Seeking visual information to support spoken language comprehension

Kyle MacDonald[1] (kylem4@stanford.edu), **Virginia Marchman**[1] (marchman@stanford.edu),
Anne Fernald[1] (afernald@stanford.edu), **Michael C. Frank**[1] (mcfrank@stanford.edu)
[1] Department of Psychology Stanford University

Abstract

Efficient language comprehension is a multisensory integration process where listeners integrate information from the visual and linguistic signal. But the usefulness of a given information source can vary across contexts – as in the case of processing speech in noise or monitoring another speaker's social cues to reference (e.g., eye gaze). How do listeners adapt decisions about visual fixation to support comprehension? Here, we report two experiments that provide evidence for an adaptive information-seeking account: that listeners alter the dynamics of gaze during real-time processing to seek additional visual information when it is useful for supporting comprehension. First, we show that adults (n=33) and children (n=40, 3-5 y.o.) delayed their eye movements away from a speaker while processing speech in noise. Intrestingly, the decision to delay resulted in a speed-accuracy tradeoff, with more accurate gaze shifts and fewer random responses (E1). Next, we present results showing one limit of this adaptive response: both adults (n=33) and children (n=54, 3-5 y.o.) did not delay eye movements to gather process a post-nominal social cue when the auditory signal was sufficient to establish reference (E2). Together, these results suggest that the dynamics of eye movements can flexibly adapt to the demands of different processing contexts, and that even very young listeners will seek additional visual information when it is useful for language comprehension.

Keywords: eye movements; language processing; information-seeking; speech in noise; social cue processing

Introduction

Real-time language comprehension is a multimodal phenomenon. As listeners, we integrate information from both the visual and the linguistic signal to reach a final interpretation. One classic demonstration of this integration process is the "McGurk effect" where a speaker's mouth movements suggest one sound while their acoustic output suggests another, resulting in the listener perceiving a third, intermediate sound (J. MacDonald & McGurk, 1978). Moreover, prominent theories of speech perception (McClelland, Mirman, & Holt, 2006) and lexical processing (M. C. MacDonald & Seidenberg, 2006; Smith, Monaghan, & Huettig, 2017) have argued that *interactive* processes – where information is combined from multple sources in parallel – are a defining feature of human language. And recent work on effortful listening shows that people will allocate additional cognitive resources to compensate for degraded information in the auditory modality (Van Engen & Peelle, 2014).

However, the usefulness of different kinds of information varies depending on features of the listener **and** on features of the context. Consider the case of processing a visual-manual language like American Sign Language (ASL). Here, the value of allocating visual fixations to the language source (i.e., the signer) is high since all of the language-relevant information is in that fixation location. In our prior work, we showed that, compared to spoken language learners, young ASL-learners prioritize information accumulation and accuracy over and above speed when deciding to seek a named referent during real-time ASL comprehension (K. MacDonald, Blonder, Marchman, Fernald, & Frank, 2017). We proposed an information-maximization account inspired by goal-based theories of vision (Hayhoe & Ballard, 2005): that signers are sensitive to the higher value of a certain fixation behavior and adapted the dynamics of gaze to avoid shifting away too quickly and miss information that could be used for comprehension.

In the work reported here, we aim to test predictions of our information-maximization account in two novel domains wherefeatures of the context modulate the value of seeking visual information: (1) processing speech in noise and (2) processing speech accompanied with a visual cue to reference (eye gaze). We chose the case of processing speech in noisy environments because we hypothesized that adding background noise would make the auditory signal less reliable, and in turn make the visual signal more useful. In fact, classic empirical work on speech perception shows that adults are better able to "recover" linguistic information in noisy contexts when they have visual access to a speaker's face (Erber, 1969). We chose social cue processing because a speaker who gazes at on object is more informative, providing visual information that completely disambiguates reference. Moreover, social-pragmatic theories of language acquisition emphasize the role of processing social cues for early language acquition (Clark, 2009) and empirical work shows that gaze following emerges in the first year of development (Brooks & Meltzoff, 2008).

We think these experiments are important for three reasons. First, they provide a confirmatory test of our information-maximization account of the findings reported in K. MacDonald et al. (2017) while providing an important control for the multitude of other population-level differences between ASL- and English-learners. Second, they inform the generalizability of the account by testing predictions across a wider variety of processing contexts that are not typically studied in work on early language comprehension. And third, this work brings together ideas from several rich research programs and theoretical accounts. For example, work on language-mediated visual attention shows that adults and children rapidly shift visual attention upon hearing the name of an object in the visual scene (Allopenna, Magnuson, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). These findings have led to debates about whether language-mediated gaze shifts are automatic as opposed to under the control of the listener. Empirical work on vision during natural tasks shows that people overwhelmingly prefer to look at *goal-relevant* locations – e.g., an upcoming obstacle while walking (Hayhoe & Ballard, 2005). These accounts make the prediction that gaze patterns during language comprehension should adapt to the value of fixation locations with respect to the goal of rapid language processing. Finally, work on effortful listening shows that listeners generate compensatory responses (e.g., increases in attention and working memory) within "challenging" comprehension contexts such as processing noisy or accented speech (Van Engen & Peelle, 2014). These accounts predict that our young listeners might compensate for the reduced quality of the auditory signal by allocating gathering additional visual information.

Experiment 1

E1 tests whether our information-maximization account of eye movements would generalize to a novel and ecologically valid language processing context – processing speech in noise. We recorded eye movements during a real-time language comprehension task where children and adults processed familiar sentences (e.g., "Where's the ball?") while looking at a simplified visual world with 3 fixation targets (see Fig 1). Using a within-participants design, we manipulated the signal-to-noise ratio of the auditory information by adding brown noise. We predicted that processing speech in noise would increase the value of fixating on the speaker to gather additional inormation before generating a shift to the named referent even after the target linguistic item began unfolding in time.

To test this prediction, we compare the Accuracy and Reaction Times (RTs) of first shifts across the two conditions. We also present two model-based analyses that link the observable behavior to underlying psychological constructs. First, we use an exponentially weighted moving average (EWMA) method (Vandekerckhove & Tuerlinckx, 2007) to categorize participants' gaze shifts as language-driven or random. In contrast to the standard RT/Accuracy analysis, the EMWA allows us to quantify differences in participants willingness to generate gaze shifts prior to collecting sufficient information to seek the named referent. Next, we use drift-diffusion models (DDMs) (Ratcliff & Childers, 2015) to ask whether the behavioral differences in Accuracy and RT are driven by a more cautious responding strategy or by more efficient information processing – a critical distinction for our theoretical account.

Method

Participants Participants were native, monolingual English-learning children (n = 39; 22 F, 17 M) and adults (n = 31; 22 F, 9. All participants had no reported history of developmental or language delay and normal vision. 14 participants (11 children, 3 adults) were run but not included in the analysis because either the eye tracker falied to calibrate or the participant did not complete the task.

Stimuli *Linguistic stimuli.* The stimuli were recorded in a sound-proof room and featured two female speakers who used natural child-directed speech and said one of two phrases: "Hey! Can you find the (target word)"" or "Look! Where's the (target word) – see panel A of Fig 1. The target words were: ball, bunny, boat, bottle, cookie, juice, chicken, and shoe. The target words varied in length (shortest = 411.68 ms, longest = 779.62 ms) with an average length of 586.71 ms.

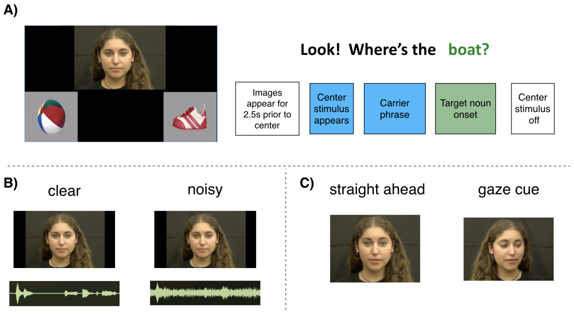

Figure 1: Stimuli for E1 and E2. Panel A shows the layout of the three fixation locations (speaker, target, and distracter), and the timecourse of a single trial. Panel B shows a visual representation of the clear and noisy waveforms used in E1. Panel C shows the social cue manipulation used in E2.

Noise manipulation. To create the noisy stimuli, we convolved the recordings with Brown noise using the Audacity audio editor. The average signal-to-noise ratio[1] in the noise condition was 2.87 dB compared to the clear condition, which was 35.05 dB.

Visual stimuli. The image set consisted of colorful digitized pictures of objects presented in fixed pairs with no phonological overlap between the target and the distracter image (cookie-bottle, boat-juice, bunny-chicken, shoe-ball). Side of target picture was counterbalanced across trials.

Design and procedure Participants viewed the task on a screen while their gaze was tracked using an SMI RED corneal-reflection eye-tracker mounted on an LCD monitor, sampling at 60 Hz. The eye-tracker was first calibrated for each participant using a 6-point calibration. On each trial, participants saw two images of familiar objects on the screen for two seconds before the center stimulus appeared (see Fig 1). Then they processed the target sentence – which consisted of a carrier phrase, a target noun, and a question – followed by two seconds without language to allow for a response. Child participants saw 32 trials (16 noise trials; 16 clear trials) with several filler trials interspersed to maintain interest. Adult participants saw 64 trials (32 noise; 32 clear).

Results and Discussion

Analysis plan First, we present behavioral analyses of First Shift Accuracy and Reaction Time (RT). RT corresponds to the latency to shift away from the central stimulus to either

[1] The ratio of signal power to the noise power, with values greater than 0 dB indicating more signal than noise.

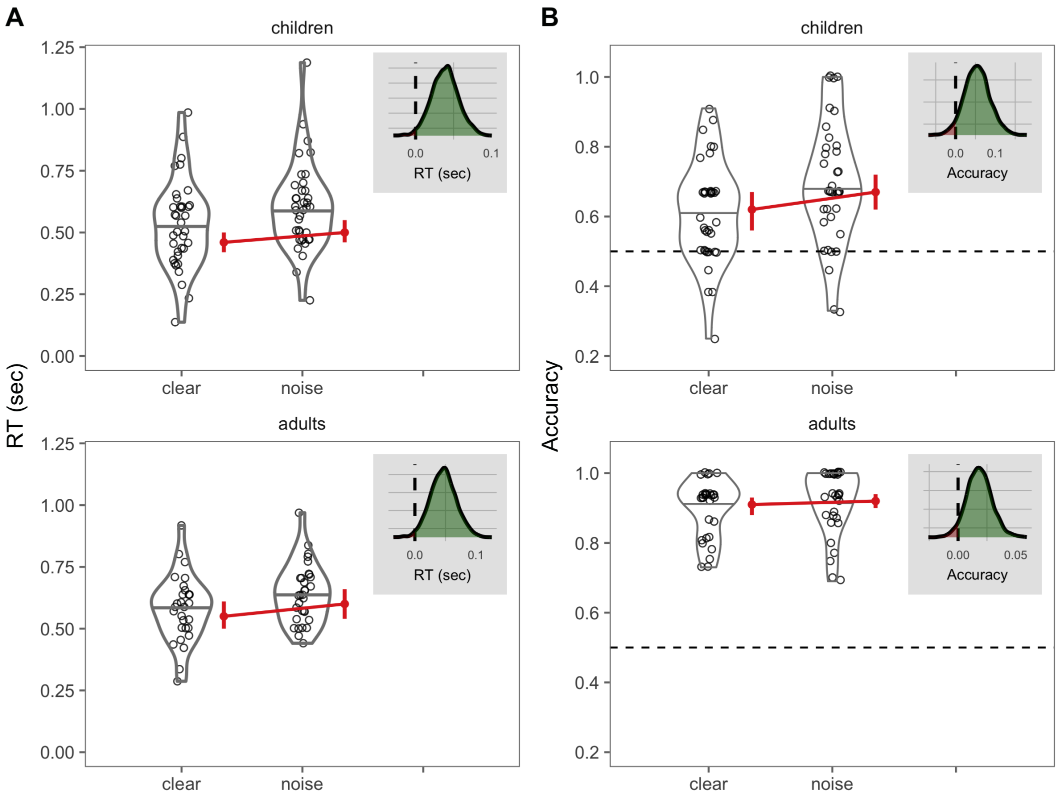

Figure 2: Behavioral results from E1. Panel A shows violin plots representing the distribution of median RTs for each participant in each condition. The dark red points represent the most likely estimate of the group mean with the error bars showing the 95% Highest Density Interval. The grey inset plot shows the full posterior distribution of plausible RT differences across conditions with the vertical dashed line representing the null value of zero condition difference. The green shading represents estimates above the null value and the red shading represents estimates below the null value. Panel B shows the same information but for First Shift Accuracy.

picture measured from onset of the target noun in the linguistic stimuli (we log transformed all RTs prior to analysis). Accuracy corresponds to whether the participant's first gaze shift landed on the target or the distracter picture. We used the `rstanarm` (Gabry & Goodrich, 2016) package to fit Bayesian mixed-effects regression models. The mixed-effects approach allowed us to model the nested structure in our data (multiple trials for each participant; and a within-participants manipulation) by including random intercepts for each participant and item, and a random slope for each item and noise condition. We used Bayesian estimation to quantify the uncertainty in our point estimates of the group means and condition differences. To communicate this uncertainty we report the 95% Highest Density Interval (HDI), which provides a range of credible values given the data and model. All analysis code can be found in the online repository for this project: `https://github.com/kemacdonald/speed-acc/R/analysis`.

Next, we present the two model-based analyses – the EWMA and HDDM – discussed in the introduction. Again, the goal of these models is to move beyond a description of the data and map behavioral differences in eye movements to underlying psychological variables. First, we use an EWMA method to model changes in random shifting behavior as a function of delays in responding (i.e., RT). For each RT, the model generates two values: a "control statistic" (**CS**, which captures the running average accuracy of first shifts) and an "upper control limit" (**UCL**, which captures the pre-defined threshold when gaze shifts would be categorized as deviating from random responding). Here, the CS is an expectation of random shifting to either the target or the distracter image (nonlanguage-driven shifts), modeled a Bernoulli process with $P(success) = 0.5$. As participants delay their response, we assume that they have gathered more information and should become more accurate, which we model a Bernoulli process with $P(success) > 0.5$. Using this model, we can quantify and compare: a) the cutoff point when the CS exceeds the UCL in the RT distribution, indicating the processing time required before participants generated language-driven shifts and b) the proportion of all gaze shifts that the model categorizes as language-driven vs. nonlanguage-driven.

Finally, we took the shifts that were categorized as language-driven by the EWMA and fit a hierarchical

Bayesian drift-diffusion model (HDDM) to quantify differences in the underlying decision process that led to different patterns of behavior. We chose to implement a hierarchical Bayesian version of the DDM using the HDDM Python package (Wiecki, Sofer, & Frank, 2013) since we had relatively few trials from the child participants and recent simulation studies have shown that the HDDM approach was better than other DDM fitting methods for small data sets (Ratcliff & Childers, 2015). The model assumes that people accumulate noisy evidence in favor of one alternative with a response generated when the evidence crosses a pre-defined decision threshold. Here, we focus on two parameters of interest that map onto meaningful decision variables that we hypothesized would vary across our conditions: **boundary separation**, which indexes the amount of evidence gathered before generating a response (higher values suggest more cautious responding) and **drift rate**, which indexes the amount of evidence accumulated per unit time (higher values suggest more efficient processing of the stimulus).

Behavioral analyses *RT.* To make RTs more suitable for modeling on a linear scale, we analyzed responses in log space using a logistic transformation, with the final model was specified as: $log(RT) \sim noise\_condition + age\_group + (sub\_id + noise\_condition \mid item)$. Panel A of Figure 2 shows the data distribution for each participant's RT, the estimates of condition means, and the full posterior distribution of the estimated difference between the noise and clear conditions. Both children and adults were slower to identify the target in the noise condition (Children M_{noise} = 0.5 ms; Adult M_{noise} = 0.6 ms), as compared to the clear condition (Children M_{clear} = 0.46 ms; Adult M_{clear} = 0.55 ms). RTs in the noise condition were 42.55 ms slower on average, with a 95% HDI from 4.88 ms to 83.34 ms that did not include the null value of zero condition difference.

Accuracy. Next, we modeled adults' and children's first shift accuracy using a mixed-effects logistic regression with the same specifications (see Panel B of Fig 2). Overall, both groups responded at rates different from a model of random behavior (null value of 0.5 falling well outside the lower bound of all group means). Adults were more accurate ($M_{adults} = 91\%$) compared to children ($M_{adults} = 62\%$). Both groups tended to be more accurate in shifting to the target image in the noise condition (Children M_{noise} = 67%; Adult M_{noise} = 92%) as compared to the clear condition (Children M_{clear} = 62%; Adult M_{clear} = 91%). Accuracy in the noise condition was 4% higher on average, with a 95% HDI from 0% to 11%.Note that while the null value of zero difference falls within the 95% HDI, 96% of the credible values fall below the null, providing evidence for higher accuracy in the more challenging noise condition.

Model-based analyses *EWMA.*

HDDM.

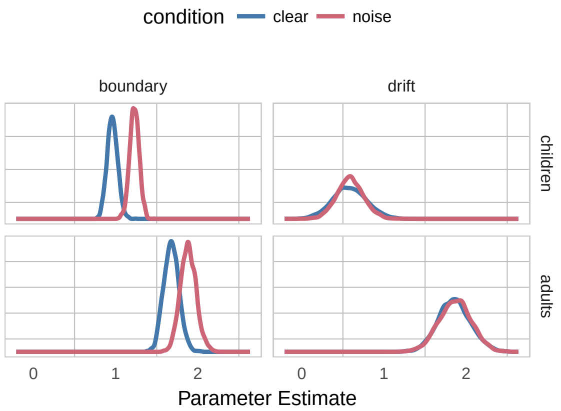

Figure 3: HDDM results E1.

Experiment 2

Method

Participants XX Stanford undergraduates participated (XX male, XX females) for course credit. All participants were monolingual, native English speakers and had normal vision.

Stimuli Audio and visual stimuli were identical to the Face and Bullseye tasks in E1. We included a new center fixation stimulus type: printed text. The text was displayed in a white font on a black background and was programmed such that only a single word appeared on the screen, with each word appearing for the same duration as the corresponding word in the spoken language stimuli.

Design and procedure The design was identical to E1. Participants saw a total of 128 trials while their eye movements were tracked using automated eye-tracking software.

Results and Discussion

Behavioral analyses *RT.*

Accuracy.

Experiment	Parameter	Contrast	Estimate (95% HDI)
E1	Cut point	age group	0.21 [0.19, 0.24]
E1	Cut point	noise	0.04 [0.01, 0.06]
E1	Guessing	age group	-0.46 [-0.49, -0.43]
E1	Guessing	noise	0.12 [0.07, 0.17]
E2	Cut point	age group	0.09 [0.03, 0.14]
E2	Cut point	gaze	0 [-0.06, 0.06]
E2	Guessing	age group	-0.45 [-0.48, -0.42]
E2	Guessing	gaze	-0.03 [-0.07, 0]

Table 1: EWMA results for E1 and E2. Cut point refers to the response time in the RT distribution when gaze shifts reliably deviated from random. Guessing refers to the proportion of gaze shifts categorized as random vs. language-driven. Estimate refers to the average difference between condition or age group.

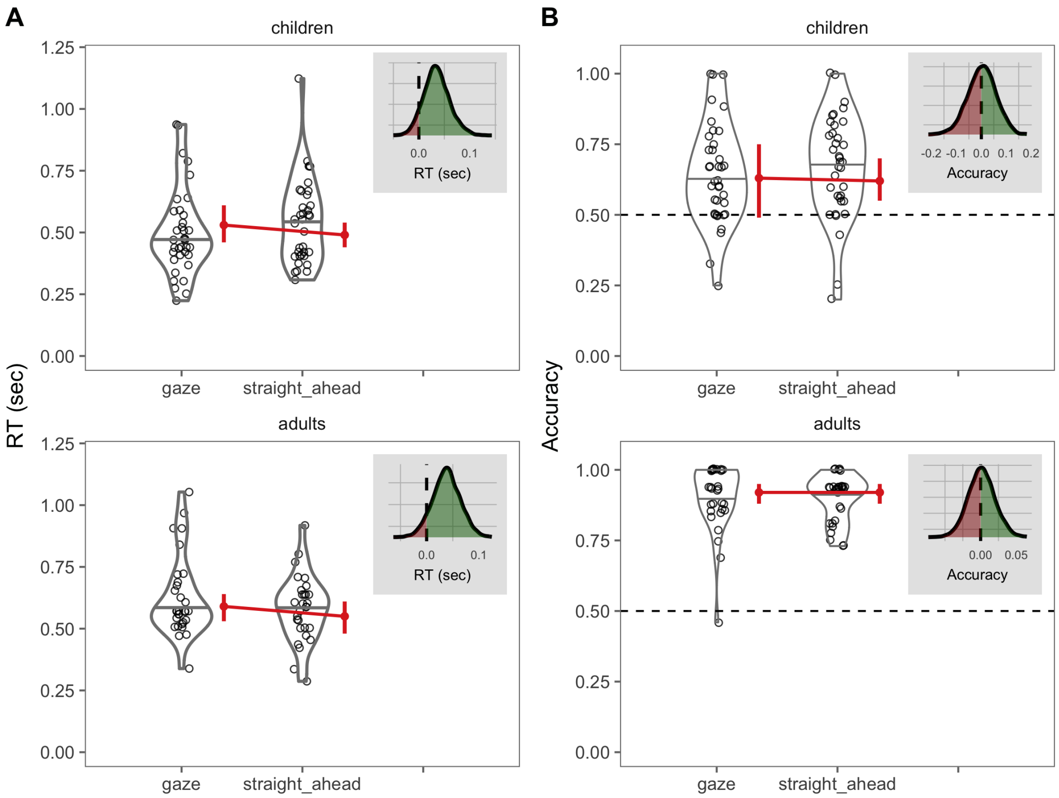

Figure 4: Behavioral results from E2. All plotting conventions are the same as in Figure 2.

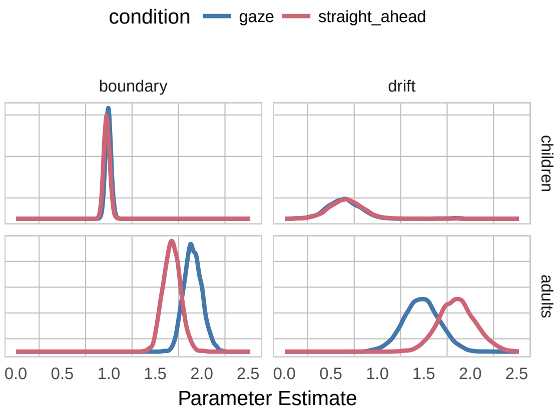

Figure 5: HDDM results E1.

Model-based analyses *EWMA.*

HDDM.

General Discussion

Language comprehension can be facilitated by fixating on relevant features of the nonlinguistic visual world or on the speaker. But how do we decide where to look? We propose that eye movements during language processing reflect a sensitivity to the tradeoffs of gathering different kinds of information. We found that young ASL-learners generated slower but more accurate shifts away from a language source and produced a smaller proportion of nonlanguage-driven shifts compared to spoken language learners. We found the same pattern of behavior within a sample of English-speaking adults processing displays of printed text compared to spoken language. These results suggest that as the value of fixating on a location to gather information about the linguistic signal increases, eye movements to the *rest* of the visual world become less useful and occur less often.

Our work here attempts to synthesize results from different populations and stimuli in a single framework, but it has several limitations that we hope will pave the way for future work. First, we have not performed a confirmatory test of the DDM findings: both ASL-learners (E1) and adults processing language from a person (E2) prioritize accuracy over speed. So these findings, while interesting, are preliminary. Second, we do not know what might be driving the population differences in E1. It could be that ASL-learners' massive experience dealing with competition for visual attention leads to changes in the deployment of eye movements during language comprehension. Or, it could be that the in-the-moment constraints of processing a visual language cause different fixation behaviors. Finally, we used a very simple visual world, with only three places to look, and very simple linguistic stimuli, especially for the adults in E2. Thus it remains an open question how these results might scale up to more complex language information and visual environments.

This work attempts to integrate top-down, goal-based models of vision (Hayhoe & Ballard, 2005) with work on language-driven eye movements (Allopenna et al., 1998). While we chose to start with two case studies – ASL and text processing – we think the account is more general and that there are many real world situations where people must negotiate the tradeoff between gathering more information about language or about the world: e.g., processing spoken language in noisy environments or at a distance; or early in language learning when children are acquiring new words and often rely on nonlinguistic cues to reference such as pointing or eye gaze. Overall, we hope this work contributes to a broader account of eye movements during language comprehension that can explain fixation behaviors across a wider variety of populations, processing contexts, and during different stages of language learning.

Acknowledgements

We are grateful to the families who participated in this research. Thanks to Tami Alade and Hannah Slater for help with data collection. This work was supported by an NSF GRFP to KM.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(01), 207–220.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425.

Gabry, J., & Goodrich, B. (2016). Rstanarm: Bayesian applied regression modeling via stan. r package version 2.10. 0.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Attention, Perception, & Psychophysics*, *24*(3), 253–257.

MacDonald, K., Blonder, A., Marchman, V. and, Fernald, A., & Frank, M. C. (2017). An information-seeking account of eye movements during spoken and signed language comprehension. In *Proceedings of the 39th annual conference of the cognitive science society*.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. *Handbook of Psycholinguistics*, *2*, 581–611.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–369.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237–279.

Smith, A. C., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. *Journal of Memory and Language*, *93*, 276–303.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632.

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*, 14.