Published in final edited form as:
Int J Psychophysiol. 2012 January ; 83(1): 56–64. doi:10.1016/j.ijpsycho.2011.10.002.

Memory strength and specificity revealed by pupillometry

Megan H. Papesh, **Stephen D. Goldinger**, and **Michael C. Hout**
Arizona State University

Abstract

Voice-specificity effects in recognition memory were investigated using both behavioral data and pupillometry. Volunteers initially heard spoken words and nonwords in two voices; they later provided confidence-based old/new classifications to items presented in their original voices, changed (but familiar) voices, or entirely new voices. Recognition was more accurate for old-voice items, replicating prior research. Pupillometry was used to gauge cognitive demand during both encoding and testing: Enlarged pupils revealed that participants devoted greater effort to encoding items that were subsequently recognized. Further, pupil responses were sensitive to the cue match between encoding and retrieval voices, as well as memory strength. Strong memories, and those with the closest encoding-retrieval voice matches, resulted in the highest peak pupil diameters. The results are discussed with respect to episodic memory models and Whittlesea's (1997) SCAPE framework for recognition memory.

In the present study, we examined the extent to which memory strength and specificity for spoken items are revealed by pupillometry across learning and recognition. Although the speech signal is characterized by idiosyncratic variations that listeners must fluently overcome, debate surrounds the necessity of *encoding* this information into memory during on-line perception. For example, people encounter little perceptual resistance when processing the same words spoken by different speakers, each of whom has a unique vocal structure, pattern of intonation, and speaking rate. Changes in context and other non-linguistic variables similarly pose little challenge to the perceptual system. Speech perception is clearly robust to idiosyncratic variations in the input signal, but do these variations get "filtered out" during perception, or are they somehow stored in a detailed memory trace, capable of affecting subsequent perception or retrieval?

Two general, and opposing, approaches to this problem have dominated the literature. According to the first, the speech signal is stripped of idiosyncratic information upon encoding, allowing the perceiver to activate abstract representations in memory (Joos, 1948). Such theories generally treat idiosyncratic variations as undesirable noise in the speech signal, a problem for the perceptual system to overcome (Pisoni, 1993). According to the second approach, surface properties of speech are stored in unique, episodic traces; surface information is not noise, but is instead utilized to aid subsequent recognition (McLennan & Luce, 2005). These theoretical approaches are denoted *abstractionist* and *episodic* theories, respectively. Both views have empirical support, and either would allow

Please address correspondence to M. H. Papesh, Department of Psychology, Box 871104, Arizona State University, Tempe, AZ 85287-1104., megan.papesh@asu.edu, Ph: 1-480-965-1377; Fax: 1-480-965-8544.

listeners to resolve variability in speech, yielding immediate and effortless mapping of speech signals to segmental and lexical representations.

Abstractionist theories posit *normalization* mechanisms that "correct" the speech signal for its idiosyncratic properties, allowing perception to operate at the level of stored, ideal representations. Normalization is proposed to explain the constancy of linguistic perception, despite variations across talkers and contexts (see Bradlow, Nygaard, & Pisoni, 1999; Magnuson & Nusbaum, 2007). For example, Marlsen-Wilson and Warren (1994; Lahiri & Marslen-Wilson, 1991) proposed an account whereby variations in spoken words are immediately resolved; lexical access occurs when a stored lexical unit receives sufficient activation from this corrected input (see McClelland & Elman, 1986). Attention is directed to the level of word meaning (as it should be); later, memory for surface information should be negligible. Abstractionist theories are both logically appealing (see Bowers, 2000; McQueen, Cutler, & Norris, 2006) and have empirical support. For example, early in word perception, priming effects appear to be mediated by abstract phonemic representations, devoid of superficial details (McLennan & Luce, 2005; see also McQueen, Cutler, & Norris, 2003).

Although the normalizing process has intuitive appeal, there are many empirical demonstrations that spoken word perception creates detailed, episodic memory traces. In theoretical terms, episodic theories have many desirable properties. For example, the correctly predict that speech perception becomes more robust as people are exposed to a wider range of exemplars (Lively, Logan, & Pisoni, 1993). Of greater importance, they provide a natural mechanism to explain *specificity effects* in perception and memory (Goldinger, 1996; McLennan & Luce, 2005). Across many experimental paradigms, performance is affected (usually improved) when items presented in a study phase are later repeated in a test phase. Such improvements are typically stronger when surface information (such as the speaker's voice) is preserved across study and test. These *voice specificity effects* are especially robust in implicit memory. For example, when the voice associated with a word changes across study and test, it reduces performance in perceptual identification (Church & Schacter, 1994; Goldinger, 1996; Pilotti, Bergman, Gallo, Sommers, & Roediger, 2000; Sheffert, 1998), naming (Goldinger, 1998), lexical decision (Luce & Lyons, 1998), and word-stem completion (Church & Schacter, 1994; Schacter & Church, 1992, but see Pilotti et al., 2000). Exposure to voice-specific tokens of words can also affect later speech production (Goldinger, 1998; Goldinger &Azuma, 2004; Pardo, 2006; Shockley, Sabadini, & Fowler, 2004).

In tests of explicit memory, word recognition is again typically best for items that are tested in their original study voices. For example, Palmeri et al. (1993; Sheffert & Fowler, 1995; Senkfor & Van Petten, 1998) observed benefits for same-voice repetitions in continuous recognition memory, relative to changed-voice repetitions. Voice effects were robust for lists including as many as 20 talkers and study-test lags up to 64 items. Goldinger (1996) further examined voice effects, manipulating the similarity among talkers, levels of initial processing, and extending the retention time out to a week. He found same-voice advantages in both implicit and explicit measures, although the same-voice advantage vanished in explicit recognition after a 1-week delay (see also Goh, 2005). In the present research, we further examined voice effects in recognition memory, complementing the standard paradigm with real-time measures of changes in pupil diameters. By doing so, we were able to assess participants' cognitive effort during word encoding and test, with special emphasis on potential voice effects: If people encode voice-specific traces of spoken words, later repetition of those same tokens should reduce cognitive effort, as indicated by the pupillary reflex.

The present paradigm also allowed us to assess whether voice changes across study and test influence memory strength. By obtaining 1–6 confidence estimates alongside recognition judgments (e.g., 1 = *very sure new*; 6 = *very sure old*), we examined subjective memory strength in two ways. First, confidence estimates allowed us to create receiver operating characteristic (ROC) curves, plotting the hit and false-alarm rates at various levels of confidence or bias (Macmillan & Creelman, 2005). ROC curves are also commonly *z*-transformed and plotted on standardized axes (*z*-ROCs). Different memory theories make different predictions regarding the shapes of these curves (Wixted, 2007; Yonelinas & Parks, 2007). For example, *dual-process* theories (e.g., Yonelinas, 1994) assume that two processes subserve memory decisions: A thresholded, all-or-none recollection process and a graded, strength-based familiarity process. These theories generally predict that ROC curves will be linear and *z*-ROCs will be curvilinear. On the other hand, *strength-based* theories (Wixted & Mickes, 2010) assume that recollection and familiarity signals are both graded and summed into a single memory strength signal, which is used as the basis for recognition decisions. Strength theories generally predict curvilinear ROCs and linear *z*-ROCs. In the present study, we collected confidence estimates, allowing us to assess the underlying strength distributions for targets and lures. We also examined pupillary reflexes during study as a function of *subsequent* confidence. If pupil dilation reflects part of the recognition memory process, as suggested by Võ et al. (2008) and Kafkas and Montaldi (in press), then pupillary changes during encoding may accurately "track" subsequent estimates of memory strength. This approach allowed us to assess whether pupillometry reveals differences in strong versus weak memories across both encoding and retrieval.

Pupillometry

Pupillary reflexes occur during all forms of visual and cognitive processing, and are hypothesized to reflect brain activity during processing (Beatty & Kahneman, 1966). Enlarged pupils are typically associated with increased cognitive demand (Porter, Troscianko, & Gilchrist, 2007) and provide sensitive indices of cognitive effort, similar to ERP waveforms (Beatty, 1982). Using the *subsequent memory paradigm*, researchers can compare neurophysiological measures across study and test to differentiate the neural activity associated with subsequently remembered versus forgotten information. Such investigations have been reported using fMRI (e.g., Ranganath, Cohen, & Brozinsky, 2005) and ERP (Cansino & Trejo-Morales, 2008; Duarte, Ranganath, Winward, Hayward, & Knight, 2004; Guo, Duan, Li, & Paller, 2006). In the current investigation, we used pupillometry to compare the effort involved in cognitive operations across study and test, with emphasis on voice effects. We had two key questions: First, does greater effort during encoding predict greater success during recognition? And second, does preservation of voice information influence processing during test?

The appeal of pupillometry to the investigation of cognitive phenomena lies in its automaticity. Pupillary reflexes are controlled by the sympathetic and parasympathetic systems, which hold reciprocal connections to central nervous system (CNS), suggesting that they may exert an influence on CNS structures relevant to cognition (Gianaros, Van der Veen, & Jennings, 2004). Pupils dilate following sympathetic system activation and/or parasympathetic system inhibition and constrict following activity of the parasympathetic system (Steinhauer, Siegle, Condray, & Pless, 2004). Although the pupils change reflexively in response to general factors, such as emotional arousal and anxiety, such *tonic* changes are independent of *phasic* changes, which arise upon the onset of stimuli for cognitive processing. Such phasic changes are known as task-evoked pupillary responses (TEPRs), and have long been used to infer cognitive effort across domains such as lexical decision (Kuchinke, Võ, Hofmann, & Jacobs, 2007), attention allocation (Karatekin, Couperus, & Marcus, 2004), working memory load (Granholm, Asarnow, Sarkin, & Dykes, 1996; Van

Gerven, Paas, Van Merriënboer, & Schmidt, 2004), face perception (Goldinger, He, & Papesh, 2009), and general cognitive processing (Granholm & Verney, 2004). In fact, Kahneman (1973) used TEPRs as his primary index of mental processing load in his theory of attention, owing to its sensitivity to variations within or between tasks, and its ability to reflect individual differences in cognitive ability.

The relationship between human memory and the pupillary reflex has seldom been investigated, but animal models suggest a potential relationship between pupil dilation and memory encoding/retrieval (Croiset, Nijsen, & Kamphuis, 2000). As has been shown with rats (Clark, Krahl, Smith, & Jensen, 1995), stimulating the vagus nerve in the parasympathetic pathway in humans (patients undergoing treatment for epilepsy) enhances memory retention, if stimulation is applied during consolidation (Clark et al., 1999). Such findings suggest a modulatory influence of autonomic activity on memory formation and retrieval. The first study examining human memory and the pupillary reflex was reported by Võ et al. (2008), who noted the similarity between pupillary and ERP waveforms, which are known to reflect memorial processes (Dietrich et al., 2000; Johansson, Mecklinger, & Treese, 2004). Võ et al. observed a "pupillary old/new effect," wherein pupils were larger during study trials leading to hits, relative to correct rejections. They interpreted this effect in a dual-process framework (Yonelinas, 2001, 2002), suggesting that enlarged pupils were observed for hits because they included recollection, which is hypothesized to be a slow, cognitively demanding process. Similar effects were reported by Papesh and Goldinger (2011), who found a pupillary old/new effect across study and test presentations of auditory low- and high-frequency words. Specifically, when participants studied words that were subsequently remembered, those trials were associated with enlarged pupils, relative to subsequently forgotten and new words. This pattern was especially strong for low-frequency words, suggesting that memorial encoding, coupled with the cognitive operations usurped in processing low-frequency words (see Goldinger & Papesh, 2009; Kuchinke et al., 2007; Papesh & Goldinger, 2008), resulted in an overall increase in cognitive demand. In prior studies (Goldinger, 1996; 1998), voice effects were stronger for low-frequency words, relative to high-frequency words, and were stronger for nonwords, relative to words. To maximize our likelihood of observing voice effects in the pupillary responses, we used all three classes of stimuli in the present design.

Related to the current study, Kafkas and Montaldi (in press) employed a modified "remember/know" task (from Montaldi, Spencer, Roberts, & Mayes, 2006) to examine the relationship between pupillary reflexes and subsequent memory strength. In their study, participants incidentally encoded a series of images, and were later asked to rate recognized items along a "strength" scale (F1 = weakly familiar, F2 = moderately familiar, F3 = strongly familiar, R = inadvertent recollections), and to call unrecognized items "new." The authors found that pupil *constriction* at encoding was associated with subsequently stronger memories. This finding was interpreted to reflect a strong influence of parasympathetic activity during the incidental encoding of visual information.[1]

Unique to our study, we recorded pupil diameters during both encoding and retrieval. Prior memory studies have employed pupillometry in a subsequent memory procedure, comparing pupil diameters during study trials that lead to eventual hits or misses. This focus allows researchers to avoid analyzing reflexes that may reflect the decision and motor processes involved in issuing a memory judgment. In the present study, however, we were interested in the entire memory process, including encoding, strength, and decision dynamics.

[1]We consider this apparent contradiction in findings in the Discussion, but presently note that differences in both stimulus materials (visual scenes vs. spoken words) and encoding procedures (incidental vs. intentional learning) make the present study difficult to directly compare to the study by Kafkas and Montaldi (in press).

Although the abstractionist/exemplar debate nicely encapsulates the encoding process, and memory strength describes confidence, we suggest that Whittlesea's (1997) SCAPE framework is well-suited to explain pupil data derived from test trials. According to SCAPE, memory decisions are made in a two-stage process, including the production (i.e., generation, or "calling to mind") of prior memory states and the evaluation of production fluency (Whittlesea & Leboe, 2000; Whittlesea & Williams, 1998; 2001). Pupillometry represents an ideal method by which to test the SCAPE framework. If people truly engage in the hypothesized generation process, it should be reflected by similar pupil diameters across encoding and retrieval. We hypothesized that pupillary changes during study trials would predict memory accuracy and confidence during test trials. We also expected to observe "pupillary voice effects" in recognition memory, allowing us to contrast predictions from episodic and abstractionist theories of lexical access.

Method

Participants

Twenty-nine Arizona State University students participated in exchange for partial course credit. All participants were native English speakers with normal (or corrected-to-normal) vision and no known hearing deficits. Four participants were dropped from analysis for having more than 6% missing fixations, and three participants were excluded for poor recognition performance (2 for extreme liberal response criteria and 1 for extreme conservative responding, as indexed by the signal-detection bias index, *C*), leaving 22 participants (12 men, 10 women, M_{age} = 18.82 years, SD_{age} = .65) for analysis.

Materials

Twelve native English speakers (6 women and 6 men in their early 20's), with no discernable accents or abnormal speech characteristics, volunteered to record items for this experiment. From each person, we recorded 160 nonwords[2] and 160 words; of the words, half were high frequency (HF) and half were low frequency (LF). Stimuli were recorded at 41000 Hz, using GoldWave software in a sound-attenuated booth, and were digitally spliced into individual wav files. WavePad Sound Editor was used to equate the mean RMS amplitudes of the sound files. All items were pseudo-randomly assigned to four study-test lists, such that word frequency and lexicality (i.e., words versus nonwords) were represented equally across lists. All stimulus items are listed in the Appendix and relevant linguistic data (obtained from the English Lexicon Project; Balota et al., 2007) are provided in Table 1.

From this initial pool of 12 speakers, only four were used in each experiment (randomly selected for each participant to avoid potential artifacts arising from any particular voice). The four voices were always selected to be half male, and of those voices, one male and one female were randomly selected to be studied voices. The remaining voices were used as novel test speakers.

Apparatus

The experiment was presented on a Tobii 1750 17-in. (43.18-cm) monitor, with stimulus presentation and data collection managed by E-Prime 1.2 software (Psychology Software Tools, 2006). Auditory stimuli were played at a comfortable listening level over Sennheisser HD-250 headphones. A chin rest maintained participants' viewing position at 60 cm, and pupil diameters were monitored binocularly at 50 Hz. The lighting in the room was set to a constant dim level for all participants.

[2]Nonwords were recorded by having the speaker shadow a pre-recorded version of the items spoken by the experimenter, to ensure that all nonwords were pronounced similarly across speakers.

Procedure

Participants were first familiarized with the experiment and the eye tracker. The chin rest was adjusted so that the position of the eyes was maintained centrally on the computer screen, and the eye tracker was calibrated. This procedure establishes a map between each participant's known gaze position and the eye tracker's coordinate estimate of that position. The routine proceeds by having participants follow a blue dot as it moves to 9 locations on the screen. If the software or the researcher identified any missed fixations, the calibration routine was repeated. All participants were successfully calibrated within two attempts.

During the study phase of the experiment, participants were presented with 80 items (40 nonwords, 20 HF words, and 20 LF words) spoken by two speakers, one male and one female. Assignment of items to voices was random, with the constraint that voices be used equally often within each item type. Although the only visible element was a 1000-ms fixation cross to initiate each trial, we encouraged participants to keep their gaze on the computer screen throughout the session by informing them that off-screen fixations would slow the trial progression.

Following the study phase, participants completed three 60-s computer mouse-tracking games, which required them to use the mouse (with an open-circle cursor) to follow a moving target around the screen. During the test trials, participants listened to 160 items and, approximately 1000 ms later, made old/new judgments by issuing confidence estimates along a 6-point scale, where 1 represented "*very sure new*" and 6 represented "*very sure old*." To ensure that participants maintained their gaze on the computer screen, confidence estimates were made by clicking one of six on-screen boxes, numbered along the scale. Test voice was manipulated across old items by presenting them in one of three voice types, the studied voice, the familiar opposite-gender study voice, or a new, completely novel male or female voice (assignment of words to voices was again random, with the restriction that each change type be used equally often with each stimulus type). New test items were spoken either by one of the studied speakers, so that the voice was familiar, or a new speaker. Participants were not given feedback. The experiment lasted approximately 45 minutes.

Results

Recognition Accuracy

For all analyses, alpha was set to .05 and multiple comparisons were Bonferroni corrected. Overall recognition accuracy was examined by computing the signal detection index for sensitivity, d'. Participants' average d' was .94, and there were no statistically reliable differences across word types (HF, LF, nonword), $F(2, 21) = .01$, $p = .98$. This equivalence is reflected in the group-level ROC and *z*-ROC, depicted in the top half of Figure 1. Note that, across word types, the lines are essentially on top of one another and, as such, all individual-level statistics were collapsed across word type. As is evident from the left panel of the figure, the group ROC is curvilinear, yet the right panel displays a linear group *z*-ROC. Recall that such an ROC profile is consistent with strength-based recognition theories, rather than dual-process theories. These subjective impression were supported by statistics obtained from individual subject ROCs and *z*-ROCs, presented in Table 2[3]. Average quadratic constants in individual ROCs ($M = .55$, $SE = .21$) were reliably above zero, $t(21) = 2.55$, $p < .05$, and a quadratic fit of the data accounted for greater variance ($M = 95\%$, $SE = .$

[3]Although ROCs are typically plotted at the group-level, statistical analyses are performed on individual ROC analyses to prevent averaging artifacts (Brown & Heathcote, 2003; Wickens, 2002; see Yonelinas & Parks, 2007, for a review of memory ROC measurement and analysis issues).

01), relative to a linear fit (M = 88%, SE = .01), $t(21) = -6.49$, $p < .05$. In the individual z-ROCs, the average quadratic constant did not differ reliably from zero ($M = -.01$, SE = .03), $t(21) = 0.38$, $p = .71$, indicating a linear progression of points. Further, although a quadratic fit to the data did accounted for a greater proportion of the variance (M = 98%, SE = .003), relative to a linear fit (M = 96.7%, SE = .006), $t(21) = -3.75$, $p < .05$, the difference was relatively small 1.3%.

The influence of voice was also examined by ROC analyses, allowing us to assess whether voice matches increased memory strength generally, or prompted more instances of recollection. Group-level ROCs and z-ROCs are presented in the bottom half of Figure 1. For ease of presentation, we collapsed the voice manipulation into familiar and unfamiliar voices. Familiar voices were those that were used in study trials, regardless of the old/new status of the item. Unfamiliar voices were those that were novel at test. As with the plots by word type, the ROC appears curvilinear, and the z-ROC appears linear. These conclusions were supported by statistics from individual ROC data. Average quadratic constants from familiar and unfamiliar voice trials were compared to each other, and to zero. Both sets produced quadratic constants reliably different from zero, $t(21) = -4.82$, $p < .05$ (familiar) and $t(21) = -2.69$, $p < .05$ (unfamiliar), but not from each other (see Table 2). The proportion of variance accounted for by linear and quadratic fits to the data were also compared in a 2 (Equation: linear/quadratic) × 2 (Voice) within-subjects ANOVA. Only the main effect of Equation was statistically reliable, revealing that the quadratic fit accounted for more of the variance (M = 97%, SE = .005), relative to the linear fit (M = 91%, SE = .01), $F(1, 21) = 21.59$, $p < .05$, $\eta^2_p = .51$, suggesting that the ROCs were predominantly curvilinear.

Individual z-ROCs were also examined by voice. Quadratic constants for familiar voices ($M = -.03$, SE = .09) and unfamiliar voices (M = .05, SE = .34) did not differ reliably from each other or zero, all ts < .5, $p > .7$. The proportion of variance accounted for was analyzed in a 2 (Equation) x 2 (Voice) within-subjects ANOVA. The quadratic fit again provided the better fit of the data (M = 96.7%, SE = .004), relative to the linear fit (M = 93.8%, SE = .01), $F(1, 21) = 20.7$, $p < .05$, $\eta^2_p = .49$. Although the benefit for the quadratic fit was relatively small (~3%), it was statistically reliable. The interaction between Voice and Equation, $F(1, 21) = 7.24$, $p < .05$, $\eta^2_p = .26$, was consistent with the main effect of Equation, suggesting that the quadratic equation fit the data better, regardless of voice. Taken together, the ROC results suggest that voice matches generally increased the strength of recognition probes, rather than creating a separate category of recollected trials.

To examine the prediction that same-voice repetitions would enhance recognition accuracy, confidence estimates were collapsed into hits, false alarms, correct rejections, and misses. Note that signal-detection analyses were not appropriate on these data because lures have no inherent voice information, leaving only one false-alarm rate. Therefore, we analyzed hit rates as a function of voice change, in a one-way ANOVA with Voice (same, familiar, new) as a within-subjects variable. Univariate tests indicated that voice affected hit rates, $F(2, 21) = 6.38$, $p < .05$, $\eta^2_p = .38$, such that same voices resulted in greater hits (M = 67%, SE = .02), relative to new voices (M = 59%, SE = .02). No differences were observed for familiar voices (M = 63%, SE = .03). No reliable differences were observed in false alarms across familiar (M = 33%, SE = .02) and new (M = 30%, SE = .03) voices, $t(22) = 1.53$, $p = .14$.

Recognition RTs

Response times were trimmed prior to analysis to remove outliers, defined as responses occurring sooner than 250 ms or longer than 3 standard deviations above the cell mean. To examine the influence of confidence, RTs were collapsed across word type to avoid missing data. A one-way ANOVA on RT by confidence revealed that participants' confidence

influenced the speed of responding, $F(5, 18) = 3.69$, $p < .05$, $\eta^2_p = .51$. This main effect was driven primarily by high-confidence decisions, which were made faster ($M = 1131$ ms, $SE = 47$) than decisions at all other confidence levels, with the exception of level 4 ($p = .06$).

To examine the influence of voice during old test trials, RT data were analyzed in a 3 (Voice: same/familiar/new) × 3 (Word Type) within-subjects, repeated-measures ANOVA. Although we predicted that changed voices would yield slower RTs, no reliable differences in RT emerged as a function of voice, $F(2, 20) = 1.55$, $p = .24$, nor did we observe any influence of word type, $F(2, 20) = 1.29$, $p = .29$. The effect of voice was also analyzed in new test trials in a 2 (Voice: familiar/new) × 3 (Word Type) within-subjects, repeated measures ANOVA. As in old test trials, there were no reliable effects of voice, $F(1, 21) = 3.26$, $p = .08$, or word type, $F(2, 20) = 1.55$, $p = .23$.

Pupil Diameters

Prior to analysis, pupil data from each participant's "better" eye (i.e., the eye with fewer missing observations) were corrected for missing observations by linear interpolation across a 100-ms window around the missing cell. This resulted in fewer than 6% corrected cells per participant. All pupil diameters were baseline-corrected on a trial-by-trial basis by subtracting the observed diameter from the average diameter during the fixation cross on that trial. Because of differences in response frequencies (e.g., some participants never responded with a confidence estimate of "1" to a studied LF word), several analyses collapsed across variables, as with the behavioral data above. Peak diameters were calculated for each trial during the window between stimulus onset and 1000 ms following the old/new RT. Although we did not tightly equate the stimulus characteristics (e.g. duration) across words and nonwords, we analyzed peak diameters to reduce the possibility that item length unduly influenced our findings.

Study Trials—Peak diameters were analyzed in a 2 (Subsequent Accuracy: hit/miss) × 3 (Word Type) within-subjects ANOVA. Although a main effect of Word Type revealed that nonwords resulted in enlarged pupils ($M = 1.17$ mm_d, $SE = .07$), relative to HF and LF words (both .99 mm_d), $F(2, 20) = 14.7$, $p < .05$, $\eta^2_p = .60$, this effect was qualified by an interaction with subsequent accuracy, $F(2, 20) = 8.94$, $p = .002$, $\eta^2_p = .47$. Pairwise comparisons revealed that differences in peak diameter by word type were only observed in trials leading to hits, $F(2, 20) = 23.1$, $p < .01$, $\eta^2_p = .70$. No reliable differences in word type were observed in trials leading to misses, $p = .49$.

To examine whether memory strength yields differences in pupil diameter during encoding, peak diameters during study were analyzed by subsequent confidence estimate, collapsed across word type in a one-way ANOVA. Results revealed that study trials leading to high-confidence hits yielded the largest peak diameters ($M = 1.31$ mm_d, $SE = .08$), relative to all other confidence estimates, $F(5, 19) = 7.02$, $p = .001$, $\eta^2_p = .65$. Although there was a trend for peak diameter to decrease with decreases in subsequent confidence (see Figure 2), pairwise comparisons were not statistically reliable.

Old test trials—Peak diameters were analyzed in a 2 (Accuracy: hit/miss) × 3 (Word Type) within-subjects ANOVA, showing that peak diameters were larger during correct ($M = 1.23$ mm_d, $SE = .08$), relative to incorrect ($M = 1.14$ mm_d, $SE = .08$), trials, $F(1, 22) = 5.55$, $p < .05$, $\eta^2_p = .21$. Further, and mirroring the RT data, a one-way ANOVA on confidence revealed an influence of confidence on peak diameter, such that high-confidence decisions yielded the largest peak diameters ($M = 1.44$ mm_d, $SE = .12$), $F(5, 19) = 4.98$, $p = .004$, $\eta^2_p = .57$. Although there appeared to be a gradual decrease in peak diameter with decreasing

confidence estimates (see Figure 3), no other pairwise comparisons were statistically reliable.

To test the SCAPE prediction that pupil diameters at retrieval should be similar to those during encoding, we examined the influences of word type and voice change in a 2 (Word Type) × 3 (Voice: same, familiar, new) within-subjects ANOVA. As in the study trials, we observed a main effect of Word Type, such that nonwords resulted in the greatest peak diameter (M= 1.46 mm_d, SE = .09), followed by LF words (M = 1.37 mm_d, SE = .10), and HF words (M = 1.29 mm_d, SE = .10), $F(2, 20) = 9.33$, $p = .001$, $\eta^2_p = .48$. As predicted by SCAPE, voice also affected participants' peak pupil diameters, $F(2, 20) = 15.04$, $p < .01$, $\eta^2_p = .60$. When words were repeated in their studied voices, participants' peak pupil diameters were reliably larger (M = 1.51 mm_d, SE = .10), relative to familiar voice trials (M = 1.33 mm_d, SE = .09) and new voice trials (M = 1.28 mm_d, SE = .10).

New test trials—Peak diameters were analyzed in a 2 (Voice: familiar/new) × 3 (Word Type) within-subjects, repeated-measures ANOVA. No reliable main effects or interactions emerged, both Fs < 1.0.

Discussion

The present results add to the literature relating episodic memory to lexical processing, replicating findings that voice matches across study and test improve recognition memory. Our results further demonstrate that the strength and specificity of memory are observable in a physiological index of cognitive effort, the pupillary reflex. Using the subsequent memory paradigm, we found that, when participants devoted greater cognitive effort (reflected by larger pupils) to encoding, they were more accurate at test (as in Võ et al., 2008). We further demonstrated that encoding effort is directly related to subsequent memory strength, as reflected by overt confidence ratings and recognition accuracy. This effect was not limited to encoding: When participants accurately recognized old items during test, their pupils were again more dilated, relative to when they missed or committed false-alarms. High-confidence decisions were reliably associated with larger peak diameters, as were same-voice repetitions. These results represent the first demonstration of voice specificity revealed by pupil dilations. In fact, the pupillary reflex was sensitive to the strength and content of memory; trials in which memory strength was strong, and those in which retrieval cues matched encoding cues, yielded the largest peak diameters.

Recognition accuracy was greatest for words repeated in their study voices, replicating previous work (Goldinger, 1996; Lively, 1994; Sheffert, 1998). Although we did not observe a reliable difference between hit rates for words repeated in studied voices versus familiar voices, hit rates numerically increased with voice familiarity, a finding that is compatible with those from Goh (2005). Subsequent signal detection analyses collapsing voice into *familiar* and *unfamiliar* revealed no d' differences across familiar and new voices (p = .08). We did, however, observe a difference in bias, c, which is centered at zero (Macmillan & Creelman, 2005). When presented with an unfamiliar voice, participants were conservatively biased (.17), relative to when they were presented with a familiar voice (.01), $t(22) = -3.46$, p = .002. Like Goh (2005), we found that voice familiarity encouraged more liberal responding.

The present experiment joins a handful of controlled studies to examine recognition memory through behavioral and physiological indices. By examining pupil sizes, we gauged the cognitive effort devoted to both learning and recognition across experimental manipulations, including lexicality, word frequency, and voice. Although word frequency did not strongly influence pupil responses, we observed a strong lexicality effect, such that participants'

pupils were larger during all nonword trials. Although previous researchers have documented that word frequency typically affects pupillary reflexes (Kuchinke et al., 2007; Papesh & Goldinger, 2008), the tasks employed to elicit this effect were predominantly perceptual in nature, without instructions to intentionally memorize or retrieve words. The present findings suggest that, in the context of intentional learning, pupils reveal extra cognitive effort devoted to memorizing nonwords, relative to HF and LF words. When participants encoded and retrieved nonwords, their pupils dilated significantly more, relative to when they encoded and retrieved both HF and LF words. Subsequent analyses suggested that these effects on peak dilation were not due to slight differences in item length, but instead reflect the difficulty of processing novel phonetic strings. Considering that pupils are typically smallest for HF words (Kuchinke et al., 2007; Papesh & Goldinger, 2008), and that LF words are often remembered best (Glanzer & Adams, 1990), this finding is consistent with the literatures on word perception and recognition memory.

The pupillary results also reflected voice effects in recognition memory, which have not been consistent in behavioral research. Whereas several studies have shown that same-voice repetitions of words are more accurately remembered, relative to changed-voice repetitions (e.g., Goh, 2005; Goldinger, 1996; Sheffert, 1998), others have found voice effects only in tests of implicit memory, such as stem completion or perceptual identification (e.g., Church & Schacter, 1994; Schacter & Church, 1992). Our results revealed both behavioral and physiological evidence for voice specificity effects. When study words were later repeated in their original voices, participants' pupils were essentially unchanged across encoding and recognition: Response confidence was high and peak pupil diameters reflected this increase in memory strength. These effects are consistent with instance-based memory models (e.g., MINERVA 2, Hintzman, 1986, 1988; see Goldinger, 1998).

Our pupil results also revealed a previously undocumented, but intuitively predictable, pattern: Participants devoted greater cognitive resources to encoding items that were subsequently remembered with higher degrees of confidence, relative to those that were remembered with less confidence or subsequently forgotten. This finding suggests that encoding strength can accurately predict whether a person will remember an item in a subsequent memory test. Although Kafkas and Montaldi (in press) observed the opposite pattern, we do not consider the results to be irreconcilable. Several differences in stimuli and in study-test procedures could have caused Kafkas and Montaldi to observe pupillary constriction with high-confidence memories, whereas we observed dilation with high-confidence memories. Importantly, their procedure involved incidental encoding of images, whereas our procedure involved intentional memorization of spoken words. In the former case, it is reasonable to hypothesize that easily identified images would trigger little effort during initial viewing, and would likely support accurate recognition during test. In our case, all items were spoken, removing any tonic changes due to visual onset. Because we required intentional memorization, it is not surprising that greater apparent effort during encoding led to greater performance during test. Whereas Kafkas and Montaldi interpreted their findings as reflecting parasympathetic activity, because incidental encoding does not impose cognitive demand or impart any emotional significance, we interpret our findings to reflect parasympathetic inhibition and/or sympathetic activity, because intentional encoding is a cognitively demanding task.

A novel finding appeared in the analyses on pupil dilations during the recognition test, wherein we observed effects of voice and confidence. Specifically, pupil diameters were larger (1) when people were highly confident and accurate and (2) when the study and test voices matched. Less confident memories were not characterized by a lack of dilation, but by progressively smaller peak diameters. This finding is compatible with Whittlesea's (1997) SCAPE framework for recognition decisions. In SCAPE, recognition memory

reflects a two-stage process. According to Whittlesea and Leboe (2000; Whittlesea & Williams, 1998; 2001), the first stage is *production* of mental states, which involves elaboration of perceptual inputs. During this stage, the person receives input, and then "fills it in" by bringing associated images or ideas to mind (Neisser, 1967). The second stage is *evaluation*, wherein the person automatically evaluates the production functions. As discussed by Whittlesea (1997; Leboe & Whittlesea, 2002), this is not direct evaluation of the *stimulus*, such as comparing memory strength to a decision criterion. Instead, evaluation indexes the relative harmony of mind.

To illustrate these processes, imagine that you encounter a coworker after she has changed her hairstyle: The production process will easily recognize your acquaintance, but the evaluation process will produce a sense of disharmony – you know that "something is different." According to Whittlesea and Williams (2001), this reflects a *discrepancy-attribution* process. In most situations, people have implicit expectations of fluent processing. You expect to see your coworker, easily recognize her, and move on. As a result, fluent processing typically creates no "feelings" of memory – you simply know your acquaintance. When processing is dysfluent, however, the evaluation process generates a mismatch signal, which is experienced as a feeling of unexplained familiarity. Despite standard usage, feelings of familiarity do not generally evoke a sense of *memory*, but rather a sense of memory failure (Mandler, 1980). In SCAPE, the evaluation process signals potential memory failures whenever perceptual inputs engage "extra" elaboration during production. In terms of the present experiment, words studied and tested new voices will generate (relative) processing dysfluency, interpreted in a recognition test as novelty.

We anticipated and observed that processing fluency at test, due to same-voice repetitions, would elicit accurate memory performance. We also observed, however, that pupils were enlarged for items that were correctly recognized, relative to those that were missed, and that same-voice trials yielded the highest peak diameters. These counterintuitive findings suggest that, when presented with test items, participants may elaborate the perceptual input, essentially recreating the cognitive processes associated with encoding (see Masson, 1989; Whittlesea & Cantwell, 1987; Was, 2010). Because more "strongly encoded" items result in larger pupils during study, the production process during test results in similar dilation. Indeed, comparing Figures 2 and 3, the similarity in peak diameters is visually evident. We also observed significant positive correlations across observations ($p < .05$ for all comparisons, with the exception of confidence estimates of 5, $p = .14$). Further research into the evaluation process is needed, however, before strong conclusions can be drawn regarding the physiological indices of SCAPE.

Taken together, the present behavioral and physiological findings suggest that spoken words are stored as rich memory traces, with idiosyncratic perceptual details intact (Pisoni, 1993). Such details are naturally retained in memory, aiding subsequent perception and recognition. When a person hears a test word that only partially matches its stored trace, as when items are repeated in new voices, recognition performance drops (Goh, 2005; Goldinger, 1996; Nygaard, Sommers, & Pisoni, 1995). Perceptual fluency appears as a key factor in such voice effects: As indicated by pupillometry, voice specificity effects strongly influence subjective feelings of memory strength and cognitive demand.

Acknowledgments

This research was supported by NIDCD Grant R01-DC04535-11 awarded to Stephen D. Goldinger. We thank Tresa Marchi, Melissa Miola, Hannah Messick, Anne Kuni, and Rachelle Friedman for their assistance with data collection. We also thank Erik Reichle for his thoughtful comments on an earlier version of this article.

References

Balota DA, Yap MJ, Cortese MJ, Hutchison KI, Kessler B, Loftis B, Neely JH, Nelson DL, Simpson GB, Treiman R. The English lexicon project. Behavior Research Methods. 2007; 39:445–459. [PubMed: 17958156]

Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin. 1982; 91:276–292.10.1037/0033-2909 .91.2.276 [PubMed: 7071262]

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

Bowers JS. In defense of abstractionist theories of word identification and repetition priming. Psychonomic Bulletin & Review. 2000; 7:83–99. [PubMed: 10780021]

Bradlow AR, Nygaard LC, Pisoni DB. Effects of talker, rate and amplitude variation on recognition memory for spoken words. Perception & Psychophysics. 1999; 61:206–219. [PubMed: 10089756]

Brown S, Heathcote A. Averaging learning curves across and within participants. Behavior Research Methods, Instruments & Computers. 2003; 35:656–661.10.3758/BF03195493

Brysbaert M, New B. Moving beyond Kuçera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved frequency measure for American English. Behavioral Research Methods. 2009; 41:977–990.10.3758/BRM.41.4.977

Cansino S, Trejo-Morales P. Neurophysiology of successful encoding and retrieval of source memory. Cognitive and Affective Behavioral Neuroscience. 2008; 8:85–98.10.3758/CABN.8.1.85

Church B, Schacter D. Perceptual specificity of auditory priming: Memory for voice intonation and fundamental frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1994; 20:521–533.10.1037/0278-7393.20.3.521

Clark KB, Krahl SE, Smith DS, Jensen RA. Post-training unilateral vagal stimulation enhances retention performance in the rat. Neurobiology of Learning and Memory. 1995; 63:213–216.10.1006/nlme.1995.1024 [PubMed: 7670833]

Clark KB, Naritoku DK, Smith DC, Browning RA, Jensen RA. Enhanced recognition memory following vagus nerve stimulation in human subjects. Nature Neuroscience. 1999; 2:94–98.10.1038/4600

Croiset G, Nijsen MJMA, Kamphuis PJGH. Role of corticotropin-releasing factor, vasopressin and the autonomic nervous system in learning and memory. European Journal of Pharmacology. 2000; 405:225–234. [PubMed: 11033330]

Dietrich DE, Emrich HM, Waller C, Wieringa BM, Johannes S, Münte TF. Emotion/cognition-coupling in word recognition memory of depressive patients: An event-related potential study. Psychiatry Research. 2000; 96:15–29. [PubMed: 10980323]

Duarte A, Ranganath C, Winward L, Hayward D, Knight RT. Dissociable neural correlates for familiarity and recollection during the encoding and retrieval of pictures. Cognitive Brain Research. 2004; 18:255–272.10.1016/j.cogbrainres.2003.10.010 [PubMed: 14741312]

Feenan K, Snodgrass JG. The effect of context on discrimination and bias in recognition memory for pictures and words. Memory & Cognition. 1990; 18:515–527.

Gianaros PJ, Van der Veen FM, Jennings JR. Regional cerebral blood flow correlates with heart period and high-frequency heart period variability during working-memory tasks: Implications for the cortical and subcortical regulation of cardiac autonomic activity. Psychophysiology. 2004; 41(4):521–530.10.1111/1469-8986 .2004.00179.x [PubMed: 15189475]

Glanzer M, Adams JK. The mirror effect in recognition memory: *Theory and data*. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1990; 16:5–16.10.1037/0278-7393.16.1.5

Goh W. Talker variability and recognition memory: Instance-specific and voice-specific effects. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31:40–53.10.1037/0278-7393.31.1.40

Goldinger SD. Words and voices: Episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22:1166–1183.10.1037/0278-7393.22.5.1166

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105:251–279.10.1037/0033-295X.105.2.251 [PubMed: 9577239]

Goldinger SD, Azuma T. Auditory episodes reflected in printed word naming. Psychonomic Bulletin & Review. 2004; 11:716–722. [PubMed: 15581123]

Goldinger SD, He Y, Papesh M. Deficits in cross-race face learning: Insights from eye-movements and pupillometry. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2009; 35:1105–1122.10.1037/a0016548

Goldinger, SD.; Papesh, MH. Pupil dilation as an index of cognitive effort. Presented at the Annual Meeting of the Society for Psychophysiological Research; Berlin, Germany. 2009.

Granholm E, Asarnow RF, Sarkin AJ, Dykes KL. Pupillary responses index cognitive resource limitations. Psychophysiology. 1996; 33:457–461.10.1111/j.1469-8986.1996.tb01071.x [PubMed: 8753946]

Granholm E, Verney SP. Pupillary responses and attentional allocation on the visual backward masking task in schizophrenia. International Journal of Psychophysiology. 2004; 52:37–52.10.1016/j.ijpsycho.2003.12.004 [PubMed: 15003371]

Guo C, Duan L, Li W, Paller KA. Distinguishing source memory and item memory: Brain potentials at encoding and retrieval. Brain Research. 2006; 1118:142–154.10.1016/j.brainres.2006.08.034 [PubMed: 16978588]

Hintzman DL. “Schema abstraction” in a multiple-trace memory model. Psychological Review. 1986; 93:411–428.10.1037//0033-295X .93.4.411

Hintzman DL. Judgments of frequency and recognition memory in a multiple-trace memory model. Psychological Review. 1988; 95:528–551.10.1037/0033-295X.95.4.528

Johansson M, Mecklinger A, Treese A. Recognition memory for emotional and neutral faces: An event-related potential study. Journal of Cognitive Neuroscience. 2004; 16:1840–1853.10.1162/0898929042947883 [PubMed: 15701233]

Joos MA. Acoustic phonetics. Language. 1948; 24:1–136.

Kafkas A, Montaldi D. Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. Quarterly Journal of Experimental Psychology. (in press). 10.1080/17470218.2011.588335

Kahneman, D. Attention and effort. New York: Prentice Hall; 1973.

Karatekin C, Couperus JW, Marcus DJ. Attention allocation on the dual task paradigm as measured through behavioral and psychophysiological responses. Psychophysiology. 2004; 41:175–185.10.1111/j.1469-8986.2004.00147.x [PubMed: 15032983]

Kučera, H.; Francis, W. Computational analysis of present-day American English. Providence: RI: Brown University Press; 1967.

Kuchinke L, Võ ML, Hofmann M, Jacobs AM. Pupillary responses during lexical decisions vary with word frequency but not emotional valence. International Journal of Psychophysiology. 2007; 65:132–140.10.1016/j.ijpsycho.2007.04.004 [PubMed: 17532075]

Lahiri A, Marslen-Wilson WD. The mental representation of lexical form: A phonological approach to the mental lexicon. Cognition. 1991; 38:245–294.10.1016/0010-0277(91)90008-R [PubMed: 2060271]

Leboe JP, Whittlesea BWA. The inferential basis of familiarity and recall: Evidence for a common underlying process. Journal of Memory and Language. 2002; 46:804–829.10.1006/jmla.2001.2828

Lively SE, Logan JS, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. Journal of the Acoustical Society of America. 1993; 94:1242–1255.10.1121/1.408177 [PubMed: 8408964]

Luce P, Lyons E. Specificity for memory representations for spoken words. Memory & Cognition. 1998; 26:708–715.

Magnuson JS, Nusbaum HC. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. Journal of Experimental Psychology: Human Perception and Performance. 2007; 33:391–409.10.1037/0096-1523.33.2.391 [PubMed: 17469975]

Mandler G. Recognizing: The judgment of previous occurrence. Psychological Review. 1980; 87:252–271.10.1037//0033-295X .87.3.252

Marslen-Wilson WD, Warren P. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. Psychological Review. 1994; 101:653–675.10.1037/0033-295X.101.4.653 [PubMed: 7984710]

Masson, MEJ. Fluent reprocessing as an implicit expression of memory for experience. In: Lewandowsky, S.; Dunn, JC.; Kirsner, K.; Kim, editors. Implicit memory: Theoretical issues. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; 1989. p. 123-138.

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18:1–86.10.1016/0010-0285(86)90015-0 [PubMed: 3753912]

McLennan CT, Luce PA. Examining the time course of indexical specificity effects in spoken word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31:306–321.10.1037/0278-7393.31.2.306

McQueen JM, Cutler A, Norris D. Flow of information in the spoken word recognition system. Speech Communication. 2003; 41:257–270.10.1016/S0167-6393(02)00108-5

McQueen JM, Cutler A, Norris D. Phonological abstraction in the mental lexicon. Cognitive Science. 2006; 30:1113–1126.10.1207/s15516709cog0000_79 [PubMed: 21702849]

Montaldi D, Spencer TJ, Roberts N, Mayes AR. The neural system that mediates familiarity memory. Hippocampus. 2006; 16:504–520.10.1002/hipo.20178 [PubMed: 16634088]

Neisser, U. Cognitive Psychology. New York: Appleton-Century Crofts; 1967.

Nygaard LC, Sommers M, Pisoni DB. Effects of stimulus variability on perception and representation of spoken words in memory. Perception & Psychophysics. 1995; 57:989–1001. [PubMed: 8532502]

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of speaker's voice and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1993; 19:309–328.10.1037/0278-7393.19.2.309

Papesh, MH.; Goldinger, SD. Pupil-blah-metry: Word frequency reflected in cognitive effort. Presented at the Annual Meeting of the Psychonomic Society; Chicago, IL. 2008.

Papesh, MH.; Goldinger, SD. Your effort is showing! Pupil dilation reveals memory heuristics. In: Higham, P.; Leboe, J., editors. Constructions of Remembering and Metacognition. Palgrave Macmillan; 2011. p. 215-224.

Pardo, J. On phonetic convergence during conversational interaction; Journal of the Acoustical Society of America. 2006. p. 2382-2393.http://dx.doi.org/10.1121/1.2178720

Pilotti M, Bergman ET, Gallo DA, Sommers M, Roediger HL III. Direct comparison of auditory implicit memory tests. Psychonomic Bulletin & Review. 2000; 7:347–353. [PubMed: 10909144]

Pisoni DB. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. Speech Communication. 1993; 13:109–125.10.1016/0167-6393(93)90063-Q [PubMed: 21461185]

Porter G, Troscianko T, Gilchrist ID. Effort during visual search and counting: Insights from pupillometry. Quarterly Journal of Experimental Psychology. 2007; 60:211–229.

Psychology Software Tools. E-Prime (Version 1.2). 2006. http://www.pstnet.com

Ranganath C, Cohen MX, Brozinsky CJ. Working memory maintenance contributes to long-term memory formation: Neural and behavioral evidence. Journal of Cognitive Neuroscience. 2005; 17:994–1010.10.1162/0898929054475118 [PubMed: 16102232]

Schacter D, Church B. Auditory priming: Implicit and explicit memory for words and voices. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1992; 18:915–930.10.1037/0278-7393.18.5.915

Senkfor A, Van Petten C. Who said what? An event-related potential investigation of source and item memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1998; 24:1005–1025.10.1037/0278-7393.24.4.1005

Sheffert S. Contributions of surface and conceptual information to recognition memory. Perception & Psychophysics. 1998; 60:1141–1152. [PubMed: 9821776]

Sheffert S, Fowler CA. The effects of voice and visible speaker change on memory for spoken words. Journal of Memory and Language. 1995; 34:665–685.10.1006/jmla.1995.1030

Shockley K, Sabadini L, Fowler C. Imitation in shadowing words. Perception & Psychophysics. 2004; 66:422–429. [PubMed: 15283067]

Snodgrass JG, Corwin J. Pragmatics of measuring recognition memory: Applications to dementia and amnesia. Journal of Experimental Psychology: General. 1988; 117:34–50.10.1037//0096-3445 . 117.1.34 [PubMed: 2966230]

Steinhauer SR, Siegle GJ, Condray R, Pless M. Sympathetic and parasympathetic innervations of pupillary dilation during sustained processing. International Journal of Psychophysiology. 2004; 52:77–86.10.1016/j.ijpsycho.2003.12.005 [PubMed: 15003374]

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

Võ ML, Jacobs AM, Kuchinke L, Hofmann M, Conrad M, Schacht A, Hutzler F. The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. Psychophysiology. 2008; 45:30–40.10.1111/j.1469-8986.2008.00745.x

Was CA. The persistence of content-specific memory operations: Priming effects following a 24-h delay. Psychonomic Bulletin & Review. 2010; 17:362–368.10.3758/PBR.17.3.362 [PubMed: 20551359]

Whittlesea, BWA. Production, evaluation and preservation of experiences: Constructive processing in remembering and performance tasks. In: Medin, D., editor. The psychology of learning and motivation. Vol. 37. New York: Academic Press; 1997. p. 221-264.

Whittlesea BWA, Cantwell A. Enduring influences of the purpose of experiences: Encoding-retrieval interactions in word and pseudoword perception. Memory & Cognition. 1987; 15:465–472.

Whittlesea BWA, Leboe JP. The heuristic basis of remembering and classification: Fluency, generation, and resemblance. Journal of Experimental Psychology: General. 2000; 129:84–106.10.1037//0096-3445 .129.1.84 [PubMed: 10756488]

Whittlesea BWA, Williams LD. Why do strangers feel familiar, but friends don't? The unexpected basis of feelings of familiarity. Acta Psychologica. 1998; 98:141–166.10.1016/S0001-6918 (97)00040-1 [PubMed: 9621828]

Whittlesea BWA, Williams LD. The discrepancy attribution hypothesis: I. The heuristic basis of feelings of familiarity. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2001; 27:3–13.10.1037/0278-7393.27.1.3

Wickens, TD. Elementary signal-detection theory. New York: Oxford University Press; 2002.

Wixted JT. Dual-process theory and signal detection theory of recognition memory. Psychological Review. 2007; 114:152–176.10.1037/0033-295X .114.1.152 [PubMed: 17227185]

Wixted JT, Mickes L. A continuous, dual-process model of remember/know judgments. Psychological Review. 2010; 117:1025–1054.10.1037/a0020874 [PubMed: 20836613]

Yonelinas AP. Components of episodic memory: The contribution of recollection and familiarity. The Philosophical Transactions of the Royal Society, Series B. 2001; 356:1363–1374.

Yonelinas AP. The nature of recollection and familiarity: A review of 30 years of research. Journal of Memory and Language. 2002; 46:441–517.10.1006/jmla.2002.2864

Yonelinas AP, Parks CM. Receiver operating characteristics (ROCs) in recognition memory: A review. Psychological Bulletin. 2007; 133:800–832.10.1037/0033-2909 .133.5.800 [PubMed: 17723031]

Appendix

Words and nonwords used in the experiment.

Words	
HF	***LF***
also	*anvil*

Words	
HF	***LF***
basis	*binder*
big	*blame*
boy	*bleed*
car	*boar*
care	*brood*
church	*burglar*
day	*calf*
door	*chose*
else	*clove*
end	*coop*
face	*cork*
fact	*fake*
feet	*fool*
fire	*glean*
force	*glove*
girl	*grapes*
good	*haze*
hand	*heal*
head	*locker*
heard	*moot*
help	*nail*
high	*propel*
hope	*repeal*
house	*rouge*
level	*slate*
like	*sneak*
made	*stamp*
man	*starch*
paper	*stove*
real	*thief*
simple	*thumb*
still	*tulip*
stood	*wade*
strong	*wallet*
table	*weld*
took	*wolf*
water	*worm*
wife	*yolk*
woman	*yore*

Nonwords	
mazz	*borse*
flazick	*lexel*
infloss	*zeat*
wurve	*squeet*
sarlin	*ashwan*
breen	*corple*
preck	*meegon*
freem	*forch*
tupe	*lapek*
tramet	*remond*
greele	*yole*
sagad	*ostrem*
goip	*sorneg*
hinsup	*rebook*
hesting	*nork*
neep	*blukin*
hine	*chark*
erbow	*brant*
manuge	*daver*
zolite	*loash*
vorgo	*reast*
swoke	*dorve*
puxil	*roaken*
fegole	*floak*
humax	*kosspow*
gurst	*vour*
bilark	*bawn*
modge	*plitch*
vasult	*tink*
yertan	*rotail*
lactain	*skave*
rensor	*yolash*
seck	*duforst*
blemin	*sleam*
natch	*yusock*
plaret	*yince*
verm	*gisto*
subar	*behick*
glane	*murch*
serp	*redent*

Highlights

- We examined recognition memory for spoken words while tracking pupillary changes.
- Words were heard in different voices; half changed in the recognition test.
- Pupil dilation during learning predicted later memory accuracy and confidence.
- Study-to-test voice changes were reflected in pupil dilation during test.
- Pupillometry reveals cognitive effort during memory encoding and retrieval.

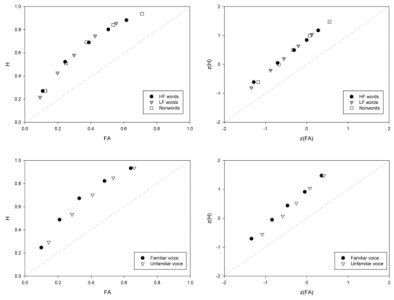

Figure 1.
Group-level receiver operating characteristic (ROC, left panel) and *z*-ROC (right panel) plots. The upper plots present data by word type, and the lower plots present data by voice status (familiar versus unfamiliar).

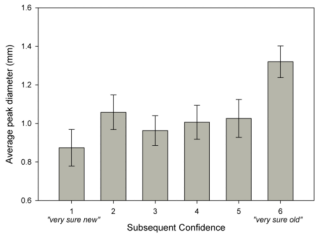

Figure 2.
Baseline-corrected peak diameters during study trials by subsequent confidence estimate. Error bars represent standard error of the mean.

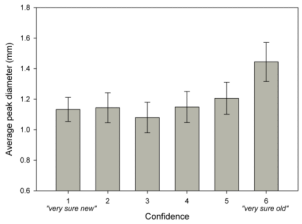

Figure 3.
Baseline-corrected peak diameters during test trials by confidence estimate. Error bars represent standard error of the mean.

Table 1

Summary Statistics for the Stimulus Items

		Means			
Item Type	**n**	**KF**†	**Subtitle Freq**‡	**Letters**	**Syllables**
High Frequency	40	414.3	518.8	4.3	1.2
Low Frequency	40	7.5	11.2	4.8	1.2
Nonword	80	n/a	n/a	5.3	1.6

†Kučera & Francis (1967)

‡Brysbaert & New (2009)

Table 2

Statistics for behavioral ROCs and z-ROCs.

	ROCs		z-ROCs	
Plot Type	**Quadratic Constant**	**Linear Slope**	**Quadratic Constant**	**Linear Slope**
Words	0.55 (1.01)	0.77 (0.12)	0.01 (0.15)	0.75 (0.26)
Voices	−2.04 (1.89)	1.25 (0.24)	−0.04 (1.01)	1.43 (0.43)

Note: Standard error is in parentheses.