ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 3aSCa: Imitation, Accommodation, and Convergence in Speech Communication

3aSCa1. The cognitive basis of spontaneous imitation: Evidence from the visual world
Stephen D. Goldinger*

*Corresponding author's address: Department of Psychology, Arizona State University, 950 S. McAllister Drive, Tempe, Arizona 85287-1104, goldinger@asu.edu

It is well-established that, when people are asked to identify and quickly repeat spoken words, they show a strong tendency to spontaneously imitate the vocal and/or phonetic characteristics of the stimulus tokens. There is mixed evidence, however, regarding the underlying basis of such imitation: Does it only reflect gestural attunement (as in Direct Realism), or does it also reflect cognitive principles of word perception and memory? The gestural attunement view has face validity, as imitation seems to require tacit knowledge of other peoples' articulatory actions. The role of memory is less obvious, although people can certainly imitate others from memory. In this talk, I will present evidence from three new experiments, pairing procedures from the "visual world" paradigm with a speech production task. Across studies, there is clear evidence that degrees of speech imitation are tightly connected to attention and memory processes that were engaged during initial exposure to spoken words. The results show clear imitation (in naming depicted objects), both with and without spoken words prompting responses, and show strong effects of competition among visual objects: Imitation increases when other potential objects have similar names, or even similar appearances. Spontaneous imitation is both a gestural and a cognitive behavior.

Published by the Acoustical Society of America through the American Institute of Physics
THE “DOUBLE LIVES” OF SPOKEN WORDS

By its most basic definition, speech is a medium of communication – a carrier signal for the words, sentences, and ideas that constitute conversation. In vision, the distinction between perceptual objects and their media was articulated by Gibson (1966): Object perception is achieved via reflected light that is uniquely structured by edges, contours, etc. People do not perceive the light; it is merely an informational medium. Gibson argued that perceptual systems evolved to efficiently use media, connecting organisms with environments. In this regard, many psycholinguistic theories 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; see Goldinger & Azuma, 2003; McNeill & Lindig, 1973). Although few people would dispute this self-evident description of communication, the speech signal is also, in itself, a potential perceptual object. Instead of focusing on a speaker’s message, a listener may attend to voice quality, dialect, etc. Thus, spoken words lead “double lives,” serving as perceptual objects, with unique acoustic characteristics, and as gateways to lexical representations.

As it happens, numerous investigations have focused on the perceptual domain, studying “surface” (typically voice) memory for spoken words (e.g., Creel et al., 2008; Creel & Tumlin, 2011; Goh, 2005; Goldinger, 1996; Nielsen, 2011). Across studies, it is clear that word perception creates episodic memory traces complete with information such as voice and context. Thus, I have previously hypothesized that detailed episodic traces are created in word perception, influencing later behavior. In particular, lexical episodes may affect not only word perception, but production as well. This episodic theory (based on Hintzman’s, 1986, MINERVA 2) was tested in shadowing experiments by Goldinger (1998). Although the typical measure in shadowing is RT, a seldom-used, secondary measure is speech output, which may be recorded and analyzed. Using two groups of participants, shadowers and listeners, I examined the influence of episodic details on the acoustic realization of word shadowing. Participants’ baseline words were first recorded as they read printed words from the computer screen. Later, they listened to different speakers repeat those words 2, 6, or 12 times as they completed a search task (locating the words within a visual matrix). During a subsequent shadowing block, participants heard the words again, but were given either immediate or delayed shadowing instructions. The shadowed and baseline words were later presented to listeners for an AXB classification task. In AXB classification, each shadowers’ baseline (A) and shadowing (B) tokens were compared to the original stimulus token (X), with listeners judging whether A or B sounded like a “better imitation” of X. Although shadowers were not instructed to imitate, listeners reliably judged the shadowing token as a better imitation (for replications and extensions, see Kim et al., 2011; Namy et al., 2002; Nielsen, 2011; Nye et al., 2003; Pardo, 2006; Pardo et al., 2010; 2012; Shockley et al., 2004).

Goldinger (1998) did not only observe spontaneous imitation (also called accommodation or convergence) as a general tendency. Instead, degrees of imitation were affected by psychological variables, such as word frequency, the number of exposures, and response timing. The MINERVA model made four key predictions: First, imitation should increase as repetitions accrue, as more matching traces join the echo (the theoretical construct hypothesized to drive production). Second, less common words should elicit more imitation than common words, as experimental traces will have a stronger recency effect. Third, frequency and repetition should interact; with each repetition, the influence of experiment-specific traces should increase. As a result, differences between low- and high-frequency words should diminish. Fourth, the experiment also included immediate and delayed shadowing conditions, as the model predicted decreasing imitation over delays: As the participant awaits a “go” signal, the prepared word cycles between WM and LTM, as a series of echoes and probes. This cycle forces regression to the mean, as each successive echo drifts closer to the central tendency of all stored traces. After several seconds, the echo (the basis for responding) will be a generic member of the lexical category. All these predictions were supported in the data, and suggested an underlying cognitive basis for spontaneous imitation.

Despite the evidence for cognitive influences on imitation, it cannot be the entire story. First, there are myriad social factors that also affect the presence or degree of linguistic accommodation (e.g., Pardo et al., 2010; 2012). Second, cognitive factors cannot easily explain the rather miraculous phonetic alignment that occurs, effortlessly and automatically, in speech accommodation. That is, even without prior exposure to stimulus speakers, listeners are able to track their articulatory gestures and rapidly approximate those same gestures while responding. Research by Fowler et al. (2003) suggested that listeners have near-instantaneous access to the speech gestures of their interlocutors, a finding predicted by Fowler’s (1986) Direct Realist theory, which shares philosophical roots with Gibson’s (1966) theory of direct perception. According to this theory, imitation occurs because speech perception is mediated by appreciation of the “common currency” of articulatory gestures: Understanding speech is isomorphic with tracking speech gestures, which makes imitation a natural and automatic behavior. This gestural attunement view has great face validity, as imitation seems to require tacit knowledge of other peoples’ articulatory actions.
The role of memory is less obvious (although people, such as professional impressionists, can clearly imitate others from memory). In this presentation, I will suggest that – while the gestural attunement account must be correct at some level – it is likely not the entire story.

The challenge in interpreting the foregoing data is that the stimulus array – the speech which elicits an imitation response – is cotemporaneous with the occurrence of imitation. The participant hears something and repeats it. The observation of hidden lexical effects (e.g., word frequency) suggests some degree of cognitive mediation, but stronger evidence is required. At least one prior study on spontaneous imitation helps illuminate the role of long-term memory in speech imitation: Goldinger and Azuma (2004) observed spontaneous vocal imitation in a task without speech shadowing. In their study, participants recorded baseline words from text, as described above. The following day, they heard with the same words, spoken by an unrelated speaker between 2 and 12 times. Crucially, participants were not asked to shadow the words, but instead searched a visual display for each word as it was spoken. One week later, the same participants were re-recorded the original set of words, again reading from a computer screen. Goldinger and Azuma (2004) again observed imitation after long response delays, as judged in a separate AXB task. This result suggests that lexical episodes have both immediate and long-term consequences for the perception and representation of spoken words, and that speech imitation can derive from a purely cognitive source, memory for speech patterns.

Controlling Levels of Imitation by Controlling Encoding Difficulty

As noted, the study by Goldinger and Azuma (2004) shows that imitation can be cognitively mediated. Note that it does not rule out a gestural attunement account, which may not be theoretically possible. (That is, one could argue that the memorized traces were structured as encoded gestural patterns, which is potentially impossible to disprove, and may very well be correct.) It is of great interest, however, to understand how such cognitive mechanisms might influence spontaneous imitation. For one thing, imitation is not universally observed, and it is affected by many social and cognitive variables (e.g., Pardo et al., 2010; Shockley et al., 2004). In studies such as Goldinger (1998), people typically had to hear the same tokens at least four times in a training session before imitation would reliably exceed chance. One possibility is that, in typical listening conditions, speech perception is so fluent and transient that listeners can effectively operate at an abstract linguistic level: They use the rich speech signal (and all its vocal and other characteristics) to support lexical, syntactic, and semantic analysis, but those signals are processed at such a shallow level that they leave little behind in memory. By way of analogy, when walking across a crowded campus, a person is easily able to perceive and avoid bumping other pedestrians, but would be hard-pressed to identify those same people afterward. From a cognitive standpoint, events are theorized to create episodic memory traces to the degree that those events (and their relevant, tested dimensions) are attended during initial processing (e.g., Logan, 1988). Thus, to better understand the roots of speech imitation, it may be fruitful to manipulate attention.


In three new experiments, I recently paired procedural elements from the “visual world” paradigm (Allopenna et al., 1988; Dahan et al., 2001; Magnuson et al., 2003) with a speech production task. The first experiment was loosely modeled after Goldinger and Azuma (2004): Participants recorded baseline items from text, then heard those same words (eight times each) being produced by a pre-recorded speaker. Three days later, they recorded the words again, and imitation was evaluated using the AXB procedure. The critical difference, however, was in the nature of the encoding task that occurred during the exposure to speech. In Goldinger and Azuma’s study, people saw words in a grid on the screen, and were asked to mouse-click each item as it was heard. There were no special relations among the words, nor were there any experimental manipulations. In Experiment 1, this was not the case.

In Experiment 1, while people were exposed to spoken words, they were asked to identify each word by locating and clicking a visual image in an array. (Participants are familiarized with the objects and names ahead of time.) Schematic versions are shown in Figure 1, although the figure shows more objects and more structure than the real arrays used (the experiment showed four icons per trial, with quasi-random locations). The manipulation of key interest was simple: Some words were consistently heard while the visual array contained no clear competition. That is, the image representing the word had no special visual or phonological relations to the other items on the screen; they were chosen randomly from a set of dissimilar distractors. However, other words were consistently paired either with other objects that either sounded similar to the target object (as in panel A), or looked similar to the target object (as in panel B). Different words were randomly assigned to these conditions for each participant, and were held consistent throughout the experiment. Moreover, when viewing these displays, any one of the
depicted items could become a target on any given trial. Therefore, participants could not merely guide attention to a likely target. Finally, three days after the identification trials, participants returned and recorded the items again, from text. Those items (and the stimulus words from the training blocks) were used to create AXB classification experiments. The results were clear and robust: When words were displayed against a background of random objects, AXB listeners were able to detect imitation better than chance, with a mean correct detection rate of 57.7%. When those same words had been encoded while their corresponding icons were surrounded by visual competitors, later imitation increased, and was detected at a rate of 64.8%. And when those words had been encoded while their corresponding icons were surrounded by phonological competitors, later imitation increased again, and was detected 79.5% of the time. Together, the results suggest that attention to the speech signal was modulated by the challenge of the search task: When people needed to carefully monitor speech to locate the appropriate targets (click RTs verified the challenge of the competition conditions), they created episodic traces that better supported imitation.

![FIGURE 1. Schematic displays to illustrate the principal stimulus manipulations in the experiment. (In the experiment proper, the arrangements were not so orderly, and there were only four items per display.) The items in panel A are phonologically similar, but visually dissimilar (they are a beetle, beater, beaker, beachball, beaver, beagle, beehive, and beeper). The items in panel B are visually similar but phonologically dissimilar (they are a cookie, button, candle, pizza, Frisbee, doughnut, quarter, and lemon). Participants encoded items while finding them either among competitors, as shown here, or among unrelated items.](image)

**Experiment 2: The “Click and Say” Procedure, Part 1**

In Experiment 2, participants were familiarized with 120 objects, seeing each with its printed name, until they could identify them all perfectly. Next, they saw displays (similar to Figure 1, but showing four objects) for one second. Then, one of the four object names appeared in the center; the participant was asked to click the object while speaking its name aloud; these baseline utterances were recorded. After a short break, the (potential) imitation stage began: Participants viewed the same displays for one second, now followed by a spoken word in one of four voices (half female). The task was again to click the appropriate icon while speaking the name aloud. Critically, as before, the displays either contained no obvious competitors for the target objects, or they contained either visual or phonological competitors. All objects were named 8 times, always shown in the same arrays and spoken in the same voice; click RTs and utterances were recorded. For all items, the last speaking trial was used in the AXB task, unless the recording was deemed unacceptable.

As outlined above, the question of interest in Experiment 2 is motivated by gestural attunement accounts of imitation, and the possibility is that peoples’ reliance on careful speech attention may wax and wane, depending upon momentary needs. Indeed, we again observed that imitation increased when competitors were present: Detected imitation levels were 60.2% in the no-competition condition, 66.9% in the visual competition condition,
and 81.0% in the phonological competition condition. When careful attention to the speech signal was required by the visual search task, imitation automatically increased.

Experiment 3: The “Click and Say” Procedure, Part 2

Finally, Experiments 3 was similar to Experiments 2, except the manipulation of competition was removed. Instead, in half the click-and-say trials, the visual icons for objects were changed. (Specifically, the same icon had been used in seven prior trials, but was changed in the critical eighth trial, used for the AXB testing.) Using a resonance metaphor, we expected this change would force people to rely less on “resonating” with familiar pictures, and to rely more on spoken words to guide behavior. For the other half of the trials, the original objects were not changed from all prior trials. The difference in observed imitation was stark, and somewhat surprising: With unchanged objects, imitation was detected 73.6% of the time. When the visual icons changed, however, imitation sharply increased, and was correctly detected 93.0% of the time. (We are currently conducting follow-up experiments, attempting to rule out artifacts, such as participants sounding hesitant.) For the time being, it appears that, once the visual array is less helpful, participants “lean” more heavily on the speech input, and naturally imitate it more faithfully.

Across studies, we observed clear evidence that degrees of speech imitation are tightly connected to the deployment of attention during initial exposure to spoken words. The results show clear imitation (in naming depicted objects), both with and without spoken words prompting responses, and show strong effects of competition among visual objects: Imitation increases when other potential objects have similar names, or even similar appearances. Spontaneous imitation is both a gestural and a cognitive behavior.

ACKNOWLEDGMENTS

This research was supported by NIDCD Grant R01-DC04535-11, awarded to S.D. Goldinger. I thank Megan Papesh for extensive help in preparing this manuscript.

REFERENCES


