Auditory word recognition

Language provides humans with the remarkable capacity to express their thoughts through a physical medium to share with others. To do so, we combine elements, words, whose form has been conventionalized within a particular language community. Thus a critical step in the process of retrieving a talker's message consists of identifying these elements in his or her speech. This entry discusses how our knowledge of the auditory forms that words take may be represented in memory, and how listeners decide, based on the auditory stimulus, which words they heard, out of all possible word combinations the talker may have said.

When we listen to someone talk, words seem to pop out of his or her speech effortlessly. This impression is misleading, however. Words are not neatly segregated from one another in speech as they are in print. How many words the utterance contains, and where they begin and end in the speech stream, are properties that the listener must establish. Moreover, the way spoken words sound varies considerably across contexts, e.g., when produced by a man or a woman, in the clear speech used in lecture halls or in the casual speech characteristic of informal conversation. Our knowledge of the form of words must accommodate this variability. Two approaches to this issue can be contrasted.

First, listeners may represent the form of a word as a compilation of the memory traces that correspond to all past exposure with the word. Each instance retains the acoustic properties resulting from the context in which the word was uttered. Such a representation is sometimes described as a cluster of observations in a multi-dimensional space. This approach assumes ever-changing word representations because new instances

of words are constantly added to the cluster. A more compact representation may also be postulated, such as one that represents the central tendency derived from past instances of a word, its prototype.

Such an approach to word-form representation is rooted in cognitive theories of categorization, and contrasts with a linguistically-grounded view where words are represented by the features that distinguish them from other words. The acoustic properties of a spoken word, such as the voice quality of the talker that utters it, are considered irrelevant to this distinction and consequently not part of the representation of the word's form. This approach assumes abstract, context-independent, and immutable representations. Normalization algorithms transform information extracted from the speech to neutralize the influence of contextual variability, in effect treating it as noise, or to model the variation and factor out its influence.

Distinguishing between the two approaches has proven difficult. For instance, some have taken the fact that people recognize words uttered by familiar talkers more readily than the same words from unfamiliar talkers as evidence supporting the instance-based approach because it demonstrates that non-distinctive properties of spoken stimuli are maintained in memory and contribute to recognition. However, the finding is also compatible with the "abstractionist" approach if one assumes that the normalization algorithms can be optimized to reflect past experience with a given talker.

Another widely discussed issue concerns the internal structure of words. Phonological theories describe words as built out of elements, the phonemes, grouped into larger units, such as syllables. Words' internal structure is known (albeit implicitly) to language users because changes that word forms undergo under the influence of

morphology or other linguistic constraints have regularities that depend on the decomposition of word forms into such a structure. The critical question here is whether people, when analyzing speech, decompose the signal into individual elements to establish which word matches the structure best. Recognizing phonemes or other units first, as opposed to analyzing the spoken word as a whole, may offer an advantage because there are fewer phonemes to discriminate than there are words. However, phoneme recognition itself has proven difficult because the acoustic realization of a given phoneme varies greatly across contexts. Furthermore, listeners can successfully retrieve which of phonemically identical strings (such as *two lips* and *tulips*) the talker said because they make use of subtle acoustic differences between the strings. This finding is difficult to explain if the speech signal was first translated into its phonemic subcomponents.

Contrary to the question of how the form of words is represented, the process by which the perceptual stimulus is compared to these representations is relatively well understood and uncontroversial. Spoken words become available to the listener over time. Because speech is a complex, transient, and rapidly changing signal and because sensory memory is limited, speech must be evaluated and interpreted incrementally rather than word by word. But the early portion of a spoken word (e.g., *cap…*) is often compatible with many different words (e.g., *cap*, *captive*, *capital, captain*). Dominant views posit that all possible interpretations of the spoken word can be simultaneously considered. For example, in William Marslen-Wilson's "Cohort" theory, the first sounds of a spoken word determine a cohort of hypotheses compatible with this early information. Subsequent information serves to prune the hypotheses that are no longer

supported by the signal. Although the "propose-then-dispose" aspect of the theory has since been falsified by evidence that words can be successfully recognized even when their first sounds are distorted, the privileged role to the early portion of a word has been maintained by assuming a form of competition among simultaneously considered hypotheses. The more evidence has accumulated in favor of a given hypothesis, the less likely its alternatives. This mechanism, in effect, favors words that match the early portion of the spoken stimulus over those that match a later portion because the latter will have been largely discounted before the stimulus supports them as possible contenders. Importantly, competition is modulated by the likelihood of encountering each hypothesis, which can be estimated by how often it has been encountered before. Frequent words are recognized more accurately and faster than rarer words, and frequent hypotheses interfere with the recognition of rare words.

Delphine Dahan

See also Visual Word Recognition; Cohort Model of Auditory Word Recognition; Frequency Effects in Word Recognition; Incremental Processing, Comprehension.

Further Readings

Dahan, D., & Magnuson, J. S. (2006). Spoken-word recognition. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (pp. 249-283). Amsterdam, The Netherlands: Elsevier.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71-102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.