Summer 8-2019

Task-evoked Pupillary Response for Completely Intelligible Accented Speech

Drew McLaughlin
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological and Brain Sciences

Task-evoked Pupillary Response for Completely Intelligible Accented Speech

by

Drew J. McLaughlin

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

August 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

This project was supported by a National Science Foundation Graduate Research Fellowship awarded to Drew J. McLaughlin (DGE-1745038).

We offer special thanks to the Washington University School of Engineering for allowing us to use their dissertation and thesis template as a starting point for the development of this document.

Lastly, I would like to thank my Master's Examination Committee: Kristin Van Engen, Mitch Sommers, and Jonathan Peelle.

Drew J. McLaughlin

Washington University in St. Louis

August 2019

Abstract

Task-evoked Pupillary Response for Completely Intelligible Accented Speech
by
Drew J. McLaughlin

Master of Arts in Psychological and Brain Sciences
Washington University in St. Louis, 2019
Dr. Kristin J. Van Engen

Speech perception under adverse conditions, such as those caused by noise in the environment or a speaker's accent, can be cognitively demanding. For second language- (L2-) accented speech, mismatches between the speech patterns of an L2-accented speaker and a listener can result in poorer understanding and reduced intelligibility (i.e., fewer words in the speech stream can be correctly identified). However, it remains unclear whether completely intelligible L2-accented speech imposes greater *cognitive load* (defined here as the degree to which cognitive resources are recruited at a given moment to meet processing demands) than native speech. In the current study, we used pupillometry to examine cognitive load during the perception of completely intelligible, Mandarin Chinese-accented English speech and standard American-accented English speech. Results from two experiments showed greater and more rapid pupillary response (indicating greater cognitive load) for L2-accented speech than native speech. Additionally, participants subjectively rated the L2-accented speaker as more effortful to understand than the native speaker. Consistent with an executive recruitment account for accented speech, these findings indicate that mismatches between the speech patterns of L2-accented speech and native listeners' representations require greater cognitive load to process—even when recognition accuracy is at ceiling.

Task-evoked Pupillary Response for Completely Intelligible Accented Speech

1.1 Introduction

Listening to second language- (L2-) accented speech is often described as an effortful process, even when L2 speakers are highly proficient (Munro & Derwing, 1995). Both L2- and regionally-accented speech are characterized by systematic segmental and suprasegmental deviations from standard pronunciations of a given language. Reduced speech intelligibility and increased listening effort are caused by mismatches between these systematic speech patterns, which require additional cognitive resources to process (Van Engen & Peelle, 2014). Under this executive recruitment account, even when an L2-accented speaker is completely intelligible (i.e., all of the words in their speech can all be correctly identified) speech processing should nonetheless require greater *cognitive load* (defined here as the degree to which cognitive resources are recruited at a given moment to meet processing demands; Pichora-Fuller et al., 2016). In particular, this increased cognitive load during speech processing may be attributable to the recruitment of executive functions including working memory (McLaughlin, Baese-Berk, Bent, Borrie, & Van Engen, 2018) and inhibition (Janse & Adank, 2012; Banks, Gowen, Munro, & Adank, 2015), which positively correlate with individual listeners' recognition of and adaptation to accented speech, respectively.

Neuroimaging evidence has also indicated that listeners recruit additional cognitive resources when processing speech under adverse conditions (for a full review, see Peelle, 2018).

For L2-accented speech specifically, Yi, Smiljanic, and Chandrasekaran (2014) found that areas associated with executive processing (including the inferior frontal gyrus, insula, and anterior cingulate; Peelle, 2017; Adank, Nuttall, Banks, & Kennedy-Higgins, 2015) were more activated for Korean-accented English speech than native English speech, as were sections of the motor and somatosensory cortices. Adank, Davis, and Hagoort (2012) also examined neural activation for unfamiliar accented speech, using a constructed (i.e., artificial) accent. The benefit of using a constructed accent is that it ensures complete novelty of the accent for all participants, however this can also limit the ecological validity of the stimuli, capturing only some aspects of natural accented speech (e.g., phonetic differences) but not others (e.g., prosodic differences). Results of their study showed greater activation of the left superior temporal gyrus and sulcus for unfamiliar accented speech, but not activation in executive processing areas (as was found by Yi et al., 2014). Together, the current neuroimaging evidence appears to be consistent with an account of accented speech perception requiring greater cognitive load than standard speech perception (Adank et al., 2015), but it is unclear how specific accent qualities, such as the difficulty and/or intelligibility of an accent, may impact cognitive load.

Our research question was the following: Does L2-accented speech that is completely intelligible impose greater cognitive load than native speech? Behavioral studies have primarily relied on performance measures such as intelligibility (e.g., the proportion of keywords correctly recognized) to determine the difficulty of L2-accented speech. However, the cognitive load associated with processing speech from highly proficient L2 speakers, including those who are completely intelligible, requires a different methodological approach. To address our research question, we used pupillometry (the measure of pupil diameter over time) to examine cognitive load during the perception of completely intelligible, Mandarin Chinese-accented English speech and standard American-accented English speech. Pupillometry has been used for many decades as a non-intrusive, temporally-sensitive psychophysiological index of cognitive load (Beatty, 1982), and more recently has been applied within the domain of speech perception for examining degraded speech (for a review, see Van Engen & McLaughlin, 2018). Of interest for cognitive research is the task-evoked pupillary response (from here forward simply *pupillary response*), which is a phasic increase of pupil diameter linked temporally to a specific cognitive event (Beatty, 1982). Larger pupillary response during a cognitive task indicates greater cognitive load. These increases in pupil diameter are an order of magnitude smaller than changes due to luminance, and are related to up-regulation of the sympathetic nervous system (Beatty & Lucero-Wagoner, 2000; Peysakhovich, Vachon, & Dehais, 2017).

To the best of our knowledge, this is the first study to use pupillometry to investigate cognitive load during accented speech perception. Using degraded speech, it has been demonstrated that pupillary response increases systematically as speech becomes less intelligible (Zekveld, Kramer, & Festen, 2010). Additionally, Winn, Edwards, and Litovsky (2015) found that even when analyses are limited to completely intelligible trials, there is a systematic increase of pupillary response (i.e., indicating greater cognitive load) as spectral quality of the speech signal decreases. Thus, pupillometry is capable of capturing cognitive load as a function of speech intelligibility, but also revealing more nuanced differences in cognitive load when intelligibility is at ceiling. The latter makes pupillometry the ideal method for addressing our research question.

In Experiment 1, we measured pupillary response for natural speech from a native, standard American-accented speaker of English and a nonnative, Mandarin Chinese-accented speaker of English. We predicted that growth curve analysis (Mirman, 2014) of the pupillary response would show a significant effect of speaker type (i.e., overall pupil diameter would be greater for the L2-accented speaker condition), and a significant interaction between speaker type and the linear time term (i.e., indicating a difference in the rate of change of pupil diameter between conditions), reflecting greater cognitive load for processing L2-accented speech than native speech (see OSF preregistration at: https://osf.io/43wqe/). Additionally, we predicted that the L2-accented speaker would be rated as more effortful to understand than the native speaker, and rated as sounding more accented than the native speaker. The latter self-report was included as a manipulation check to confirm that the highly proficient L2 speaker was recognized as accented. All analysis scripts and raw data are available at: https://osf.io/7dajv/files/.

1.2 Experiment 1

1.2.1 Methods

Participants. Young adult participants (*N* = 30, 20 female and 10 male, ages 18-27) were recruited from Washington University's Psychology Subjects Pool. Eight additional participants were excluded and replaced because of data loss and/or equipment malfunction, or reports that they were not native speakers of American English, had parents who were native speakers of American English, or had extensive exposure to Mandarin Chinese. An audiogram was taken of each participant to confirm that they were of normal hearing. Participation in the study was in exchange for class participation credit.

Stimuli. Stimuli were recordings of a native speaker of standard American English and a Mandarin-Chinese accented speaker of American English reading simple sentences. Sentences were taken from Van Engen, Chandrasekaran, and Smiljanic (2012), and were originally designed to be easy for nonnative speakers to read aloud. All sentences were six words long, with four keywords (e.g., "the *gray mouse ate* the *cheese*"). Recordings of the Mandarin

Chinese-accented speaker used in the current study—along with two additional Mandarin Chinese-accented speakers—were piloted with ten native English participants. The goal of this pilot was to identify an accented speaker who was both extremely intelligible and notably accented. The intelligibility of the sentences was measured based on the proportion of keywords correctly recalled, and the degree of accentedness was determined by ratings on a scale of 1 to 9 (where 1 = "native" and 9 = "extremely foreign-accented"). For the selected accented speaker, mean sentence intelligibility was 99% and the mean accentedness rating was 5.82 (*SD* = 1.92). Root mean squared amplitudes for all stimuli were equalized using the program Praat.

Procedure. Participants provided written informed consent prior to participation in the study. Following the consent process, participants filled out a language background questionnaire and demographic questionnaire. Next, an audiogram was taken of participants in a soundproof booth. Participants were seated on one side of the booth wearing over-ear headphones and holding a response-clicker. Tones were presented at the following frequencies: 250Hz, 500Hz, 1000Hz, 2000Hz, 4000Hz, and 8000Hz. Thresholds were determined by decreasing tone intensity in 10dB intervals until participants could not detect a tone, and then increasing in 5dB intervals until participants could detect the tone again. This process was repeated three times for each frequency in each ear, and the lowest intensity at which participants responded two or more times was recorded. Values at 500Hz, 1000Hz, 2000Hz were used to establish a pure-tone average for each participant.

Next, participants were seated in a quiet room facing a monitor and EyeLink camera. All equipment was set at distances following EyeLink specifications. During the experiment, participants rested their chins on a head-mount. The present experiment was administered prior to two additional pupillometry experiments for another study, which will not be reported here.

Prior to beginning the practice section and the main section of the experiment, the EyeLink camera was calibrated to eye movement. Individuals' right and left pupils have been found to dilate similarly (Winn, Wendt, Koelewijn, & Kuchinsky, 2018); thus, during the task pupil diameter was recorded at 500 Hz from the left eye for all participants. Six stimuli (three from each speaker condition) were presented in the practice session and 60 stimuli (30 from each speaker condition) were presented in the main session. Stimuli were blocked by speaker condition, such that the practice section was comprised of two short blocks (i.e., three stimuli each) and the main section was comprised of four blocks (i.e., 15 stimuli each). All blocking was counterbalanced across subjects. After each block, participants were asked to rate the previous speaker on a scale of 1 to 9 for how effortful they were to understand (where 1 = "not effortful", 5 = "moderately effortful", and 9 = "extremely effortful") and how accented they sounded (where 1 = "native English speaker", 5 = "moderately accented", and 9 = "extremely accented"). Ratings were recorded by a researcher in the room and by a recording device.

The trial-by-trial design is depicted in Figure 1A. Participants were instructed to fixate on a cross located in the center of the screen at all times. Pupil diameter was recorded during all periods in which the color of the cross was red, and participants were asked to reduce their blinking during this window. Each stimulus was preceded by a 2000ms quiet baseline period (used during analysis to establish pupil size prior to input) and followed by a 2000ms delay period. After the recording window, the color of the cross changed to green, cuing the participant to repeat what they heard aloud for the researcher in the room and the recording device. All participants were debriefed after concluding the experiment.

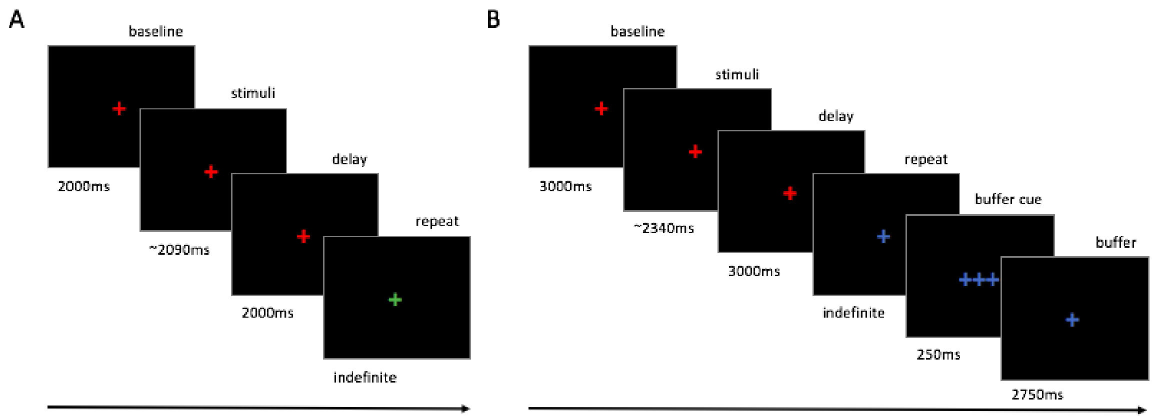

Figure 1. Timing of trial events for Experiment 1 (A) and Experiment 2 (B). In Experiment 1, participants were asked to reduce blinking while the fixation cross was red, and blink freely when the cross was green. In Experiment 2, the contrast color was changed from green to blue to make the task color-blind friendly. Additionally, a buffer period was added between trials to allow more time for the pupil diameter to stabilize.

1.2.2 Analyses

Intelligibility. Intelligibility was calculated as a proportion of keywords correctly identified. Differences in plurality and tense were scored as correct. Trials in which participants did not recognize all of the keywords were excluded from the analyses. For Experiment 1 this amounted to 13 trials total, with no more than one trial for any given participant.

Subjective Responses. For each participant, there were two responses to the effortfulness and accentedness probes for each condition. The mean of these responses was used as a composite rating score. For both probes, quantile-quantile plots of the scores demonstrated that they were not normally distributed. Thus, Wilcoxon rank-sum tests were used in place of t-tests, which we had planned to use (see OSF preregistration: https://osf.io/43wqe/).

Pupil Data Maintenance. A custom script written in R was used to manage pupil data. This maintenance pipeline began by aligning data from each trial at sentence offset, with the intent of aligning the peak responses across trials regardless of stimuli length (Klingner, Tversky,

& Hanrahan, 2011). Next, pupil diameter during the 500ms preceding each target sentence was averaged to determine baseline for each trial. This baseline value was subtracted from pupil diameter throughout the rest of the trial to create a measure of absolute change, and to control for the potential effect of fatigue across the experiment. To deblink the data, intervals of missing values were identified and then expanded to remove noisy readings caused by the eyelids moving prior to (by 30ms) and following (by 160ms) blinks (Winn et al., 2015; Zekveld et al. 2010). This expanded blink interval was then extrapolated across using linear approximation. Lastly, the pupil data was smoothed using a 10Hz moving average filter (Winn et al., 2018). Trials that required more than 50% interpolation were excluded from analyses. Across all participants, less than one trial was excluded on average for this reason, and no participant exceeded 20% trial loss.

Growth Curve Analysis. Growth curve analysis (GCA) was used to model the pupil data in R (version 3.5.1) using the lme4 package. GCA is similar to polynomial regression, but controls for potential collinearity issues by orthogonalizing the polynomial time terms (Mirman, 2014). The window of data used for model fitting began at the departure of pupillary response from baseline (-2327ms) and ended after the approximate peak dilation of both conditions (1000ms). Due to large file sizes, the pupil data was time-binned, reducing the sampling rate from 500Hz to 50Hz. Table 1 shows the full model and its specifications. The time-course of pupillary response was modeled up to the third-order (cubic) orthogonal polynomial. For the speaker condition manipulation, dummy coding specified the native speaker as the reference (i.e., zero) group. The largest random effect structure that would converge included random intercepts and slopes for the subject-by-condition term on each of the orthogonal polynomial terms. Correlations between these slopes and intercepts were included in the random effects (as recommended by Mirman, 2014). This random effect structure captures individual differences in the effect of the condition manipulation on each subject (p.68, Mirman, 2014), and, thus, should capture more variance in the model than the subject random effect would alone. In comparing models, we found the model reported in Table 1 to be the most conservative.

Table 1. Growth curve model code and summary output from Experiment 1. "ot1", "ot2", and "ot3" refer to the linear, quadratic, and cubic polynomial time terms, respectively. Data was dummy coded such that the native speaker condition was the "Intercept" term.

R Code:

```
lmer(Diameter ~    (ot1+ot2+ot3)*Condition +                   # fixed effects
                   (1 + (ot1+ot2+ot3) | Subject:Condition),    # random effects
                        data = PRA_E1_TB50_AOI.df)
```

Term	Estimate	SE	DF	*t*	*p*
Intercept (Native)	78.14	20.95	57.98	3.731	0.000436 ***
ot1	92.50	248.68	58.03	0.372	0.711279
ot2	-924.45	220.37	57.90	-4.195	9.48e-05 ***
ot3	-1299.26	145.03	57.98	-8.959	1.56e-12 ***
Condition (Nonnative)	75.51	29.62	57.98	2.549	0.013468 *
ot1:Condition (Nonnative)	1109.12	351.68	58.03	3.154	0.002552 **
ot2:Condition (Nonnative)	-168.05	311.64	57.90	-0.539	0.591794
ot3:Condition (Nonnative)	982.81	205.09	57.97	4.792	1.19e-05 ***

1.2.3 Results

Pupil Data. Table 1 contains output from the full model, and Figure 2 shows the raw data with the model fit overlaid. As predicted, there was a significant effect of condition (*ß* = 75.51, *SE* = 29.62, *p* = 0.013468), reflecting greater mean pupil diameter for the L2 speaker condition than the native speaker condition. The interactions between condition and the linear and cubic terms were also significant (*ß* = 1,109.12, *SE* = 351.68, *p* = 0.002552; and, *ß* = 982.81, *SE* =

205.09, p = 11.90e-6, respectively). The interaction between condition and the linear term—while consistent with our hypotheses—is likely attributable to systematic difference in stimuli length between the two conditions, and not due to a difference in the rate at which pupil diameter increased (see Figure 2). Similarly, while the interaction between condition and the cubic term may reflect differences in the steepness of the inflection points between conditions (i.e., in the native condition there is a quicker recovery after the pupil diameter peaks; Figure 2), this effect may also be driven by a systematic difference in stimuli length. Thus, interpretation of the differences in the shape of the pupillary response is reserved for Experiment 2, in which speaking rate was controlled for between speakers.

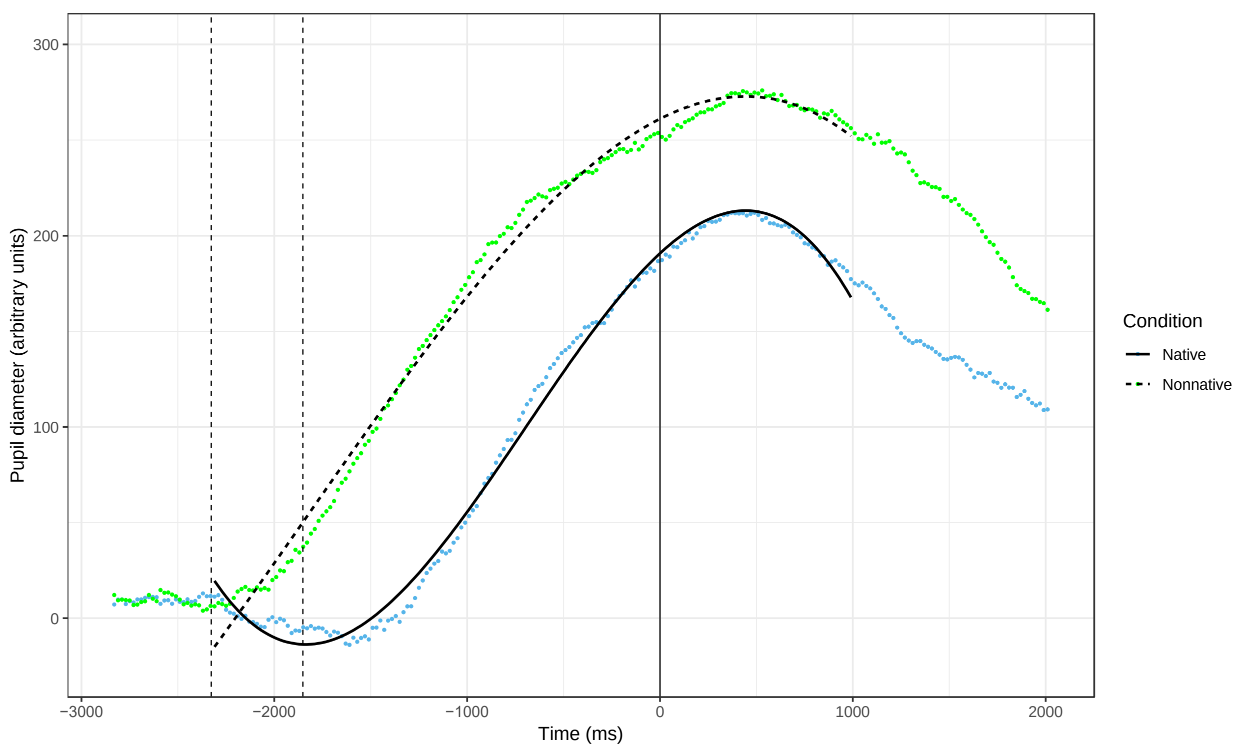

Figure 2. Data from Experiment 1 show larger pupillary response for L2-accented than native speech. Raw data means and standard errors are represented as points and ribbons, and model fit is overlaid with black lines. The solid vertical line represents the aligned offset time for all sentences, and the two dashed vertical lines represent average sentence onset times for the native (rightmost) and L2-accented (leftmost) conditions. The gray box indicates the selected area of interest used for modeling the pupillary response. Pupil diameter is measured in EyeLink's arbitrary units on the y-axis.

Subjective Responses. As shown in Figure 3, the L2 speaker was rated as more effortful to understand than the native speaker (p = 17.45e-7, 95% CI = [-3.00, -2.00]), and as sounding more accented (p = 15.60e-7, 95% CI = [-4.50, -3.75]). Thus, participants recognized the L2 speaker as accented, and they believed them to be more effortful to understand than the native speaker—despite being fully intelligible.

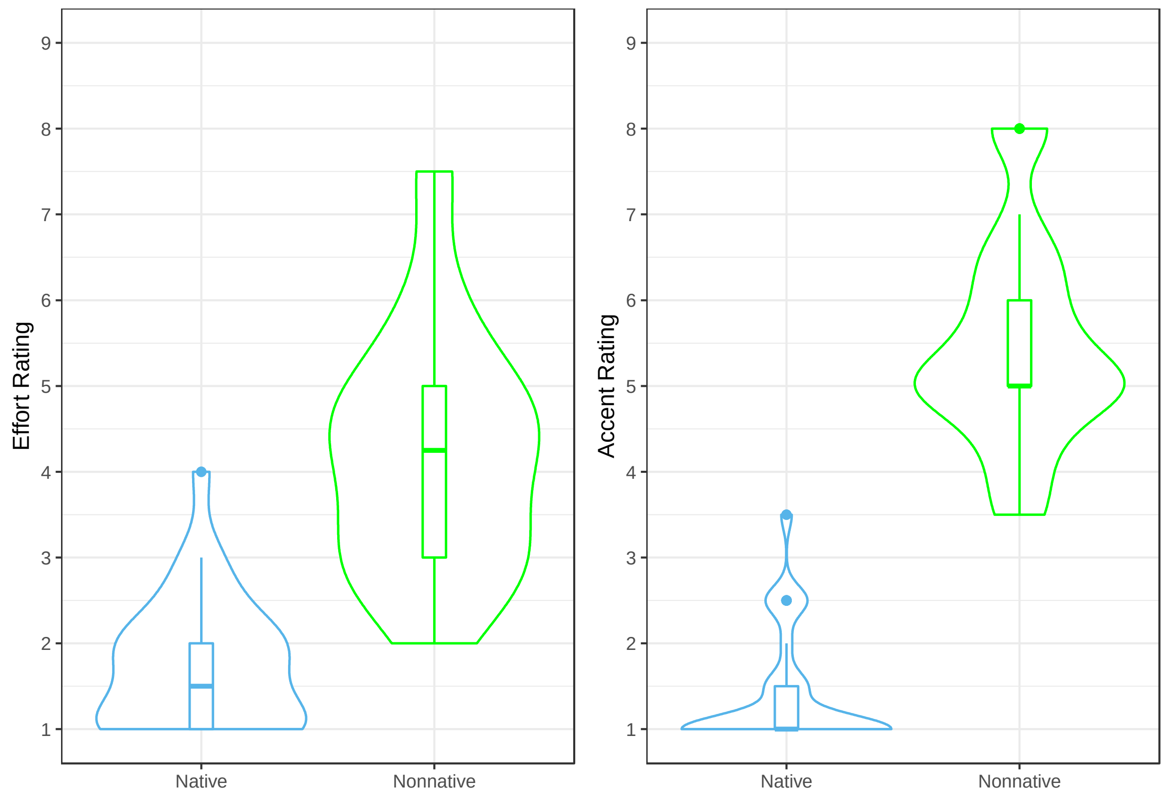

Figure 3. Subjective response data from Experiment 1 are represented using a combination of box-and-whisker and violin (i.e., density) plots. Effortfulness and accentedness probes were presented on a scale of 1 to 9 where: 1 = "not effortful", 5 = "moderately effortful", and 9 = "extremely effortful"; and, where 1 = "native English speaker", 5 = "moderately accented", and 9 = "extremely accented," respectively.

1.3 Experiment 2

In Experiment 2, we conducted a conceptual replication of Experiment 1 in which the native speaker stimuli were acoustically lengthened to match the length of the L2 speaker stimuli – thus controlling for natural differences in speaking rate. This allowed us to rule out alternative explanations for the results obtained in Experiment 1 (i.e., a confound with speaking rate) and replicate the effect found with a larger sample size determined using a post-hoc power analysis of Experiment 1. For Experiment 2, we again hypothesized that pupillary response would be greater for listening to L2-accented speech than native speech. Specifically, we predicted greater steepness of the inflection point following sentence offset (i.e., indicating a difference in the rate of change of pupil diameter between conditions) and greater overall mean pupil diameter for the L2-accented speaker condition than the native speaker condition. Additionally, we predicted that subjective ratings of listening effort would be greater for the L2-accented speaker condition (see OSF registration at: https://osf.io/bgcx7/). All analysis scripts and data are available at: https://osf.io/7dajv/files/.

1.3.1 Methods

Participants. Participants were young adults (*N* = 52, ages 18-22, 39 female and 13 male) recruited from Washington University's Psychology Subjects Pool. Hearing thresholds were measured for each participant to confirm that they were of normal hearing. An additional 11 subjects participated and were excluded and replaced due to data loss, program malfunction, or because they did not meet the eligibility criteria. All qualifications remained the same as in Experiment 1. The target sample size was increased to 50 participants based on a post-hoc power analysis of Experiment 1. Power analyses for GCA require a simulation-based approach, and our simulation was limited due to convergence issues. However, the results did indicate that power for the effect of condition on the quadratic term could possibly benefit from increased sample size.

Stimuli. The same stimuli set described in Experiment 1 was used in Experiment 2. To control for natural differences in speaking rate, the native speaker files were acoustically lengthened. Adjustments were made using the Stretch function in Adobe Audition CC, which was set to adjust file lengths while controlling for pitch and speech characteristics. Files were all lengthened by a constant amount, so that the average length of the native speaker files would equal the average length of the L2 speaker files. The L2 speaker files were not altered.

Procedure. The general trial-by-trial design of Experiment 2 matched Experiment 1 (Figure 1B). However, between each trial a buffer period was added to allow additional time for phasic pupillary response to recover. The color of the fixation cross during the response and buffer period was also changed to blue, making the color contrast with red color-blind friendly.

Stimuli were presented in a randomized order in an attempt to remove any possible blocking effects. For the subjective ratings of listening effort, participants were probed every three trials about the last speaker they had listened to and used the keyboard to respond, resulting in approximately 10 responses per condition. The practice section of the experiment was replaced with a visual walkthrough. Lastly, the location of the eye-tracker was changed to a soundproof booth for Experiment 2, and an audio recorder was used to record the sentences repeated aloud (i.e., no researcher was present in the room).

1.3.2 Analyses

All subjective response and intelligibility data were handled in the same manner as in Experiment 1.

Pupil Data Maintenance. The custom R script used to manage pupil data in Experiment 1 was altered to align data at sentence onset, but otherwise remained the same. Across all participants, approximately one trial was excluded on average and no one participant exceeded 20% trial loss.

Growth Curve Analysis. Experiment 2 followed the same method of analysis described in Experiment 1. The window of data used for model fitting began at stimuli onset time (0ms) and ended after the approximate peak dilation of both conditions (3344ms). Table 2 shows the R code used in the full model, which included orthogonal polynomials up to the third-order (cubic). The random effect structure included random intercepts and slopes for the following: item, subject, and subject-by-condition.

Table 2. Experiment 2 code for growth curve analysis and summary output. "ot1", "ot2", and "ot3" refer to the linear, quadratic, and cubic polynomial time terms, respectively. Data was dummy coded such that the native speaker condition was the "Intercept" term.

R Code:

```
lmer(Diameter ~    (ot1+ot2+ot3)*Condition +                      # fixed effects
                   (1 + (ot1+ot2+ot3) | Subject) +                # random effects
                   (1 + (ot1+ot2+ot3) | Item) +                   # random effects
                   (1 + (ot1+ot2+ot3) | Subject:Condition),       # random effects
                   data = PRA_E2R_TB50_AOI.df)
```

Term	Estimate	SE	DF	*t*	*p*
Intercept (Native)	181.31	17.37	91.89	10.438	< 2e-16 ***
ot1	1045.80	112.15	97.88	9.325	3.57e-15 ***
ot2	-569.64	65.55	117.58	-8.690	2.50e-14 ***
ot3	-156.15	34.91	123.22	-4.472	1.73e-05 ***
Condition (Nonnative)	62.43	15.31	105.54	4.079	8.81e-05 ***
ot1:Condition (Nonnative)	408.28	102.29	106.20	3.992	0.000121 ***
ot2:Condition (Nonnative)	29.70	69.52	83.26	0.427	0.670334
ot3:Condition (Nonnative)	17.67	42.21	89.33	0.419	0.676559

1.3.3 Results

Pupil Data. The summary output from the full model is reported in Table 2, and the model fit is shown in Figure 4. We predicted that there would be greater overall pupillary response and a steeper increase of pupillary response (i.e., a greater rate of change) for the L2 speaker condition than the native speaker condition. The results confirmed both of these predictions, with a significant main effect of condition (*ß* = 62.43, *SE* = 15.31, *p* = 8.81e-05) and a significant interaction between the linear polynomial term and condition (*ß* = 408.28, *SE* = 102.29, *p* = 0.000121). The interactions between condition and the quadratic and cubic terms were not significant (*ß* = 29.70, *SE* = 69.51, *p* = 0.670334; and, *ß* = 17.67, *SE* = 42.21, *p* = 0.676559, respectively), indicating that the shape of the pupillary response for each condition was similar. Overall, the pattern of results in the full model clearly demonstrate larger pupillary response—and, thus, greater cognitive load—for L2-accented speech.

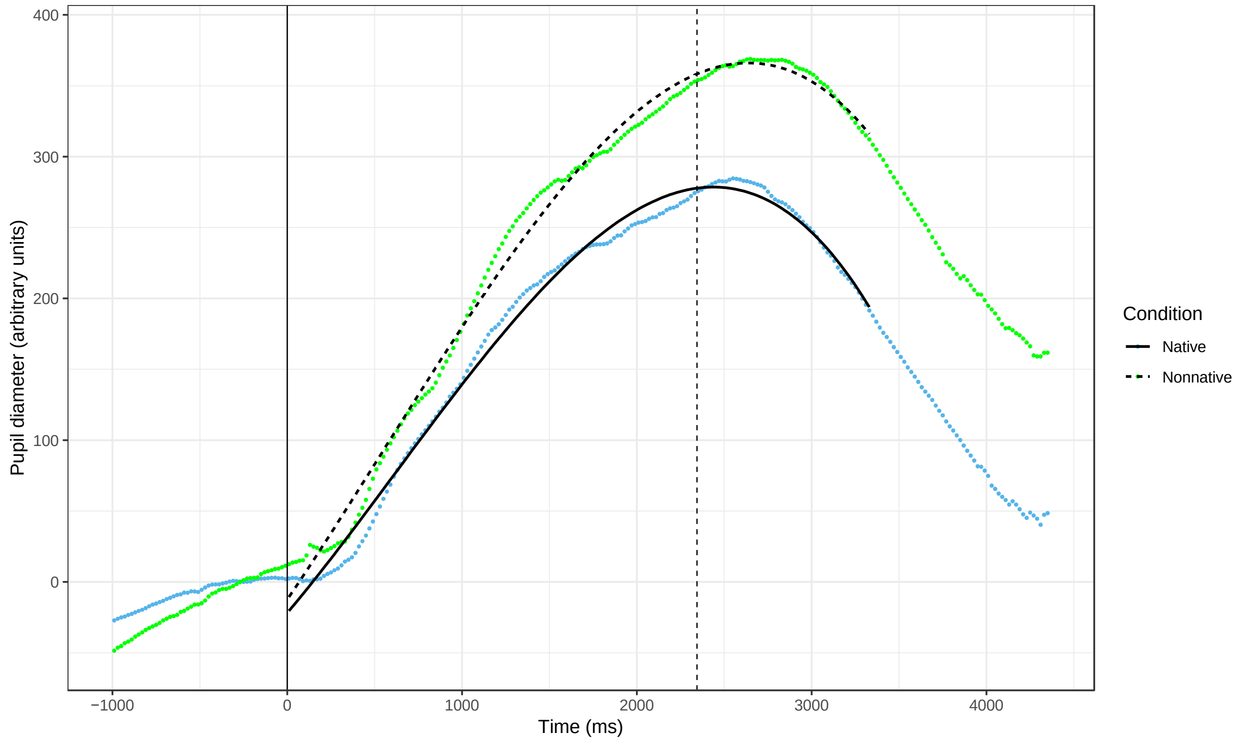

Figure 4. Data from Experiment 2 replicate the findings from Experiment 1 while controlling for speaking rate between conditions. Points and ribbons are used to represent the raw data means and standard errors, and model fit is overlaid with black lines. In Experiment 2, the sentences were aligned at onset time (i.e., 0ms) for analysis, represented here as a solid black line. The dashed vertical line represents the average sentence offset time for both conditions. The gray box indicates the selected area of interest used for modeling the pupillary response. Pupil diameter is measured in EyeLink's arbitrary units on the y-axis.

Subjective Responses. As was found in Experiment 1, in Experiment 2 the L2 speaker was rated as more effortful to understand than the native speaker (p = 3.417e-09, 95% CI = [-2.03, -1.31]; Figure 5).

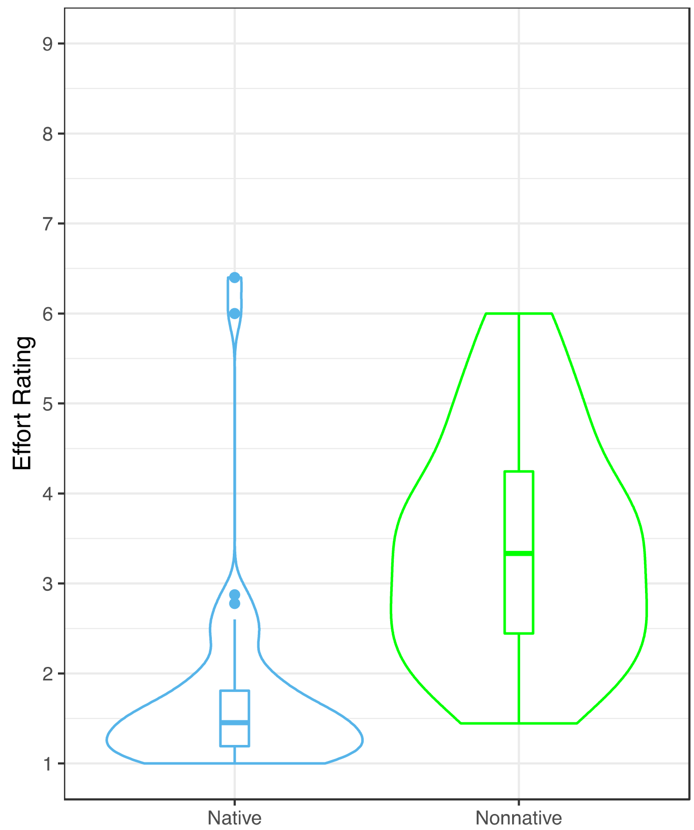

Figure 5. In Experiment 2 participants were probed every three trials to rate the last speaker they had listened to on a scale of 1 to 9 for how effortful they were to understand, where: 1 = “not effortful”, 5 = “moderately effortful”, and 9 = “extremely effortful.” Here, the subjective response data are represented using a combination of box-and-whisker and violin (i.e., density) plots.

1.4 Discussion

The results of the current study indicate that processing completely intelligible L2-accented speech requires greater cognitive load than native speech. These findings are consistent with an executive recruitment account for accented speech, under which we would predict that mismatches between the systematic speech patterns of a speaker and a listener will require greater cognitive load to process—even when recognition accuracy is at ceiling. In particular, working memory may be key for resolving perceptual ambiguities in accented speech. Evidence indicates that working memory capacity is positively correlated with intelligibility for accented speech perception in young and older adults (McLaughlin et al., 2018; Janse & Adank, 2012). Current speech perception models, such as the Ease of Language Understanding model (Rönnberg, Rudner, Foo, & Lunner, 2008; Rönnberg, et al., 2013), propose a dual role of working memory during speech perception, emphasizing that it is key for storage of the unfolding speech stream and for top-down processing of phonetic and semantic ambiguity. A recent study by Chan, Chiu, Dailey, and Jalil (2019) found evidence in support of this hypothesis, demonstrating that serial recall for lists of L2-accented words is poorer than recall for native-accented words. Notably, the authors found that the amount of errors for L2-accented word lists was reduced by increasing the inter-stimulus interval of the task, indicating that additional time for top-down processing of ambiguous L2-accented words helps to reduce pressure on the shared pool of working memory. Our results indicate that even processing completely intelligible L2-accented speech may exhaust working memory resources more than native speech, though this remains to be investigated directly.

Our findings also indicate that listeners may be aware that processing L2-accented speech is more cognitively demanding than processing native speech. In both experiments, participants rated the highly proficient L2 speaker as significantly more effortful to understand than the native speaker, as was found by Munro and Derwing (1995). However, while these subjective ratings match the physiological evidence, it is also possible that they reflect participants' expectations—and/or biases—regarding accented speech, and not their awareness of increased cognitive load. In future experiments, the relationship between subjective ratings of listening effort and pupillary response for each trial could be examined to try and tease this apart.

Pupillometry could also be a valuable tool in the future for investigating perceptual adaptation to accented speech. Behavioral evidence has shown that listeners rapidly adapt to L2-

accented speech, resulting in improved intelligibility over time for speakers with the same accent (Bradlow & Bent, 2008), and even for speakers with different accents (i.e., accent-independent adaptation; Baese-Berk, Bradlow, & Wright, 2013). Clarke and Garrett (2004) also found that adaptation could occur within just two to four sentences, at which point processing delays imposed by L2-accented speech were attenuated. Cognitive load, however, has yet to be examined as a function of perceptual adaptation to accented speech. If native listeners' speech representations are being shifted during perceptual adaptation to accommodate a given accent, then the mismatch between the accented speech input and these representations should be reduced over time, resulting in reduced cognitive load. Experiments utilizing pupillometry not only have the potential to fill this gap in the literature, but could also open an entirely new avenue for researching adaptation to highly proficient accented speakers.

In summary, our findings indicate that processing L2-accented speech requires greater cognitive load than processing native speech, as evidence by greater and more rapid pupillary response. Most notably, recognition accuracy for the L2-accented speech stimuli used in the current study was at ceiling, indicating that the increased cognitive load for L2-accented speech was not related to intelligibility. Rather, we propose that mismatches between the speech patterns of L2-accented speech and native listeners' representations require greater support from executive resources (such as working memory) to reconcile. Additionally, while the accented stimuli in the present study were recordings of a Mandarin-Chinese accented speaker of American English, we would predict these findings to replicate with other L2 and regional accents as well. If—as we have proposed—increased cognitive load stems from the systematic mismatches between the speech patterns of a speaker and a listener, then there should be greater pupillary response for other accent types also. Testing the boundaries of this prediction will be an important next step for this area of inquiry.

References

Adank, P., Davis, M. H., & Hagoort, P. (2012). Neural dissociation in processing noise and accent in spoken language comprehension. *Neuropsychologia, 50*(1), 77-84.

Adank, P., Nuttall, H. E., Banks, B., & Kennedy-Higgins, D. (2015). Neural bases of accented speech perception. *Frontiers in human neuroscience*, *9*, 558.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America, 133*(3), EL174-EL180.

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, *137*(4), 2015-2024.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin, 91*(2), 276.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of psychophysiology, 2*, 142-162.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Chan, K. Y., Chiu, M. M., Dailey, B. A., & Jalil, D. M. (2019). Effect of Foreign Accent on Immediate Serial Recall. *Experimental psychology*.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647-3658.

Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *The Quarterly Journal of Experimental Psychology, 65*(8), 1563-1585.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48*(3), 323-332.

McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., & Van Engen, K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. . *Attention, Perception, & Psychophysics, 80*(6), 1559-1570.

Mirman, D. (2014). *Growth curve analysis and visualization using R.* Boca Raton, FL: CRC Press.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning, 45*(1), 73-97.

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204.

Peysakhovich, V., Vachon, F., & Dehais, F. (2017). The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load. *International Journal of Psychophysiology, 112*, 40-45.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., … & Naylor, G. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing, 37*, 5S-27S.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., ... & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in systems neuroscience, 7*, 31.

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology, 47* (sup2), S99-S105.

Van Engen, K. J., & McLaughlin, D. J. (2018). Eyes and ears: Using eye tracking and pupillometry to understand challenges to speech recognition. *Hearing research*.

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in human neuroscience, 8*, 577.

Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PLoS One, 7*(9), e43753.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and hearing, 36* (4), e153.

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in hearing, 22*, 2331216518800869.

Yi, H. G., Smiljanic, R., & Chandrasekaran, B. (2014). The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in human neuroscience, 8*, 768.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and hearing, 31*(4), 480-490.