Linking cognitive and neural models of audiovisual processing to explore speech perception in autism

Grace C. Brown (grcbrown@stanford.edu)
Department of Linguistics, Stanford University, Stanford, CA 94305 USA

Naomi H. Feldman (nhf@umd.edu)
Department of Linguistics and UMIACS, University of Maryland, College Park, MD 20742 USA

Abstract

Autistic and neurotypical children do not handle audiovisual speech in the same manner. Current evidence suggests that this difference occurs at the level of cue combination. Here, we test whether differences in autistic and neurotypical audiovisual speech perception can be explained by a neural theory of sensory perception in autism, which proposes that heightened levels of neural excitation can account for sensory differences in autism. Through a linking hypothesis that integrates a standard probabilistic cognitive model of cue integration with representations of neural activity, we derive a model that can simulate audio-visual speech perception at a neural population level. Simulations of an audiovisual lexical identification task demonstrate that heightened levels of neural excitation at the level of cue combination cannot account for the observed differences in autistic and neurotypical children's audiovisual speech perception.

Keywords: speech perception; autism; Bayesian neural network; multisensory integration

Introduction

The brain instantiates the mind, and in principle, it should be possible to use theories from neuroscience to explain cognitive phenomena. One place where such a connection is likely to be fruitful is in understanding differences in audiovisual integration between autistic and neurotypical children. On the one hand, there is behavioral evidence that autistic and neurotypical children do not handle audiovisual input in the same manner during perception, due to differences in audiovisual integration (Foxe et al., 2015; Baum, Stevenson, & Wallace, 2015). On the other hand, there is a neural theory that attributes differences in autistic and neurotypical perceptual behavior to an excitatory/inhibitory (E/I) imbalance in the direction of hyper-excitability in autistic people (Rubenstein & Merzenich, 2003). Our study focuses on a specific extension of this theory, which predicts that E/I imbalance is implemented through divisive normalization, a canonical neural computation that inherently reflects an E/I balance by normalizing an excitatory drive of a neuron using the activity of the surrounding neural population (Rosenberg, Patterson, & Angelaki, 2015). It is not yet known whether the neural theory of E/I imbalance via divisive normalization can account for the observed differences in audiovisual integration at the behavioral level.

Testing whether Rosenberg et al.'s E/I imbalance theory can account for behavioral data on audiovisual integration is challenging because it requires a linking hypothesis between the cognitive and neural levels. At a cognitive level, there is a widely accepted mathematical model of how perceivers integrate auditory and visual cues with differing reliabilities (Landy, Maloney, Johnston, & Young, 1995; Jacobs, 1999; Ma, Zhou, Ross, Foxe, & Parra, 2009), but it is not obvious how and why representations of cue reliability would differ among autistic and neurotypical people. At the neural level, it is straightforward to characterize differences in neural activity that would result from differences in divisive normalization, but it is less clear how such changes in activity would impact the cognitive representations involved in audiovisual integration. Explaining audiovisual integration behavior based on neural activity requires us to integrate these different levels of analysis (Marr, 1982).

In this paper, we construct a model that bridges the gap between the neural theory of sensory processing in autism and higher-level probabilistic theories of cue combination. We integrate a cognitive model of cue reliability into a neural model of audiovisual integration, which allows us to simulate how perceivers' estimates of cue reliability are propagated through the neural model and how they are impacted by divisive normalization. We find that altered divisive normalization during multisensory integration predicts higher overall lexical identification accuracy relative to unaltered divisive normalization; this finding directly contradicts the behavioral observation that neurotypical children perform with greater accuracy than autistic children in an audiovisual lexical identification task. Our results call into question whether an Excitatory/Inhibitory imbalance of neural activity, implemented via divisive normalization, can robustly predict the perceptual behavior associated with autism.

Our paper is organized as follows. We first describe the data and theory that we aim to link, then detail the model components that form the basis for our linking hypothesis. The next section outlines our model and presents simulations of the lexical identification task in neurotypical and autistic children. We conclude by discussing the implications of our results and suggesting directions for future research.

Behavioral Data on Audiovisual Integration

Autistic and neurotypical differences in processing are evidenced by audiovisual speech perception tasks. Foxe et al.'s (2015) study of autistic and neurotypical children (ages 7-11) reveals that autistic children do not identify audiovisual speech in noise as accurately as neurotypical children. In this

study, they asked these children to identify common monosyllabic words in two different tasks: an audio-only task, where the children were exposed to single words in various levels of pink noise (ranging from no noise to -15 dBA sound pressure level), and an audiovisual task, where they received the same auditory stimuli with an added visual stimulus (i.e., a woman producing the auditory stimulus, presented as a video). Children's accuracy during the lexical identification task was measured as the percentage of auditory stimuli correctly reproduced. Foxe et al. (2015) found that autistic and neurotypical children did not perform significantly differently in the audio-only task. Both groups demonstrated significantly improved performance on the lexical identification task when visual stimuli were present. However, neurotypical children improved significantly more than autistic children (Figure 1).

Foxe et al. (2015) sought to determine if the difference in autistic and neurotypical children's performance on the audiovisual task was attributable to each group's respective attention to the face/mouth of the speaker. Using eye-tracking data obtained during the audiovisual trials, Foxe et al. (2015) compared the autistic and neurotypical groups' average fixations to the mouth of the speaker. They found that both neurotypes exhibited similar average fixations to the mouth throughout the audiovisual task. This similarity of attention, taken together with their similar audio-only performance, suggests that the differences in autistic and neurotypical children's performance on the audiovisual lexical identification task is attributable to differences in multisensory integration, rather than differences in unisensory processing.

We aim to capture the two main qualitative findings from Foxe et al. (2015): the relationship between accuracy and noise, and the higher accuracy of neurotypical children relative to autistic children in the audiovisual condition.

Excitatory/Inhibitory Imbalance in Autism

Our investigation is concerned with the explanatory power of the Excitatory/Inhibitory (E/I) Imbalance theory of autism, as characterized by Rosenberg et al. (2015). This theory posits that perceptual symptoms of autism (e.g., altered visual spatial suppression, sensory hypersensitivities) arise from an E/I imbalance, altered in the direction of hyper-excitable neural populations. According to Rosenberg et al. (2015), this E/I imbalance is caused by an alteration in the canonical neural computation known as divisive normalization,

$$R_i = \frac{E_i}{v + c\sum_j E_j} \tag{1}$$

Specifically, the context sensitivity, c, a neurobiological constant thought to scale the suppressive field, $\sum_j E_j$, is reduced in autism, resulting in a greater neural response, R_i.

We attempt to use this model to replicate the behavioral findings of Foxe et al. (2015), wherein neurotypical children perform audiovisual speech perception in noise with greater accuracy than neurotypical children. We find that a reduced context sensitivity, c, in divisive normalization produces the opposite results.

Components of a Linking Hypothesis

Our approach in connecting the E/I imbalance hypothesis to the behavioral data brings together two different previously proposed models for capturing multisensory integration. The first, at the neural level, incorporates divisive normalization, the key component of the E/I hypothesis (Ohshiro, Angelaki, & DeAngelis, 2011). The second, at the cognitive level, provides a normative account of how cues of different reliabilities should be integrated (Jacobs, 1999). We use probabilistic population codes (Ma, Beck, Latham, & Pouget, 2006; Beck, Latham, & Pouget, 2011) to create an integrated framework that includes both of these components.

Multisensory Integration via Divisive Normalization

To test the E/I imbalance theory in the context of multisensory integration, we need a model that can perform multisensory integration using divisive normalization. Ohshiro et al. (2011) provide us with a starting framework. Under their model of multisensory spatial integration, each unisensory input is denoted by its spatial position in Cartesian coordinates, $\theta = (x_\theta, y_\theta)$, and its intensity, d. The receptive field of each unisensory neuron T is a two-dimensional Gaussian, which we denote as $G_T(\theta) = N(\mu_T, \sigma_T^2)$. Here, $\mu_T = (x_T, y_T)$ represents the center location of the receptive field of neuron T. Each receptive field represents the stimulus preference (e.g., a specific auditory frequency) of a given neuron within a unisensory neural population. The response of each unisensory neuron is assumed, under the Ohshiro et al. (2011) model, to scale linearly with stimulus intensity, d:

$$d * G_T(\theta) \tag{2}$$

They assume that two inputs, presented at positions θ_{1a}, θ_{1b} and with respective intensities of d_{1a}, d_{1b}, interact linearly. Furthermore, Ohshiro et al. assume that each unisensory neural response is transformed by a sub-linearly increasing function, $h(x) = \sqrt{x}$. This transformation is intended to account for the sublinear intensity response functions typically observed in sensory neurons. Taken together, we get a weighted sum of unisensory inputs (Equation 3). This summation is performed by the multisensory neuron to give us its linear response, E:

$$E = h(d_{1a} * G_T(\theta_{1a})) + h(d_{1b} * G_T(\theta_{1b})) \tag{3}$$

The activity of the multisensory neural population is defined by a version of divisive normalization:

$$R_i = \frac{{E_i}^n}{\alpha^2 + \frac{1}{N}\sum_j {E_j}^n} \tag{4}$$

Here, E_i is the linear response of multisensory neuron i, α is the semisaturation constant, N is the total number of multisensory neurons, and n is the exponent of a power-law nonlinearity that represents the relationship between a neuron's

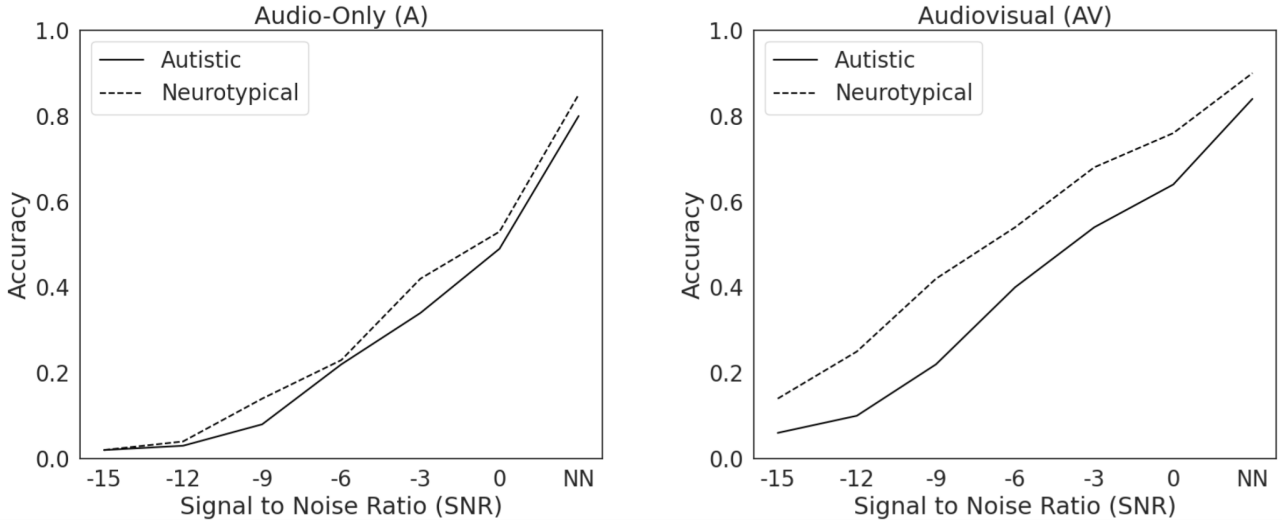

Figure 1: Lexical Identification Accuracy (Foxe et al., 2015)

membrane potential and its firing rate. Ohshiro et al. set this exponent (n) to 2 in their simulations, and they set the semisaturation constant to 1. This model offers a neural-level framework for cue integration that accounts for the classical empirical findings of multisensory integration, such as the principle of inverse effectiveness; however, it has no defined relationship with optimal cue combination.

Optimal Cue Combination

Bayes-optimal cue combination is the dominant cognitive model for thinking about sensory integration, whether it's on a unisensory level (Landy et al., 1995) or a multisensory level (Battaglia, Jacobs, & Aslin, 2003; Jacobs, 1999; Ma et al., 2009). At its core, optimal cue combination describes the inferential process wherein the integration of two cues, c_a and c_v, allow a perceiver to infer something about a stimulus, s. Given c_a and c_v, the posterior over s is derived from Baye's rule,

$$P(s|c_a, c_v) \propto P(c_a, c_v|s)P(s) \tag{5}$$

If we assume that the two cues, in our case auditory and visual cues, are conditionally independent given some stimulus, we can decompose the above equation into

$$P(s|c_a, c_v) \propto P(c_a|s)P(c_v|s)P(s) \tag{6}$$

Critically, if we assume a flat prior and that the likelihood functions, $P(c_a|s)$ and $P(c_v|s)$, are Gaussian with means μ_a and μ_v and variance σ_a^2 and σ_v^2, respectively, the mean, μ_{av} and variance, σ_{av}^2, of the posterior, $P(s|c_a, c_v)$, is given as:

$$\mu_{av} = \frac{\sigma_v^2}{\sigma_a^2 + \sigma_v^2}\mu_a + \frac{\sigma_a^2}{\sigma_a^2 + \sigma_v^2}\mu_v \tag{7}$$

$$\frac{1}{\sigma_{av}^2} = \frac{1}{\sigma_a^2} + \frac{1}{\sigma_v^2} \tag{8}$$

Because optimal cue combination relies on Gaussian representations, cues can be implemented using Probabilistic Population Codes (PPCs). PPCs are representations of neural activity that can encode probability distributions because of the variation inherent in neural populations (Ma et al., 2006). The response of a neuron in a population with Poisson-like noise is a function of its tuning curve. A tuning curve is akin to the receptive field parameter used in Ohshiro et al. (2011); a neuron's tuning curve is a Gaussian, $N(\mu_T, \sigma_T)$, along a set of possible stimuli (e.g., acoustic values), and the center of this Gaussian, μ_T, is representative of said neuron's preference for a particular stimulus value. If a stimulus aligns with a neuron's tuning curve, that neuron will be highly active relative to the surrounding population. The strongest neural response in a given population is the center of the PPC's distribution, and when the neural population is sufficiently large, its response distribution converges to a Gaussian. This allows us to encode and decode PPCs via Bayes' theorem.

Linking the Neural and Cognitive Levels

To assess how well E/I imbalance via divisive normalization explains autistic and neurotypical children's lexical identification, we integrate the neural model of multisensory integration using divisive normalization with the cognitive-level principles of cue combination via PPCs. We aim to combine the auditory and visual percepts of the acoustic characteristics of a spoken word to get an audiovisual percept. This audiovisual percept, which is represented as a probability distribution over audiovisual stimuli, is then sampled from to provide the perceiver's "guesses" at the audiovisual stimulus, given their auditory and visual percepts.

Our model takes a single stimulus value for each sensory modality, s_a and s_v, as input. Each unisensory input is transformed into a unidimensional PPC. The intensity of the neural

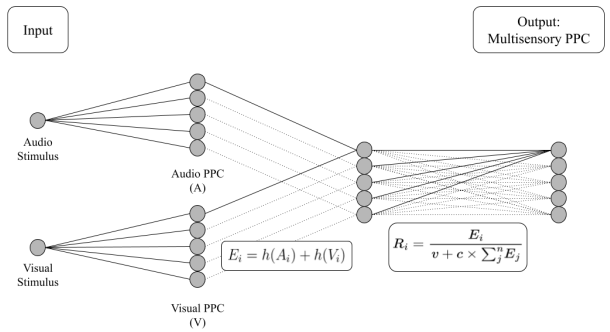

Figure 2: Model Structure

activity (i.e., the height of neural responses) of each PPC is determined by its gain, g, a free parameter that is introduced during the input and corresponds to its reliability, which we assume is inversely related to external noise. Ma et al. (2006) showed that we can perform optimal cue combination by simply summing PPCs; however, we opt to incorporate the neural model of multisensory integration proposed by Ohshiro et al. (2011) in order to introduce divisive normalization into the cue combination process. Under the divisive normalization model, the response, r_i, of neuron i in each unisensory population is sublinearly transformed and summed in a modification of Equation 3,

$$E_i = h(r_{a_i}) + h(r_{v_i}) \tag{9}$$

Here, $h(x) = \sqrt{x}$, which is consistent with Ohshiro et al. (2011). We deviate from Ohshiro et al. (2011) in that we do not include an explicit weight term, as the reliability of each PPC, encoded by g, is implicitly encoded. The summed activity of the unisensory PPCs, E, is then divisively normalized (Eq 4).

To obtain the mean and variance of the divisively normalized multisensory PPC, R, we use the properties of that population's tuning curves,

$$\mu_{av} = \frac{\sum_i \frac{\mu_{T_i}}{\sigma^2_{T_i}} \cdot R_i}{\sum_i \frac{1}{\sigma^2_{T_i}} \cdot R_i} \tag{10}$$

$$\sigma^2_{av} = \frac{1}{\sum_i \frac{1}{\sigma^2_{T_i}} \cdot R_i} \tag{11}$$

Here, μ_T is a vector of all tuning curve centers and σ^2_T is a vector of the tuning curve variances of the multisensory neural population. We confirmed that this model, when unaltered, performs near-optimal cue combination by comparing the mean and variance obtained by Equations 10 and 11 to the mean and variance of Equations 7 and 8. We observed that we derive the same audiovisual percept means from Equations 10 and 7, and we found a near-zero difference in the audiovisual percept variances generated by Equations 11 and 8.

After obtaining a probability distribution over the audiovisual percept by decoding the output layer of the network, we can sample from that distribution to obtain a perceiver's guesses at the audiovisual percept.

In summary, this model takes a visual input and an auditory input and transforms these into unisensory PPCs, which provide us with a representation of the activity of a unisensory neural population in response to a given input. We linearly combine these PPCs to represent the summed activity of the unisensory populations. This representation is then divisively normalized to form the output of the model: a multisensory PPC. This structure is useful because it allows us to model E/I imbalance within the stage of divisive normalization while remaining Bayes-optimal.

Modeling Lexical Identification

Children in the Foxe et al. (2015) experiment performed a lexical identification task, deciding which word they heard based on their audiovisual percept. To model the lexical identification task, we use a simple Bayesian model of a choice between two words.

Our generative model assumes two words, whose acoustics are represented as unidimensional Gaussians, with means μ_{w_1} and μ_{w_2} and variances $\sigma^2_{w_1}$ and $\sigma^2_{w_2}$. The stimulus is taken to be an instance of w_1 sampled from the corresponding Gaussian. Here, w_1 represents the target lexical category, and w_2 is the lexical competitor. Although we are using a two-choice format, this model can in principle be expanded to any number of competitor words.

The audiovisual stimulus, s, is assumed to be generated from the Gaussian associated with the target word, w_1,

$$P(s|w_1) = N(\mu_{w_1}, \sigma^2_{w_1}) \tag{12}$$

The experimental participant hears the stimulus, recovers the audiovisual percept, $\hat{s}$, using the neural model described above, and then needs to infer which word was spoken. We assume that the participant uses Bayes' rule to identify which word was spoken, with equal prior probabilities for each of the two words,

$$p(w_1|\hat{s}) = \frac{P(\hat{s}|w_1)}{P(\hat{s}|w_1) + P(\hat{s}|w_2)} \tag{13}$$

The participant then guesses the word's identity by sampling from their posterior distribution over words, so that their accuracy is equal to the value of the posterior probability of the target word. For a given set of auditory and visual percepts, we obtain a mean lexical identification accuracy by averaging over 50 samples taken from the Gaussian derived from that combination's multisensory PPC.

Simulation

Our simulation is designed to replicate the audiovisual performance of autistic and neurotypical children on a lexical iden-

	Auditory	Visual	Multisensory
Gain (g)	10, 20, 30, 40, 50, 60, 70, 80	50	N/A
Tuning Curve Variance (σ^2)	10	10	10

Table 1: Free Parameters in Unisensory Populations

tification task in varying levels of noise (Foxe et al., 2015). Using the divisive normalization stage of our base model, we can create two models to represent the hypothesized differences in neurotypical and autistic listeners. The neurotypical model will maintain a standard suppressive field gain term, c, which is proposed by Ohshiro et al. (2011) to be $\frac{1}{N}$, where N is the number of neurons in the multisensory neural population. The autistic model will have a reduced suppressive field gain term, $\frac{1}{4*N}$ (Rosenberg et al., 2015).

We predict that if the E/I imbalance resulting from a reduced suppressive field gain term sufficiently explains differences in multisensory integration between autistic and neurotypical children, the model with altered divisive normalization in the form of a reduced suppressive field gain term would be less accurate at identifying lexical items in noise than the model with unaltered divisive normalization.

Parameters

We chose to abstractly represent possible auditory and visual stimuli as integer values ranging from 0 to 100. The size of the neural populations of both unisensory layers and the multisensory layer, N, is also 100. The target lexical item is represented as a Gaussian with $\mu_{w_1} = 45$ and $\sigma^2_{w_1} = 3$, and the competitor lexical item is a similar Gaussian with $\mu_{w_2} = 55$ and $\sigma^2_{w_2} = 3$. These parameters were selected because they reliably encoded and decoded PPCs, but they were not optimized in any way for the behavioral data.

A neurotypical (unaltered divisive normalization) and autistic (altered divisive normalization) model each receive the same input (Table 1). The auditory stimulus, s_a, and visual stimulus, s_v, are each created by taking a sample ($n = 10$) from the Gaussian representation of the target word, $N(\mu_{w_1}, \sigma^2_{w_1})$. Each auditory stimulus is paired with each gain value for a total of 80 stimulus-gain combinations, or 80 unique auditory inputs. The visual stimuli are also sampled from a target lexical item, but they are given a static gain value, as the externally induced noise for visual stimuli in Foxe et al. (2015) does not vary.

Results

We tested two hypotheses with our simulated lexical identification task. First, we predicted that if our model accurately incorporates noise, we would observe a positive relationship between inverse noise, represented as gain, and lexical identification accuracy. Across both the neurotypical (unaltered divisive normalization) and autistic (altered divisive normalization) models, we reproduced the same general effect of noise observed in Foxe et al.'s (2015) audio-only and audiovisual lexical identification tasks. As gain increases, our audiovisual model performed lexical identification with increasing accuracy (Figure 3). This suggests that we are appropriately capturing noise with the gain parameter. As such, we can assume that we can use gain as a representation of noise within our audio-visual simulations.

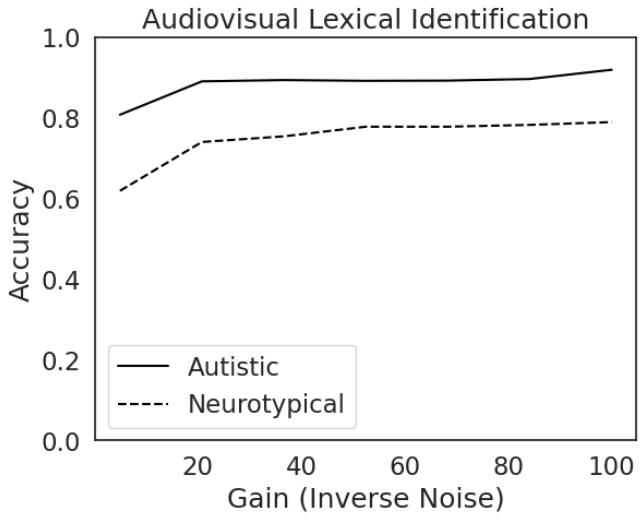

Figure 3: Performance of Autistic and Neurotypical Models on Lexical Identification Task

Second, we predicted that if an E/I imbalance implemented via divisive normalization accounts for differences in autistic and neurotypical perception, we would observe a higher overall lexical identification accuracy from the neurotypical model in comparison to the autistic model. Our simulation demonstrates the opposite pattern of results. The autistic model, which differs from the neurotypical model in only its suppressive field gain term, performs with greater accuracy on the lexical identification task (Figure 3).

Note that because we did not fit our models to the behavioral data, both performed with an overall greater accuracy than the children observed in Foxe et al. (2015). Furthermore, the gain values we used in our simulation resulted in a weaker relationship between noise and lexical identification accuracy than observed behaviorally. Despite these differences, our qualitative results clearly illustrate that the altered divisive normalization model performs with greater accuracy than the unaltered divisive normalization model, which is the opposite of what was found behaviorally.

We reason that this result is due to the E/I imbalance in the autistic model. There are no differences between the autistic and neurotypical models before multisensory integration via divisive normalization. When combining auditory and neural PPCs, the neurotypical model produces a multisensory PPC with a larger variance than that of the autistic model.

Because we are sampling from the resulting Gaussians from each respective multisensory PPC, the samples obtained from the Gaussians produced by the autistic model are more likely to lie within the distribution of the target lexical item than the samples from the neurotypical model. As such, any time the auditory or visual inputs are closer to the target lexical item than the competitor lexical item, which is generally true of the stimuli used in our simulation, the autistic model will perform with greater accuracy than the neurotypical model. In other words, greater excitation under the conditions outlined in our simulation will result in increased lexical identification accuracy.

Discussion

In this study, we tested an extended interpretation of the theory that altered sensory processing in autism is the result of an E/I imbalance caused by altered divisive normalization. We applied this theory in a speech perception context by implementing a simulation that used divisive normalization to perform probabilistic cue integration of auditory and visual speech cues. Through a lexical identification task, we measured the performance of an autistic model, with altered divisive normalization, to a baseline (neurotypical) model (no altered divisive normalization). We found that the altered divisive normalization model had greater accuracy in the lexical identification task than the baseline mode.

The relative performance of the altered and unaltered divisive normalization models suggests that altered divisive normalization during multisensory integration does not account for the differences in audiovisual speech perception observed behaviorally. Our results are incompatible with the theory that an E/I imbalance is caused by altered divisive normalization and that this imbalance alters audiovisual speech perception in autism. Under the assumption that Bayes-optimal cue combination is the correct framework for multisensory integration, there are two possible ways this theory could be wrong. The first is that an E/I imbalance does not affect multisensory processing at the stage of integration. This interpretation does not rule out that an E/I imbalance could be relevant in other aspects of perception, but the E/I imbalance theory would need to be extended to specify why an E/I imbalance introduced by divisive normalization does not alter perception in a manner consistent with behavioral accounts of autism. The second possible explanation for our results is that multisensory integration via divisive normalization is not the correct implementation of cue combination. Although probabilistic population codes and divisive normalization can implement Bayes-optimal cue combination in principle, this interpretation of our results suggests that it is not the implementation that human listeners use. In light of the inconsistencies between our model's performance and behavioral accounts of audiovisual speech perception in autistic and neurotypical children, it is reasonable to argue that our adaptation of the multisensory integration model proposed by Ohshiro et al. (2011) does not adequately capture the mechanism by which unisensory cues are integrated during speech perception. In other words, altered divisive normalization and Bayes-optimal cue combination are incompatible under this interpretation of our results.

A third explanation for the performance of our model on the lexical identification task is that Bayes-optimal cue combination is not the correct framework for thinking about these differences in multisensory integration. Our probabilistic cue combination made incorrect predictions about the relationship between E/I imbalance and lexical identification accuracy. Similar issues have been observed with Bayesian cue combination models in other settings. For example, Bates and Jacobs (2021) found that models that use Rate Distortion Theory (RDT) to derive optimal attentional allocation predictions in uncued and cued visual search better align with human behavior than Bayesian models. Namely, the Bayesian model cannot explain why people's performance on visual search tasks decreases as the amount of information in the scene increases, whereas RDT attributes this difference to resource allocation due to the capacity limits of a noisy perceptual channel. Given these findings, the fixed internal variance imposed by a Bayesian cue combination model may fail to capture the source of the observed differences in perception between autistic and neurotypical listeners. Rather, divisive normalization may be impacting the capacity of the channel.

Our model is limited in the simplicity of its lexical identification task, which represents lexical items unidimensionally and with equal priors. However, even if we created a more complex model of the lexical identification task, we expect the same direction of results obtained from our current model. This is because the altered and unaltered divisive normalization models are identical until the stage where the combined auditory and visual percepts are divisively normalized. This results in two multisensory percepts with the same mean and different variances. If we were to give our lexical representations varying priors, for instance, we would still expect higher lexical identification accuracy for the altered divisive normalization model because it consistently represents multisensory percepts with smaller variances.

Through testing a computational model of speech perception that incorporates a neural theory of autism, this study acts as a first step in developing explicit linking functions for theoretical models of perception in autism. A natural next step is to continue augmenting cognitive models to account for neural data in an effort to explain differences in the performance of autistic and neurotypical perceivers on audiovisual lexical identification tasks.

Acknowledgments

This work was supported by a University of Maryland Baggett fellowship, NSF grant BCS-2120834, and NSF GRFP grant DGE-2146755. We thank Wei Ji Ma and three anonymous reviewers for helpful discussion and feedback. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not

necessarily reflect the views of the National Science Foundation.

References

Bates, C. J., & Jacobs, R. A. (2021). Optimal attention allocation in the presence of capacity constraints in uncued and cued visual search. *Journal of Vision*, *21*(5), 1-23.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America*, *20*(7), 1391-1397.

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.

Beck, J. M., Latham, P. E., & Pouget, A. (2011). Marginalization in neural circuits with divisive normalization. *The Journal of Neuroscience*, *31*(43), 15310–15319.

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H.-P., Russo, N. N., Blanco, D., ... Ross, L. A. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cereb Cortex*, *25*(2), 298–312.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, *39*, 3621-3629.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*(3), 389-412.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

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*, *3*(4), e4638.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. *Nature Neuroscience*, *14*(6), 775–782.

Rosenberg, A., Patterson, J. S., & Angelaki, D. E. (2015). A computational perspective on autism. *Proceedings of the National Academy of Sciences*, *112*(30), 9158–9165.

Rubenstein, J. L. R., & Merzenich, M. M. (2003). Model of autism: Increased ratio of excitation/inhibition in key neural systems. *Genes, Brain and Behavior*, *2*(5), 255-267.