Published in final edited form as:
Psychophysiology. 2013 January ; 50(1): 23–34. doi:10.1111/j.1469-8986.2012.01477.x.

Pupil size varies with word listening and response selection difficulty in older adults with hearing loss

Stefanie E. Kuchinsky[1], **Jayne B. Ahlstrom**[1], **Kenneth I. Vaden Jr.**[1], **Stephanie L. Cute**[1], **Larry E. Humes**[2], **Judy R. Dubno**[1], and **Mark A. Eckert**[1]

[1]Medical University of South Carolina

[2]Indiana University Bloomington

Abstract

Listening to speech in noise can be exhausting, especially for older adults with impaired hearing. Pupil dilation is thought to track the difficulty associated with listening to speech at various intelligibility levels for young and middle-aged adults. This study examined changes in the pupil response with acoustic and lexical manipulations of difficulty in older adults with hearing loss. Participants identified words at two signal-to-noise ratios (SNRs) among options that could include a similar-sounding lexical competitor. Growth Curve Analyses revealed that the pupil response was affected by an SNR-by-lexical competition interaction, such that it was larger and more delayed and sustained in the harder SNR condition, particularly in the presence of lexical competition. Pupillometry detected these effects for correct trials and across reaction times, suggesting it provides additional evidence of task difficulty than behavioral measures alone.

Nearly two-thirds of people 70 years of age and older have hearing loss (Lin, Thorpe, Gordon-Salant, & Ferrucci, 2011) and the proportion of older adults in the U.S. population is expected to double over the next 20 years. Hearing loss increases the difficulty of speech recognition, particularly in background noise (e.g., Plomp, 1994). Even when speech recognition for hearing-impaired and normal-hearing adults is similar, the ease with which the task is accomplished may differ substantially among individuals (Fraser, Gagne, Alepins, & Dubois, 2010; Gatehouse & Gordon, 1990; Hällgren, Larsby, Lyxell, & Arlinger, 2005). Indeed, for older adults with hearing loss, the experience of feeling tired or frustrated listening to speech in noisy environments is all too familiar (Edwards, 2007). Characterizing differences in the task demands experienced during speech recognition may help explain why certain individuals are more likely to discontinue participation in hearing loss interventions and may help in tailoring individualized training programs or hearing aid fittings.

Increased effort and fatigue have been associated with both hearing loss (Hetu, Riverin, Lalande, Getty, & St-Cyr, 1988; Hicks & Tharpe, 2002; McCoy et al., 2005; Nachtegaal et al., 2009) and aging (Anderson Gosselin & Gagne, 2011; Piquado, Isaacowitz, & Wingfield, 2010), based on purported metrics of effort that included self-report, slowed reaction times on dual-task interference paradigms, reduced memory for speech, and physiological measures. Older adults with clinically normal hearing appear to exert more task-related effort than younger adults perhaps because of declining cognitive (Salthouse, 1996;

Corresponding Authors: Stefanie E. Kuchinsky and Mark A. Eckert, Address: 135 Rutledge Ave, MSC 550, Charleston, SC 29425, Phone: 843-792-2774 (SEK); 843-792-5914 (MAE), kuchins@musc.edu; eckert@musc.edu.
Address reprint requests to: Stefanie E. Kuchinsky (kuchins@musc.edu) or Mark A. Eckert (eckert@musc.edu), Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, 135 Rutledge Ave., MSC 550, Charleston, SC 29425.

Salthouse, Babcock, & Shaw, 1991; Wingfield, Stine, Lahar, & Aberdeen, 1988) and neural resources (Dennis & Cabeza, 2008; Eckert, 2011).

Previous studies that examined subjective measures of listening effort have typically included self-assessment scales (e.g., Hetu et al., 1988; Nachtegaal et al., 2009; see also Humes, 1999 for examples). However, there are at least two limitations to self-report measures of effort. First, people may vary in their ability to introspect and be unbiased in responding. Second, surveys are filled out after task completion, thus increasing the likelihood that perceived effort may not accurately be recalled and that responses will reflect an average of perceived effort across many trials and many different conditions. These factors may contribute to the lack of consistent correlations between subjective ratings and task demands and with other purported metrics of effort (e.g., Hicks & Tharpe, 2002, Mackersie & Cones, 2011).

In contrast, physiological measurements can be unobtrusively obtained throughout a task, are free from subjective biases, and have been shown to relate to task demand (e.g., skin conductance and electromyography, Mackersie & Cones, 2011). In particular, research has shown that pupils dilate with increasing task demands (Kahneman & Beatty, 1966; Piquado et al., 2010; Zekveld, Kramer, & Festen, 2010, 2011) and response selection difficulty (Laeng, Orbo, Holmlund, & Miozzo, 2011), as the pupillary response is controlled in part by the locus coeruleus (LC) attention system (Aston-Jones & Cohen, 2005; Gilzenrat, Cohen, Rajkowski, & Aston-Jones, 2003; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Rajkowski, Kubiak, & Aston-Jones, 1993). While more evidence is needed to clarify the link between how demanding a task is and how much cognitive effort is exerted to complete it (cf. “evaluation of demands on capacity” and “allocation policy” of Kahneman’s capacity model for attention; Kahneman, 1973), task related pupil dilation is thought to reflect variation in cognitive control and attention (Jepma & Nieuwenhuis, 2011; Moresi et al., 2008; Verney, Granholm, & Marshall, 2004).

Recently, pupillometry has been used to study the pupil response during measures of speech recognition in difficult listening environments. Zekveld and colleagues (2010) examined the pupillary response of younger adults with normal hearing who listened to sentences presented in noise at three signal-to-noise ratios (SNRs). With decreasing SNR, mean pupil dilation, peak pupil dilation, and the time-to-peak latency increased. Self-reported ratings of effort also increased with decreasing intelligibility. In a separate study, Zekveld and colleagues (2011) tested speech recognition in middle aged adults and found that pupil size decreased with increasing sentence intelligibility, but to a lesser degree for hearing impaired individuals. This was interpreted to indicate that adults with hearing loss experienced greater task demands even in relatively easy listening conditions.

While manipulations of SNR, memory load, and response selection difficulty have been shown to influence pupil size, pupil size also varies with the lexical properties of words. For example, correctly recognizing words with lower lexical frequency is associated with larger pupil dilation than correctly recognizing words with higher lexical frequency in younger adults (Kuchinke, Vo, Hofmann, & Jacobs, 2007; Papesh & Goldinger, 2012). Additionally, behavioral studies have shown that the likelihood of correctly identifying a word decreases as its neighborhood density (the number of words that sound similar to it) increases, for both listeners with impaired and normal hearing (Dirks, Takayana, & Moshfegh, 2001; Dirks, Takayanagi, Moshfegh, Noffsinger, & Fausti, 2001). However, the influence of lexical competition on speech recognition performance has been shown to vary with age. Older adults are more negatively affected by lexical competition than younger adults (Sommers, 1996; Sommers & Danielson, 1999; Taler, Aaron, Steinmetz, & Pisoni, 2010). For example, Ben-David, Chambers, Daneman, Pichora-Fuller, Reingold, and Schneider (2011) used eyetracking measures to demonstrate that older adults were less likely to preferentially fixate a target word compared to a rhyming word, particularly in background noise. This was true, even though mean levels of performance were equivalent across age. These results suggest that response selection difficulty associated with lexical competition may compound the listening difficulty experienced by older adults.

Extending previous work, our study of older adults examined variation in listening difficulty by measuring percent correct word identification, reaction time (RT), and pupil size in response to both acoustic listening difficulty and lexical response selection difficulty. Pupillometry data were sampled across the duration of each trial, allowing for the examination of the impact of task difficulty manipulations across the time course of the pupil response. Using an orthographic version of the Visual World Paradigm (VWP), we examined changes in the shape and timing of the pupil response associated with the difficulty of identifying an aurally-presented word at different SNRs as well as with selecting that word from a set of orthographically-presented options that could include a lexical competitor.

To the extent that behavioral measures are sensitive to acoustic and lexical response selection manipulations of task difficulty, increased difficulty (poorer SNR or the presence of a lexical competitor) was predicted to result in decreased word identification and increased RT. We assessed the relative sensitivity of these measures, given the evidence that RT decreases during easier listening conditions (aided vs. unaided speech), even when speech recognition is similar (Gatehouse & Gordon, 1990; though cf. Hällgren et al., 2005).

We also examined the extent to which pupillometry would provide additional evidence of task difficulty beyond word identification and reaction times. The benefit of collecting pupil size metrics in addition to behavioral measures is that dilation is an involuntary physiological response that can be recorded across an entire experiment and is not under direct control (compared to behavioral measures such as reaction time). This kind of metric, along with the appropriate statistical analyses, provides a means to examine changes in difficulty throughout task processing.

The current study paradigm (orthographic VWP) had participants listen to words at two SNRs, which were selected to be easier and harder than a baseline SNR that elicited a moderate level of speech recognition performance. Participants selected the word they heard from a set of options, which either did or did not contain a competitor word, from a touchscreen computer. The harder relative SNR (rSNR) was predicted to result in greater average pupil size that was delayed to peak and remained elevated following the peak (i.e., more sustained). Because response selection difficulty has been shown to influence the pupil response (Laeng et al., 2011), lexical competition (phonological overlap among response options) was also predicted to result in a larger average pupil size, increased delay to peak, and a sustained, elevated response following the peak. An interaction between rSNR and lexical competition was hypothesized to the extent that listening during a harder rSNR condition increases the difficulty of distinguishing among similar sounding words. To evaluate these hypotheses, we used Growth Curve Analysis (GCA), a statistical technique that is appropriate for analyzing changes in metrics over time (Magnuson, Dixon, Tanenhaus, & Aslin, 2007; Mirman, Dixon, & Magnuson, 2008). We examined these GCA results with respect to behavioral measures of task difficulty (percent correct word identification and RTs) and to traditional pupil size measures that involve conducting analysis of variances (ANOVAs) on amplitudes and latencies obtained from peak-picked data within a particular time window of interest.

In sum, we examined the extent to which the pupil response, a physiological response, indexes both acoustic listening difficulty and lexical response selection difficulty, independently of differences in behavior in a sample of older adults with hearing loss. In this way, we assessed the extent to which the difficulty experienced in acoustic conditions is compounded by lexical competition in hearing-impaired older adults.

Method

Participants

Twenty-one healthy older adults (10 female) age 61 to 88 [*M* = 73.14, *SD* = 8.22 years] participated in this experiment as part of a larger study assessing a speech perception training protocol. All were monolingual native speakers of American English and reported normal or corrected-to-normal visual acuity. Average IQ (FSIQ) as measured by the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999) was 118.90 [*SD* = 11.21] and degree of handedness of the Edinburgh handedness questionnaire (Oldfield, 1971) was 95.00 [*SD* = 8.06] on a scale of 100 (maximally right-handed) to −100 (maximally left-handed). Participants provided written informed consent before participating in the Medical University of South Carolina Institutional Review Board approved study.

Audiometry

Pure-tone thresholds at standard intervals between 0.25 and 8.0 kHz were measured with a Madsen OB922 clinical audiometer calibrated to appropriate ANSI standards (American National Standards Institute, 2004) using TDH-39 headphones. Participants all had mild to moderate high-frequency sensorineural hearing loss (Figure 1) and could be considered candidates for hearing aids. However, no participant had previous experience with hearing aids. Asymmetry between the ears was limited to 15 dB at each frequency. For most (16) participants, stimuli were presented to the right ear. If sufficient audibility (to be discussed below) could not be assured for the right ear, stimuli were presented to the left ear (5 participants).

Equipment

Participants sat 12″ away from a 19″ Elo touchscreen monitor with a 60 Hz refresh rate, allowing them to comfortably reach the screen to respond. The illumination of the room was kept constant across the experiment for all individuals, dimmed to 160 lux. Stimulus presentation was controlled by MATLAB 2009b running the Psychophysics Toolbox v.3.0.8 on a Dell Optiplex 780 computer. Auditory stimuli were presented through a Tucker Davis Technologies (TDT) System III Real-Time Processor (RP2.1), attenuated by a headphone buffer (TDT HB7), and delivered to one of a pair of ER-3A insert earphones (though both earphones were inserted).

Following eyetracker setup and calibration, an Arrington Research PC-60 ViewPoint EyeTracker® using software v.2.8.5 sampled pupil size and position data at 60 Hz using the dark pupil method (www.ArringtonResearch.com). Because of individual variations in the distance of the camera to the eye, pupil measurements were recorded as proportions of the screen size rather than in millimeters. Participants sat with their heads in a fixed position and a slip-correction phase, in which participants refixated one of the calibration points, occurred periodically to allow for the correction of slight head movements.

Pupil height and width were measured with better than .03 mm resolution with no averaging. The ViewPoint software fits an ellipse to the pupil in real time, for each sample (every 16.67 ms) outputting the length of the major axis (longest length, which at central fixation is the pupil width) and minor axis (shortest length, which at central fixation is the pupil height).

While the minor axis becomes small as the eye rotates away from the camera, the major axis does not, resulting in accurate pupil size measurements during eye movements. Calculations confirmed this result, as the average Pearson correlation across the study participants between pupil size and proximity to the lower right side of the screen (the eyetracker camera location) was .03 for the major axis [two-tailed z-test for Fisher's converted r-to-z scores, $z = .47$, n.s. different from 0] but .25 for the minor axis [$z = 5.39$, $p < .001$]. Thus, only the major axis was used as the measurement of pupil size. Additionally, we observed that the average number of fixations did not vary across any of the conditions [rSNR, $F(1, 20) < 1$, lexical competition, $F(1, 20) < 1.45$, $p > .05$; interaction, $F(1, 20) < 1$; all post-hoc paired t-tests, $p > .10$], thus precluding variations in fixation frequency as the source of rSNR and lexical competition effects on the pupil response.

Materials

Speech and background noise were spectrally shaped for each individual (i.e., levels of speech and noise at each one-third octave band frequency were at least 20 dB higher than threshold) to ensure audibility through 4.0 kHz, thus simulating a well-fit hearing aid (Burk & Humes, 2008; Humes, Burk, Strauser, & Kinney, 2009). The noise consisted of two-talker International Collegium for Rehabilitative Audiology (ICRA) vocoded noise (Dreschler, Verschuure, Ludvigsen, & Westermann, 2001), which is spectrally shaped like speech but completely unintelligible. Each word was presented at three rSNRs: baseline, easier (2 dB better than baseline SNR) and harder (2 dB worse than baseline SNR). Baseline SNR was determined through an initial open-set word recognition task in which participants repeated aloud words presented with ICRA noise. The default baseline SNR was −2 dB and was adjusted if percent correct recognition for an individual fell outside the 30–70% range. This procedure minimized floor or ceiling effects and provided a moderate level of task difficulty, such that individuals would be unlikely to not try or give up. Thus, the baseline SNR was −2 dB for 15 participants, −1 dB for 5 participants, and 0 dB for 1 participant. Open-set word recognition scores at this individually tailored baseline SNR ranged from 36% to 53% ($M = 47\%$, $SD = 5\%$). Only the easier and harder rSNR conditions were analyzed in this study because participants were exposed to the baseline SNR prior to the eyetracking session.

Each trial required participants to make a four-item forced-choice response between the target word and three alternatives presented orthographically onscreen. In addition to the target word, for lexical competitor trials, the four items contained a foil word that shared the vowel and the initial or final consonant cluster with the target, and two filler items that did not overlap with the target or foil word. The no-competitor trials contained only the target and three filler words. For half of the no-competitor trials, two of the three filler items shared the vowel and a consonant cluster with one another so that the presence of two similar sounding words in the response option set was not a cue that one of those two options was the target. Words were presented in rSNR blocks of 8 words, with an equal number of trials (32) in each of the four possible rSNR-by-lexical competition conditions, for a total of 128 trials that were entered into these analyses.

The order of rSNR presentation was balanced within participants, such that each rSNR condition block appeared before and after each other rSNR block the same number of times across the experiment for each participant. Thus, while each word was presented exactly once in each rSNR (3 times total), the order of the rSNR presentation varied from item to item. In particular, there were six possible orderings of the three rSNR conditions (e.g., easier then baseline then harder versus easier then harder then baseline, etc.). Words were evenly distributed across each of these six possibilities within each participant. No target word was presented within 60 trials of its previous presentation. The order of the lexical competition manipulation was pseudo-random, such that neither lexical competition condition (competitor, no-competitor) appeared more than twice in a row. No individual word appeared in the same rSNR or spatial location on the touchscreen as it had appeared in previously. Additionally, the spatial location of the correct response was never the same more than two trials in a row. Because the design was balanced within participants, each participant received the same order of events.

The 256 words that made up these four-item response sets were selected from a list of the 600 most frequent words in English on which participants would be trained in the speech-perception training program. All words were spoken by a young adult male talker who was a native speaker of Midland North American English (see Humes et al., 2009 for additional recording details). All words presented in a given four-item response set had the same number of syllables (one or two) and the same number of letters (two to seven letters). Target and foil words were balanced for (log) SUBLEX$_{US}$ word frequency (Brysbaert & New, 2009), orthographic and phonological word density, and the frequency of orthographic and phonological neighbors (Balota et al., 2007), such that there were no significant differences on these measures between target and foil words (all $p > .05$). Following standard VWP design, two stimuli lists were created so that the target word on the first list became the foil word for a competitor trial or the filler word for a no-competitor trial on the second list (and vice versa). This further ensured that lexical properties of the target words did not drive the pattern of results. Twelve participants were run on list one and nine participants on list two.

Procedure

The study paradigm comprised the sequence of events presented in Figure 2. Using the orthographic VWP (McQueen & Viebahn, 2007; Salverda & Tanenhaus, 2010), each trial began with the participants fixating a small black circle (.80° × .80° of visual angle) in the center of the display. Background noise began to play at 1000 ms, while the circle remained on the screen. At 1800 ms into the trial, the array of orthographic options appeared for 200 ms prior to word onset, as with previous VWP studies, allowing participants to preview the location of the words onscreen, but preventing them from strategically reading the words and accessing their phonological codes prior to hearing the target. Words were printed in 24 point Courier font in all caps so that each letter subtended the same amount of visual space (1.19° × 1.19° of visual angle per letter). They were positioned on the screen such that the center of each word (which were all the same length on a given trial) was 4″, or 18.93° of visual angle away from the central fixation circle. This configuration required that participants make an eye movement away from central fixation in order to read the words, as they appeared well outside the 2 – 6° of visual angle that make up the parafovea. At 2000 ms into the trial, a word was presented in the ongoing background noise. Participants were instructed to start by fixating the circle and then to select the word they heard from the visual array as quickly and accurately as possible by touching the screen. The next trial began immediately following their response, thus allowing for 2000 ms in between response selection and the subsequent word presentation onset. Four practice trials were completed before the task began.

Analyses

Behavioral—Trials for which the RT was more than 2.5 SD greater than that individual's mean were excluded from all analyses. This resulted in an average of 3.09% excluded trials [*SD* = 1.79%] per participant.

Separate 2 × 2 repeated measures ANOVAs were performed to examine the impact of rSNR and lexical competition on word identification scores and log-transformed reaction time (log RT) measures (Huynh-Feldt corrections of repeated measures ANOVAs yielded identical statistical results throughout the paper). Two-tailed t-tests were computed to clarify significant interaction effects.

Pupillometry

Preprocessing: Pupil data preprocessing methods were similar to those previously reported in the literature (Piquado et al., 2010; Zekveld et al., 2010, 2011). The output of the eyetracker data files included the major axis length of the pupil at each time sample (i.e., every 16.67 ms for this 60 Hz eyetracker) as well as a data quality code that indicated whether the pupil was successfully located. Samples for which the pupil was not accurately detected or for which the pupil aspect ratio was less than one standard deviation below the mean for that trial (signaling a blink; Piquado et al., 2010) were removed and interpolated across. Trials for which more than 50% of the data required interpolation were removed from pupillometry analyses (Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003), which was on average 13.57 trials [*SD* = 21.97] per person (approximately 11% of trials), though on average only 25% [*SD* = 11%] of trial time points required interpolation. The number of time points requiring interpolation was not correlated with the average number of fixations per trial [$r(19) = .11$, $p > .05$] nor the number of fixations within any particular experimental condition [all $p > .05$]. Data were then smoothed with a five-point moving average (across an 83.35 ms window).

Finally, pupil data were rescaled and baseline corrected. Mean pupil size across trials was on average positively scaled with its standard deviation within individuals (*M r* = .61). The strength of this relation differed across participants (*SD r* = .23), though was not related to individual variation in age, hearing loss, number of fixations, or number of interpolated data points [all $p > .05$]. Among the approaches that have controlled for scaling variability in pupillometry analyses (e.g., Engelhardt, Ferreira, & Patsenko, 2010; Hess & Polt, 1964; Nassar et al., 2012; Payne, Parry, & Harasymiw, 1968; Piquado et al., 2010), we selected a within-trial mean scaling method to ensure consistent scaling of pupil size values across both trials and participants. Each trial data point was divided by the mean of the entire trial time series (from trial onset to RT) for each individual. Baseline correction (e.g., Zekveld et al., 2010, 2011) was then performed to examine how much greater the pupil response was following word presentation in noise compared to a noise-only baseline. Mean pupil size was calculated within the 1000 ms epoch in which background noise played prior to word onset, and then was subtracted from each data point of that trial.

The preprocessed time series data were averaged within each participant and within each condition. For each average, a total of 113 time samples (1883.71 ms) following word onset were entered into pupil size analyses. This was the maximum number of time points for which average pupil size values were obtained for every condition and every participant. This cutoff was well below the average RT across individuals, even in the easiest condition [easier rSNR, no lexical competition, *M* = 2408.67 ms, *SD* = 343.20]. Thus, limiting the analyzed data to this period reduced the likelihood that changes in the fit of the tail of the curve would solely be driven by variation in RTs across conditions.

Growth Curve Analysis (GCA): GCA, which allows for the analysis of changes in pupil size over time, is a type of multilevel regression model (Bryk & Raudenbush, 2002; Hofmann, 1997) that fits orthogonal polynomial terms to time series data to model variations in the curve shape across conditions and individuals. GCA has become an increasingly adopted method for analyzing eyetracking data, particularly in the VWP (Kalénine, Mirman, Middleton, & Buxbaum, 2012; Magnuson et al., 2007; Mirman et al., 2008; Mirman, Yee, Blumstein, & Magnuson, 2011), as it is a more powerful analysis tool for understanding how metrics changes over time compared to ANOVAs on time-binned data.

Key to the approach is that orthogonal parameters are used to describe the shape of the curve. This allows for the independent analysis of linear and non-linear changes in pupil size over time, rather than potentially collinear measures such as average pupil size and peak pupil height and latency. In addition to analyses of logistic fixation data, GCA also has been applied to analyzing changes in continuous ocular metrics, such as fixation duration preferences across time (Trueswell & Papafragou, 2010). In the current study, GCA is extended to the analysis of pupil size data that are recorded during the VWP. For a clear and detailed explanation of the theory and application of GCA, including sample code, see Mirman et al. (2008) and the related author website (www.danmirman.org/gca).

The statistics program R (R Core Development Team, 2011) and its associated nlme package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Development Team, 2009) were used for GCA computations. Pupil size is modeled as the dependent measure that is predicted by a series of fixed (regressors) and random effects (error terms). Orthogonalized time vectors capture independent aspects of the shape of the pupil response across time. In this study, level-1 of the hierarchical model included an intercept and five orthogonal polynomial shape parameters (linear, quadratic, cubic, quartic, and quintic). For detailed depictions of how these parameters affect curve shape, see Kalenine et al. (2012). To summarize, the intercept term represents the overall mean of the pupil response function (greater parameter estimates signify overall greater amplitude) and the linear term represents the slope of the entire pupil response (more positive parameter estimates signify greater pupil size towards the end versus the beginning of the pupil response). The quadratic term reflects the shape of the primary inflection point of the curve (more positive signifying a more linear, flatter distribution). For the cubic term, more positive values indicate an earlier shifted peak response, while more negative values indicate a later shifted peak response. Quartic and quintic terms describe the shape of secondary peaks in the tails of the pupil response.

The canonical pupil response includes an initial increase in pupil size in response to processing a stimulus (around a 700 ms peak latency) that falls back to baseline over time (Kahneman & Beatty, 1966). As shown in the Results, examination of the pupil response across all trials revealed that a fifth order polynomial best captured the overall shape of the pupil response. Polynomial shape parameters were allowed to vary across individuals by modeling participants as fixed-effect modulators of the shape parameters and by random effects. Of theoretical interest in this study was how rSNR and lexical competition affect the timing and amplitude of the pupil response peaks and tail and thus we examined the extent to which interactions between rSNR and lexical competition modulated the intercept, linear, quadratic, cubic, quartic, and quintic terms.

To test whether an rSNR-by-lexical competition interaction significantly altered various aspects of the shape and timing of the pupillary response, a *base* model which contained participant fixed and random effects, as well as all main effects (e.g. rSNR, linear term) and lower-order interactions (e.g., rSNR × linear term) was constructed (as the inclusion of these terms is necessary to anchor and evaluate higher order interactions). The higher order interaction terms (i.e., rSNR × lexical competition × intercept, linear, quadratic, cubic, quartic, and quintic) were added through stepwise regression model building as level-2 effects to the base model. The Likelihood Ratio of the models (the difference between models' −2LL, minus two times the log-likelihood) was examined to determine the extent to which the addition of each parameter significantly improved the model fit. If a parameter did not significantly improve model fit, it was removed to examine if the lower order interactions or main effects contributed to fit or also should be removed. Parameter beta (β) weights and their associated *p*-values from submodels (e.g., the effect of rSNR when competition is held constant) were examined to clarify the strength and direction of significant effects.

Trials entered into GCA analyses were limited to those for which participants correctly identified the target word in order to examine the extent to which the pupillary response was sensitive to task difficulty even with correct performance across conditions. Correct trials were the focus because of the potentially unknown source of error for incorrect trials; a lapse in attention or misperceiving the word could contribute to an incorrect response, but was unlikely to have occurred for correct responses. A total of 1213 correct trials out of the 2605 correct and incorrect trials that met preprocessing criteria were entered into GCA models. While the experimental design manipulated task difficulty such that there were different numbers of correct trials in each condition, baseline rSNR determination prevented ceiling and floor effects that could result in any one condition being under-powered. Finally, to further examine the relative sensitivity of pupil size and RT metrics, the average log RT within each condition for each participant was added as a covariate to the model.

Peak-picking analyses: To relate the results of the GCA to previous results, traditional analyses were conducted on amplitudes and latencies obtained from peak-picked pupillometry data, which had been preprocessed in the same way as the data for GCA. A time window of interest was defined as beginning at word onset and lasting 1500 ms. Based on a visual inspection of the data, this range captured the entire rise and initial fall of the pupil response across conditions and individuals. The average maximum pupil size and its corresponding latency value within this window were entered as dependent variables into two, 2 (rSNR) × 2 (lexical competition) repeated measures ANOVAs. Correlations between peak amplitude and latency with the parameter estimates for each of the GCA polynomial terms were calculated.

Results

Sensitivity of Behavioral Measures to Task Difficulty

Two-by-two repeated measures ANOVAs were performed to examine the effect of rSNR (easier vs. harder) and lexical competition (competitor vs. no-competitor) on percent correct word identification scores and RTs. There was a significant interaction between rSNR and lexical competition for correct word identification [$F(1, 20) = 28.93$, $p < .001$, $\eta_p^2 = .59$]. As shown in the top panel of Figure 3, word identification was poorer overall with lexical competition than without competition in both the easier [$t(20) = −16.70$, $p < .001$, $d = 3.64$] and harder [$t(20) = −7.12$, $p < .001$, $d = 1.55$] rSNR conditions. While word identification was significantly greater with the easier than the harder rSNR without lexical competition [$t(20) = −6.11$, $p < .001$, $d = 1.33$], there was no difference in the presence of lexical competition [$t(20) = −.74$, $p = .47$, $d = .16$].

To examine whether RT provided additional information regarding task difficulty beyond performance alone, variation in RT was examined specifically for correct word identification trials (Figure 3). A 2 × 2 repeated measures ANOVA on the log-transformed RTs for correct trials demonstrated a main effect of rSNR [$F(1, 20) = 29.35$, $p < .001$, $\eta_p^2 = .60$], a main effect of lexical competition [$F(1, 20) = 5.01$, $p < .05$, $\eta_p^2 = .20$], but no interaction between them [$F(1, 20) < 1$, $\eta_p^2 =.00$]. The weaker effect size for lexical competition compared to rSNR, suggests, however, that RT may be more sensitive to effects of rSNR than lexical competition. Individuals were slower to respond on harder than easier rSNR trials both when there was lexical competition [$t(20) = 3.33$, $p < .01$, $d = .73$] and when there was not lexical competition [$t(20) = 4.99$, $p < .001$, $d = 1.09$]. However, within just the harder or easier rSNR conditions, there was only a numerical trend for participants to respond more quickly on trials without competition than with competition [within the harder rSNR, $t(20) = 1.80$, $p = .09$, $d = .39$; within the easier rSNR $t(20) = 1.50$, $p = .15$, $d = .33$]. There was no evidence that this pattern of results was related to individual differences in speed-accuracy tradeoff criteria. The correlation between average percent correct word identification and average log-transformed RT across all trials was non-significant within each rSNR and lexical competition condition (all $p > .10$).

Sensitivity of the Pupil Response to Task Difficulty

GCA demonstrated an impact of the task difficulty manipulations (rSNR and lexical competition) on the shape and timing of the pupil response. Across all correct trials (Table 1A) an rSNR × lexical competition interaction affected each polynomial shape term, except for the linear term with which only an effect of rSNR was observed. The mathematical equation summarizing the best-fitting GCA model is found in the caption for Table 1. Table 1B-D presents the beta values (parameter estimates) associated with each polynomial term in four submodels that were created to interpret the interaction terms: easier rSNR trials only, harder rSNR trials only, no-competitor trials only, and competitor trials only. For these categorical predictors, beta values represent a contrast estimate, such that they signify how much greater the effect was for the more difficult condition (harder rSNR, competitor) compared to the easier condition (easier, no-competitor). Figure 4 displays the pupil response data within each rSNR by lexical competition condition (white lines) and the fitted responses of the best fitting model (colored lines) including the standard error of the mean for these fits (error bars). The upper panels of Figure 4 demonstrate the extent to which the predicted model fits the shape of the data across the time domain. The bottom panel of Figure 4 allows for comparison of model fits across the four conditions. The following sections clarify these patterns of results through submodel tests (Table 1B–D).

Acoustic and lexical selection difficulty affect average pupil size—The average size of pupil response (intercept) was largest in the most challenging condition (harder rSNR with lexical competition) compared to the other conditions, which did not significantly differ from one another (Figure 4, Table 1B-D). Specifically, mean pupil amplitude in the most challenging condition was larger than in the harder rSNR with no lexical competition condition (Table 1C) and than in the easier rSNR with competition condition (Table 1E). There was no difference between the least challenging condition (easier rSNR with no competition) compared to these two conditions (Table 1B, 1D).

Acoustic and lexical selection difficulty affect pupil response delay to peak—The pupil response was more delayed to peak (cubic term) for the harder rSNR compared to the easier rSNR conditions, regardless of lexical competition (Table 1D, 1E). There was a trend for the pupil response to be more delayed for competitor compared to no-competitor trials, though only within the easier rSNR condition (Table 1B). Thus, more difficult conditions generally resulted in a delayed pupil response peak though in the harder rSNR, the pupil peak was equally delayed across competition conditions (Table 1C).

Acoustic and lexical selection difficulty affect pupil size sustainment—In both the most (Table 1C, 1E) and least (Table 1B, 1D) challenging conditions, the pupil response rose steeply (quadratic term). However, while the peak response in the most challenging condition was followed by sustained, elevated pupil dilation, following the peak response in the least challenging condition, pupil size fell towards baseline. Specifically, pupil size remained elevated compared to baseline in the most challenging condition (linear term), was less elevated for the harder rSNR without lexical competition condition, and was least elevated whenever the rSNR was easier (linear term, Table 1B-E).

Generally, more challenging conditions also resulted in a larger secondary peak (quartic and quintic terms), with the exception again of there being no competition difference within the harder rSNR condition (Table 1C).

Additive contribution of RT—Including a RT covariate on each polynomial term of the model (intercept - quintic) improved model fit [Likelihood Ratio = 59.19, $p < .001$], such that longer RTs were associated with overall flatter (more linear) primary [quadratic β = 2331, $p < .0001$] and secondary peaks [quartic β = −2055, $p < .0001$; quintic β = −1134, $p = .01$]. However, even with the inclusion of these covariates, the pattern of significance of the condition effects (rSNR on linear term, rSNR × lexical competition on all other terms) remained unchanged. The average percent change in beta values for these effects was 0.50% (*SD* = 1.19%). Thus the pupil shape parameters captured additive information regarding the effects of rSNR and lexical competition across time with RT and correct word identification.

Relation of peak-picking analyses to GCA—To compare the results of the GCA to previous pupillometry analyses, individual difference fixed-effect parameter estimates were correlated with individual differences in peak pupil amplitude and latency obtained through traditional peak-picking. Peak-picked pupil amplitude [*M* = .40, *SD* = .24] was significantly correlated with the intercept [$r(19) = .84$, $p < .001$], cubic [$r(19) = .52$, $p < .05$] and quartic [$r(19) = -.44$, $p < .05$] terms of the GCA model. Peak-picked latency [*M* = .85 sec, *SD* = .28] was significantly correlated with the linear [$r(19) = .53$, $p < .05$] and cubic [$r(19) = -.50$, $p < .05$] terms of the model. Thus, greater traditional peak pupil size was related to greater GCA average pupil size across the trial as well as an earlier peaking response. Greater traditional peak delay was related to GCA measures of elevated pupil size in the response tail as well as greater peak delay.

Traditional ANOVAs performed on peak-picked data revealed only a trend of an rSNR effect on both the peak amplitude [$F(1, 20) = 4.03$, $p = .058$, $\eta_p^2 = .17$] and latency [$F(1, 20) = 3.71$, $p = .069$, $\eta_p^2 = .16$], with pupils tending to peak to a greater degree and later in the harder [amplitude *M* = .41; latency *M* = .92 sec] than the easier rSNR condition [amplitude *M* = .39; latency *M* = .78 sec]. No main effect or interaction with lexical competition was observed within the window of interest for either the pupil peak amplitude [main effect $F(1,20) < 1$, $\eta_p^2 = .02$; interaction $F(1,20) < 1$, $\eta_p^2 = .02$] or the latency [main effect $F(1, 20) < 1$, $\eta_p^2 = .02$; interaction $F(1, 20) < 1$, $\eta_p^2 = .04$]. Thus, there was relatively greater sensitivity for the GCA than peak-picking approach for capturing condition effects on the pupil response.

Summary—More difficult task conditions were generally associated with independent measures of greater average pupil size, a more delayed peak, and elevated, sustained pupil size following the peak. Specifically, the most challenging condition (harder rSNR with lexical competition) was associated with the greatest average pupil response amplitude. The pupil response was more delayed in the harder rSNR condition, regardless of lexical competition, though lexical competition tended to increase delay within the easier rSNR condition. The most challenging condition was also associated with a more steeply peaking response which was followed by sustained, elevated pupil size. Because these results were obtained for correct trials despite variation in RTs, pupillometry additively characterizes the effects of rSNR and lexical competition on task difficulty in older adults with hearing loss.

Discussion

The results of the current study revealed information about task demands that were not indexed by word identification or RT measures alone, and thus provide a richer characterization of task difficulty experienced in challenging listening conditions. This study extends previous findings that pupil size is sensitive to acoustic manipulations of listening difficulty in younger and middle-aged normal-hearing and hearing-impaired adults (Zekveld et al., 2010, 2011) in three key ways. First, the current study was conducted with older adult participants (mean age 73 years old) who had mild to moderate sensorineural hearing loss. Second, following from findings that older adults are more negatively affected by lexical competition (Sommers, 1996; Sommers & Danielson, 1999; Taler et al., 2010), the current paradigm demonstrates that the pupil response is sensitive to an interaction between both lexical competition in response selection and acoustic features of speech during listening. Third, the current study employed GCA in order to quantify the influence of these task difficulty manipulations across time, thus enhancing the sensitivity of pupillometry analyses.

Analysis of the behavioral data revealed that word identification was affected by lexical competition during response selection across listening conditions. However, acoustic listening difficulty effects were only observed in the absence of response selection competition. Additional behavioral analyses for correct trials revealed that RT was sensitive to manipulations of rSNR, following from previous literature showing that RT reflects variation in acoustic listening difficulty even when word identification is similar (Gatehouse & Gordon, 1990). Thus, behavioral data suggest that word identification scores may be more reflective of response selection difficulty while RTs on correct trials may be more reflective of listening difficulty. Both performance measures provide additive information about the task difficulty experienced by older adults with hearing loss.

Following from findings that pupil size decreases with increasing intelligibility for young and middle-aged adults with normal and impaired hearing (Zekveld et al., 2010, 2011), we found a similar relation between rSNR and the shape of the pupil response in a sample of older adults, even when examining only correct trials. More difficult rSNR resulted in greater average pupil size, a more delayed peak, and sustained pupil size following the peak. Extending previous work, we found that lexical competition interacted with rSNR such that when both were challenging, the pupil response was largest overall and specifically in the tail following the peak (i.e., most sustained). Importantly, these patterns of results were obtained when participants correctly identified the target word and when including an RT covariate in the pupil response model, indicating that the pupil response provides additive measures of task difficulty compared to behavioral measures alone.

Though previous research has suggested that pupil size in older adults is not sensitive to small changes in cognitive load (Van Gerven, Paas, Van Merrienboer, & Schmidt, 2004), the current manipulations of task difficulty (4 dB difference in SNR and the presence a competitor word), along with the use of GCA, were sufficient to observe a significantly modulated pupil response in our sample of older adults. Indeed, the effect of lexical competition on the pupil response may have been more easily observable in this study because older adults appear more negatively affected by lexical competition than younger adults (Ben-David et al., 2011; Sommers, 1996; Sommers & Danielson, 1999; Taler et al., 2010).

Traditional peak-picking measures (i.e., peak amplitude, latency) were related to multiple parts of the pupil response curve: traditional peak amplitude was related to the mean and delay of the pupil response, as well as the shape of the extremities of the response as measured by GCA. Peak latency was related to the delay to peak and pupil size in the tail of the response. This indicates that peak amplitude may be influenced by multiple aspects of the pupil response shape and that peak-picked measures are likely to be collinear in some studies (i.e., both traditional peak amplitude and latency were related to the cubic term). GCA orthogonalizes each shape parameter term and therefore allows for independent assessment of the aspects of the pupil response affected by task manipulations. Traditional peak-picking analyses on the same data revealed only a trend of a main effect of rSNR on peak amplitude and latency similar to previous effects (Zekveld et al., 2010, 2011), and no effect of lexical competition on either measure. This suggests that GCA is a particularly sensitive analysis technique for detecting changes in the shape of the pupil response over time, similar to what has been described in fixation pattern analyses (Mirman et al., 2008). Future work using deconvolution analysis on pupillometry data (Wierda, Rijn, Taatgen, & Martens, 2012) in combination with GCA may increase the ability to identify and distinguish the onsets and durations of different cognitive events (e.g., speech perception versus response selection). Separately analyzing each event type may then limit the complexity of the GCA models needed to describe these distinct pupil responses in empirically determined windows of interest.

A limitation of many studies of listening effort is that manipulations of task difficulty are assumed to be equivalent to manipulations of cognitive effort, although this may not necessarily be true. A task could be so easy or challenging that the participant exerts relatively little effort, thus resulting in a nonlinear relationship between performance or RT measures and the construct of cognitive effort. For example, in the current study, log RT across all trials was not correlated with correct word identification [$r(19) = -.22$, $p > .05$], and thus faster responding (potentially an indication of less effort) occurred for both correct and incorrect responses. Previous memory research has demonstrated that faster RTs are generally associated with both greater subjective ratings of confidence and reduced subjective ratings of effort (Robinson, Johnson, & Herndon, 1997). The extent to which RT taps into listening effort versus response confidence and the relation between confidence and listening effort thus requires careful examination.

Pupil size measures aim to bridge the gap between task difficulty and cognitive effort, as the current and previous studies have shown that pupil size is not directly tied to behavioral performance. However, more work is needed to validate the relationship between task demands and effort, and to understand the cognitive and neural processes that underlie various measures of effort. For example, while Zekveld et al. (2011) found that pupil size was related to self-report measures of effort, the same relationship has been shown to not hold for other effort metrics, such as dual-task measures (Sarampalis, Kalluri, Edwards, & Hafter, 2009), skin conductance (Mackersie & Cones, 2011), and in some cases performance and RT (Hällgren et al., 2005). The current study employed moderate levels of task difficulty (rSNRs were tailored to the individual to avoid ceiling and floor effects). Studies in which task difficulty is systematically manipulated in order to capture a wide range of levels of difficulty while accounting for individual differences in personality metrics such as motivation, will further characterize the potentially nonlinear relation between cognitive effort and task difficulty. For example, research in the memory domain has found that the pupil size of younger adults tends to plateau (Cabestrero, Crespo, & Quiros, 2009; Peavler, 1974) or even decrease (Granholm, Asarnow, Sarkin, & Dykes, 1996) when participants are asked to remember strings of digits that far exceed their working memory capacities. Thus, pupil size metrics may best inform theories of cognitive processing when they do not vary in a linear fashion with manipulations of task difficulty and with behavioral metrics, as when multiple sources of task difficulty and/or individual differences in cognition interact.

The results of this study set the foundation for assessing whether pupillometry may be a useful assessment tool for clinical populations. Indeed, hearing-loss intervention success is a multidimensional construct that may be assessed more completely by obtaining objective and subjective measures of speech recognition and subjective measures of listening effort (Humes, 1999). However, the extent to which the pupil response may be used to assess how effort changes with 1) individual differences in hearing and/or cognitive skills, or 2) within-individual changes following hearing-loss interventions has yet to be determined. Future research will evaluate whether pupillometry can provide an empirically based metric of hearing loss intervention success that is additive to current behavioral assessments, thus improving the clinical care that individuals with hearing loss receive (Kuchinsky, Eckert, & Dubno, 2011).

Acknowledgments

We thank the study participants. Work was supported in part by a Hearing Health Foundation Centurion Clinical Research Award, the National Institute on Deafness and other Communication Disorders (P50 DC00422), and was conducted in a facility constructed with support from Research Facilities Improvement Program (C06 RR14516) from the National Center for Research Resources, NIH. This project was also supported by the South Carolina Clinical and Translational (SCTR) Institute, with an academic home at the Medical University of South Carolina, NIH/NCRR (UL1 RR029882).

References

American National Standards Institute. American National Standard Specifications for Audiometers. New York: Author; 2004.

Anderson Gosselin P, Gagne J-P. Older adults expend more listening effort than young adults recognizing speech in noise. Journal of Speech, Language, and Hearing Research. 2011; 54(3):944–58.10.1044/1092-4388(2010/10-0069)

Aston-Jones G, Cohen JD. An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. Annual Review of Neuroscience. 2005; 28:403–50.10.1146/annurev.neuro.28.061604.135709

Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, Loftis B, Neely JH, et al. The English Lexicon Project. Behavior Research Methods. 2007; 39(3):445–59.10.3758/BF03193014 [PubMed: 17958156]

Ben-David BM, Chambers CG, Daneman M, Pichora-Fuller MK, Reingold EM, Schneider BA. Effects of aging and noise on real-time spoken word recognition: Evidence from eye movements. Journal of Speech, Language, and Hearing Research. 2011; 54(1):243–62.10.1044/1092-4388(2010/09-0233)

Bryk, AS.; Raudenbush, SW. Hierarchical linear models: Applications and data analysis methods. 2. Thousand Oaks, CA: Sage; 2002.

Brysbaert M, New B. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods. 2009; 41(4):977–90.10.3758/BRM.41.4.977 [PubMed: 19897807]

Burk MH, Humes LE. Effects of long-term training on aided speech-recognition performance in noise in older adults. Journal of Speech, Language, and Hearing Research. 2008; 51(3):759–71.10.1044/1092-4388(2008/054)

Cabestrero R, Crespo A, Quiros P. Pupillary dilation as an index of task demands. Perceptual and Motor Skills. 2009; 109(3):664–678.10.2466/pms.109.3.664–678 [PubMed: 20178265]

Dennis, NA.; Cabeza, R. Neuroimaging of Healthy Cognitive Aging. In: Salthouse, TA.; Craik, FEM., editors. Handbook of Aging and Cognition. 3. New York: Psychological Press; 2008. p. 1-54.

Dirks DD, Takayana S, Moshfegh A. Effects of lexical factors on word recognition among normal-hearing and hearing-impaired listeners. Journal of the American Academy of Audiology. 2001; 12(5):233–244. [PubMed: 11392435]

Dirks DD, Takayanagi S, Moshfegh A, Noffsinger PD, Fausti SA. Examination of the neighborhood activation theory in normal and hearing-impaired listeners. Ear and Hearing. 2001; 22(1):1–13.10.1097/00003446-200102000-00001 [PubMed: 11271971]

Dreschler WA, Verschuure H, Ludvigsen C, Westermann S. ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Journal of Audiology. 2001; 40(3):148–157.10.3109/00206090109073110

Eckert MA. Slowing down: Age-related neurobiological predictors of processing speed. Frontiers in Neuroscience. 2011; 5:25.10.3389/fnins.2011.00025 [PubMed: 21441995]

Edwards B. The future of hearing aid technology. Trends in Amplification. 2007; 11(1):31–45.10.1177/1084713806298004 [PubMed: 17301336]

Engelhardt PE, Ferreira F, Patsenko EG. Pupillometry reveals processing load during spoken language comprehension. Quarterly Journal of Experimental Psychology. 2010; 63(4):639–45.10.1080/17470210903469864

Fraser S, Gagne J-P, Alepins M, Dubois P. Evaluating the effort expended to understand speech in noise using a dual-task paradigm: the effects of providing visual speech cues. Journal of Speech, Language, and Hearing Research. 2010; 53(1):18–33.10.1044/1092-4388(2009/08-0140)

Gatehouse S, Gordon J. Response times to speech stimuli as measures of benefit from amplification. British Journal of Audiology. 1990; 24(1):63–68.10.3109/03005369009077843 [PubMed: 2317602]

Gilzenrat MS, Cohen JD, Rajkowski J, Aston-Jones G. Pupil dynamics predict changes in task engagement mediated by locus coeruleus. Society for Neuroscience Abstracts. 2003:515:19.

Gilzenrat MS, Nieuwenhuis S, Jepma M, Cohen JD. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. Cognitive, Affective & Behavioral Neuroscience. 2010; 10(2):252–69.10.3758/CABN.10.2.252

Granholm E, Asarnow RF, Sarkin AJ, Dykes KL. Pupillary responses index cognitive resource limitations. Psychophysiology. 1996; 33(4):457–61.10.1111/j.1469-8986.1996.tb01071.x [PubMed: 8753946]

Hällgren M, Larsby B, Lyxell B, Arlinger S. Speech understanding in quiet and noise, with and without hearing aids. International Journal of Audiology. 2005; 44(10):574–583.10.1080/14992020500190011 [PubMed: 16315448]

Hess EH, Polt JM. Pupil size in relation to mental activity during simple problem-solving. Science. 1964; 143(3611):1190–2.10.1126/science.143.3611.1190 [PubMed: 17833905]

Hetu R, Riverin L, Lalande N, Getty L, St-cyr C. Qualitative analysis of the handicap associated with occupational hearing loss. British Journal of Audiology. 1988; 22(4):251–264.10.3109/03005368809076462 [PubMed: 3242715]

Hicks CB, Tharpe AM. Listening effort and fatigue in school-age children with and without hearing loss. Journal of Speech, Language, and Hearing Research. 2002; 45(3):573–584.10.1044/1092-4388(2002/046)

Hofmann DA. An overview of the logic and rationale of hierarchical linear models. Journal of Management. 1997; 23(6):723–744.

Humes LE. Dimensions of hearing aid outcome. Journal of the American Academy of Audiology. 1999; 10(1):26–39. [PubMed: 9921723]

Humes LE, Burk MH, Strauser LE, Kinney DL. Development and efficacy of a frequent-word auditory training protocol for older adults with impaired hearing. Ear and Hearing. 2009; 30(5):613–27.10.1097/AUD.0b013e3181b00d90 [PubMed: 19633564]

Jepma M, Nieuwenhuis S. Pupil diameter predicts changes in the exploration–exploitation tradeoff: Evidence for the adaptive gain theory. Journal of Cognitive Neuroscience. 2011; 27(3):1587–1596. [PubMed: 20666595]

Kahneman, D. Attention and Effort. Englewood Cliffs, NJ: Prentice-Hall Inc; 1973.

Kahneman D, Beatty J. Pupil diameter and load on memory. Science. 1966; 154(3756):1583–1585.10.1126/science.154.3756.1583 [PubMed: 5924930]

Kalenine S, Mirman D, Middleton EL, Buxbaum LJ. Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts. Journal of Experimental Psychology: Learning, Memory, and Cognition. 201210.1037/a0027626

Kuchinke L, Vo ML-H, Hofmann M, Jacobs AM. Pupillary responses during lexical decisions vary with word frequency but not emotional valence. International Journal of Psychophysiology. 2007; 65(2):132–40.10.1016/j.ijpsycho.2007.04.004 [PubMed: 17532075]

Kuchinsky SE, Eckert MA, Dubno JR. The eyes are the windows to the ears: Pupil size reflects listening effort. Audiology Today. 2011; 23:56–59.

Laeng B, Orbo M, Holmlund T, Miozzo M. Pupillary Stroop effects. Cognitive Processing. 2011; 12(1):13–21.10.1007/s10339-010-0370-z [PubMed: 20865297]

Lin FR, Thorpe R, Gordon-Salant S, Ferrucci L. Hearing loss prevalence and risk factors among older adults in the United States. The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences. 2011; 66(5):582–90.10.1093/gerona/glr002

Mackersie CL, Cones H. Subjective and psychophysiological indexes of listening effort in a competing-talker task. Journal of the American Academy of Audiology. 2011; 22(2):113–22.10.3766/jaaa.22.2.6 [PubMed: 21463566]

Magnuson JS, Dixon J, Tanenhaus M, Aslin R. The dynamics of lexical competition during spoken word recognition. Cognitive Science. 2007; 31(1):133–156.10.1080/03640210709336987 [PubMed: 21635290]

McCoy SL, Tun PA, Cox LC, Colangelo M, Stewart RA, Wingfield A. Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology. 2005; 58(1):22–33.10.1080/02724980443000151

McQueen JM, Viebahn MC. Tracking recognition of spoken words by tracking looks to printed words. Quarterly Journal of Experimental Psychology. 2007; 60(5):661–71.10.1080/17470210601183890

Mirman D, Dixon JA, Magnuson JS. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of Memory and Language. 2008; 59(4):475–494.10.1016/j.jml.2007.11.006 [PubMed: 19060958]

Mirman D, Yee E, Blumstein SE, Magnuson JS. Theories of spoken word recognition deficits in aphasia: Evidence from eye-tracking and computational modeling. Brain and Language. 2011; 117(2):53–68.10.1016/j.bandl.2011.01.004 [PubMed: 21371743]

Moresi S, Adam JJ, Rijcken J, Van Gerven PWM, Kuipers H, Jolles J. Pupil dilation in response preparation. International Journal of Psychophysiology. 2008; 67(2):124–30.10.1016/j.ijpsycho. 2007.10.011 [PubMed: 18067982]

Nachtegaal J, Kuik DJ, Anema JR, Goverts ST, Festen JM, Kramer SE. Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. International Journal of Audiology. 2009; 48(10):684–91.10.1080/14992020902962421 [PubMed: 19863354]

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI. Rational regulation of learning dynamics by pupil-linked arousal systems. Nature Neuroscience. 2012:1–9.10.1038/nn.3130

Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia. 1971; 9(1):97–113.10.1016/0028-3932(71)90067-4 [PubMed: 5146491]

Papesh MH, Goldinger SD. Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation. Attention, Perception & Psychophysics. 201210.3758/s13414-011-0263-y

Payne DT, Parry ME, Harasymiw SJ. Percentage of pupillary dilation as a measure of item difficulty. Perception & Psychophysics. 1968; 4(3):139–143.10.3758/BF03210453

Peavler WS. Pupil size, information overload, and performance differences. Psychophysiology. 1974; 11(5):559–566.10.1111/j.1469-8986.1974.tb01114.x [PubMed: 4415394]

Pinheiro, J.; Bates, B.; DebRoy, S.; Sarkar, D. R Development Core Team. nlme: Linear and nonlinear mixed effects models. 2009. Retrieved from http://cran.r-project.org/web/packages/nlme

Piquado T, Isaacowitz D, Wingfield A. Pupillometry as a measure of cognitive effort in younger and older adults. Psychophysiology. 2010; 47(3):560–9.10.1111/j.1469-8986.2009.00947.x [PubMed: 20070575]

Plomp R. Noise, amplification, and compression: Considerations of three main issues in hearing aid design. Ear and Hearing. 1994; 15(1):2–12.10.1097/00003446-199402000-00002 [PubMed: 8194676]

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011. Retrieved from http://www.r-project.org

Rajkowski J, Kubiak P, Aston-Jones G. Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. Society for Neuroscience. 1993; 19:19:974.

Robinson MD, Johnson JT, Herndon F. Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. The Journal of Applied Psychology. 1997; 82(3): 416–25. [PubMed: 9190148]

Salthouse TA. The processing-speed theory of adult age differences in cognition. Psychological Review. 1996; 103(3):403–428.10.1037//0033-295X.103.3.403 [PubMed: 8759042]

Salthouse TA, Babcock RL, Shaw RJ. Effects of adult age on structural and operational capacities in working memory. Psychology and Aging. 1991; 6(1):118–127.10.1037//0882-7974.6.1.118 [PubMed: 2029360]

Salverda AP, Tanenhaus MK. Tracking the time course of orthographic information in spoken-word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2010; 36(5): 1108–17.10.1037/a0019901

Sarampalis A, Kalluri S, Edwards B, Hafter E. Objective measures of listening effort: effects of background noise and noise reduction. Journal of Speech, Language, and Hearing Research. 2009; 52(5):1230–40.10.1044/1092-4388(2009/08-0111)

Siegle GJ, Steinhauer SR, Stenger VAA, Konecky R, Carter CS. Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. Neuroimage. 2003; 20(1):114–124.10.1016/S1053-8119(03)00298-2 [PubMed: 14527574]

Sommers MS. The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. Psychology and Aging. 1996; 11(2):333–341.10.1037//0882-7974.11.2.333 [PubMed: 8795062]

Sommers MS, Danielson SM. Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. Psychology and Aging. 1999; 14(3):458–472.10.1037//0882-7974.14.3.458 [PubMed: 10509700]

Taler V, Aaron GP, Steinmetz LG, Pisoni DB. Lexical neighborhood density effects on spoken word recognition and production in healthy aging. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences. 2010; 65(5):551–60.10.1093/geronb/gbq039

Trueswell JC, Papafragou A. Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. Journal of Memory and Language. 2010; 63(1):64–82.10.1016/j.jml.2010.02.006

Van Gerven PWM, Paas F, Van Merrienboer JJG, Schmidt HG. Memory load and the cognitive pupillary response in aging. Psychophysiology. 2004; 41(2):167–174.10.1111/j.1469-8986.2003.00148.x [PubMed: 15032982]

Verney SP, Granholm E, Marshall SP. Pupillary responses on the visual backward masking task reflect general cognitive ability. International Journal of Psychophysiology. 2004; 52(1):23–36.10.1016/j.ijpsycho.2003.12.003 [PubMed: 15003370]

Wechsler, D. Wechsler abbreviated scale of intelligence. San Antonio, TX: The Psychological Corporation; 1999.

Wierda SM, Rijn HV, Taatgen NA, Martens S. Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. Proceedings of the National Academy of Sciences. 2012; 109(22):8456–8460. www.pnas.org/cgi/doi/10.1073/pnas.1201858109.

Wingfield A, Stine EA, Lahar CJ, Aberdeen JS. Does the capacity of working memory change with age? Experimental Aging Research. 1988; 14(2–3):103–107.10.1080/03610738808259731 [PubMed: 3234452]

Zekveld AA, Kramer SE, Festen JM. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. Ear and Hearing. 2010; 31(4):480–90.10.1097/AUD.0b013e3181d4f251 [PubMed: 20588118]

Zekveld AA, Kramer SE, Festen JM. Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. Ear and Hearing. 2011; 32(4):498–510.10.1097/AUD.0b013e31820512bb [PubMed: 21233711]

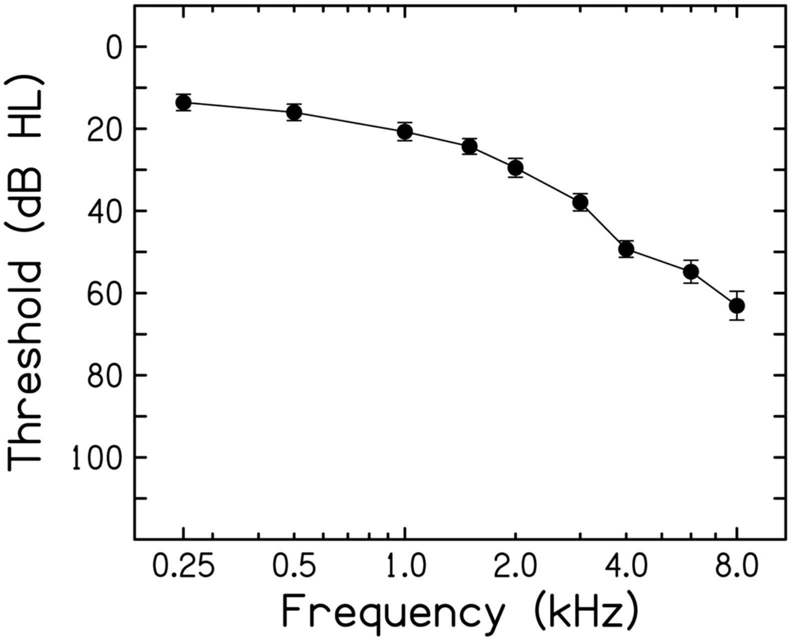

Figure 1.
Participants had mild to moderate high-frequency sensorineural hearing loss as depicted in the average audiogram for the tested ear. Error bars represent the standard error of the mean.

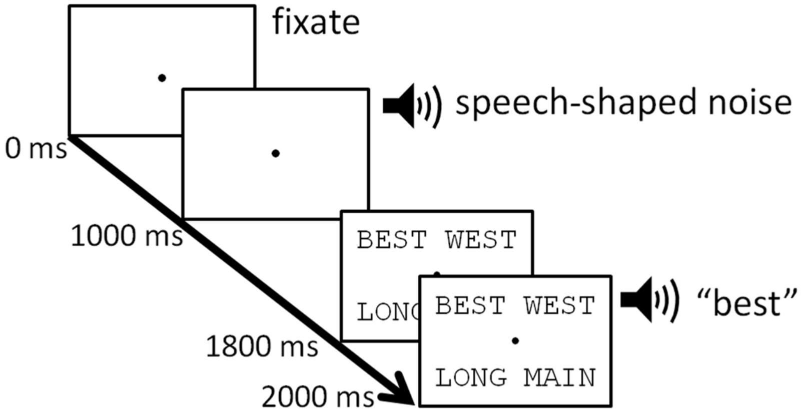

Figure 2.
Participants began each trial by fixating a central circle. ICRA noise was presented in the background beginning at 1000 ms into the trial. The four orthographic options appeared on screen beginning 200 ms before the word to be identified was presented 2000 ms into the trial. The trial ended when the participant touched the screen to make a response. Screen display is not depicted to scale.

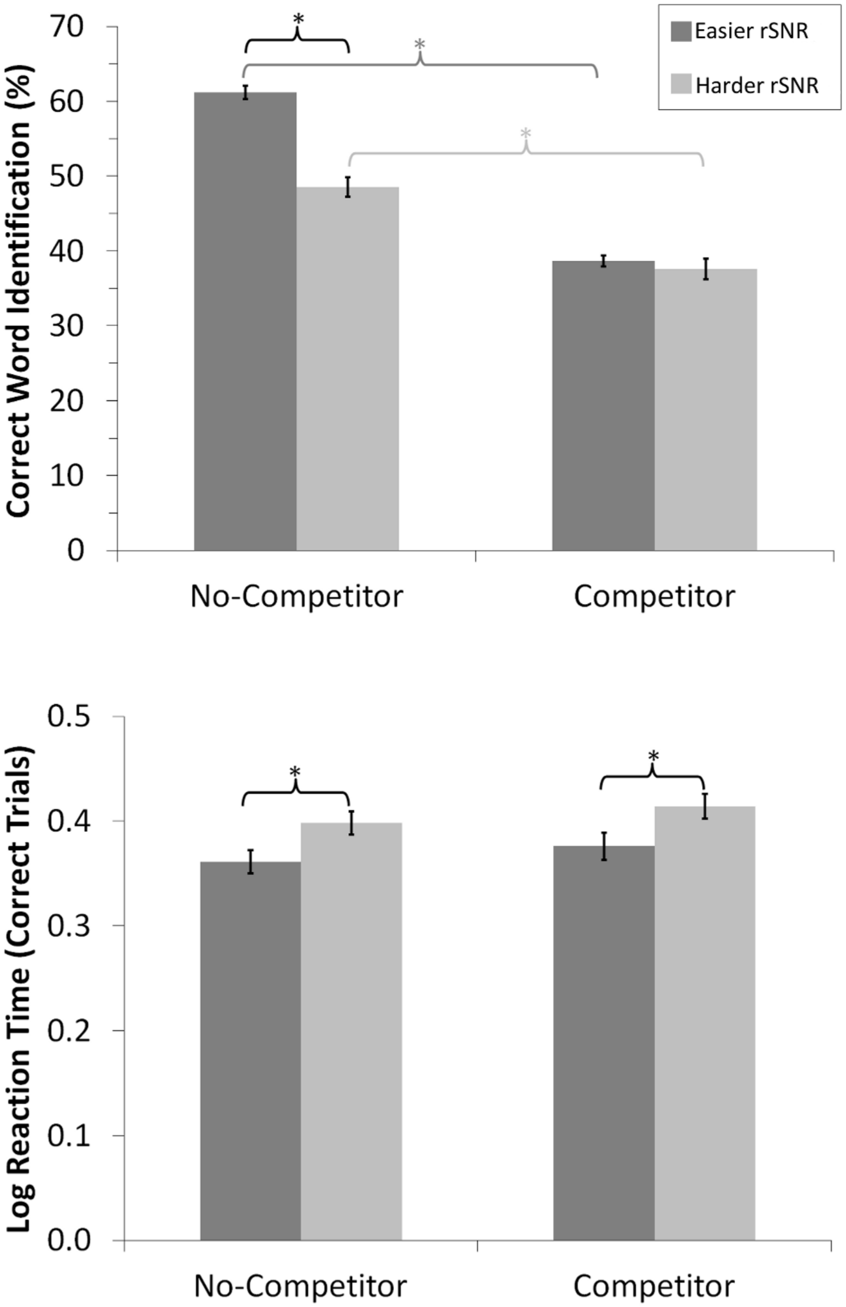

Figure 3.

Top: Word identification scores were sensitive to changes in lexical competition but only to rSNR without lexical competition. Bottom: When examining only correctly identified word trials, log RT was sensitive to changes in rSNR, but did not reliably exhibit differences due to lexical competition within each rSNR. *$p < .05$

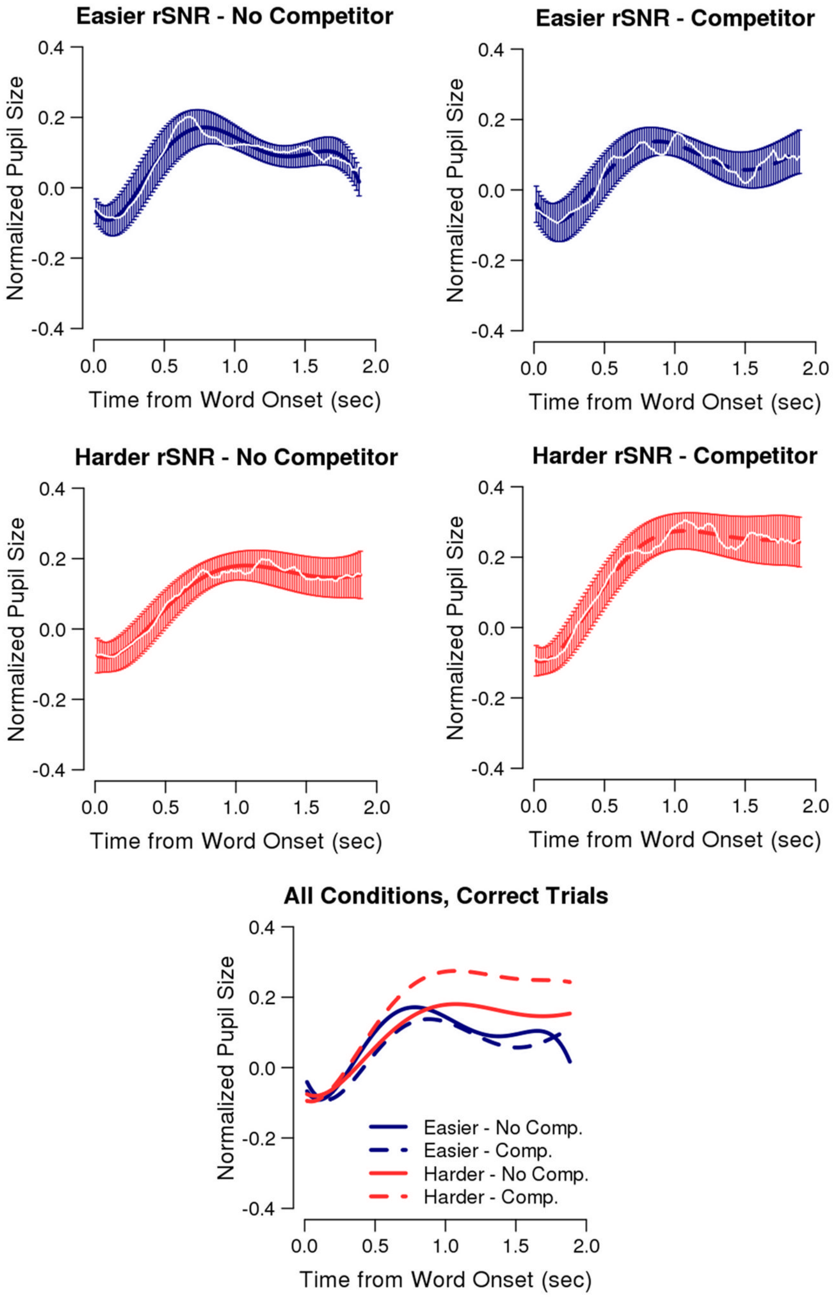

Figure 4.

Pupil size data relative to baseline (white lines) and fitted model responses (colored lines) are plotted as a function of time from word onset. The upper graphs also plot the standard error of the mean for the fitted model responses. The bottom graph shows the fitted model responses for each condition across trials for which participants correctly identified the word. Note that mean peak-picked values from traditional analyses will necessarily be greater than the highest amplitude in this figure: peak-picked amplitudes occurred at different latencies for each condition for each individual, and thus for a given time point in the above pupil response curves, peak values were averaged with non-peak values.

Table 1

Results of the rSNR × Lexical Competition Interaction Model

	(A) Model Fit			(B) Easier rSNR C > NC			(C) Harder rSNR C > NC			(D) No Competitor H > E rSNR			(E) Competitor H > E rSNR		
Term	**−2LL**	**L Ratio**	***p***	**β**	***t***	***p***	**β**	***t***	***p***	**β**	***t***	***p***	**β**	***t***	***p***
Intercept	116178	3.09	.08	−24.32	−0.94	.36	72.58	2.91	.008	22.51	0.84	.41	119.41	3.55	.002
Linear†	116177	0.90	.34	87.56	0.51	.61	339.90	2.27	.02	365.23	2.20	.03	617.57	4.12	<.0001
Quadratic	116122	55.53	<.0001	151.70	5.11	<.0001	−193.13	−6.16	<.0001	95.25	3.41	.0007	−249.58	−8.43	<.0001
Cubic	116117	5.19	.02	−54.44	−1.53	.07	50.81	1.62	.11	−187.33	−6.70	<.0001	−82.08	−2.77	.006
Quartic	116099	17.59	<.0001	201.97	6.80	<.0001	8.31	0.27	.79	85.96	3.08	.002	−107.70	−3.64	.0003
Quintic	116087	11.67	.0006	116.66	3.93	<.0001	−40.93	−1.31	.19	209.66	7.50	<.0001	52.07	1.76	.08

−2LL = −2*Log Likelihood; L Ratio = Likelihood Ratio

†No significant rSNR × Lexical Competition interaction, but significant main effect of rSNR on the linear term. Including the main effect of SNR on the linear term [−2LL = 116178] improves fit compared to a model without the term [−2LL = 116191; L Ratio = 12.61, *p* = .0004]. The same was not true for the main effect of competition on the linear term [−2LL = 116181; L Ratio = 2.54, *p* = .11].

Note: The best-fitting model is summarized following the notation used in Mirman et al. (2008):

$$Y_{ij}=Time_{ij}*\left(\sum_{P=[0:5]}\beta_{P0}+\sum_{P=[0:5]}\beta_{Pi}+\sum_{P=[0:5]}\beta_{P0S}*S+\sum_{P=[0,2:5]}\beta_{P0L}*L+\sum_{P=[0,2:5]}\beta_{P0SL}*S*L\right)+\varepsilon_{ij}.$$

The dependent measure, pupil size, is denoted by Y_{ij}, with subscripts that refer to individual participants, i = [1:21], and measurements, j = [1:113]. Time_{ij} and the fixed effects error term, ε_{ij}, use these same subscripts. Parametric effects are denoted β. The first subscript on the beta parameters, P, indexes the order of polynomial effects (0 = intercept, 1 = linear, 2 = quadratic, etc.), while the second subscript, i, indexes individual participants or denotes effects across individuals with i = 0, as in β_{P0}. S and L index rSNR and lexical competition conditions, respectively. The intercept and linear terms that varied by individual, β_{0i} and β_{1i}, were computed by partitioning random effects related to individuals and experiment conditions, ξ_{PiSL}, from the fixed effects, γ_{Pi}, such that $\beta_{Pi} = \gamma_{Pi} + \xi_{PiSL}$.