Models and Theories of Speech Production and Perception

After reading this chapter, you will

1. Understand the differences between theories and models.

2. Be able to describe issues in speech production, such as serial order, degrees of freedom, and context sensitivity.

3. Be familiar with some important theories of speech production, such as target models, feedback and feedforward models, dynamic systems models, and connectionist models.

4. Be aware of issues in speech perception, such as linearity, segmentation, speaker normalization, basic unit of perception, and the specialization of speech perception.

5. Appreciate the different categories of speech perception theories, such as active versus passive, bottom-up versus top-down, and autonomous versus interactive.

6. Be acquainted with some important theories of speech perception, such as motor theory, acoustic invariance theory, direct realism, connectionist theories, logogen theory, cohort theory, fuzzy logical model, and native language magnet theory.
Models and theories are tools that help us to understand and predict the behavior or functioning of a system or some part of a system. Once we understand the various parts of a system, how they are related to each other, and how the system works as a whole, we can predict how the system will work under different conditions, and we can apply this knowledge to understanding how the system or any of its parts breaks down under various conditions. This type of understanding, in turn, can lead to the development of procedures to repair the system.

For example, the systems of speech production and perception are very complex and include many separate subsystems that must work together in a coordinated fashion, such as the respiratory system and the phonatory system. Each of these systems, in turn, is composed of many structures that must also cooperate to achieve the goals and targets of the system. Respiration, for instance, is a system comprised of numerous muscles and other structures that work in a coordinated manner to achieve the goal of inhaling and exhaling. The phonatory system consists of the larynx with all its cartilages, muscles, vocal folds, and so on, all acting to achieve the goal of voice production. And all the systems work together to achieve the final goal, the rapid, meaningful production and perception of speech sounds to encode and decode our thoughts, ideas, and feelings in a linguistically appropriate manner.

The speech production and perception systems can break down at any point within any of the subsystems. Because these systems and subsystems are so complex, it is necessary to have theories that integrate current knowledge about them and to make models of the systems to test and extend our knowledge about how they work, how they break down, and how breakdowns can best be compensated for or repaired. The terms theory and model are often used interchangeably, but there are differences between the two. Very often, models are used to test theories.

**Theories**

A theory is a statement about a particular phenomenon, incorporating the underlying principles and assumptions. A theory is a way of interpreting facts about the phenomenon in an integrated manner. Theories help to explain observed data and information and can be used to make predictions about events related to the phenomenon in question. Because theories are based on incoming information and new research, they are always subject to change. Indeed, as Kent (1997a) has remarked,
Progress in science is marked by the development of new theories, by the rejection of theories that are inadequate to account for the data, and by the modification of theories that account for some but not all of the data. Theories should explain data that have been gathered and should help to guide the collection of important new data. (p. 401)

Theories are not just interesting from a research point of view, they also help to institute changes in practice. These changes can have far-reaching effects in the clinical domain. For example, prior to the late 1970s, most clinicians and researchers in speech pathology believed that spasmodic dysphonia (SD) had a psychological origin, based on certain aspects of the disorder, such as its resistance to traditional voice therapy, its often close association with stressful periods in a person’s life, and the fact that many people with this disorder had great difficulty in normal conversational speech, but were able to sing, or whisper, or talk at a higher than normal pitch without trouble. Therefore, the treatment for this disorder was often focused on a combination of speech therapy and psychological counseling. However, in the late 1970s and 1980s, data started accumulating suggesting that spasmodic dysphonia is neurogenic. Many different speech laboratories around the United States gathered experimental data about the neurological aspects of this disorder. By about the mid-1980s, the theory of causation changed from psychogenic to neurogenic.

Based on a neurogenic theory, current treatment techniques have a neurological, rather than a psychological, basis. For instance, injections of botulinum-A toxin (Botox), which affect the neurological functioning of the vocal folds, are often prescribed. This kind of treatment has been shown to be extremely effective in alleviating the symptoms of SD in most patients, whereas previous traditional voice and psychological treatments had not, in general, been successful. Thus, research and theory interact with each other to generate new research, new theories, and new practices.

Models

A model is a simplification of a system or any of its parts. Models are constructed to represent the system in some way that can then be manipulated in a controlled manner. Modeling a system can be done in numerous ways, such as fashioning a physical or mechanical version of the system to be tested, using specimens for physiological modeling, or applying mathematical and computer algorithms to the system.

An example of a mechanical model is one developed by Georg von Bekesy, who proposed a theory of hearing based on the traveling wave along the basilar membrane in the cochlea. In his model of the cochlea,
a sheet of rubber of varying thickness levels simulated the basilar membrane and was placed in a tank of water representing the endolymph within the cochlea. Waves of different frequencies were introduced into the tank. As von Békésy’s theory predicted, at high frequencies the thin part of the rubber sheet vibrated with the highest amplitude, and at low frequencies the thicker part of the sheet responded with the greatest amplitude of vibration. Thus, von Békésy’s model of the traveling wave provided support for his theory of hearing.

Physiological models often use specimens taken from animal or human cadavers to determine how a particular structure responds under different conditions. For example, a current theory relating to phonation states that the hydration level of the vocal folds affects vocal fold tissue viscosity, which is proportional to phonation pressure threshold ($P_{th}$), the minimum subglottal pressure required to set the vocal folds into vibration. This theory has been the basis for clinical management of some laryngeal problems in singers, in which lack of hydration of the vocal folds has been the assumed cause of increased effort in initiating vibration. Based on this assumption, patients are encouraged to increase the humidification of their environments and to drink large amounts of water.

Jiang, Ng, and Hanson (1999) used excised canine larynges to test this theory. The larynges were mounted on a special holder and dehydrated with warm dry air. Air was blown through the vocal folds, causing vibration and sound production to occur, and $P_{th}$ was measured. Within about five minutes of dehydration, vibration and sound production stopped. The larynges were then immersed in a saline solution for 30 minutes, and $P_{th}$ was determined again. $P_{th}$ decreased after the rehydration, allowing sound production to occur with increased efficiency. These findings confirmed clinical impressions that hydration is critical in the physiology of normal phonation.

Mathematical models are an effective means of explaining phenomena and testing theories. For instance, neurological, biomechanical, and aerodynamic factors interact to generate jitter in the human voice. Titze (1991b) was interested in teasing out the effects of neurological sources, such as the number of motor units contributing to muscular contraction and the firing rate of the motor units, on jitter. The application of mathematical models to the activity of the thyroarytenoid muscle allowed Titze to hold certain parameters of muscle function constant, vary other parameters in a systematic manner, and thus isolate the effects on the resulting jitter values. In this way, Titze demonstrated that around 0.2 to 1.2 percent of jitter in the human voice seems to result from neurological sources.

Models and theories abound in trying to explain various aspects of speech production and perception. We will focus on only a few models and theories that attempt to account for the overall processes of production and perception. Kent (1997a) noted that the current understanding of speech behavior is represented by a “mosaic of theories,
most of which pertain to only certain levels of observation or conceptualiza-
tion. He also pointed out that a major divide exists between theories of speech
production and speech perception, which have had largely separate theoreti-
cal developments.

Speech Production

Numerous different categories of models and theories of speech production
exist, such as target models, feedback-feedforward models, dynamic systems
theory, and connectionist theories. However they are categorized, most theo-
ries of speech production try to address three major issues related to the orga-
nization and regulation of speech motor control. These include how speech is
ordered serially, the problem of degrees of freedom, and the question of con-
text sensitivity.

The Serial-Order Issue

Although the output of speech is a continually varying waveform, the linguis-
tic elements that make up speech are produced in a serial order. The order is
important for meaning: Although the phonemes /k/, /t/, and /æ/ are used in
the words cat, tack, and act, the order in which they are produced determines
how the word will be perceived and recognized. Speech is thus a sequence of
elements. The question is precisely which elements are serialized. The ele-
ments could be specific features of a sound (e.g., voicing or nasality),
phonemes, syllables, parts of syllables, or other larger or smaller elements. Re-
search on this issue continues, with evidence so far favoring a phoneme- or syl-
lable-sized unit of speech organization.

Degrees of Freedom

When we speak, we need to control a huge number of muscles, including those
of the respiratory, laryngeal, and articulatory systems. In addition, many struc-
tures in these systems can move in different ways, at different speeds, and in
different combinations. For example, the lower lip and jaw can move in phase
with each other (have the same relative timing) and in the same direction; out
of phase with each other (different relative timing), but in the same direction;
or in phase with each other, but in opposite directions (Kent, 1997b). Each dif-
ferent potential muscular contraction of each muscle in each system constitutes
what is known as a degree of freedom, so the total number of possible number
of contractions, or degrees of freedom, is enormous. The speech motor system
must somehow regulate all the muscular contractions of all the speech subsys-
tems to ensure that the appropriate structures are moving rapidly and in the correct sequences to generate the target sounds and words.

Many theories have been put forward to explain how the speech motor system achieves this level of control. Some theories propose that the speech motor system “programs” separate neuromuscular signals for each required muscle contraction. Another class of theories organizes muscular control in a hierarchy, with upper levels of the system controlling lower levels. Still other theories suggest that the speech motor system uses various strategies to reduce the total number of degrees of freedom to a smaller number, for instance, by combining muscles into functional groupings that work in a coordinated fashion to achieve a desired goal. In this way, muscles are controlled in groups, rather than individually.

Context-Sensitivity Problem

Theories of speech production need to take into account the fact that sounds vary with the context in which they are produced and are influenced by speaking rate, stress, clarity of articulation, and other factors. Coarticulation is an integral aspect of speech production that results in enormous variability in the production of a target sound. A given speech sound often can be produced in several different ways, and this variability in production is a central factor in speech motor regulation.

Theories of Speech Production

Target Models

Target models describe speech production as a “process in which a speaker attempts to attain a sequence of targets corresponding to the speech sounds he is attempting to produce” (Borden et al., 1994). Some theorists have suggested that these targets are spatial. Spatial target models posit that there is an internalized map of the vocal tract in the brain that allows the speaker to move his or her articulators to specific regions within the vocal tract. The speaker can achieve the targets no matter from what position the articulator(s) begin(s) the movement. The fact that articulators must reach a particular position from different starting points is important, because it means that the movements of the articulator for a specific sound cannot be invariant, but must change depending on the starting point.

For example, to achieve the velar target for the /k/ sound, the tongue would have to move in a different trajectory depending on the preceding vowel. The path of the tongue would vary, for instance,
Feedback is used to detect and correct errors in speech output; feedforward signals are used to make articulatory adjustments online.

During which the signals from the periphery (i.e., vocal tract) return to a central processor in the brain for comparison of the intended and the actual movements, feedforward signals make adjustments at the periphery so that the system is primed to move in an efficient, coordinated manner. Feedforward is therefore a much faster process, which may help to explain why disruptions such as bite blocks do not have much of an effect on speech production. It is likely that both feedback and feedforward are used in the control and regulation of speech production.

**Dynamic Systems Models**

In this kind of theory, the degrees of freedom problem is addressed by positing that groups of muscles link up together to perform a particular task. These linkages between muscles are not fixed: A muscle might be grouped with a particular set of muscles in what is called a **synergy** or a **coordinative structure** to achieve one particular goal and with a different set of muscles in a different synergy to achieve a different goal. Instead of individual muscles being controlled, an appropriate coordinative structure is selected, comprising a set of muscles that essentially form an integrated unit focused on achieving a particular motor activity. The different muscular responses in a synergy can be adjusted to meet the requirements of a particular task under different conditions.

For example, the lip and jaw muscles function as a coordinative unit in bilabial closure. Typically, the lip and jaw muscles cooperate to close the lips. However, if the person has a bite block clamped between the teeth to immobilize the lower jaw, the upper and lower lips can compensate by increasing the force or extent of their movement. If the lips cannot move for some reason, the jaw by itself can bring the lips together. This theory thus reduces the degrees of freedom required for control, while allowing a great deal of flexibility in the organizational linkages between muscles and structures.

**Connectionist Models**

Computer models have been developed that simulate the neural processing of the human brain. These models are also known as **spreading activation models** and **parallel-distributed processing models** (PDP). PDP models are based on a way of processing signals that is nonhierarchical. In other words, rather than finishing one step in the process before moving on to the next step, steps are processed more or less in parallel. This kind of processing is somewhat akin to the way that the brain processes information. Indeed, the performance of steps in parallel, or at least with much temporal overlap, is typical of speech produc-
Speech Perception

As with speech production, many issues in speech perception give direction to the theories attempting to explain how we analyze and perceive the spoken word. Some of these issues are linearity, segmentation, speaker normalization, and the basic unit of speech perception.

Linearity and Segmentation

The linearity principle asserts that a specific sound in a word corresponds to a specific phoneme. The sounds that make up the word are distinct from each other and occur in a particular sequence. The segmentation principle is based on the notion that the speech signal can be divided into discrete units that correspond to specific phonemes.

These two principles suggest that speech perception is based on a linear correspondence between the acoustic speech signal and the linguistic phonemic units. An abundance of research, however, has established that this is not the case. For example, acoustic characteristics of a given phoneme vary depending on context. The /k/ in /ku/, for instance, will have somewhat different acoustic cues than the /k/ in /ki/. The same has been shown to be true for the /d/ in /di/ versus /du/.
Chapter 14 / Models and Theories of Speech Production and Perception

A rising $F_2$ transition in /di/ and a falling one in /du/ both serve as acoustic cues for /d/. See Figure 14.1.

The frequency of stop bursts also signals different places of articulation depending on context. Research has shown that when a burst is centered at 1440 Hz it is heard as /pi/ before /i/ and /u/, but as /k/ before /a/. Also, because of coarticulation, information is present about the acoustic properties of a specific phoneme, as well as the phonemes that precede and follow it. Consequently, although we perceive speech as a series of separate and distinct phonemes and words, the acoustic boundaries between phonemes are blurred.

Speaker Normalization

As well as the lack of invariance and clear segmentation of phonetic units, theories of speech perception try to take into account other important factors. One of the most puzzling questions in speech perception is that of speaker normalization. That is, how is it that listeners are able to recognize sounds and words despite the large variations in the way that speakers produce them? Speakers vary speech sounds in terms of pitch, loudness, stress, and rate of speech. Speaker age and sex also influence sound production. For example, the same vowel can be produced with very different ranges of formant frequencies from speaker to speaker, and with considerable overlap with other vowels, yet most listeners have no difficulty in vowel recognition. Even infants are able to discriminate speech sounds produced by different talkers with relative ease. Somehow we normalize speech; that is, we seem to be able to ignore irrelevant differences between productions of a given sound, while focusing...
on the features that characterize it as a member of that particular phoneme family.

Basic Unit of Perception

Much research effort has also been expended on the question of the basic unit of perception. Do we perceive acoustic-phonetic features, allophones, phonemes, syllables, or larger units of speech? This question remains an issue, although each of these units has some evidence in its support. It may be that individuals rely more on one kind of unit than another depending on the situation and context of the speech to be perceived. For example, in a quiet listening environment with no competing noise, such as background party noise or work-related noise, listeners may pay more attention to smaller units and rely less on the syllabic structure of speech to perceive meaningful information. In a noisy situation, the person may be forced to process at a higher level of linguistic information to supplement and facilitate acoustic-phonetic recognition.

This issue also has a developmental aspect. There is evidence that infants and young children may process auditory information using larger units such as syllables, whereas older children and adults may rely more on smaller units such as phonemes. Children appear to go through a developmental weighting shift, in which their perceptual strategies change with increasing linguistic experience (Nittrouer, 1996a). Very young children seem to focus more on syllables to process speech information, because with their limited lexicons they do not need to process auditory information at a more detailed level. As the child's lexicon increases, so too does the need for more detailed acoustic-phonetic representation of words to discriminate between the greater number of similar words in the child's memory.

Nittrouer (1996a) suggested that mature language users appear to weight the various components of the signal in ways that they know will lead to correct decisions about phoneme identity, that is, by focusing on the spectral characteristics of the sound wave. Young children focus more on formant transitions between syllables to divide the acoustic signal into syllabic units. As the child matures, he or she begins to weight the more static and more detailed components of the signal more heavily.

Specialization of Speech Perception

Specialization of speech perception is another theoretical question that has been widely studied and that continues to remain an unsolved issue. Speech perception is viewed by some theorists as a specialized process unique to humans. Early findings about categorical perception seemed to corroborate this view. Categorical perception is considered by some scientists to be a mechanism that
Some theorists view speech perception as a specialized human process; others assert that speech signals are perceived in the same way as nonspeech sounds. Humans have developed by which their perceptual system has become specially adapted for the perception of speech. However, experiments with animals such as chinchillas, monkeys, dogs, and birds have shown that these animals are also capable of categorical perception, which therefore cannot be a uniquely human ability. Instead, general auditory perceptual mechanisms common to many mammals may be involved in categorical perception, rather than some specialized speech process.

Nittrouer (1996a) supported the view that speech signals are perceived in the same way as other acoustic events. She noted that “Most sound-producing events result in multicomponent signals. The way in which information from the separate components is combined, and the weight is assigned to each, is determined by what the listener wishes to learn from that signal. Thus, perceptual strategies appear to be specific to the perceptual decisions to be made.”

On the other hand, evidence for some specialization of speech perception comes from studies of a phenomenon called the perceptual magnet effect. This effect has to do with prototypes of speech stimuli. A prototype is the most representative instance of a category (Hawkins, 1999c). It has been found that certain examples of a sound seem to be better prototypes than others. In other words, individuals are able to identify the best or prototype examples of a vowel compared to poorer or nonprototype examples of that same vowel from a series of acceptable instances of the vowel. The vowels that the person selects as prototypes depend on the person’s native language. Infants as young as 6 months old demonstrate this effect, but only for their own language. Research has shown that American and Swedish babies show this perceptual magnet effect for the vowels of their ambient language, but not for the vowels of the non-native language.

The perceptual magnet effect is so called because, figuratively, the prototype “pulls” sounds that are acoustically similar to it toward itself, resulting in poorer discriminability for sounds clustering closely around the prototype. In other words, differences between good representations of a sound are reduced, thus helping individuals to ignore irrelevant differences between members of a category. What makes this effect a possible candidate as evidence of some unique human speech perception abilities is that, unlike categorical perception, this effect is not demonstrated by animals.

Categories of Speech Perception Theories

Numerous theories of speech perception have evolved over the years, and new theories continue to be developed as new information becomes available. The theories have been grouped into three major categories: active versus passive, bottom-up versus top-down, and autonomous versus interactive.
Active versus Passive

Active theories stress the links between speech perception and speech production, which share common properties. In these theories, the individual's knowledge of how sounds are produced is an integral factor in facilitating sound recognition. Passive theories, on the other hand, emphasize the sensory aspects of speech perception. They stress the filtering mechanism of the listener, with knowledge of speech production and vocal tract characteristics playing a minor role and used only in difficult listening situations.

Bottom-Up versus Top-Down

Bottom-up theories are based on the premise that all the information necessary for the recognition of sounds is contained within the acoustic signal. Therefore, the listener does not need to involve linguistic and cognitive processes in decoding sounds. By contrast, top-down theories emphasize higher level linguistic and cognitive operations as playing a crucial role in the identification and analysis of sounds. Most theories are neither completely top-down nor bottom-up, but place more or less weight on acoustic versus linguistic-cognitive contributions to speech perception.

Autonomous versus Interactive

Speech perception involves many stages of processing of the acoustic signal. The signal is initially processed at the individual's ear and then goes through many further transformations as it travels to the brain, eventually to be interpreted and understood. Interactive theories posit that information and knowledge from many sources available to the listener are involved at any or all stages of the processing of the signal on its journey through the speech perception system. Autonomous theories propose that the signal is processed in a serial manner, from phonetic to lexical stages, to syntactic stages, to semantic stages, and so on. Other sources of linguistic and cognitive knowledge and contextual information are not brought into play in autonomous theories; rather, the output of one stage of processing provides the input to the next stage.

Motor Theory

An early and still very influential theory of speech perception is the motor theory, developed at the Haskins Laboratory at Yale University. Motor theory is
Motor theory emphasizes the link between speech production and speech perception in terms of articulatory gestures that individuals are innately able to perceive. An active theory that stresses the link between perception and production of speech. In essence, according to this theory, an individual perceives speech because he or she produces speech. Because listeners have experience in producing speech sounds themselves, they are aware, at some level, of the relationship between movements of the articulators, vocal tract configurations, and the acoustic consequences of these articulator movements and positions. Motor theory assumes that speech perception is unique, relying on a special processor located somewhere in the brain to decode speech.

The older version of the theory assumes that there are unchanging, invariant motor commands in the form of neural signals to the articulators to produce the same phoneme in different phonetic contexts. The motor theory has recently been revised. In the revised theory the acoustic signal is thought to be perceived in terms of articulatory gestures that individuals are innately able to perceive, such as tongue backing and lip rounding. The listener does not, however, perceive the actual movements but some kind of abstract articulatory plan that controls the vocal tract movements that would, ideally, result in a perfect production of the utterance (Hawkins, 1999b).

In other words, the abstract gestures produce particular constrictions in the vocal tract, with each constriction being appropriate for a specific phonetic place and manner of articulation (Hawkins, 1999b). These abstract articulatory plans are known as gestures. A gesture is, according to Hawkins, one of a family of movement patterns that all achieve the same goal, such as a particular constriction in the vocal tract (e.g., bilabial closure). Gestures for speech can be thought of as the basic units of speech production, which control and coordinate the cooperative activity of the articulators. The gestures are neuromotor commands that are specific to speech and therefore to humans. They are phonetic in character, invariant, and accessible only in the specialized phonetic module in the brain. The theory proposes that listeners somehow retrieve the intended gesture for the underlying phoneme from the variable acoustic signal by compensating for the effects of coarticulation.

Research, however, does not support many of the assumptions of motor theory. For instance, Peter Eimas and colleagues (e.g., Eimas, Miller, & Jusczyk, 1987) carried out research demonstrating that prelingual infants up to about 10 months of age are able to discriminate not only the phonemes of their ambient language, but, indeed, the phonetic contrasts between most of the sounds in most of the world’s languages. Clearly, infants do not have experience in producing phonemes and are not aware of how phonemes are produced by the articulators. The fact that infants are not able to produce the sounds that they can perceive contradicts the basic tenet of motor theory.

Research has shown that prelingual infants without any productive experience are able to discriminate the phonetic contrasts between most of the sounds in most of the world’s languages.
Acoustic Invariance Theory

The acoustic invariance theory assumes that for each distinct phoneme there is a corresponding set of acoustic features. Whenever the phoneme is produced, a core of acoustic properties are always present, regardless of coarticulation and other contextual effects. This core can be thought of in terms of a template against which the listener compares the incoming sound. Stevens and Blumstein (1978) applied this theory to the perception of place of articulation of stops, based on the spectrum of their bursts. The burst for bilabial stops is diffuse and falling. That is, the acoustic energy in the spectrum is mainly concentrated in a few frequency locations across the frequency range, with the amplitude of successive peaks decreasing toward the higher frequencies. Alveolar stops have a diffuse but rising spectrum. Velar stops are characterized by a compact spectrum, with acoustic energy concentrated in one relatively narrow frequency region. See Figure 14.2. Based on a kind of matching process, the listener makes a decision about the similarity in the burst spectrum between the incoming sound and the stored template.

Other theorists have suggested that, rather than templates, the basis of acoustic invariance incorporates features such as voicing and nasality. The use of features allows the more than forty phonemes of American English to be broken down into approximately seven essential minimal acoustic contrasts, thus facilitating the process of speech perception (Ryalls, 1996). According to this theory, listeners abstract the essential features of an incoming sound to make a decision about its identity.

Direct Realism

The direct realist theory was developed in the 1980s. This theory is based on the notion that speech perception does not rely on specialized and unique processes, but is similar to other types of perception, such as visual perception. The rationale for this view is that we perceive objects and events directly, rather than reconstructing or interpreting the object or event from the sensory input to the brain. The example often used to illustrate this concept is that of seeing and recognizing a chair. When we perceive a chair, we do so directly, rather than seeing its different angles and lines and patterns of light and shade. This theory is interactive, because it posits that direct knowledge does not stem only from the object or event itself, but also from the experience and activities of the individual doing the perceiving.

Hawkins (1999c) provides the example of a child looking at a box of cookies. She notes:
Fig. 14.2

Spectral patterns for stop consonants.
If the child knows what it is, has handled something like it before, it is experienced as a cookie box, as a complete object, knowing its shape even if not all of it can be seen from one position, knowing how much force to use to open it, and so on. If you have never seen a cookie box or its like before, you experience it differently: you may guess but do not know what the shape of the invisible part is, you do not know how heavy it is, and so on. The inexperienced person has a different percept from the experienced person, even though exactly the same set of lines and colors strike both their retinas. (p. 235)

According to direct realism, patterns of light energy or sound energy are only useful in terms of their relationship to the physical environment. In visual terms, we directly perceive an object such as a book or a piano, rather than the different sets of lines, colors, and shadows that make up the stimulus. In terms of speech, we directly perceive the acoustic signal, rather than the gestures of the vocal tract, that is, the place and degree of constriction of the vocal tract by the articulators.

TRACE Model

The TRACE model is a connectionist model that tries to account for the integration and parallel processing of multiple sources of information in speech perception. The basic assumption of connectionist networks is that behavior can be modeled. The network of units that characterizes connectionist models includes phonetic features, phonemes, and words, and there are feedback and feedforward links between the units. The links allow perceptual processing to occur within as well as across the different levels of the system. The connections between upper and lower units allow information to flow in both directions, that is, both top-down and bottom-up. However, the connections within a level are inhibitory, so if a feature or word is present, it will be activated, while competing features or words within that level will be suppressed. The word that is perceived is the one with the greatest amount of activation. The TRACE model is an integrative and interactive system that attempts to explain speech perception and word recognition without relying on specialized mechanisms.

Logogen Theory

Logogen theory is an interactive theory that focuses on word recognition, rather than on acoustic-phonetic aspects of speech perception. A logogen is assumed to be some kind of neural processing device associated with each word in a person's vocabulary. All information about the word—its meaning or meanings, its phonetic and orthographic
structure, its syntactic functions, and so on—is contained in the logogen. The logogen monitors speech production to detect any information indicating that its particular word is present in the speech signal. If the information is detected and confirmed by appropriate neural activity, the logogen may become activated and result in word recognition.

Cohort Theory

This is another theory focusing on word recognition. Cohort theory differs from logogen theory because it proposes two stages in spoken word recognition: an autonomous stage and an interactive stage. In the autonomous stage, acoustic-phonetic information at the beginning of a word activates all the words in the person’s memory having the same word initial information. The words that are activated based on word initial information make up the cohort. For example, a word that begins with may would activate all words with that beginning, such as maybe, mayhem, and mayonnaise. The interactive stage involves eliminating inappropriate words in the cohort, based on the listener’s linguistic and cognitive knowledge, as well as the context of the conversation. For instance, if the conversation revolves around lunch, the appropriate word is more likely to be mayonnaise, rather than mayhem.

Fuzzy Logical Model of Perception

This model assumes that there are three operations in phoneme identification. First, features are evaluated to determine their presence in an interval of sound. However, rather than being rated dichotomously as being present or absent, features are assigned continuous “fuzzy” values ranging from 0 to 1, indicating the degree of certainty that the feature appears in the signal. Zero means that the feature is absent, and 1 means that the feature is definitely present. A value of 0.5 indicates that the signal is completely ambiguous in terms of the feature. The clearest and least ambiguous information has the greatest impact on classification decisions.

The second operation is called prototype matching. At this stage the features that were determined to be present at the first stage are compared with prototypes of phonemes stored in the person’s memory. The final operation, pattern classification, determines the best match between candidate phonemes and the input.

This theory rejects the notion of specialized processes in speech perception. The model argues that speech perception is not necessarily categorical, but can be explained by the integration of continuously evaluated features. Thus, continuous information remains available in speech perception, despite the categorical identification and discrimination functions obtained in typical studies.
Native Language Magnet Theory

Native language magnet theory (NLM) has been influential over the past decade in guiding research on speech perception (Frieda, Walley, Flege, & Sloane, 2000). The critical element of this theory is that phonetic categories of a language are organized in terms of prototypes. This organization starts to occur in early infancy. During the first ten months or so of life, infants are able to distinguish between most phonemes of most languages in the world. By age 10 or 11 months, the infant has gone through a reorganization of phonetic categories based on the child's ambient language. By this age the infant can no longer make the distinctions in phonetic contrasts of languages other than its own. NLM theory proposes that these prototypes function as perceptual magnets that assimilate other members of the same phonetic category. Irrelevant perceptual distinctions between members of the same category that are close to the prototype are reduced and can therefore be ignored. Perceptual distinctions between category boundaries, on the other hand, become even more distinct, so the boundaries between phonemes become clearly demarcated. Kuhl and her colleagues, who developed this theory based on their extensive research in infant perception, posited that these perceptual prototypes also serve as speech production targets for infants and young children, thus emphasizing the link between perception and production (Kuhl et al., 1997; Kuhl & Meltzoff, 1996).

In sum, theories of speech perception emphasize different levels of processing of the speech signal. Decoding a spoken message involves the analysis of various-sized components of the signal, including acoustic, phonetic, phonological, lexical, suprasegmental, syntactic, and semantic components. The process of speech perception was nicely summed up by Nygaard and Pisoni (1995), who noted:

Listeners are able to flexibly alter their processing strategies based on the information available to them in the signal and the information available from their store of linguistic knowledge. On one hand, if the physical signal provides rich, unambiguous information about a phonetic contrast, listeners appear to attend primarily to the physical signal when making their phonetic identifications. On the other hand, if the physical signal is noisy, impoverished, or degraded, as synthetic speech often is, listeners may shift their attention to different levels of processing to assist in phonetic categorization. Thus speech perception appears to be a highly adaptive process in which listeners flexibly adjust to the demands of the task and to the properties of the signal. (p. 82)
A theory is a statement about a particular phenomenon, incorporating underlying principles, facts, and assumptions. Theories change based on incoming research and effect practice. A model is a simplification of a system that can be manipulated in a controlled manner. Models can be mechanical, physiological, mathematical, or computer based. Speech production theories attempt to account for serial order, degrees of freedom, context sensitivity, and other issues. Numerous theories of speech production have been proposed, including target models, feedback and feedforward models, dynamic systems models, and connectionist models.

Speech perception theories attempt to account for linearity, segmentation, speaker normalization, basic unit of perception, specialization of speech perception, and other issues. Theories of speech perception can be categorized as active versus passive, bottom-up versus top-down, and autonomous versus interactive. Most theories of perception focus on acoustic-phonetic or phonemic aspects, including motor theory, acoustic invariance theory, direct realism, fuzzy logical models, and connectionist theories; recent theories also attempt to explain word recognition, including logogen theory and cohort theory.

Summary

1. Differentiate between models and theories and give an example of each to illustrate your points.
2. Explain what is meant by the problem of degrees of freedom and give an example from an area of speech production.
3. List the critical elements in Perkell’s articulatory and acoustic target framework of speech production.
4. Describe the reasons why feedback models cannot adequately account for speech production.
5. Compare and contrast dynamic systems models and connectionist models of speech production.

Review Exercises

6. Identify and explain the major issues that theories of speech perception try to account for.
7. Give an example of an active theory, a passive theory, an autonomous theory, and an interactive theory of speech perception.
8. Explain the main aspects of motor theory, and compare them with those of acoustic invariance theory.
9. Describe how connectionist models can be applied to both speech production and speech perception.