 2014). The present finding adds to previous work showing that Bayesian perceptual models can explain how human speech perception is modified by different kinds of disparity between the auditory and visual speech streams, including temporal asynchrony (Magnotti *et al*., 2013), content disparity (Magnotti and Beauchamp, 2017), and are flexible enough to manage interindividual, interstimulus, and intergroup differences (Magnotti and Beauchamp, 2015; Stropahl *et al*., 2015).

4.1. Sources of Individual Differences

A natural question is the source of the individual variation that leads VIS subjects to report visual percepts and AUD subjects to report auditory percepts when they do not perceive the McGurk effect. When viewing static faces, some subjects look more at the mouth of the talkers and others look more at the eyes of the talker (Mehoudar *et al*., 2014; Peterson and Eckstein, 2013). The same holds true for subjects viewing talking faces, with the additional result that subjects who preferentially fixate the talker's mouth are more susceptible to the McGurk effect (Gürler *et al*., 2015). In the absence of eye tracking data in the present study, we speculate that AUD subjects correspond to subjects who mainly fixate the eyes, placing the talker's mouth in the visual periphery and lowering the quality of the visual speech information available to the subject. Conversely, VIS subjects mainly fixate the mouth, resulting in higher precision visual speech information. The playback rate manipulation enhanced the subject's dominant, or preferred modality, leading to increased visual percepts in the VIS group and increased auditory percepts in the AUD group.

Although grouping subjects based on their preferred non-McGurk response is not common in studies of the general population, studies with clinical groups have often reported such differences. For instance, individuals with cochlear implants report far more visual percepts than individuals with typical hearing for McGurk stimuli (Rouger *et al*., 2008; Stropahl *et al*., 2015). Such group differences are consistent with Bayesian integration, as more weight is given to the modality with the higher precision; for most subjects this will be the auditory percept (Basu Mallick *et al*., 2015; McGurk and MacDonald, 1976; Strand *et al*., 2014; van Wassenhove *et al*., 2007).

4.2. Role of Asynchrony

One possible explanation for our results is that slowing the playback rate created a noticeable mismatch between the visual mouth movements and the auditory speech sounds (as in a dubbed foreign movie) reducing integration and the McGurk effect. To test this idea, we measured the temporal asynchrony between the auditory and visual speech onsets in the slow playback condition, finding a mean of 184 ms (visual ahead of auditory). However, a previous study by van Wasenhove *et al*. (2007) reported no effect on McGurk perception for visual-leading asynchronies up to 267 ms. Additionally, slowing the playback rate in our study resulted in an increase in *visual* responses (from 19% to 31%) rather than the increase in *auditory* responses observed with large asynchronies (Magnotti *et al*., 2013). Finally, we found no significant correlation ($r = -0.01$) between a stimulus' asynchrony and the reduction in McGurk responses caused by slowed playback. Taken together, these findings suggest that the effect of playback rate is not attributable solely to changes in asynchrony.

4.3. Relationship to Previous Studies

Two previous studies have considered how speech rate impacts McGurk responses (Fixmer and Hawkins, 1998; Munhall *et al*., 1996). In these studies, speech rate was manipulated by instructing talkers to alter their speech rate by talking clearly, talking naturally, or talking quickly. Fixmer and Hawkins reported greater McGurk responses when talkers were instructed to talk clearly compared with talking quickly (70% *vs*. 59%; their Fig. 5). Munhall *et al*. also found greater McGurk responses when talkers were instructed to talk clearly, compared with talking naturally or talking quickly (48.2% *vs*. 44.7% *vs*. 36.1%; their Table 3). At first glance, these results seem discordant with the results of the present study, in which slowing speech *decreased* McGurk responses. However, instructing talkers to talk clearly can change a variety of speech properties in addition to speech rate. For instance, Munhall *et al*. show that when instructed to talk clearly, talkers open their mouth about 30% wider (1996; their Fig. 3). This contrasts with our stimuli, in which the physical properties of the stimulus (such as mouth aperture) are identical between conditions: only the timing of the changes differs.

4.4. Role of Speech Rate in Unisensory Perception

The model we developed fit the data well under the assumption that changing playback rate changed only the relative sensory noise in each unisensory modality, keeping category boundaries fixed. Are the data also consistent with playback rate affecting category boundaries? Port (1979) found that increased speech rate (asking people to speak more quickly) can shift the auditory-

only perceptual boundary between voiced /b/ and unvoiced /p/ (faster /b/ can be perceived as /p/) using words embedded in a fast/slow sentence context, consistent with other findings that speech rate can have strong effects on perceived voice onset time (Miller, 1981; Summerfield, 1981). In the current study, stimulus differences in voice onset time were not predictive of the playback rate effect, likely because we used only voiced syllables (ba, da, and ga) and our forced-choice response setup required subjects to choose amongst voiced syllables. Any changes in perceived voicing could not be assessed. As in the previous studies of speech rate on audiovisual speech perception, these auditory-only studies manipulated speech rate by asking individuals to speak more quickly or more slowly, which leads to changes in the acoustic signal beyond just its duration. Unlike these auditory-only studies, we found striking differences across subjects in how they responded to the playback rate manipulation — some perceiving more visual 'ga' and some perceiving more auditory 'ba'. Combined with the lack of relationship between congruent-trial performance and playback rate effect, we think it is unlikely the playback rate effect is causing a shift in category thresholds, although explicit category-threshold testing would be needed to definitively rule out this possibility.

4.5. Conclusions and Future Directions

Our study shows that, in isolation, slow speech rates decrease McGurk perception. The studies of Fixmer and Hawkins (1998) and Munhall *et al*. (1996) suggest that this effect can be countered by the exaggerated mouth movements and other speech modifications that occur when talkers consciously slow their speech. An important natural example of this is infant-directed speech. When adults talk to language learners, they slow their talking speed and exaggerate both visual and auditory speech features (Bortfeld *et al*., 2013). This exaggeration can lead to different behavior on the part of the perceiver. For instance, Lewkowicz and Hansen-Tift (2012) reported that infants look longer at the talker's mouth when viewing infant-directed speech but look longer at the talker's eyes when viewing adult-directed speech.

A counter-example of isolated changes in speech rate (without changes in any other speech features) is the ability to change the playback rate in online video services (e.g., YouTube) and computer-based learning environments. This ability is popular with content viewers because it allows them to dynamically calibrate the speed at which material is presented: slowing to allow for easier following of materials, or speeding to increase the amount of content viewers can watch in a fixed amount of time. Our results are relevant to this situation because the playback rates used for our stimuli are similar to those available in online environments. Our results suggest that if content viewers slow the content considerably, this may decrease some of the benefits

of multisensory integration for speech perception. Conversely, we did not find any impairment in multisensory integration for speeded playback (although of course comprehension and retention may be affected).

References

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration, *Curr. Biol.* **14**, 257–262.

Andersen, T. S. (2015). The early maximum likelihood estimation model of audiovisual integration in speech perception, *J. Acoust. Soc. Am.* **137**, 2884–2891.

Angelaki, D. E., Gu, Y. and DeAngelis, G. C. (2009). Multisensory integration: psychophysics, neurophysiology, and computation, *Curr. Opin. Neurobiol.* **19**, 452–458.

Basu Mallick, D., Magnotti, J. F. and Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type, *Psychonom. Bull. Rev.* **22**, 1299–1307.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4, *J. Stat. Softw.* **67**, 1. DOI:10.18637/jss.v067.i01.

Bejjanki, V. R., Clayards, M., Knill, D. C. and Aslin, R. N. (2011). Cue integration in categorical tasks: insights from audio–visual speech perception, *PLoS One* **6**(5), e19812.

Bortfeld, H., Shaw, K. and Depowski, N. (2013). The miracle year: from basic structure to social communication, in: *Theoretical and Computational Models of Word Learning: Trends in Psychology and Artificial Intelligence*, L. Gogate and G. Hollich (Eds), pp. 153–171. IGI Global, Hershey, PA, USA.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion, *Nature* **415**(6870), 429–433.

Fixmer, E. and Hawkins, S. (1998). The influence of quality of information on the McGurk effect, in: *Proceedings of the Auditory–Visual Speech Processing Conference*, Terrigal, Australia.

Golinkoff, R. M., Can, D. D., Soderstrom, M. and Hirsh-Pasek, K. (2015). (Baby) Talk to me the social context of infant-directed speech and its effects on early language acquisition, *Curr. Dir. Psychol. Sci.* **24**, 339–344.

Gürler, D., Doyle, N., Walker, E., Magnotti, J. and Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements, *Atten. Percept. Psychophys.* **77**, 1333–1341.

Jiang, J. and Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects, *J. Exp. Psychol. Hum. Percept. Perform.* **37**(4), 1193–1209.

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B. (2016). lmerTest: tests in linear mixed effects models. https://CRAN.R-project.org/package=lmerTest.

Lewkowicz, D. J. and Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech, *Proc. Natl Acad. Sci. USA* **109**, 1431–1436.

Ma, W. J., Beck, J. M., Latham, P. E. and Pouget, A. (2006). Bayesian inference with probabilistic population codes, *Nat. Neurosci.* **9**, 1432–1438.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J. and Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space, *PLoS One* **4**, e4638. DOI:10.1371/journal.pone.0004638.

MacDonald, J. and McGurk, H. (1978). Visual influences on speech perception processes, *Percept. Psychophys.* **24**, 253–257.

Magnotti, J. F. and Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect, *Psychonom. Bull. Rev.* **22**, 701–709.

Magnotti, J. F. and Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech, *PLoS Comput. Biol.* **13**, e1005229. DOI:10.1371/journal.pcbi.1005229.

Magnotti, J. F., Ma, W. J. and Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech, *Front. Psychol.* **4**, 798. DOI:10.3389/fpsyg.2013.00798.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**(5588), 746–748.

Mehoudar, E., Arizpe, J., Baker, C. I. and Yovel, G. (2014). Faces in the eye of the beholder: unique and stable eye scanning patterns of individual observers, *J. Vis.* **14**, 6. DOI:10.1167/14.7.6.

Miller, G. A. (1981). *Language and Speech*. W. H. Freeman and Co., San Francisco, CA, USA.

Munhall, K. G., Gribble, P., Sacco, L. and Ward, M. (1996). Temporal constraints on the McGurk effect, *Percept. Psychophys.* **58**, 351–362.

Olasagasti, I., Bouton, S. and Giraud, A. L. (2015). Prediction across sensory modalities: a neurocomputational model of the McGurk effect, *Cortex* **68**, 61–75.

Peterson, M. F. and Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation, *Psychol. Sci.* **24**, 1216–1225.

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place, *J. Phon.* **7**, 45–56.

R Core Team (2016). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rouger, J., Fraysse, B., Deguine, O. and Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects, *Brain Res.* **1188**, 87–99.

Seilheimer, R. L., Rosenberg, A. and Angelaki, D. E. (2014). Models and processes of multisensory cue combination, *Curr. Opin. Neurobiol.* **25**, 38–46.

Strand, J., Cooperman, A., Rowe, J. and Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: links with lipreading and detecting audiovisual incongruity, *J. Speech Lang. Hear. Res.* **57**, 2322–2331.

Stropahl, M., Schellhardt, S. and Debener, S. (2016). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: the Oldenburg Audio Visual Speech Stimuli (OLAVS), *Psychonom. Bull. Rev.* **24**, 863–872.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception, *J. Exp. Psychol. Hum. Percept. Perform.* **7**, 1074–1095.

van Wassenhove, V., Grant, K. W. and Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception, *Neuropsychologia* **45**, 598–607.