THE ROLE OF PERCEPTUAL EPISODES IN LEXICAL PROCESSING

S.D. Goldinger
Arizona State University

ABSTRACT

Nearly all theories of spoken word perception presume a lexicon with singular entries corresponding to each word. In turn, the perceptual system is presumed to operate by matching entries to the variable signals that speakers produce, requiring either normalization or sophisticated guessing. In contrast, episodic theories assume that people store multiple entries, in the form of detailed perceptual traces, for each known word. Such episodic theories are robust to variation, and they provide a natural account of extra-linguistic learning, such as learning voices. This paper presents a new experiment on episodic effects in word perception, and it reviews the application of a multiple-trace model to the data. Theoretical issues are briefly discussed, including the critical role of selective attention, and the relation of episodic theories to other burgeoning views.

1. INTRODUCTION

By its most basic definition, speech is a *medium* of communication – a carrier signal for the words, sentences, and ideas that constitute conversation. In theories of perception, a strong distinction between perceptual objects and their media was applied by Gibson [1]. When an observer gazes upon an object, it is perceived via reflected light that is uniquely structured by physical characteristics, such as edges and contours. The observer does not perceive the *light*; it is merely an informational medium. In this regard, most theories of word perception resemble Gibson's view; speech signals are not considered true perceptual objects. Instead, the important objects in speech are phonemes, syllables, words, or phrases (depending on context [2]). Although few people would dispute this self-evident description of speech, the medium is also, in itself, a potential object of perception. Instead of focusing on the message, listeners may primarily attend to tone of voice, dialect, etc. Thus, spoken words lead "double lives," serving as both perceptual objects (with unique voice characteristics), and as "gateways" to linguistic representations.

As it happens, numerous investigations have focused on the perceptual domain, primarily by studying "surface memory" for printed and spoken words. Although the literature contains some null results [3, 4], an impressive collection of positive findings exists. With respect to printed words, font memory is often observed [5], and similar effects arise with spoken words: Voices are reliably stored in long-term memory as a side-effect of lexical access, affecting both direct and indirect memory tests [6]. Detailed episodic traces are apparently created in spoken word perception, affecting later perceptual and memorial tests. Therefore, it has been suggested that the mental lexicon may consist of stored episodes, rather than abstract units [6, 7].

1.1. An Episodic Model – MINERVA 2

To demonstrate this approach to word perception, I have applied a formal model to voice-memory data [6], replicating the qualitative patterns. The model, called MINERVA 2 [8], takes episodic storage to a logical extreme, assuming that all experiences create separate, detailed memory traces. During perception, however, aggregates of traces combine to create behavior. Consider word perception: For every known word, a potentially vast collection of traces resides in memory. When a test word is presented, a *probe* is communicated (in parallel) to all traces, which are activated in proportion to similarity. An aggregate of activated traces constitutes an *echo,* sent to working memory from long-term memory. Echoes may contain information not present in the probe, such as conceptual knowledge, thus associating the stimulus to past experience.

Echoes have two key properties in MINERVA 2: Echo *intensity* reflects the total activity in memory created by the probe. Intensity grows with greater similarity of the probe to existing traces, and with greater numbers of such traces. Thus, it estimates stimulus familiarity, and can be used to simulate recognition memory. Echo *content* is the "net response" of memory to the probe. Because all stored traces respond in parallel, each to its own degree, echo content reflects a unique combination of the probe and the activated traces. For example, a common word in a familiar voice will activate many traces, creating a strong but fairly generic echo. If a rare word is presented in an unfamiliar voice, fewer traces will (weakly) respond. Thus, if a perfect match to the probe exists in memory, it will strongly contribute to echo content. Therefore, voice effects should be greater for rare words, or for words presented in unusual contexts.

Goldinger [9] recently showed that MINERVA 2 qualitatively replicates the voice effects observed in recognition memory [6]. Moreover, the model correctly predicted stronger voice effects for lower-frequency words. As noted, high-frequency words activate many traces, so the details of any particular trace (even a perfect match to the probe) are obscured in the echo. In other words, higher-frequency words inspire "abstract" echoes, relative to the more "episodic" echoes for lower-frequency words. A post-hoc analysis on the recognition data confirmed that voice effects were stronger among lower-frequency words.

1.2. Episodes in Perception and Production

Beyond episodic effects in perception and memory, Goldinger [9] also applied MINERVA 2 to *single-word*

shadowing data. In this task, participants quickly repeat spoken words, with response time as the main dependent measure [10]. A seldom-used, secondary measure is the speech output itself, which can be analyzed in various ways. For example, my research tested the degrees to which shadowing participants *imitated* the stimulus words, in terms of general auditory dimensions. Imitation was examined because it provided a unique test of MINERVA 2. Because echoes constitute the model's only basis to respond, it is simplest to hypothesize that shadowers will generate a "readout" of the echo content when speaking. Indeed, by specifying both echo content and intensity, MINERVA2 has a unique ability to predict both imitation and shadowing RT. Moreover, the model also makes predictions about the *strength* of imitation, based on experimental factors such as word frequency and training exposures.

The present experiment extends my earlier research on vocal imitation, changing the previously used method in two major regards. First, the shadowing task was no longer used. In the previous experiments, participants recorded baseline tokens while reading aloud, then later produced test tokens while shadowing. Now, to provide a stronger test of the episodic theory, participants read words aloud in both sessions. Thus, imitation could not be an artifact of the shadowing task (see also [9]). The second change was to add a recognition memory test to the procedure. In MINERVA 2, both expressions of memory – explicit recognition and (implicit) imitation – must be generated from a common set of stored traces, as opposed to separate memory systems (compare [11] to [12]). Collecting both explicit and implicit data creates a stronger test of the model.

2. METHODS

The experiment was conducted with two groups of participants. Twelve students (six men, six women) at Arizona State University were recruited for the *training group*; each received $50.00. Each member of the training group completed four experimental stages over a two-week period.

On Day 1, the participant recorded *baseline tokens* of 160 common English words. The words were evenly divided into four classes according to their frequencies of occurrence, with 40 high, medium-high, medium-low, and low-frequency words (see [9]). The participant was seated in a sound-attenuated booth equipped with an IBM Aptiva computer, a Beyer-Dynamics microphone, and a Marantz PMD-333 cassette recorder. Words were shown on the computer in random order; participants were asked to "speak each word clearly, in order to make a good recording." After these tokens were recorded, they were digitized and stored for a later part of the experiment.

On Day 2, the participant identified a set of auditory *training tokens*, chosen from an existing database. For this stage, tokens from 2 men and 2 women were used. The set of 160 words was divided into 4 training sets of 40 words (10 per frequency class). For each training set, the participant saw a grid of 40 cells (8 x 5) on the computer screen, each containing one word. Spoken words were presented over headphones, with 10 words randomly assigned to each of the four voices. Upon hearing each word, the participant's job was to find it in the grid, clicking it with the mouse. *Number of exposures* was the key manipulation: One training set was never presented (zero exposures), another was presented twice, another was presented six times, and another was presented 12 times (assignment of words to training sets and voices was counterbalanced across participants). Grids were randomly re-drawn after each pass through a training set, but all repetitions of each word were in a consistent voice.

On Day 7, participants returned to the lab to record *test tokens*, using procedures identical to those from the Day 1 baseline phase – all 160 words were presented visually and were carefully spoken for recording. As before, those tokens were digitized and stored for later use. Finally, on Day 14, participants returned for a visual recognition memory test in which they tried to discriminate the original 160 words from 160 frequency-matched foils. Recognition accuracy was the primary measure; decision times were also recorded although speed was not emphasized in the instructions.

For the imitation measure, 300 students comprised 12 *AXB classification* groups (one per training-group participant), each with 25 students. Each training-group participant's Day 1 baseline (*A*) and Day 7 test (*B*) utterances were juxtaposed with the training utterances (*X*) heard on Day 2. Half the trials presented the baseline token first; half presented it third. The AXB participants heard all three words successively and judged which utterance, first or third, was a "better imitation of the middle word." The AXB testing was done with groups of 4-7 students in a sound-attenuated room. Each trial began with a 500-ms warning (***), followed by 2 response boxes, labeled "first" and "third." After 500 ms, three words were played, separated by 750-ms intervals. The participant indicated whether *A* or *B* sounded more like *X* by clicking either box with the left mouse key. AXB "accuracy" (i.e., selecting the test token, rather than the baseline token) was the dependent measure of imitation.

3. RESULTS

The results are shown in Figure 1, with recognition and imitation data shown in the upper and lower panels, respectively. In recognition, clear effects of word frequency and repetition were observed, with a strong two-way interaction (all Fs > 114, $p < .0001$). As shown, recognition was generally better for lower-frequency words, and it steadily improved with increasing repetitions. However, the frequency effect diminished over repetitions, as all the words approached asymptotic recognition levels.

In imitation, similar main effects of frequency and repetition were observed, with another two-way interaction (all Fs > 63, p < .0001). Imitation was greater for lower-frequency words, and it generally increased for words that the participants heard more often in training. Unlike recognition, the frequency effect increased with repetitions, primarily due to rising imitation among the lower-frequency words. Because the interactions in recognition and imitation were in opposite directions, a strong three-way interaction was observed [F(1,11)= 159.8, p < .0001].

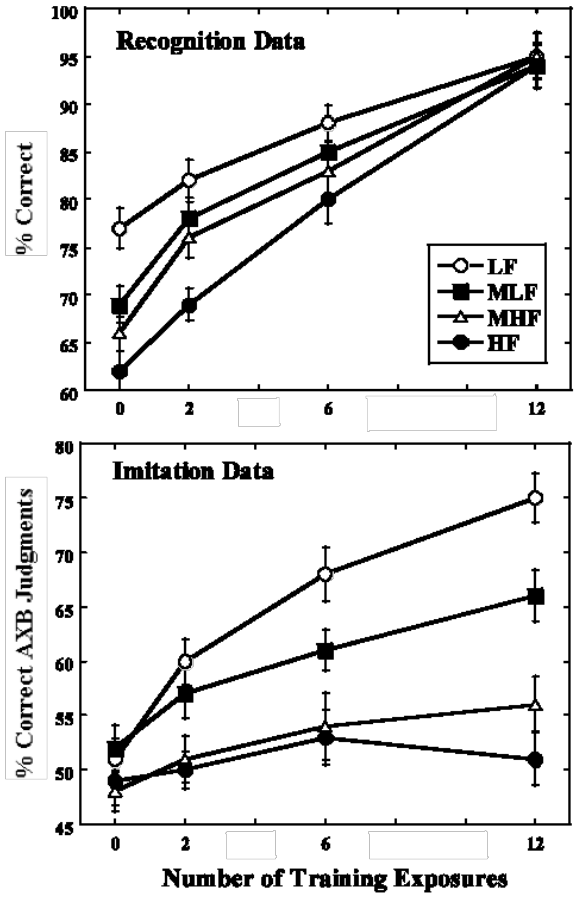

Figure 1. Recognition memory and imitation data. Means (and standard errors) are shown over word frequency and training exposures. LF = low-freq. MLF = medium-low freq. MHF = medium-high freq. HF = high-freq.

Several aspects of these data are noteworthy: First, the imitation data verify that the contents of memory can be reflected in the sound of a person's voice, as shown previously [9]. The present data are unique, showing the imitation effect in printed word naming. This finding suggests that reading aloud involves more than simple print-to-sound conversions; it also taps into memory for prior perceptual episodes [5]. Second, imitation was systematically affected by word-frequency, which is critical to interpretation: To be theoretically relevant, imitation must be a *spontaneous* response, rather than a frivolous or general tendency. Word frequency is a purely abstract variable, which rules out demand characteristics or other trivial accounts of imitation. Third, opposite patterns of interaction emerged, suggesting that recognition and imitation may have different underlying bases.

Behavioral dissociations of implicit and explicit memory are well-documented and have been explained by two general theories. One theory proposes that separate brain systems underlie the implicit-explicit distinction, giving rise to independent data patterns [11]. This account seems well-suited to explain cases in neuropsychology, such as patient H.M. The other theory assumes a single memory system, with differences in task-specific processes generating dissociations [12]. The present data show a partial dissociation of imitation (implicit memory) from explicit memory, but do not approach true stochastic independence.

This partial dissociation is reinforced by comparing the imitation levels of words that eventually produced hits and misses, respectively, in the recognition test. Figure 2 shows that imitation was generally higher (F= 12.5, p < .01) for words that generated hits. However, the imitation patterns were quite similar, regardless of participants' later recollective success or failure. Both panels of Figure 2 show reliable frequency and repetition effects (both Fs > 25, p < .001), with no interactions involving later hits/misses. Thus, it seems that imitation and recognition are correlated, but are not complete reflections of one another. Indeed, among the words that produced recognition hits, degrees of imitation were negatively correlated with recognition time (r = -.41, p < .01). Words that strongly engendered imitation were slightly (2.25%) more likely to be recognized later, and they were recognized faster.

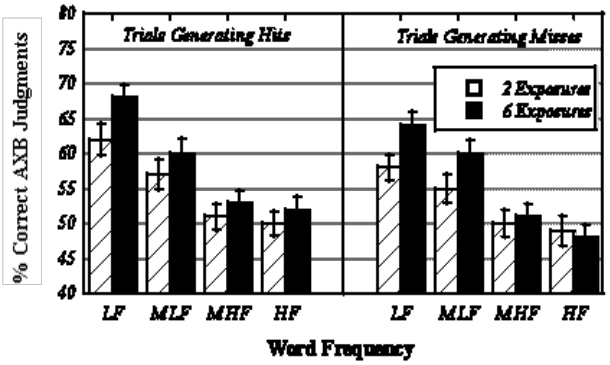

Figure 2. AXB discrimination data for words that led to recognition hits (left panel) and misses (right panel). Means (and standard errors) are shown as a function of word frequency and number of training exposures.

4. DISCUSSION

Taken together, these data suggest that imitation and recognition tap a single set of memory traces, with task-specific processes creating the dissociations. They also suggest that such traces are highly detailed, preserving

information such as voice and environmental context (a necessity to explain recognition memory for common words).

Simulations of MINERVA 2 replicate the qualitative data patterns shown in Figure 1, with echo intensity and echo content used to estimate recognition and imitation, respectively. Considering recognition, recall that echo intensity reflects the total activity in memory created by the probe. To correctly perform recognition, the model must detect that a known word was previously encountered in the specific context of the experiment, using echo intensity as a guide. This is a signal-detection problem, with the experiment-specific traces as "signal" and prior traces as "noise" [13]. Low-frequency words are initially easier to detect because they spawn weaker echoes (less noise). If a recent trace perfectly matches the probe, it significantly boosts echo intensity, relative to this baseline. As word frequency increases, this benefit decreases. However, as more "study traces" accumulate in the model's memory, this frequency effect is reduced, until all words are recognized equally.

Considering imitation, recall that echo content is a combination of the probe plus a weighted average of previously-stored traces. In the data, lower-frequency words engender more imitation, an effect that increases with repetitions. In simulations, frequency differences are enacted by varying the numbers of stored traces for different words, with all traces having random elements corresponding to voices, contexts, etc. Because higher-frequency words excite many traces, they spawn generic echoes, with study traces obscured. By contrast, echoes for lower-frequency words are strongly influenced by old traces resembling the probe. Thus, imitation is stronger for lower-frequency words. Moreover, this general difference increases over repetitions, as the "central tendencies" of lower-frequency words change faster with each added trace.

Despite the current success of MINERVA 2, there are important issues to consider, including the necessary role of *selective attention* in shaping echo content [9]. Many data now suggest that detailed traces are created in perception, and are involved in later perception. They do not, however, suggest a "capacity-free" system, with all dimensions equally coded, regardless of attention. Our own data (from [6] and new unpublished work) reinforce the common assumption of dimension-specific encoding, based on processing demands at study. In response, new modeling efforts have enacted an attention parameter. As a result, the model conforms better to common sense, and it accounts for more data. Although MINERVA 2 seems remote from other theories of word perception, the exemplar concept is central to many views, including connectionism and adaptive-resonance [14]. Episodic effects in "laboratory perception" may reveal the basic processes of lexical representation and access.

5. ACKNOWLEDGEMENTS

Support provided by NIDCD grant R29-DC02629-05 to Arizona State University. Address correspondence to S.D. Goldinger, Department of Psychology, Box 871104, Tempe, AZ, 85287-1104, USA. Email address: ***goldinger@asu.edu***.

6. REFERENCES

[1] Gibson, J. (1966). *Senses considered as perceptual systems*. Boston: Houghton-Mifflin.

[2] McNeill, D., & Lindig, K. (1973). Perceptual reality of phonemes, syllables, words, and sentences. *Journal of Verbal Learning and Verbal Behavior,* **12**, 419-430.

[3] Brown, J., & Carr, T. (1993). Limits on perceptual abstraction in reading: Asymmetric transfer between surface forms differing in typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **19**, 1277-1296.

[4] Jackson, A., & Morton, J. (1984). Facilitation of auditory word recognition. *Memory & Cognition,* **12**, 568-574.

[5] Jacoby, L.L., & Hayman, C. (1987). Specific visual transfer in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **13**, 456-463.

[6] Goldinger, S.D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166-1183.

[7] Goldinger, S., Kleider, H., & Shelley E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition,* **27**, 328-338.

[8] Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review,* **93**, 411-428.

[9] Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

[10] Radeau, M., Morais, J., & Dewier, A. (1989). Phonological priming in spoken word recognition: Task effects. *Memory & Cognition*, **17**, 525-535.

[11] Tulving, E., & Schacter, D. (1990). Priming and human memory systems. *Science*, **247**, 301-306.

[12] Roediger, H.L., Weldon, M., & Challis, B. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. Roediger & F. Craik (Eds.), *Varieties of memory and conciousness* (pp. 3-41). Hillsdale, NJ: Erlbaum.

[13] Hintzman, D.L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, **95**, 528-551.

[14] Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review*, **87**, 1-51.