through the visual world by manipulating a joystick or a mouse. Such conditions provide allothetic but not idiothetic cues to self-motion and can rapidly cause the user to become lost in the video-game world. In contrast, adding idiothetic self-motion cues can make navigation much more natural. Virtual environments have proven to be an ideal tool for isolating idiothetic and allothetic cues to spatial updating, as participants can navigate through the exact same virtual environment by physically walking and turning or by manipulating a joystick. These experiments highlight the importance of idiothetic cues to spatial updating and the insufficiency of allothetic cues.

With sufficient self-motion cues, body-to-object spatial relations in sensorimotor spatial memory are updated continually when moving through the environment. Similar to findings on long-term spatial memory, where the reference direction organization is unaffected by learning modality, the body-based nature of sensorimotor spatial memory is also unaffected by learning modality. As such, imagined perspectives aligned with the body are facilitated for object layouts learned through vision, touch, audition, or even language.

Spatial Orientation

In order to stay oriented with respect to a known environment, the navigator must match represented features from the sensorimotor spatial memory with those same features in the long-term spatial memory. In some cases, this can be accomplished by matching identifiable landmarks, like the student who uses an identifiable building to stay oriented to campus. In other cases, geometric properties of the surrounding environment, like the shape of a rectangular room, can be used to perform this match. This matching process is a critical step to staying oriented to a remembered environment and underscores the importance of coordinating long-term and sensorimotor spatial memories.

Jonathan W. Kelly and Timothy P. McNamara

See also Action and Vision; Navigation Through Spatial Layout; Self-Motion Perception

Further Readings

Gallistel, C. R. (1990). *The organization of learning.* Cambridge: MIT Press.

May, M. (2004). Imaginal perspective switches in remembered environments: Transformation versus interference accounts. *Cognitive Psychology, 48,* 163–206.

McNamara, T. P. (1986). Mental representations of spatial relations. *Cognitive Psychology, 18,* 87–121.

McNamara, T. P. (2003). How are the locations of objects in the environment represented in memory? In C. Freksa, W. Brauer, C. Habel, K. F. Wender (Eds.), *Spatial cognition III: Routes and navigation, human memory and learning, spatial representation and spatial reasoning,* Lecture Notes in Artificial Intelligence (LNAI) 2685 (pp. 174–191). Berlin, Germany: Springer-Verlag.

Mou, W., McNamara, T. P., Valiquette, C. M., & Rump, B. (2004). Allocentric and egocentric updating of spatial memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 142–157.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*(6), 1157–1165.

Speech Perception

Speech perception refers to the processes involved in identifying and understanding the meaningful patterns of spoken language. The speech signal originates from the concerted actions of the speaker's lungs, larynx, jaw, tongue, lips, and soft palate (soft tissue in the back of the roof of the mouth) to generate sounds that are shaped in particular ways. A fundamental problem in speech perception is understanding how a listener recognizes the complex acoustic pattern of sound waves as being composed of meaningful linguistic units (vowels, consonants, syllables, words, sentences, etc.). This problem becomes strikingly apparent when one realizes that there is no simple one-to-one mapping between the acoustic speech signal and our perception of what the talker said. This entry examines attributes of the human voice and speech signal, some of the major experimental findings, and several prominent theories that

attempt to shed light on the basic processes involved in speech perception.

The Speech Signal

The source of energy that drives the speech signal is the stream of air that originates from our lungs when we exhale. This air stream passes through our *vocal cords,* which cause the air stream to vibrate. When we talk, the vocal cords open and close rapidly, separating the air stream into a sequence of puffs of air. This sequence of puffs sounds like a "buzzing" noise, which changes in pitch as the cords vibrate faster or slower. The supralaryngal *vocal tract,* the part of the throat and mouth that lies above the vocal cords, further modifies the speech sound depending on its particular shape and size. Furthermore, by moving the soft palate, tongue, lips, and jaw (collectively referred to as *the articulators*), we can further alter the shape of the vocal tract and thus create a wide range of speech sounds.

One of the most basic characteristics of a human voice is its *fundamental frequency* (f_0), which corresponds to the perceived pitch of the speaker's voice (i.e., whether a person's voice sounds "deep" or "high"). Fundamental frequency is determined by the rate at which the vocal cords vibrate. Speakers have control over modifying the vibration rate while they talk, resulting in transient changes to f_0. These changes in f_0 can play a major role in the perception of different aspects of speech. For example, changes to f_0 can be used to emphasize one or more words (*word stress*). As another example, when an English utterance ends with a high pitch, this often signifies a question.

The articulators change the shape of the air stream and the frequency composition of the resulting speech waveform, which forms many of the common speech sounds of our language (e.g., vowels and consonants). The air stream can be wide open (resulting in vowels), redirected partially through the nose (resulting in the nasal consonants m and n), changed in shape over time (resulting in the gliding consonants j, w, and y), or momentarily stopped completely (resulting in the stop consonants b, d, g, p, k, and t). An important acoustical property of stop consonants is *voice onset time* (VOT), the delay between the burst of sound caused by quickly releasing a set of articulators (such as the lips) and the beginning of vocal fold vibration. For example, producing a syllable like "pah" requires the lips to release a burst of air to produce the /p/, and then a delay until the "ah" sound is made. "Pah" has a relatively long VOT. On the other hand, the syllable "bah," which involves more or less the same articulations, has a shorter VOT. As these examples illustrate, VOT can serve to distinguish speech sounds from one another. Even though the /b/ and /p/ are produced with the lips in similar ways before the vowel sound is made, it is the difference in VOT that distinguishes them.

Thus, speech sounds are generated through a complex combination of different vocal organs working together. Is it possible to identify basic units of speech from looking at the acoustic signal itself? Although most people are familiar with the idea of syllables and words in language, there is an even more elementary building block of speech: the *phoneme.* Phonemes are defined as the smallest units of sound that can distinguish one meaningful word from another. For example, consider the words *bat* and *bit.* These two words have identical sounds of /b/ and /t/ at the beginning and endings, but differ in the middle vowel sound. Thus, it is the middle elements—the phonemes (vowels in this case)—that distinguish these two words. Phonemes can be either consonants or vowels and can be further combined into larger units, including syllables, which generally consist of vowels surrounded by one or more consonants. Finally, groups of phonemes and syllables can be combined together to form meaningful words in a given language.

Despite the everyday assumption that speech can be broken into context-free discrete symbolic units, in reality the speech signal is not as tidy as the previous paragraph might lead one to believe. In fluent speech, the articulators begin to move into position to generate the next speech sound even while the current sound is still being produced. This property of speech production is called *coarticulation.* Coarticulation refers to the overlap that exists in speech production between the articulatory activity of adjacent phonemes. For example, the way in which you make a /k/ sound depends on the vowel that follows it. Because of the shape of the vocal tract necessary to produce different vowels, production of the consonant /k/

as in the word *key* requires placing the tongue farther forward in the mouth, compared to the /k/ in the word *coo,* where the tongue is placed farther back in the mouth. This results in differences in the acoustic speech signal for the two /k/ sounds, despite the fact that we perceive them as the same. These observations illustrate a general property of speech: The acoustic and articulatory features of a given phoneme are highly context dependent and are conditioned by the phonemes that precede and follow it. Strict context-free discrete perceptual units, such as vowels, consonants, phonemes, and even words, do not exist in the raw acoustic signal as they do in printed text. Rather, they are linguistic abstractions resulting from perceptual analysis. The speech sound is a continuous time-varying acoustic signal rather than a series of distinct units arranged sequentially in time. Thus, any given percept of speech may have many possible different ways of being represented at the physical level. Understanding the neural and cognitive processes involved in perceiving the abstract, idealized linguistic units of speech from the complex, context dependent, acoustic signal is one of the fundamental goals of speech perception research.

Some Important Findings

Early research in speech perception focused almost exclusively on the perception of phonemes in isolated syllables or nonsense words. One of the important findings observed in phoneme perception is called *categorical perception.* Categorical perception refers to the phenomenon of perceiving items from a large and varied stimulus set in terms of only a small number of discrete perceptual categories. For instance, consider a hypothetical example in vision. Suppose there is a wall painting colored red on the one end and yellow on the other, with the red color morphing very slowly and gradually into yellow, having no apparent distinct boundaries between the two. Generally, a viewer would look at the painting and notice the very smooth and continuous transition of colors, recognizing the myriad of color combinations between the two end points of red and yellow (e.g., "bright red," "reddish-orange"). But suppose a different viewer looked at this same painting and instead was convinced that she saw half of the painting as completely red and half as yellow, with no other colors in between. This is in essence the phenomenon of categorical perception: A physical stimulus that varies continuously is perceived as having a very clear-cut, well-defined, small number of categories.

Returning to the case of speech perception, consider the two syllables mentioned previously that differ only in terms of their VOT ("bah" and "pah"). Using modern speech synthesis techniques, it is possible to create artificial speech sounds of these two syllables and then continuously vary the VOT of these two stimuli to create a range of intermediate sounds. Using such techniques in the early 1950s, Alvin Liberman and colleagues at Haskins Laboratories found that when listeners are presented with these stimuli, they do not perceive the speech sounds continuously; rather, they identify either a clear "bah" or "pah" sound—with no intermediate percepts—even though the physical stimulus varies incrementally and gradually between the two sounds. It is as if speech is perceived like our second observer of the colored fields previously described. Thus, categorical perception may be a useful way for the brain to sort out a large, potentially confusing amount of variation in the speech signal into a limited number of more manageable, discrete perceptual categories.

Another important finding is that speech perception is significantly improved by visual access to the speaker's face. Especially under noisy listening conditions, being able to see the talker's mouth and articulators along with the acoustic speech signal significantly improves speech perception. For example, anecdotally, a common complaint among the elderly is that they are unable to understand what other people are saying unless they have their glasses on. However, it is not only under degraded auditory conditions that visual information has a functional impact on speech perception. A classic paper published by Harry McGurk and John MacDonald in the 1970s reported that auditory recordings of "ba" that were dubbed onto films of a person saying "ga" often led to reports of "da"—a "fused" utterance that was never actually presented. The *McGurk effect* is a multimodal perceptual illusion based on an unnatural co-occurrence of inputs. This illusion demonstrates that speech perception is susceptible to the influence of visual information even when

the auditory signal is not degraded. This finding has led to interest recently in just how much information the visual channel can provide during speech perception, as well as what the underlying perceptual, cognitive, and neural mechanisms are that are involved in the integration of visual and auditory speech information.

The McGurk effect demonstrates how the acoustic signal of speech may be perceived differently given different visual contexts. However, similar phenomena have been shown to occur naturally in the auditory channel alone. For example, the existence of coarticulation effects demonstrates that a single acoustic cue can be perceived differently depending on other sounds in the immediate context. For instance, the same burst of sound can be perceived as a /p/ before a vowel such as "ee," but as a /k/ before the vowel "ah." Similarly, one's knowledge of the words of language can significantly influence the interpretation of spoken sounds. For example, when part of a familiar word is artificially obliterated with a patch of noise, listeners often report hearing the part of the word that was eliminated in addition to the noise. This is known as the *phoneme restoration effect*. Another example of this is that listeners often do not notice mispronunciations of sounds within highly familiar words. These findings illustrate how one's long-term memory and knowledge of language—"top-down" effects—can influence the interpretation of "bottom-up" perception of the acoustic speech signal. Thanks to the combination of top-down and bottom-up processes, our perceptual systems are remarkably robust and highly adaptable to both the wide variability that exists in the raw acoustic speech signal and the limitless range of contexts in which speech perception occurs.

The robustness of speech perception in the face of enormous physical variation gave rise to the traditional assumption that some kind of a *normalization* process occurs, with unnecessary redundant information being stripped away as a result of early perceptual analysis. That is, consistent with the findings of categorical perception, many researchers have implicitly assumed that the speech signal is reduced to an abstract, idealized linguistic message and that signal variability that is not directly related to the linguistic message is eliminated. According to the traditional abstractionist or symbol processing view of speech, perceiving speech is akin to perceiving printed letters on a page, with speech consisting of a linear sequence of discrete, idealized symbols.

Although much initial research was consistent with the idea that speech perception involved processes of normalization and abstraction, in recent years, it has become more apparent that much of the seemingly nonlinguistic information in the speech signal is retained and used to modify perception. For example, it is now known that so-called *indexical features* of speech—aspects of the speech signal that provide information regarding the speaker's identity and physical condition—can have an effect on speech perception. For example, indexical information in a particular person's voice may allow us to identify that the speaker is John, and that he is sad or tired. Recent studies have shown that indexical information in speech is not, in fact, discarded through a normalization process but instead may interact with memory and attention processes to affect how linguistic messages are perceived. As one example of how indexical information is used in speech perception, it has been found that familiarity with a talker's voice facilitates the accuracy of identification in noise of novel utterances spoken by that same talker. In sum, although this is a new direction of speech research compared to traditional methods of inquiry, it is now widely accepted that speech perception involves encoding both indexical information and the "symbolic" linguistic message, and that speech perception is necessarily influenced by both kinds of information in the signal.

Major Theoretical Perspectives

Several theories have been proposed over the years to explain the various phenomena and findings of speech perception. Three theories reviewed in this entry are the motor theory, direct realism, and the general auditory account of speech perception.

To begin, consider again the phoneme /k/ in *key* and *coo*. At the level of the acoustic signal, the /k/ sound is different for these two syllables, due to coarticulation effects involved in producing the two different vowels. However, listeners perceive both /k/ sounds as being perceptually equivalent, despite the differences in the actual speech signal. How listeners are able to perceive highly variable

acoustic differences as equivalent sounds has been a major hurdle in our understanding of the processes of speech perception. The *motor theory of speech perception* (MTSP) attempts to bypass this hurdle by proposing that listeners unconsciously articulate the speech sounds they hear and then use their own articulation to perceive and understand what they heard. The reason we perceive equivalence despite the underlying physical differences in the speech signal, proponents of MTSP argue, is that it is the same essential articulatory gesture or motor command that is used to produce /k/ in both cases. That is, our brains attempt to map what we hear onto something that it knows how to produce, using its motor categories to define what is perceived.

If this version of MTSP is correct, then the underlying articulations producing speech ought to be less variable than the actual acoustic signal. However, it was soon discovered that low-level articulatory motor activity is no less variable than the actual acoustic signal. For example, as already discussed with *key* and *coo*, there are slight differences in how we produce the /k/ sound. Both within and across individual speakers, articulatory variability is high, even when the perceptual result is relatively stable. Motor theory was subsequently revised, such that the proposed motor correspondence no longer referred to externally measurable articulatory motions, but rather to the recovery of abstract sets of motor commands. Some recent evidence has shown that the brain contains *mirror neurons*—neurons that are active both when a person produces a particular action and when the person observes someone else producing that same action—which could provide a neural account of MTSP. Even so, with the move away from external articulations to internal motor commands, MTSP's main hypothesis has become extremely difficult to test. Another problem with MTSP is that it was found that chinchillas appear to show categorical-like perception for human speech sounds, despite the inability to produce speech themselves. Although MTSP may be less tenable than originally thought, researchers continue to explore the idea that speech production and speech perception are closely linked.

The *direct realist* (DR) approach to speech perception is based on the legacy of James Gibson's ideas of "direct perception," a theory of perception that argues that the senses provide us with direct awareness of the external world, rather than having our perceptions based on some internal representations of the world. Direct realism is similar to MTSP in that there is an emphasis on articulatory events rather than the acoustic signal. However, rather than relying on an internal, abstract motor command, proponents of this view argue that there is no need to examine the internal contents of the perceiver to explain speech perception. In other words, there is no need for positing intermediate perceptual or cognitive mental representations. Direct realism also differs from MTSP by proposing that speech perception involves domain-general mechanisms of perception that are also used in non-speech domains, such as vision, whereas MTSP argues for speech-specific perceptual mechanisms. The DR perspective provides a valuable reminder of the need to step back and consider speech perception in relation to the larger environment. The DR approach has also been an important framework that has led to a better understanding of the sources of information that are available in the acoustic signal itself. However, it has been difficult to design new methods to test the basic claims of DR. Furthermore, the DR view runs counter to the current mainstream perspective in the psychological sciences that emphasizes perception and cognition as consisting of stages of information processing and the manipulation of internal representations. Possibly for these reasons, DR currently represents a minority view in speech perception research.

Finally, the *general auditory account* (GAA) proposes that speech perception can be explained by general-purpose mechanisms and processes that are common to audition more generally, not just specific to speech perception. In this way, GAA is similar to DR. However, GAA differs from both of the previous two theories in that it assumes that speech perception relies on the acoustic signal itself rather than on the perceiver's underlying motor or articulatory gestures. How does GAA explain perceptual equivalence of speech sounds despite the large amount of underlying variability in the speech signal? The proposal is that perceptual equivalence is due to a general ability of the perceiver to learn and make use of multiple acoustic cues or sources of information in the signal to

narrow in on a single perceptual category. Perceptual equivalence thus arises from the integration of multiple cues in the speech signal, where any single cue alone is imperfect, but the combination of many cues together can be predictive and reliable. The advantage of GAA is that it does not rely on specialized speech-specific mechanisms in order to explain the basic phenomena in speech perception; thus, findings from any perceptual domain, including from nonhuman animals, can provide theoretical insight. However, a possible disadvantage of GAA at this time is that it is considered to be vague and unspecified, rather than being a coherent theory, and thus presently may be limited in the explanatory power it can currently provide relative to other theories.

Several prominent theories have been proposed to explain speech perception, but it appears that no single theory at present can adequately explain all findings.

Christopher M. Conway, Jeremy L. Loebach, and David B. Pisoni

See also Audition; Computer Speech Perception; Perceptual Development: Speech Perception; Speech Perception: Physiological; Speech Production; Word Recognition

Further Readings

Cleary, M., & Pisoni, D. B. (2001). Speech perception and spoken word recognition: Research and theory. In E. B. Goldstein (Ed.), *Blackwell handbook of perception* (pp. 499–534). Malden, MA: Blackwell.

Denes, P. B., & Pinson, E N. (2007). *The speech chain: The physics and biology of spoken language.* New York: W. H. Freeman.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55,* 149–179.

Speech Perception: Physiological

Speech, as our primary means of communication, is perhaps the most important sound in our daily lives. Current theories of how the brain perceives speech rely on more than a century of investigation that has included patients with brain damage, microelectrode recordings in nonhuman animals, electric fields measured on the human scalp, and, most recently, neuroimaging. The current consensus, as described in this entry, holds that speech proceeds along parallel pathways or streams in the cortex. A "what" stream is dedicated to speech comprehension, and a "where" or "how" stream is more important for learning speech and holding it in mind (as when you remember a phone number in your head; color insert, Figure 6). The left hemisphere of the brain tends to be dominant for many aspects of speech perception, such as understanding sentences, though the reasons why are actively debated.

Neural Measures

Numerous techniques have been used to study how the brain processes speech. One of the oldest is lesion analysis, where functional neuroanatomy is inferred from patients with localized brain damage (lesions, e.g., from stroke or trauma) who exhibit a particular language deficit or aphasia. Until the middle part of the 20th century, much of our understanding of speech and the brain came from lesions. With the widespread use of microelectrode recordings in nonhuman animals, researchers began to characterize moment-by-moment representations of sounds. Microelectrode recordings, though, are surgically invasive and cannot be performed in healthy people. Most early studies of real-time speech perception in humans instead used electroencephalography (EEG), which measures electrical fields from neural activity with electrodes resting on the scalp. Numerous characteristic deviations or oscillations in electrical waves have been identified in speech and language processing. However, early EEG studies did not use many electrodes on the scalp and could not identify *where* in the brain this speech processing occurs.

Recent decades have witnessed a flood of speech studies using neuroimaging, which aims to localize brain function. Functional magnetic resonance imaging (fMRI) for instance, developed in the 1990s, measures changes in the blood supply to infer where and roughly when (within about a second) neural