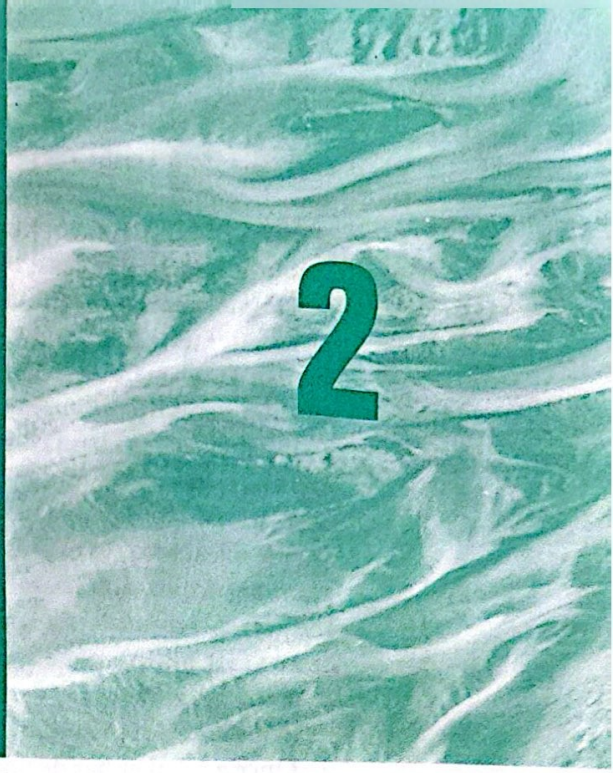
Speech Production and Comprehension



People don't just turn into a Scotsman for no reason at all.

GRAHAM CHAPMAN

In the Monty Python television series, one episode presented a sketch in which the entire population of England, women and children included, turned into Scots. ("The over-crowding was pitiful. Three men to a caber."[1]) In the skit, space aliens from the planet Skyron did the dirty work. Strangely, in the real world, there is a neurological syndrome that can turn you into a Scot. Well, you don't literally turn into a Scot, but you do end up talking like a Scot. That is, you can acquire an accent that sounds Scottish as the result of experiencing brain damage. One such adult patient spoke English with a London accent (southern English) before her injury, but after suffering a stroke "three native Scottish speakers reported that her post-stroke accent did sound Scottish to them" (Dankovičová et al., 2001, p. 213). Other kinds of accent changes have also been reported: English to Spanish or Norwegian; Belgian Dutch to French and Moroccan; Norwegian to German (caused, ironically, by injury from German shrapnel; Moen, 2000). This neurological condition is called, appropriately, *foreign accent syndrome* (FAS). It is rare, but dozens of cases have been documented starting in the early 1900s. Why do people acquire foreign-sounding accents? It has to do with the way brain injury changes the mental and motor

Introduction to Psycholinguistics: Understanding Language Science, Second Edition.
Matthew J. Traxler.

Companion Website: www.wiley.com/go/traxler/psycholinguistics2e

processes that are involved in speech production (talking). One of the chief goals of this chapter is to describe these planning and output processes. (The other is to describe how speech sounds are perceived.) Let's return to FAS and its causes once you have a bit of speech production theory under your belt. Then we'll tackle how a brain injury can turn you into a Scot, or at least make you sound like one.

Human communication occurs most frequently via speech, so understanding speech production (talking) and comprehension lays the foundation for an understanding of human language abilities. Contemporary theories of speech production take as their starting point the moment in time where the speaker has an idea she wishes to convey. Thus, they focus primarily on how speakers convert ideas into a form that can be expressed in speech, and take for granted that speakers have ideas to convey. (A separate branch of cognitive psychology focuses on how people come up with ideas, and how people select ideas to express; see e.g. Goldstein, 2007.) While the basic process of speech planning seems simple—you have an idea, you pick words to express the idea, you say the words—research on speech planning and production shows that the mental processes that intervene between thinking of an idea and producing the physical movements that create speech are quite complex. One of the main goals of this chapter is to describe some of the hidden complexity of the speech production system.

Once a speaker has decided what to say and how to say it, she produces a set of behaviors that change her immediate physical environment, chiefly by creating a pattern of sound waves—an *acoustic signal*—that is available to listeners. The listener's chief task is to somehow analyze the acoustic signal so that the speaker's intended meaning can be recovered. This, too, seems like a simple task. The listener recognizes the words that the speaker produced, matches those words to concepts, and, hey presto! understands what the speaker meant to say. However, acoustic analysis of speech shows that speakers produce wickedly complex sound waves and a great deal of mental work needs to be done after sound waves hit the ear drum before the listener can recover the speaker's intended meaning. This chapter will explain why analyzing the physical properties of speech is tricky and review current theories that explain how listeners overcome obstacles created by the peculiar acoustic properties of speech.

Speech Production

To explain how people produce speech, a theory must describe the mental representations that support the translation between ideas, which are mentally represented in a nonlanguage form, and the mental plans that cause muscles to move.[2] After all, speech requires physical action—a process called *articulation*. In fact, speech is more complicated than many other physical actions that we perform because it requires exquisitely tight control over more than 100 muscles moving simultaneously (Meister et al., 2007). Theories of speech production try to answer questions like: Once you have an idea that you wish to convey, what steps must you take to retrieve the linguistic representations you need to express your idea? How do you organize those representations? How do you translate those representations into a form that the motor system can use to generate the actual, physical gestures that create speech sounds?

Speech production requires three kinds of mental operations (Griffin and Ferreira, 2006). First, you have to think of something to say—*conceptualization*. Once you have something to say, you must figure out a good way to express that idea given the tools that your language provides—*formulation*. Finally, you need to actually move your muscles to make a sound wave that a listener can perceive—*articulation*.

Willem Levelt's production theory, which has been adapted as a mathematical model called WEAVER++ (Levelt et al., 1999; Jescheniak and Levelt, 1994; Levelt, 1989; Roelofs et al., 2007), explains the cognitive processes involved in speech production. An overview of the WEAVER++ production system appears in Figure 2.1. Take a moment to have a look at it, but don't panic! Let's break it down, step by step. The first important thing to realize about speech production is that activating an idea does *not* automatically lead to activation of the speech sounds you need to express the idea. That is, thinking of the concept "cat" does not automatically lead to activation of the speech sounds /k/, /a/, and /t/. One of the goals of WEAVER++ is to describe the intermediate mental steps between activating an idea and activating the sounds that you need to express the idea. Speech production involves a sequence of mental processes. Each mental process accomplishes a subgoal, and the output of one mental process provides the information needed for the next mental process.

Each box in the model in Figure 2.1 indicates a mental process. For example, "conceptual preparation in terms of lexical concepts" boils down to this: choose the idea(s) that you want to express, but make sure that your idea lines up with words that you have in your language. The output of this process, a *lexical concept*, is an idea for which your language has a label (Levelt et al., 1999).

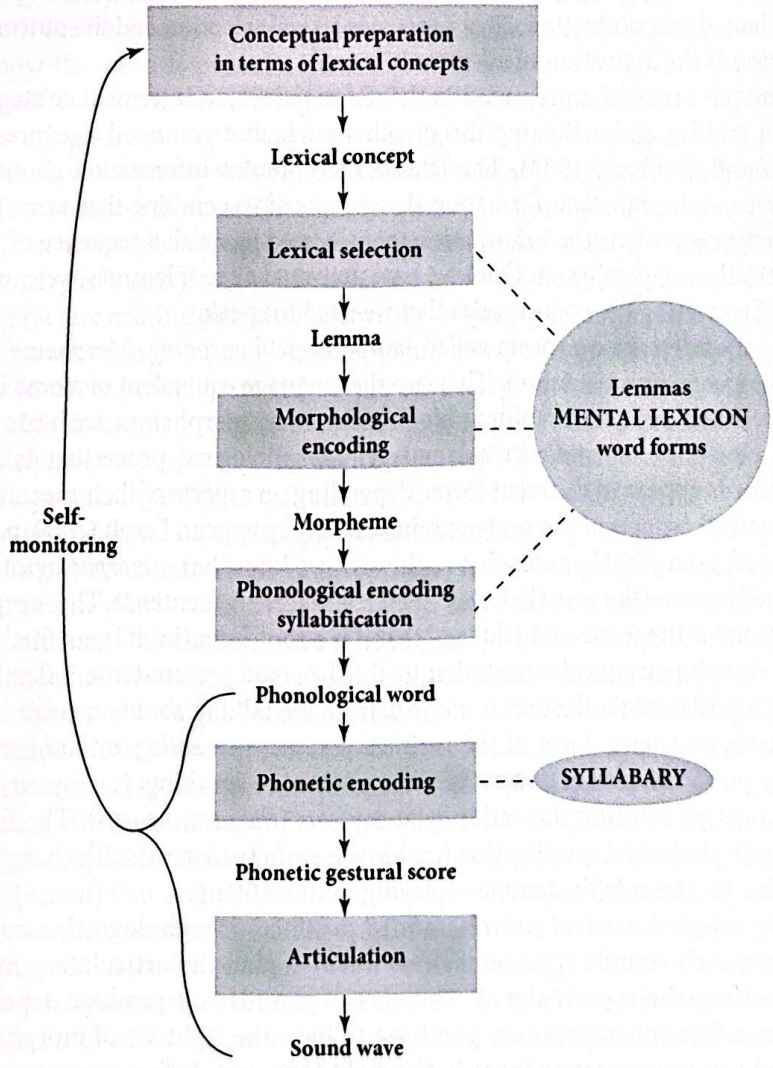


Figure 2.1 A schematic of Levelt and colleagues' speech production model. *Source:* Levelt et al. (1999), Cambridge University Press, p. 3

You may have had the experience where you have an idea, but you have trouble putting your idea into words. That could happen because your (nonlinguistic) idea does not neatly line up with any of the ideas for which your language has a pre-existing word. In that case, you need to come up with some combination of lexical concepts to express the idea for which your language does not have a single term.

Here is an example: English has a word that expresses the concept *female horse*. That word is *mare*. If you want to express the concept *female horse*, you can activate the lexical concept *mare*. But English does not have a single word that can express the concept *female elephant*. To express that idea, you need to select and combine two different lexical concepts (female, elephant). (For a PG-13 rated example from German, see the following note.[3]) Because ideas do not always line up neatly with individual words, we need a stage of processing that takes our (nonlinguistic) ideas and finds the lexical (linguistic) forms that we can use to express those ideas. The *lexicalization process* therefore serves as the interface between nonlanguage thought processes and the linguistic systems that produce verbal expressions that convey those thoughts.

When your language does have a word for the idea you wish to express, the activation of a *lexical concept*, an idea that the language can express in a word, will lead to *lexical selection*. Oftentimes, a language will have a number of different words that are close in meaning to the idea that you wish to express. In that case, a number of different representations in memory will become activated and you must choose which representation will be selected for production. That process is lexical selection and the outcome of lexical selection is the activation of a *lemma*.

A *lemma* is a mental representation that takes part in an intermediate stage between activating an idea and activating the speech sounds that you need to express the idea (Kempen and Huijbers, 1983). The lemma incorporates information about what the word means and *syntactic* information that you need to combine that word with other words. Lemmas provide the information that we need to speak a sequence of words in a grammatically acceptable way. Once we have activated a set of lemmas, we can begin the process of activating the sound codes that we need to speak.

First, we undertake a process called *morphological encoding*. Morphemes are basic units of language representation. They are the language equivalent of atoms in physics. Atoms are basic units and building blocks of matter; morphemes are basic units and building blocks of meaning in a language. Morphological processing is important because words appear in different forms depending on aspects of their meanings as well as grammatical aspects of the sentences in which they appear. Levelt (1989, p. 182) provides a good example. He notes that each word we know has a *morphological specification* that tells us how the word behaves when it is placed in a sentence. The morphological specification for the word *eat* includes, "that it is a root form (i.e. it is not further analyzable into constituent morphemes), that its third-person present tense inflection is *eats*, and that its past-tense inflection is *ate*." So, if we are talking about an event in the past, we will use a past-tense form of the verb *eat* (i.e. *ate, was eating*, or *had eaten*). If two people are performing an action at the moment you are speaking, you say *eat*, but if only one person is performing the action, you say *eats* (*they eat; he eats*). The form of the word, its morphological specification (*ate* vs. *was eating* vs. *eat* vs. *eats*), changes depending on what precise role the lemma is playing in the sentence.

Having selected a set of morphemes to produce, morphological encoding activates the speech sounds (phonemes) we need to plan the articulatory movements that will create the speech signal. The speech sounds you produce depend on the morphemes that you activate, so you have to have the right set of morphemes activated, and you must arrange them in the right sequence, before you can activate the speech sounds.

To sketch the production process so far: Concepts point you to lemmas. Lemmas point you to the morphological information you need to combine lemmas into larger phrases. Morphological encoding points you to the speech sounds (phonemes) you need to express specific sets of lemmas in specific forms.

Having the right set of morphemes activated in the right sequence gets us a step closer to moving our speech muscles, but it does not get us all the way home. Once you have the morphemes slotted into the right positions you can activate the individual speech sounds (*phonemes*), but speaking involves more than just saying a sequence of phonemes. In contemporary speech production theory, the lemma represents abstract information about a word, such as its grammatical class, its meaning, and the way it may combine with other lemmas; while what we normally think of as a word (the collection of sounds) is referred to as a *lexeme*. To produce the lexeme, we need to activate a set of phonemes (speech sounds) and organize them into groups for production.

Evidence for the lexeme as a psychologically real level of representation comes from studies involving the production of *homophones* (Jescheniak and Levelt, 1994; Jescheniak et al., 2003; see Lohman, 2018 for a different take). A homophone is a word that has more than one meaning. A lexeme like /but/ has two spellings (*butt* and *but*) and more than one distinct meaning. In English, the *but* version of the lexeme occurs very frequently, whereas the *butt* version (as in *I was often the butt of her sick practical jokes*) occurs very rarely. According to current production models (Dell, 1986; Levelt et al., 1999), both the *but* and *butt* versions activate the same lexeme, because the lexeme represents how the word is pronounced, and both versions of /but/ are pronounced the same way. If so, lexemes should experience the *frequency inheritance* effect. That is, if a word has a high-frequency twin (*but* is the high-frequency twin of *butt*), you should produce the low-frequency version (*butt*) about as fast as you produce the high-frequency version (*but*) because the overall lexeme frequency is high. By contrast, if a word has two versions, but both are low frequency, then it should take a relatively long time to respond to the word (*flecks* and *flex*, for example, are both low frequency forms). Experiments involving picture naming provide evidence for frequency inheritance, as do experiments involving translation from one language to another. In both cases, low-frequency versions of words are produced quickly if they have a higher-frequency twin (i.e. if the frequency of their lexeme is high). Thus, the time it takes you to produce a word is not based solely on how frequently that word's meaning is used. Instead, it depends partly on how often you use a particular collection of sounds (a lexeme).

When we speak, we do not simply produce a string of phonemes. Those phonemes need to be organized into larger units because, when we speak, we speak in syllables. Producing each syllable requires a coordinated set of actions, and we need to plan each set of coordinated actions. Before we start to speak, we need to figure out which speech sounds (phonemes) we need, but we also need to figure out how to map the set of activated phonemes onto a set of syllables. This latter process is called *syllabification*.[4] Syllabification involves two subcomponent processes: activating a *metrical structure* and inserting individual speech sounds (phonemes) into positions in the metrical structure. The metrical structure consists of a set of syllable-sized units. In addition to specifying the number of syllables needed, the metrical structure indicates the relative emphasis or loudness (*accent*) that each syllable should receive. The word *banana*, for example, has an accent on the second syllable. The word *Panama* has accent on the first syllable. So, the metrical structure for *banana* would be represented as "σ σ′ σ," and the metrical structure for *Panama* would be represented as "σ′ σ σ." Each σ symbol stands for a syllable, and the ′ mark indicates which syllable in the string should be accented. Once the metrical structure has been laid down, individual phonemes can be inserted into positions within each syllable.

Evidence that syllabification is a real mental process that intervenes between morphological processing and articulation can be found in studies of the way people speak. For example, consider the word *escorting* (Levelt et al., 1999, p. 5). It has two morphemes, the root *escort* and the suffix *-ing*. When people actually speak the word *escorting*, they usually produce it in three segments, which sound something like, "ess," "core," and "ting" (*ess-core-ting*, rather than *ess-cort-ing*). That means that the syllabification processes in production have placed the /t/ phoneme together with the *-ing* morpheme, rather than with the root morpheme *escort*. So, we do not simply activate morphemes, activate the phonemes that go with each morpheme, and produce them in sequence. Instead, after the morphemes are activated, we calculate the best way to organize the sequence of phonemes into syllables, and the syllables serve as the basis of production. That is true even when the processes responsible for calculating syllables lump together phonemes from different words. If you were going to speak a sentence that included the phrase *He will escort us*, you would most likely take the /t/ phoneme from the word *escort* and stick it into a syllable along with the word *us*. So, you would actually say something that sounds like "es-core-tuss" (rather than "es-cort-us"). (Can you think of a situation where someone might naturally produce the "escort-us" version?)

To sum up, while we need morphemes and words to plan what to say, speech does not simply involve activating the speech sounds in individual words. Instead, the speech-planning system activates a set of morphemes or words, and then it figures out the best way to organize those morphemes and words into a set of syllables. Sometimes the syllables respect morpheme and word boundaries, but oftentimes they do not. In the words of Levelt and Wheeldon (1994, p. 243), "Speakers do not concatenate citation forms of words, but create rhythmic, pronounceable metrical structures that largely ignore lexical word boundaries."[5]

The output of the syllabification process is a set of *phonological words*. A phonological word is a set of syllables that is produced as a single unit. So, while "escort" and "us" are two different lemmas and two different words, when they are actually spoken, they come out as a single phonological word, /ess-core-tuss/. According to the WEAVER++ model, you can begin to speak as soon as you have activated all of the syllables in a given phonological word.

Further evidence that we speak in phonological words, rather than in morphemes and (lexical) words, comes from colloquial (informal) speech and dialects. If you lived in America in the 1990s, you probably found the comedian Jeff Foxworthy endlessly entertaining. One of Foxworthy's comedy bits involved an utterance that is pronounced *wichadidja. Wichadidja* is a phonological word that is composed of four lexical words, *with, you, did,* and *you,* as in *You didn't bring your varmint gun wichadidja?* If people spoke in lexical words ("dictionary" words or *citation forms*), expressions like *wichadidja* would not exist.

While you may plan each utterance by activating a number of lemmas and morphemes simultaneously, you plan the actual speech movements (*articulation*) one phonological word at a time; and you plan the movements you need to produce each phonological word one syllable at a time, in a "left-to-right" fashion. That is, you activate the phonemes for the syllable that you will need first (e.g. "ess" in *escort us*) before you activate phonemes for syllables that you will need later on.

Evidence for left-to-right activation of phonemes in phonological words comes from studies involving *phoneme monitoring* in picture-naming experiments. In these experiments, people look at a picture and try to say a word that describes the picture as quickly as possible. If you were looking at a picture of a floppy, furry animal, you would say *rabbit* as quickly as possible. In a secondary task, you would be given a target phoneme and would be asked to press a button as quickly as possible if the target phoneme occurred in

the picture's name. So, if you were asked to monitor the target phonemes /r/ or /b/, you should press the button when you see the picture of the floppy-eared animal. If you were asked to monitor the target phoneme /k/, you should refrain from responding. People can do this phoneme-monitoring task very accurately, and they respond faster if the target phoneme comes from the beginning of the word than if it comes from the middle or the end of the word (Wheeldon and Levelt, 1995).

To summarize how the WEAVER++ model works, production begins with a set of ideas that the speaker wishes to express. In the next step, those ideas are tied to lexical concepts, because the language may have specific words for some of the ideas but may require combinations of words to express other ideas; or because the language provides more than one word for a given idea. After a set of lexical concepts has been activated, lemmas that correspond to those lexical concepts become activated. Activating lemmas provides information about the morphological properties of words, including information about how words can be combined. After a set of morphemes has been activated and organized into a sequence, speech sounds (phonemes) can be activated and placed in a sequence. Phonological encoding involves the activation of a metrical structure and syllabification (organizing a set of phonemes into syllable-sized groups, whether the specific phonemes come from the same morpheme and word, or not). The outcome of this process is a set of phonological words, each of which consists of a sequence of syllable-sized frames. During phonetic encoding, the speech production system consults a set of stored representations of specific syllables. The system activates the appropriate syllable representations and places them in the appropriate positions in the frame. This representation is used by the motor system to create a *phonetic gestural score*, which is the representation used by the motor system to plan the actual muscle movements (articulation) that will create sounds that the listener will perceive as speech.

Evidence supporting models of speech production like the WEAVER++ model can be found in three kinds of studies: speech errors, tip-of-the-tongue experiences, and reaction time studies involving picture naming, which often use a version of the picture–word interference task. Mathematical modeling using computer programs to simulate what happens in speech errors, tip-of-the-tongue experiences, and picture-naming experiments is also used to test ideas about how information flows through the speech production system.

Speech errors

The analysis of speech errors has a long and glorious history in psychology in general and psycholinguistics in particular. Sigmund Freud viewed speech errors as a window into the unconscious mind. He believed that speech errors reveal our true inner thoughts—thoughts that we suppress in order to be polite. Modern psycholinguistic theories view speech errors as reflecting breakdowns in various components of the speech production process (Dell, 1986; El-Zawawy, 2021; Garrett, 1975, 1980; Levelt, 1983; Levelt et al., 1999; Postma, 2000). We can use speech errors to inform our understanding of speech production processes because speech errors are *not* random. In particular, *slips of the tongue* occur in systematic patterns, and those patterns can be related back to aspects of the speech production process. As Dell (1986, p. 286) notes, "Slips of the tongue can be seen as products of the productivity of language. A slip is an unintended novelty. Word errors create syntactic novelties; morphemic errors create novel words; and sound errors create novel, but phonologically legal, combinations of sounds."

Each of these different kinds of errors provides information about how different components of the production system work. For instance, people sometimes substitute one word for another when they are speaking. If people are placed under time pressure, and they are asked to name a picture of a cat, they will sometimes say *rat* or *dog*. This type of *semantic substitution* error likely reflects the *conceptual preparation* or *lexical selection* component of the speech production process. Semantic substitutions could reflect conceptual preparation if an individual mistakenly focuses on the wrong (nonlinguistic) concept. Alternatively, semantic substitutions can reflect the way (nonlinguistic) concepts are related to one another, and how the activation of (nonlinguistic) concepts is tied to activation of lemmas (Dell et al., 1997; Levelt et al., 1999; S. Nooteboom, 1973). According to WEAVER++, concepts are stored in long-term memory in networks or collections of concepts. Within these networks, concepts that have similar meanings are connected to one another. As a result, when you think of the concept "cat," activation will spread to, or spill over onto, closely related concepts, such as "rat" and "dog." To select the correct lemma, you need to ignore related concepts and focus on the *target* concept (e.g. "cat").

Semantic substitutions can also reflect lemma-selection errors (rather than concept selection errors) because activating a (nonlinguistic) concept will feed activation to lemmas that are associated with that concept. So, activating the "cat" concept will activate associated concepts ("rat" and "dog"), and those associated concepts will activate associated lemmas. When it comes time for speakers to choose a lemma for further processing, they will choose the target lemma (*cat*) most of the time, but every once in a while, they may be fooled because alternative lemmas for *rat* and *dog* will also be activated. These kinds of behaviors are classified as speech errors, or "slips of the tongue," because people clearly did not use the commonly accepted term for the picture, even though they do know the appropriate term (as evidenced by frequent self-corrections in this kind of study).

Other types of speech errors may reflect breakdowns in other components of the speech production system. Sometimes, the correct set of phonemes is produced, but some phonemes appear in the wrong positions in the utterance. These *sound exchanges* are thought to reflect a stage of processing after a set of lemmas and morphemes has been activated, but before an articulatory plan (plan to move the speech muscles) has been compiled. In a sound exchange, you might be hoping to say *big feet*, but instead you say *fig beet*. These kinds of errors can be elicited in the lab by putting experimental subjects under time pressure. Researchers set up the experiment so that subjects get used to producing a specific pattern of sounds, and then they switch the pattern (Baars and Motley, 1974). Subjects might be asked to say *bid meek*, *bud muck*, and *big men*, all of which have a /b/ in the first position in the first syllable and an /m/ in the first position in the second syllable. Then, right after that, subjects might have to say *mad back*. About 10% of the time subjects make an error and say *bad mack* or *bad back*.[6] (Try this with your friends!)

Sound exchange errors almost always occur when sounds are exchanged between words in the same phrase, and the vast majority involve movement of only a single phoneme from each word (Nooteboom, 1969). You are more likely to say *That guy has fig beet*, where the two target words are in the same noun phrase, than you are to say *These beet are really fig* (target: *These feet are really big*), where one word appears in a subject–noun phrase and the other appears as part of the following verb phrase. In addition, sound exchanges almost always respect the *positional constraint*. That is, when sounds trade places, they almost always come from the same part of the word, usually the first phoneme. You would almost never say *tig feek* by mistake.

In Dell's (1986) production model, the positional constraint reflects the way individual phonemes are activated and inserted into *frames* (syllable-length mental representations, possibly, as in Levelt's model). According to the model, a number of frames can be activated simultaneously, so when you are planning for *big feet*, you activate two syllable frames, and you activate the phonemes you need to fill in those frames. Each of those phonemes is marked with an *order tag*, which tells the production system which phoneme comes first, which comes second, and so on. Because two syllable frames are activated simultaneously, and two phonemes that have "first" order tags are also activated simultaneously, sometimes the production system will confuse the two, and select the wrong phoneme for each of the two available "first" phoneme slots. Normally, the activation levels of the two "first" phonemes will differ at different points in time (generally, the /b/ phoneme will have more activation early in the planning process and the /f/ phoneme will have more activation later), and so mistakes will be relatively rare. But sometimes, if the activation levels of the two "first" phonemes are close enough, they will get reversed. Most errors respect the positional constraint because the production system will not jam a phoneme with a "first" positional tag into the slot labeled "last," and vice versa. Further, most sound exchanges involve two phonemes from the same phrase, indicating that the articulatory plan is built for no more than one phrase at a time.

For phonemes, the production process generates a set of labeled slots and activates units with tags that match available slots. *Word exchange* errors happen for similar reasons. A word exchange happens when a word that should have appeared in one position is produced in a different position. You might want to say *My girlfriend plays the piano*, but say *My piano plays the girlfriend* by mistake. In that case, *girlfriend* and *piano* participated in a word exchange. The majority of word exchange errors respect the *category* constraint (Dell, 1986; Postma, 2000). *Category* refers to parts of speech, such as *noun, verb, adjective*, and so on. Most of the time, when two words participate in an exchange, they come from the same category (hence, *category* constraint). According to frame-and-slot models (e.g. Garrett, 1975; Mackay, 1972), speech involves a degree of advance planning. Rather than planning a word at a time, we can lay out the frame for an entire clause or sentence as we are looking for a particular set of words and the precise forms we need to produce those words. This frame consists of a set of slots (places for individual words to go), and each slot is labeled for the kind of word that has to appear there (noun, verb, adjective, and so on). As with sound exchange errors, word exchanges happen when more than one candidate is activated simultaneously, more than one candidate has the same tag (e.g. noun), and the production system assigns the wrong candidate to an open slot. Because the slots are labeled, however, the production system does not get the categories wrong. Verbs do not appear in noun slots; nouns do not appear in preposition slots; prepositions do not appear in verb slots.

Tip-of-the-tongue experiences

Overt speech errors provide us with insights into the way the speech production system operates, but they are not the only game in town. *Tip-of-the-tongue experiences* (TOTs for short) also yield evidence about speech production. A TOT happens when you are trying to retrieve a word, you have a strong subjective impression that you know the word, but you are temporarily unable to consciously recall and pronounce the word. According to contemporary production theories (e.g. Dell, 1986; Levelt et al., 1999; Roelofs et al., 2007), TOT states occur when you have accessed the correct lemma, but you have been unable to fully activate the phonological information that goes along with

that lemma. TOT experiences are taken as evidence for the distinction between semantic (meaning) activation and phonological (sound) activation that plays a role in all current accounts of speech production.[7] But why not simply view TOT experiences as evidence for the failure of meaning-related semantic processes? Why view TOT experiences as reflecting the temporary failure of phonological processes? A variety of results point to phonological encoding, rather than semantic processes, as being the culprit (Brown, 1991; Brown and McNeill, 1966; Rubin, 1975; Schwartz and Pournaghdali, 2020).

But first, how do language scientists study the TOT? There are a number of ways to do this (Brown, 1991). Sometimes, researchers ask people to carry around a diary and record all of their TOT experiences in a given time period (a few weeks or months, usually). Those kinds of studies indicate that people experience TOTs about once or twice a week, with the frequency of TOT experiences increasing as people get older. TOT experiences can be triggered by providing people with the definitions of rare, but familiar, words. For example, can you think of the words that go with the following definitions?

1. The first name of the character "Scrooge" in Charles Dickens' *A Christmas Carol.*
2. A small boat of the Far East, propelled by a single oar over the stern and provided with a roofing of mats.
3. A secretion of the sperm whale intestine used to make perfume; an ingredient in the perfume sent by Dr. Hannibal Lecter to FBI agent Clarice Starling in the movie *Hannibal.*
4. A one-word name for a person who collects stamps. (Spoiler alert: The answers appear in note 8.[8])

For fun, see how many of your friends can come up with the appropriate terms, and find out whether any of them experience a TOT. The interesting question is not whether they know the word. The interesting question is: If they know the word, are they able to access the appropriate sounds straight away, or do they experience a TOT? This method of measuring TOT experiences is called *prospecting.* If you test a large enough group of people, many of them will report having a TOT experience when they try to think of the words that go with the preceding definitions. By asking about detailed aspects of the experience, researchers can figure out how much information people have about the target word (Do they really know it? Can they think of any of the sounds in the word? How many syllables does it have?), they can determine whether the retrieval failure is temporary, and they can pinpoint the source of the problem.

TOT experiences do not reflect failures of semantic activation or lemma retrieval because people who are experiencing a TOT are able to predict accurately how likely it is that they will be able to come up with the correct word in the near future (Nelson, 1984). If the correct meaning were *not* activated much of the time during the TOT experience, then people would not be able to predict their own future successful retrieval of the target word. People can activate the correct lemma during a TOT experience, but do they activate any phonological (sound) information at all? The evidence suggests that they do. People who are experiencing a TOT state are likely to report the correct number of syllables in the (temporarily inaccessible) word, they are likely to correctly report the first phoneme in the word, and when asked to produce similar words to the target, they mostly come up with words that sound like the target word (Lovelace, 1987). People experiencing a TOT are more likely to accurately report the first and last letters in the target word, and less likely to accurately report letters from the middle, suggesting that substantial information about the overall form of the word as well as its component sounds are activated during the TOT experience.

The likelihood of a TOT experience may reflect the strength of the relationship between the conceptual, lemma, and phonological levels of representation. Words that we encounter infrequently are more likely to produce TOT experiences than words that we encounter more frequently, so we will have associated sound and meaning less often for words that produce TOTs. About 40% of laboratory-induced TOTs are resolved within a few seconds or a few minutes of the onset of the TOT, which further supports the idea that TOTs reflect temporary failure of phonological activation, rather than some other aspect of the production process.

Picture naming and picture–word interference studies

Picture-naming studies provide evidence about speech production because they offer a window into very basic aspects of speech: How do you find the word you need to express a concept, and how do you activate the sounds that make up the word? Early studies in picture recognition and picture naming showed that people activate different concepts at about the same speed, but concepts that were used less frequently in speech or writing led to longer response times (Oldfield and Wingfield, 1965; Wingfield, 1968). In these experiments, participants looked at pictures and performed one of two tasks. In one task, they simply stated whether they had seen the object before (recognition test). In the other, they named the object in the picture. There were very small differences in the amount of time it took people to recognize less familiar versus more familiar objects. There were much larger differences in the amount of time it took people to name less familiar versus more familiar objects. Thus, the amount of time it takes people to plan a spoken response appears to be affected more by how often they produce the collection of sounds that labels a concept, and less by how often they think about a specific concept.

Additional research addresses how concepts are organized and how they are related to one another in long-term memory. The way concepts are organized can affect how easy it is to retrieve the specific concept you need in a particular situation. Do you activate just the concept you need right when you need it? Or do you need to sift through a set of activated concepts before you can select the one you need?

Picture-naming research suggests that concepts do compete with one another for selection during the process of speech production (Bürki et al., 2020; Dell et al., 1997; Garrett, 1975; Griffin and Ferreira, 2006; but see Gauvin et al., 2018).[9] In experiments that use the *picture–word interference* task, participants look at pictures that have words printed on top of them (see Figure 2.2). Experimenters can manipulate the relationship between the picture and the word. Sometimes, the word refers to the object in the picture—the *identity* condition. The identity condition leads to faster naming responses, most likely because both the word and the picture stimulus point toward the same

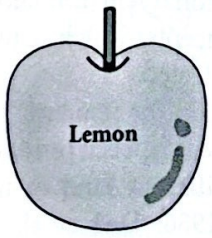


Figure 2.2 An example stimulus from a picture–word interference experiment. *Source:* Arieh and Algom (2002), American Psychological Association, p. 222

lexeme. So, the target sounds are activated by two different sources. Sometimes, the word refers to an object related to the object in the picture (the *semantic* condition). Other times, the word refers to an object whose name is similar to the object in the picture (the *phonological* condition). For instance, if the picture were of a house, the word might be *mouse* in the phonological condition. The question these kinds of experiments address is: How will presentation of a potentially competing stimulus affect access to and production of the picture name? In general, the *semantic* condition produces interference effects: People are slower to name pictures when the overlapping word has a meaning similar to the object in the picture (Cutting and Ferreira, 1999).[10] However, when the overlapping word has a similar-sounding name to the picture, people name the object in the picture faster. Because the semantic (meaning) relationship between the word and the picture produces one pattern (it slows people down), while the phonological (sound) relationship between the word and the picture produces another pattern (it speeds people up), picture–word interference experiments reinforce the distinction between conceptual/semantic activation processes and phonological encoding processes in speech production. These two aspects of speech production appear to be controlled by semi-independent processors (semi-independent because the sounds you activate depend on the concepts you activate).[11]

The spreading activation model of speech production

Production models like WEAVER++ describe a set of mental representations that is involved in speaking—concepts, lemmas, lexemes, syllabified metrical representations, gestural scores—and many researchers in production agree that those representations, or a similar collection, underlie spoken language. However, WEAVER++ also assumes a specific kind of information flow as people go from activated concepts to activated lemmas to activated sets of syllabified phonemes. In particular, the model assumes a strict *feed-forward* pattern of activation and no mutually inhibitory links between representations at a given level of representation (*mutual inhibition* means that as one mental representation gains activation it sends signals that reduce the activation of other representations). According to WEAVER++, production begins with a set of activated concepts, which leads to activation of a set of lemmas. Before phonological (sound) information can be activated, one of those lemmas must be selected for further processing. No matter how many lemmas are activated, and no matter how much activation any alternative lemmas enjoy, the phonological encoding system only works on the one lemma that gets selected. WEAVER++ falls within the *feed-forward* class of processing models because information only moves in one direction in the system, from concepts to lemmas to lexemes to phonemes. But the system does not allow activation to feed back in the opposite direction. Lexemes may not feed back and influence the activation of lemmas, and lemmas may not feed back and influence the activation of concepts. According to this account, the occasional semantic substitution error happens because a target concept activates related concepts, which activate their associated lemmas, so sometimes the wrong lemma gets selected.

But this is not the only explanation for semantic substitution errors. Accounts like Gary Dell's *spreading activation* model of speech production differ from the WEAVER++ model primarily in proposing a different kind of information flow throughout the speech production system (Dell, 1986; Dell et al., 1997; Nozari and Pinet, 2020). According to Dell, information is allowed to flow both in a feed-forward direction (as in WEAVER++) and in a feedback direction (opposite to WEAVER++). However, unlike WEAVER++, in the spreading activation account, activation is allowed to *cascade*

through the system. In WEAVER++, selection has to take place at one level of the system before activation starts to build up at the next. No phonemes get activated until lemma selection is complete. In the spreading activation account, by contrast, as soon as activity begins at one level, activation spreads to the next level. Thus, selection does not necessarily occur at one level before activity is seen at the next. The spreading activation model also assumes feedback between levels of representation. So, if the lemma for *cat* gains some activation, it will feed back to the concept layer and reinforce activation of the "cat" conceptual representation. If the phonological information associated with the pronunciation /kat/ begins to be activated, it will feed back and reinforce the activation of the "cat" lemma.

Proposing that information flows both forward and backward through the language production system in a cascade helps to explain a number of things that happen when people speak. For example, feedback connections from the phonological (sound) processors to the lemma (abstract word form) level help explain the *lexical bias* effect. The lexical bias effect refers to the fact that, when people produce sound exchange errors, more often than not, the thing that they actually produce is a real word. If speech errors simply reflected random errors in the phonological units, there is no reason why sound exchange errors would result in an actual word being produced. If errors were purely based on hiccups in phonological output processes, then you would be just as likely to get an error such as *bnip* or *tlip* or just random gibberish as any other kind of error. However, real speech errors almost never violate *phonotactic constraints* (rules about how phonemes can be combined) and they create real words more often than they should purely by chance (an error such as *slip* in place of the target *blip* is far more likely than *tlip* or *blep*). Likewise, a speaker is more likely to make an error by reversing the beginnings of *big feet* than *big horse*. In the former case, *fig* and *beet* are both words. In the latter case, neither *hig* nor *borse* is a word.

Interactive spreading activation accounts (e.g. Dell, 1986; Dell et al., 1997) explain the lexical bias effect by appealing to feedforward and feedback links between lemmas and phonological output mechanisms. Figure 2.3 shows how these two sub-processors might

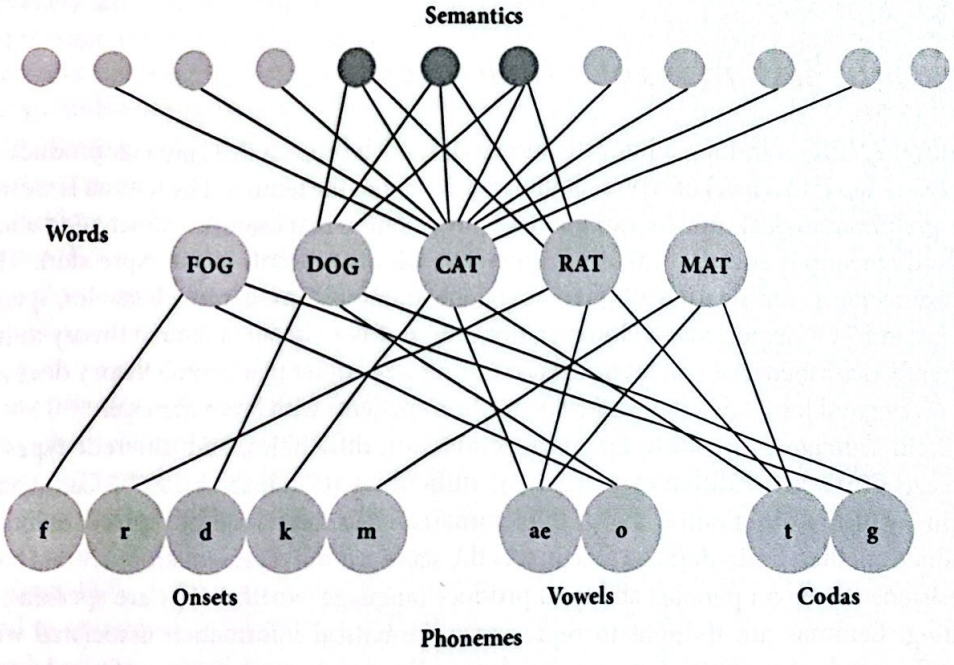


Figure 2.3 Representation of an interactive, spreading activation model for speech production. *Source:* Dell et al. (1997), American Psychological Association, p. 805

be connected (Dell et al., 1997, p. 805). In this kind of model, phonological activation begins as soon as lemmas ("words" in this diagram) begin to be activated, but before a final candidate has been chosen. As individual phonemes begin to be activated, they send feedback to the lemmas that they are connected to, increasing the activation of the lemmas. Because real words have representations at the lemma level, and nonwords do not, it is likely that mistaken activation among the phonemes will reinforce the activation of a word that sounds like the intended target word. It is less likely that a nonword error will result, because any set of phonemes that would lead to a nonword being produced will not enjoy any reinforcing activation from the lemma level. Thus, on average, sets of phonemes that produce nonwords will be less activated than sets of phonemes that produce real words.

Interactive activation accounts also help to explain *mixed errors*. In a mixed error, the word that a person produces by mistake is related in both meaning and sound to the intended word. So, a person is more likely to say *lobster* by mistake when they mean to say *oyster* than they are to say *octopus*, because *lobster* both sounds like and has a similar meaning to the target. Further, these types of mixed errors occur more frequently than they should if errors were purely random (Baars et al., 1975; Dell and Reich, 1981; Levelt et al., 1991). Spreading activation accounts of speech production view the relatively high likelihood of mixed errors as resulting from cascading activation and feedback processes between levels. Thinking about *oysters* will activate semantically related items, such as *lobsters* and *octopi*, which will lead to activation of the *oyster* lemma, but also *lobster* and *octopus* lemmas. Activating the *oyster*, *lobster*, and *octopus* lemmas will cause feedforward activation of the sounds that make up those words. Because the *ster* set of phonemes is being driven by both the target and an active competitor lemma, those sounds are highly likely to be selected for eventual output. Sounds that occur only in the target, or only in a competitor, are less likely to be selected. If there were *no* cascading activation, then either *octopus* or *lobster* would have about an equal chance of outcompeting the target (*oyster*) at the conceptual and lemma layers, and there is no reason why mixed errors should be more common than any other kind of error. Thus, Dell and colleagues interpret the relatively high frequency of mixed errors as being evidence for cascading activation.[12]

Potential limitations of lemma theory

Both WEAVER++ and spreading activation style models propose that language production processes tap into a level of representation equivalent to the lemma. The lemma is viewed as a pre-phonological (pre-sound) mental representation that captures information about a word's meaning and the way it can interact with other words in an expression. This theory accounts for a variety of phenomena, including picture-naming behavior, speech errors, and TOT experiences. However, not everyone is a big fan of lemma theory as it is currently described. For instance, Alfonso Caramazza argues that lemma theory does not do a very good job of dealing with evidence from patients with brain damage.

Brain damage can lead to language production difficulties, and different types of damage can lead to different patterns of difficulties (Caramazza, 1997). Caramazza begins with the observation that if the lemma is a necessary level of representation in production, then brain damage that affects the set of lemma representations should have consistent effects on people's ability to produce language, whether they are speaking or writing. Lemmas are thought to represent grammatical information associated with specific words. So, if the lemmas are damaged, grammatical aspects of production should be affected.

In fact, some patients do have difficulty with just some types of words. Some patients have difficulty with *content* words (semantically rich words like *cat, table*, and *Hannibal Lecter*) but little difficulty with *function words* (semantically "light" grammatical markers like *the, of*, and *was*). Lemma theory would explain a patient like that by proposing that the lemmas for content words are selectively damaged, while lemmas representing function words are intact. But, as Caramazza notes, some patients have the opposite pattern of deficits, depending on how they are producing a given word. One pattern of problems can occur in speech, while the opposite pattern can occur in written language production, *within the same patient*. A given patient could have trouble with function words (but not content words) in writing and trouble with content words (but not function words) while speaking. If both processes tap into the same set of lemmas, it should not be possible for this pattern of problems to appear. If the spoken production problem for content words is based on broken content-word lemmas, then the same problem should occur in written-language production.

Further evidence against the lemma hypothesis comes from semantic substitution errors in brain-damaged patients. Some patients when asked to name pictures out loud will consistently use the wrong word. For example, one patient consistently said the word *dish* when asked to name a picture of a cook. When the same patient was asked to write (rather than say) the name of the picture, she wrote *forks*. These errors were not random, as the patient consistently produced one word (*dish*) while speaking, and the other (*forks*) while writing. Caramazza proposes that the solution to the problem lies in having two separate sources of word-form information, one for spoken language and one for written language. He proposes further that grammatical information is stored separately from lemma representations, as this can account for different patterns of function-word and content-word deficits within the same patient that depend on whether the patient is speaking or writing.

Self-monitoring and self-repair

Speakers make errors sometimes when they talk, but they also commonly succeed in expressing their meaning, despite the complexity that speech production involves. Speakers can monitor their own output to fix mistakes after they happen (*self-repair*), but they also deploy mental machinery internal to the speech production system to prevent internal, temporary problems from creating overt errors in the output (Gauvin and Hartsuiker, 2020).

Evidence for internal, pre-output monitoring comes from studies showing that when speakers make an error, they can replace the incorrect word with the correct one with almost no time elapsing between the error and the correction. Because it takes upwards of half a second to prepare a spoken response for a concept, "instant repair" means that both the monitoring and the repair planning must take place as the errored response is being produced. As Postma (2000, p. 105), notes, "There is ample evidence that speakers are capable of anticipating forthcoming mistakes, i.e., that they can inspect their speech programs prior to articulation … Speakers in a number of cases react without delay to an overt error. In addition, the correction is executed without further waiting (suggesting it must have been ready before the interruption was made)."

Incredibly, these internal monitoring processes are able to assess whether an error will lead to an embarrassing result, and both the galvanic skin response (a measure of how much resistance the skin offers to electrical current) and the likelihood of particular kinds of errors reflect the operations of an internal monitoring system (Hartsuiker and

Kolk, 2001; Levelt, 1983; Motley et al., 1983, 1982; Wheeldon and Levelt, 1995). If you were asked to produce the phrase *toolkits*, and you committed a sound exchange error, the result could be embarrassing. (Think what would happen if you went to the hardware store and tried to say, *Could I grab one of your toolkits?* and made a sound exchange error.) Sound exchange errors under those circumstances are less likely than sound exchanges that do not produce taboo words. Also, when participants are placed under time pressure, their galvanic skin response is higher for stimuli such as *toolkits* than pairs of words that do not produce taboo words when the initial sounds are exchanged, such as *poolkits*.[13]

Although some aspects of self-monitoring are carried out before overt production, they do not come for free. The ability to self-monitor depends on sufficient mental resources being available to carry out both speech planning processes and the monitoring itself. Further, speech planning and speech monitoring compete for the same mental resources. The more resources you dedicate to speech planning, the less you have for self-monitoring. Error detection is more robust at the ends of phrases and clauses, because the great majority of the utterance has already been planned and the planning load is at its lowest level (Blackmer and Mitton, 1991; Postma, 2000).

Articulation

The ultimate goal of speech planning as laid out in accounts like WEAVER++ and Dell's spreading activation model is to make the speech muscles move to produce sound. This process is called *articulation*. To speak, we configure our *vocal tract*, which consists of everything from the vocal folds upwards and outwards to our lips and noses. Articulation is both the end point of speech planning and production and the starting point for speech comprehension. Some accounts of articulation in production classify speech sounds (phonemes) according to the way the articulators move (Browman and Goldstein, 1989, 1990, 1991). The articulators include the lips, the tongue tip, the tongue body, the velum (the part of the soft palate toward the back of your mouth), and the glottis (a structure in your throat that houses the vocal folds). These different articulators can be moved semi-independently to perturb or stop the flow of air coming out of your lungs. These disturbances of the smooth flow of air set up vibrations which are modified by the movement of the articulators and create the sound waves that characterize human speech.

According to the articulatory phonology theory, the outcome of the speech planning process is a gestural score, which creates a *contrastive gesture*—a gesture that creates a noticeable difference between the current speech signal (sound) and other signals that the language employs. The *gestural score* tells the articulators how to move. More specifically, it tells the motor system to "(1) [move] a particular set of articulators; (2) toward a location in the vocal tract where a constriction occurs; (3) with a specific degree of constriction; and (4) ... in a characteristic dynamic manner" (Pardo and Remez, 2006, p. 217). The movement of the articulators produces a set of speech sounds (phonemes) that can be classified according to their *place of articulation*, *manner of articulation*, and *voicing*. English, for instance, has six stop consonants (/k/, /g/, /t/, /d/, /p/, /b/) that differ in place of articulation. /p/ and /b/ are *labial* because they are made by pressing the lips together. /t/ and /d/ are *dental* or *alveolar* stops because they involve stopping the flow of air behind the teeth (rather than behind the lips or elsewhere). /k/ and /g/ are velar because they involve stopping the flow of air with the back of the tongue pressed up against the velum. Each of the three gestures (lips together, tongue against teeth, tongue against velum) can be accompanied by vibration of the vocal folds or not. Simultaneous

release of air with vocal fold vibration leads to a *voiced* stop (as in /b/, /d/, and /g/). A delay between releasing the pent-up air and the beginning (onset) of vocal fold vibration leads to an *unvoiced* stop (as in /p/, /t/, and /k/). Manner of articulation refers to how much the flow of air is disturbed. Maximum blockage of the air flow leads to a stop consonant, squeezing the air flow without stopping it leads to a *fricative* (as in /z/ and / sh/ sounds), while keeping the air flowing relatively freely creates vowel sounds.

Savvy observers of language will have noticed that we do not produce isolated phonemes. We produce whole gangs of them when we talk, with an average of about one phoneme every 100 milliseconds (ms) in conversational speech. (Much higher rates can be obtained if you really try.) Because we produce many phonemes in a short period of time, we have to figure out ways to transition from producing one to the next smoothly and efficiently. To do that, we *coarticulate*. That is, the gestures for one phoneme overlap in time with the gestures for the preceding and following phoneme.[14] Coarticulation affects both the production and the perception of speech. For example, the way you move your articulators for the phoneme /p/ changes depending on which phoneme needs to come next. Say the word *pool*. Now say the word *pan*. Think about how your lips are placed just *before* you say the word. (Repeat as necessary until you notice the difference between how your lips are positioned before each word. Use a mirror or, better still, ask a friend to help.)

***

Intermission while you practice *pool* and *pan*.

***

Seriously. Try it.

***

You should notice that, before you start to say the word *pool*, your lips are in a rounded shape and they stick out a little bit. Before you say *pan*, your lips are drawn slightly back and are not rounded. Why the difference? It's a function of what phoneme comes next. The "oo" sound in pool is a rounded vowel. It's also a vowel that has a relatively low tone. When you round your lips, that matches the rounded nature of the vowel. Poking your lips out lengthens the resonant chamber formed by your vocal tract, which lowers the *fundamental frequency* (the lowest of the steady-state vibrations that makes up the sound wave), and makes the "oo" have a nice, deep tone. The rounded, poking-out characteristics of the "oo" vowel are anticipated by the speech-planning process, and so they assign some aspects of the vowel (roundness, poking-out-ness) to the preceding consonant gesture. The "a" sound in *pan* is a *back vowel* because it is formed by a constriction of the air toward the back of your mouth. To anticipate that movement, the speech-planning system programs a flattened out, slightly drawn back version of the preceding /p/ phoneme.

Coarticulation not only affects the way you shape your articulators when you speak different combinations of phonemes, it also affects the sound waves that are produced as you speak them (Liberman and Mattingly, 1985). Before we explore the effects of coarticulation on the sound waves that occur when you speak, let's take a brief detour into the physical, acoustic characteristics of speech sounds.

Moving the articulators is a physical event that, like many other physical events, creates sound waves that travel through the air. Therefore, speech, like other forms of sound, can be treated as a physical, acoustic signal. Acoustic signals, however complex, can be

analyzed with respect to two properties: frequency and amplitude (Goldstein, 2006). All acoustic signals are created when an acoustic medium (air, normally, but wood, water, steel, and other substances can also be used) is subjected to physical forces that alternately compress it (make it denser) and rarify it (make it less dense). One episode of compression and rarefaction is referred to as a *cycle*, and the amount of time it takes to complete a cycle determines the *frequency* of the sound wave. More cycles in a given period means higher frequency; fewer cycles means a lower frequency. The standard measure of frequency in acoustics is Hertz (Hz), which is the number of cycles per second. We subjectively experience differences in frequency as differences in *pitch*. High-pitched sounds (Minnie Mouse's voice, tea kettles whistling) result from high-frequency vibrations. Low-pitched sounds (foghorns, the roar of the surf) result from low-frequency vibrations. Amplitude refers to the change in pressure between the peak and the minimum pressures in the sound wave. We experience increases in amplitude as increases in volume or loudness. The standard measure of amplitude is Decibels (dB). High-amplitude sounds are loud; low-amplitude sounds are quiet.

Foreign Accent Syndrome Revisited

It's time to cash in your newly acquired knowledge of speech production. Foreign accent syndrome (FAS) occurs when "speech takes on characteristics normally associated with a dialect that is not [one's] own, or it resembles the performance of a non-native speaker of the language" (Moen, 2000, p. 5). Standard explanations of FAS appeal to theoretical models like Levelt's and Gary Dell's to explain why people can develop foreign-sounding accents after brain injury (Blumstein et al., 1987; Kurowski et al., 1996; Mariën et al., 2019, 2009). Speech sounds are created when articulators are moved toward specific targets at specific velocities creating specific degrees of closure with specific timing of voicing. Before these motor movements begin, there has to be a gestural score that specifies how and when the articulators are going to move. To create this gestural score, speakers have to undertake syllabification to divide the planned output into syllable-sized chunks. Once the output is syllabified, speakers have to craft a prosodic contour that extends over multiple syllables. A foreign-sounding accent can arise because the prosodic contour is disrupted, the process of syllabification is disrupted, or the articulation of individual phonemes is disrupted. For example, some patients with FAS show smaller than normal changes in pitch when they ask questions, abnormal patterns of accents (LOUDER vs. softer words and syllables), abnormal lengthening of vowels, and abnormal pausing. All of these could result from problems computing a prosodic contour. In addition, the long pauses between utterances suggest that patients are having some difficulty coming up with an articulatory plan or gestural score. Problems articulating individual phonemes may also contribute to the foreign flavor of the patient's speech. In some cases, speech sounds that should be articulated toward the back of the mouth are produced by closing off the air flow at more anterior (forward) locations. Patients may add or delete phonemes, especially from consonant clusters (e.g. *spl*, *rtr*), because they have difficulty making individual gestures. Syllabification may also be affected, as some patients produce syllables more like isolated units. *Escort us* might be produced more like "*Ess*," "*cort*," "*us*," than the usual way with the "t" syllabified with the "us." So, different aspects of speech planning (syllabification, prosody) and the (mis) execution of specific speech gestures can turn you into a Scot.

Speech Perception

I can't understand a word you've said the whole time.
RICKY BOBBY

Take a moment to look at Figure 2.4, which provides a visual representation of the sound waves produced when someone says, *to catch pink salmon* (Cooper et al., 1952; Liberman et al., 1952). The name for graphs like those shown in Figure 2.4 is *sound spectrogram.*[15] Frequency (in Hz) is represented on the *y*-axis, and time is represented on the *x*-axis.

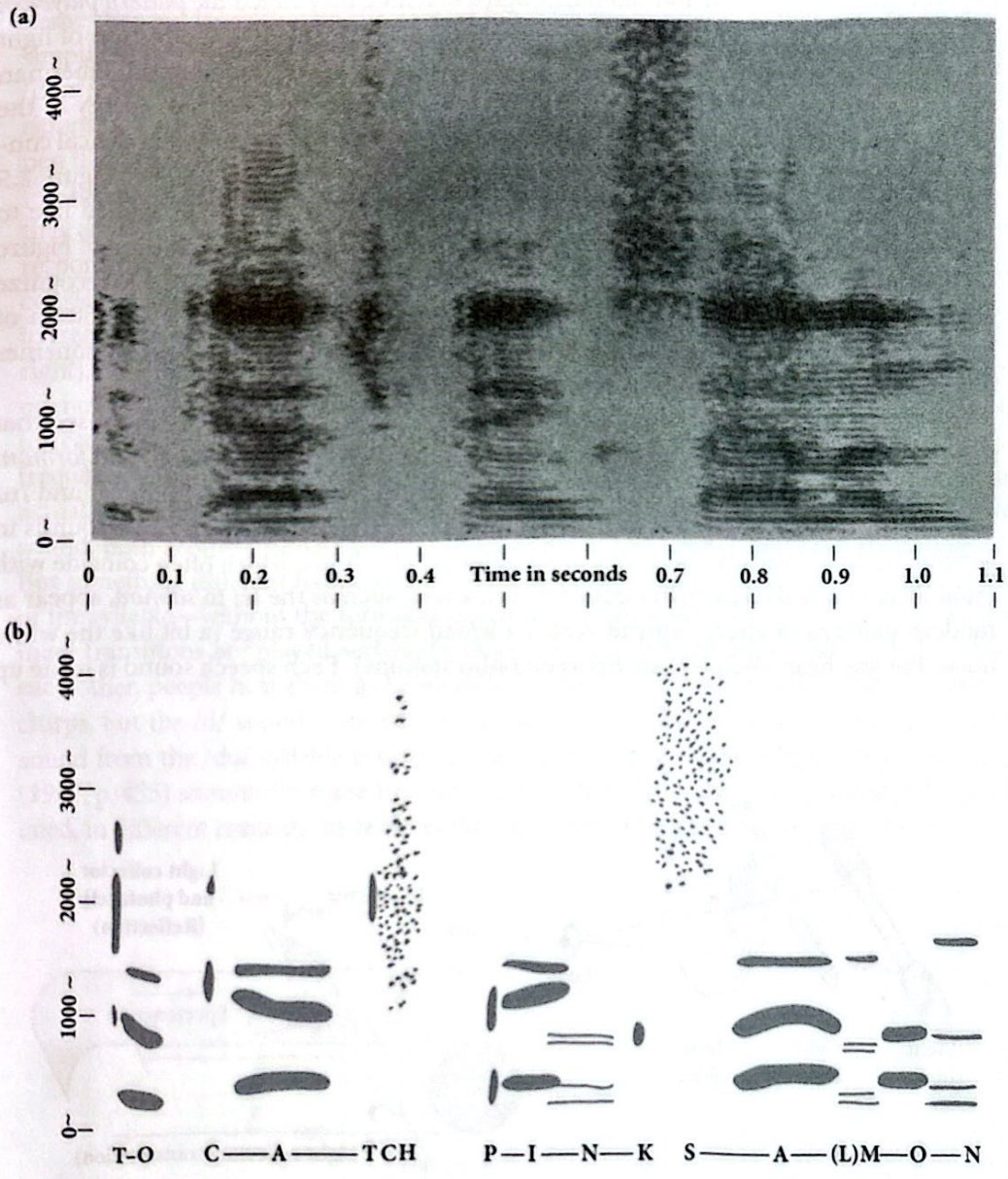


Figure 2.4 Sound spectrograms of the phrase *to catch pink salmon* created from real (top) and simplified, artificial speech (bottom). *Source:* Liberman, A. M (1952), University of Illinois Press

Consider first the top half of the figure, which represents real speech. The dark areas of the graph represent frequencies at which sound energy is present in the signal. Going vertically there are alternating bands of sound energy (represented as dark patches) and bands where no sound energy is present (represented as light patches). Over time, the pattern of energy changes. For instance, there is a lot of activity in the low-frequency part of the spectrum when someone is saying the /a/ in *catch*, and again when the /a/ sound in *salmon* is produced. But when the /ch/ sound in *catch* and the /s/ sound in *salmon* are produced there is very little energy in the low-frequency range, and much more energy at higher frequencies.

Now have a look at the bottom half of Figure 2.4. You will notice that the range of frequencies is the same as in the top half. But you will also notice that the pattern of dark and light patches is much simpler. Liberman et al. (1952) painted the pattern in the bottom of Figure 2.4 by hand and ran it through a machine they called the *pattern playback machine* (see Figure 2.5). The pattern playback machine converted the pattern of light and dark in the artificial, hand-painted spectrogram into a set of sound waves. Liberman and colleagues discovered that they could greatly simplify the pattern of energy in the sound wave without destroying the perceiver's ability to recognize the phonological content of the stimulus. That is, when they pushed the pattern in the bottom of Figure 2.5 through the pattern playback machine, their subjects reported that it sounded like *to catch pink salmon*. Thus, while the full pattern of energy represented in the top of Figure 2.4 may be necessary for the speech to sound fully natural, or for the listener to recognize whose voice created the signal, the stripped-down, simplified version at the bottom of Figure 2.4 carried all of the information necessary for people to perceive the phonemes that the signal was meant to convey (see also Remez et al., 1994).

Liberman and colleagues (Cooper et al., 1952; Liberman et al., 1952) proposed that the phonological content of speech could be described in terms of *formants* and *formant transitions*. Formants are steady-state, stable patterns of vibrations, as in the /a/ and /u/ sounds in *to catch pink salmon*, and in fact formants are associated with vowel sounds in general. Formant transitions consist of short bursts of sounds, which often coincide with rapid increases or decreases in frequency. Fricatives, such as the /s/ in *salmon*, appear as random patterns of energy spread across a broad frequency range (a bit like the white noise that you hear when you are between radio stations). Each speech sound is made up

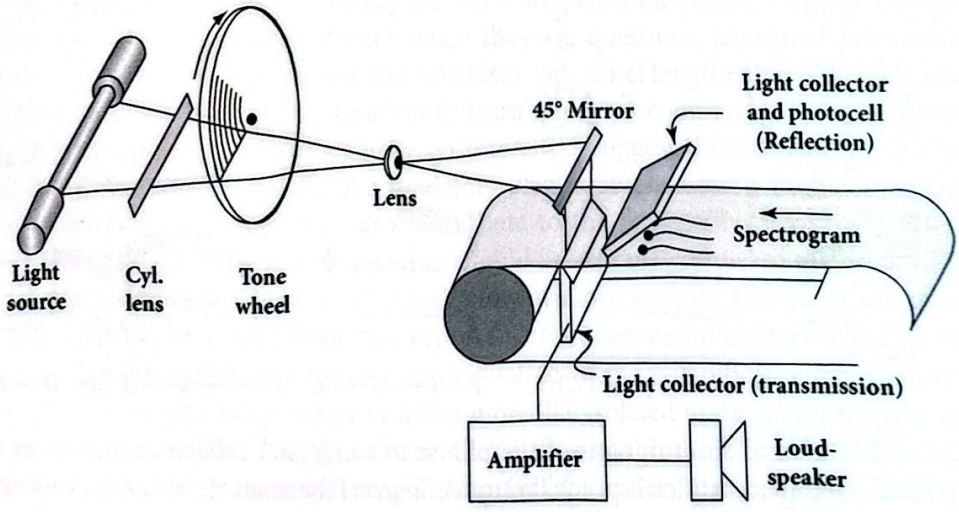


Figure 2.5 The pattern playback machine. *Source:* Liberman, A. M. (1952). University of Illinois Press

of a set of formants and/or formant transitions. The formants and transitions are classified according to their average frequency. The lowest frequency component of a phoneme is the first formant (for vowels) or first formant transition (for consonants). The next highest component is the second formant, or formant transition, and so on. Initially, speech researchers believed that they would be able to find a unique acoustic pattern for each phoneme, but they rapidly discovered that reality is more complicated than that. Which brings us back to coarticulation.

Coarticulation effects on speech perception

As noted previously, the way you move your articulators when you produce a given phoneme changes depending on the context in which the phoneme occurs. Figure 2.6 shows simplified spectrograms for two syllables, /di/ (pronounced "dee") and /du/ (pronounced "doo"). Each syllable consists of two formants (the horizontal bars) and two formant transitions (the slanty bars representing rapid changes in frequency). Notice what would happen if you split off the /d/ part of the signal from the /i/ part (on the left) and the /u/ part (on the right). Although the two /d/ parts of the syllables sound exactly the same when they are followed by two different vowels (/i/ and /u/) the actual physical signals that correspond to the /d/ phonemes are very different. In the /di/ syllable, the /d/ part of the signal consists of two formant transitions, and both of them are characterized by rapid increases in frequency over time (the bars for the two formant transitions both slant upwards to the right). Now notice the pattern for the /d/ sound in the /du/ syllable. Not only is the frequency of the second formant transition much lower for /du/ when compared to /di/, but instead of increasing in frequency over time (slanting up and to the right) it decreases in frequency over time (slanting down and to the right). Despite large differences in the actual, acoustic signal, when the two patterns are played in the context of a following vowel sound, both acoustic signals are perceived as being the same—they both sound like /d/. But something different happens if you play just the formant transitions, without the rest of the syllable—without the formants that make up the vowel sounds. When the two formant transitions are played separately from the rest of the syllable, and separately from each other, people hear them as being different sounds. They both sound like whistles or chirps, but the /d/ sound from the /di/ syllable sounds like a rising whistle, and the /d/ sound from the /du/ syllable sounds like a lower-pitched falling whistle. Liberman et al. (1967, p. 435) summarize these findings thus: "What is perceived as the same phoneme is cued, in different contexts, by features that are vastly different in acoustic terms."

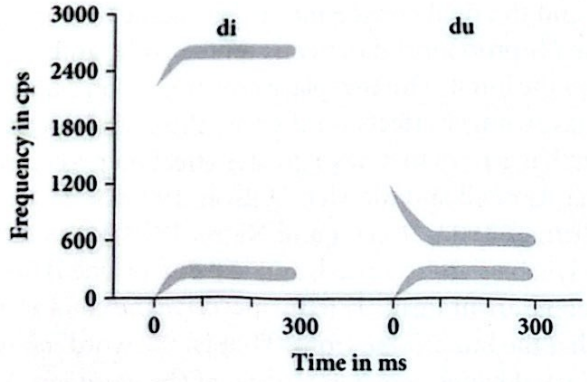


Figure 2.6 Artificial spectrogram for the syllables /di/ and /du/. *Source:* Liberman et al. (1967), American Psychological Association

A couple of important things are happening here. First, when you produce a /d/ sound, the way you make the /d/ sound and its physical form is different when the /d/ sound is followed by different vowels (that's coarticulation). Second, despite the differences in the way you make the /d/ sound and the actual physical properties of the sound waves, you perceive the two signals as being the same phoneme (this is a form of *perceptual constancy*—different physical patterns are perceived as being the same). Finally, that you perceive the two signals as being the same does not reflect insensitivity or inability to detect a difference between the formant transitions. When the two transitions are presented in isolation, you hear them as being different sounds.

Another aspect of coarticulation concerns the way speech sounds are spread out over time. When we write down speech, one letter occupies one position, the next letter occupies a separate, following position, and so on. But in speech, there are no clean breaks between phonemes, and the acoustic parts of one phoneme may overlap partially or entirely with other phonemes. Carol Fowler likens phonemes to Easter eggs, and describes speech production as being like pushing Easter eggs through a set of rollers that squish them all together. As Liberman and Mattingly (1985, p. 4) note, "the movements for gestures implied by a single symbol (phoneme) are typically not simultaneous, and the movements implied by successive symbols often overlap extensively." The listener is therefore faced with the task of going through the mass of squished-up eggs and rebuilding the original set of Easter eggs.

In one way, though, this "smearing" of information about a phoneme throughout an extended period of time can be beneficial, as demonstrated by the perception of *silent center vowels* (Liberman and Whalen, 2000). Silent center vowels are perceived when researchers edit a recording to remove part of the acoustic signal for an utterance. For example, they might erase the middle portion of the acoustic signal for *bag*. When you say the word *bag*, coarticulation means that information about the vowel starts to show up as you are pronouncing the /b/, and continues through production of the /g/ sound. When the middle part is erased, the word does not sound entirely normal, but people still correctly identify which phoneme was present in the original utterance (i.e. they hear something that sounds like *bag*, rather than something that sounds like *big*, *bug*, or *bog*). So, as long as the preceding and the following consonants carry information that results from coarticulation, listeners can accurately identify the missing vowel sound.

Evidence for coarticulation effects on speech perception also comes from studies involving *cross-spliced stimuli*. These are kind of like "Franken-stimuli," where parts of one spoken word have been chopped off and bolted onto a different word (like parts of different bodies were put together to make Frankenstein's monster). More specifically, single-syllable words can be divided into *onsets* (the "burst" of sound that starts the syllable, like the /p/ in *pink* or the /pr/ in *press*) and *codas* (the end of the syllable). The coda includes the vowel and the final consonant or consonant cluster. Coarticulation means that the way the burst is pronounced depends on the coda, and the way the coda is pronounced depends on the burst. The two place constraints on each other, but the way the burst is perceived has stronger effects on the way the coda is perceived than vice versa (i.e. the information that arrives first has a greater effect on perception than the information that arrives later; Gaskell and Marslen-Wilson, 1997; Marslen-Wilson and Warren, 1994; Martin and Brunell, 1982; Streeter and Nigro, 1979; Warren and Marslen-Wilson, 1987, 1988). If two syllables are recorded, and the end of one is spliced onto the beginning of the other, people are more likely to misperceive the coda as matching the original unspliced version that the burst came from. That is, the word *job* is likely to be misperceived as *jog* if the *jo* part came from a recording of the word *jog*. Also, if the /g/ and /b/ phonemes are presented without their initial bursts, people are likely to mistake them for one another. Thus, eliminating information that comes from coarticulation makes the

perceiver's job harder, suggesting that listeners routinely use information that appears "early" or "late" (where "early" means that the information appears during the articulation of a preceding phoneme, and "late" means that the information spills over into articulation of a following phoneme) to help identify which phoneme the speaker actually intended to produce.

Paradoxically, while coarticulated signals help the listener identify the phonemes that the speaker intended to produce, coarticulation is a major factor that makes it difficult to formally analyze the acoustic properties of speech. Ideally, we would like to know how acoustic signals line up with phonemes. In a simple world, there would be a one-to-one relationship between acoustic signals and phonemes. That way, if you had a phoneme, you would know what acoustic signal to produce. Likewise, if you had an acoustic signal, you would know what phoneme to look up (because there would only be one matching candidate in your long-term memory). Unfortunately, coarticulation as well as intra- and inter-speaker variability renders this simple system unworkable. As Liberman and Mattingly (1985, p. 12) note, "There is simply no way to define a phonetic category in purely acoustic terms." But if the speech signal can be decoded and its component phonemes identified (it can be and they are), there must be some way to untie the knot. The next sections will summarize the chief contenders for explaining how acoustic signals are analyzed so that people can recover the sets of phonemes they express. The *motor theory* of speech perception and the *general acoustic* approach represent two viable alternatives that are the focus of speech perception research today (Diehl et al., 2004; Kluender and Kiefte, 2006; Pardo and Remez, 2006).

The motor theory of speech perception

The motor theory of speech perception proposes that gestures, rather than sounds, represent the fundamental unit of mental representation in speech (Cooper et al., 1952; Liberman et al., 1952, 1954, 1967; Liberman and Mattingly, 1985; Liberman and Whalen, 2000; see also Fowler, 1986, 2008; Galantucci et al., 2006; Stokes et al., 2019).[16] That is, when you speak, you attempt to move your articulators to particular places in specific ways. Each of these movements constitutes a gesture. The motor part of the speech production system takes the sequence of words you want to say and comes up with a *gestural score* (movement plan) that tells your articulators how to move. According to the theory, if you can figure out what gestures created a speech signal, you can figure out what the gestural plan was, which takes you back to the sequence of syllables or words that went into the gestural plan in the first place. So by knowing what the gestures are, you can tell what the set of words was that produced that set of gestures. Going back to the /di/ versus /du/ example of coarticulation, the "core" part of that gesture is tapping the tip of your tongue against the back of your teeth (or your alveolar ridge, possibly). Other parts of the gesture, lip position, for example, are affected by coarticulation (flat lips for /di/, poking-out, rounded lips for /du/), but the core component of the gesture is the same regardless of the phonological context. Thus, rather than trying to map acoustic signals directly to phonemes, Alvin Liberman and his colleagues proposed that we map acoustic signals to the gestures that produced them because there is a closer relationship between gestures and phonemes than there is between acoustic signals and phonemes. In their words (Liberman et al., 1952, p. 513), "The relation between perception and articulation will be considerably simpler (more nearly direct) than the relation between perception and acoustic stimulus." Further, "Perceived similarities (and differences) will correspond more closely to the articulatory than the acoustic similarities among the sounds." Thus, differences between

two acoustic signals will not cause you to perceive two different phonemes as long as the gestures that created those two different acoustic signals are the same.

Motor theory also seeks to explain how a person can perceive an acoustic stimulus as a phoneme in one context (e.g. the formant transitions in Figure 2.6) but as a chirp or a buzz in another context. To explain this, motor theory proposes that speech perception is accomplished by a naturally selected *module* (Fodor, 1983). This speech perception module monitors incoming acoustic stimulation and reacts strongly when the signal contains the characteristic complex patterns that make up speech. When the speech module recognizes an incoming stimulus as speech, it *preempts* other auditory processing systems, preventing their output from entering consciousness. So, while nonspeech sounds are analyzed according to basic properties of frequency, amplitude, and timbre, and while we are able to perceive those characteristics of nonspeech sounds accurately, when the speech module latches onto an acoustic stimulus, it prevents the kind of *spectral analysis* (figuring out the pattern of frequencies and amplitudes in a stimulus) that general auditory processing mechanisms normally carry out for nonspeech auditory stimuli (Liberman and Mattingly, 1985, 1989; Liberman and Whalen, 2000). This *principle of preemption* explains why formant transitions are perceived as chirps or high-pitched whistles when played in isolation, but as phonemes when played in the context of other speech sounds. When transitions are played in isolation, they are not recognized as speech, so the spectral analysis dominates perception, and they sound like chirps. "When transitions of the second formant ... are presented in isolation, we hear them as we should expect to—that is, as pitch glides or as differently pitched 'chirps.' But when they are embedded in synthetic syllables, we hear unique linguistic events, [bæ], [dæ], [gæ], which cannot be analyzed in auditory terms" (Mattingly et al., 1971, p. 132).

This preemption of normal auditory perceptual processes for speech stimuli can lead to *duplex perception* under special, controlled laboratory conditions (Liberman and Mattingly, 1989; Whalen and Liberman, 1987). To create their experimental stimuli, researchers constructed artificial speech stimuli that sounded like /da/ or /ga/ depending on whether the second formant transition decreased in frequency over time (/da/) or increased (/ga/). Next, they edited the stimuli to create separate signals for the transition and the rest of the syllable, which they called the *base* (see Figure 2.7). They played the two parts of the stimulus over headphones, with the transition going in one ear and the base going in the other. The question was, how would people perceive the stimulus? Would chopping up the stimulus make it sound like gibberish? Or would it still be perceived as speech? It turned out that people perceived two different things at the same time. At the ear that the transition was played into, people perceived a high-pitched chirp or whistle. But at the same time, they perceived the original syllable, just as if the entire, intact stimulus had been presented.[17]

Liberman and colleagues argued that simultaneously perceiving the transition in two ways—as a chirp and as a phoneme—reflected the simultaneous operation of the speech perception module and general-purpose auditory processing mechanisms. Duplex perception happened because the auditory system could not treat the transition and base as coming from the same source (because two different sounds were played into two different ears). Because the auditory system recognized two different sources, it had to do something with the transition that it would not normally do. That is, it had to analyze it for the frequencies it contained, and the result was hearing it as a "chirp." But simultaneously, the speech processing module recognized a familiar pattern of transitions and formants. As a result, the auditory system reflexively integrated the transition and base, and produced the experience of hearing a unified syllable, despite the fact that it was working with two spatially distinct stimuli. In the early days of duplex perception research, speech was the only kind of stimulus known to produce such effects, which was

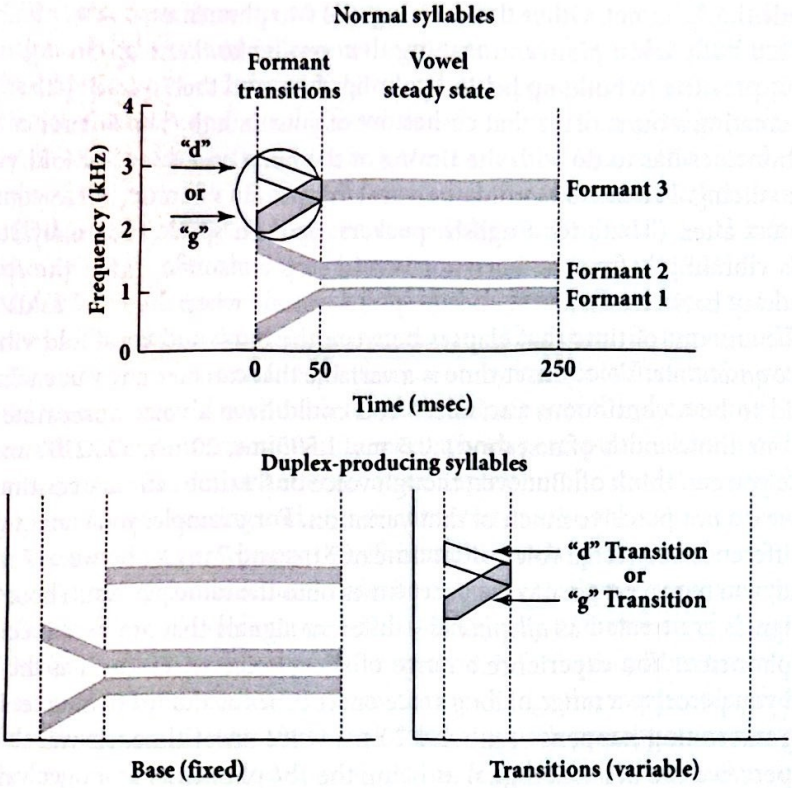


Figure 2.7 Simplified acoustic stimuli that are perceived as /da/ or /ga/. Researchers edited the stimuli so that a formant transition would be played to one ear, while the "base" (the rest of the signal) was played to the other ear. People perceived the stimulus as consisting of a "whistle" or a "chirp" at one ear and the complete syllable (/da/ or /ga/, depending on which formant transition was played) at the other ear. *Source:* Whalen and Liberman (1987), American Association for the Advancement of Science

taken as evidence that speech was "special" and subject to its own principles, separate from other kinds of auditory stimuli.

According to the motor theory, *categorical perception* is another product of the speech perception module. Categorical perception happens when a wide variety of physically distinct stimuli are perceived as belonging to one of a fixed (usually fairly small) set of categories. For example, every vocal tract is different from every other vocal tract. As a result, the sound waves that come out of your mouth when you say *pink* are different than the sound waves that come out of my mouth when I say *pink*, and those two stimuli are different than the sound waves that come out of Arnold Schwarzenegger's mouth when he says *pink*. Nonetheless, your phonological perception system is blind to the physical differences and perceives all of those signals as containing an instance of the category /p/. You may notice that your voice has slightly (or greatly) different qualities than Arnold Schwarzenegger's, but you categorize the speech sounds he makes the same way you categorize your own or anybody else's. All of those different noises map to the same set of about 40 phonemes (in English). In addition, although the acoustic properties of speech stimuli can vary across a wide range, your perception does not change in little bitty steps, with each little bitty change in the acoustic signal. You are insensitive to some kinds of variation in the speech signal, but when the speech signal changes enough, you perceive that change as the difference between one phoneme and another (Liberman et al., 1957). An example may help to illustrate.

Recall that the difference between some stop consonants and others is whether they are voiced or not. The difference between /b/ and /p/, for example, is that the /b/ is

voiced while the /p/ is not. Other than voicing, the two phonemes are essentially identical. They are both *labial plosives*, meaning that you make them by closing your lips, allowing air pressure to build up behind your lip-dam, and then releasing that pressure suddenly, creating a burst of air that rushes out of your mouth. The difference between the two phonemes has to do with the timing of the burst and the vocal fold vibrations that create voicing. For the /b/ sound, the vocal folds begin vibrating while your lips are closed or just after. (That's for English speakers. Spanish speakers actually start their vocal folds vibrating before the burst for voiced stop consonants.) For the /p/ sound, there is a delay between the burst and the point in time when the vocal folds begin to vibrate. The amount of time that elapses between the burst and vocal fold vibration is called *voice onset time*. Voice onset time is a variable that can take any value whatsoever, so it is said to be a continuous variable.[18] You could have a voice onset time of 0 ms (which is one thousandth of a second), 0.5 ms, 1.895 ms, 20 ms, 50.22293 ms, or any other value you can think of. But even though voice onset time can vary continuously in this way, we do not perceive much of that variation. For example, you cannot generally hear the difference between a voice onset time of 2 ms and 7 ms, or between 7 ms and 15 ms. Instead, you map a range of voice onset times onto the same percept. Those different acoustic signals are treated as *allophones*—different signals that are perceived as being the same phoneme. You experience a range of short voice onset times as the /b/ phoneme; and you perceive a range of long voice onset times as the /p/ phoneme. However, something interesting happens at around 20 ms voice onset time. At values less than that, you perceive the acoustic signal as being the /b/ phoneme; at longer values than that, you perceive the signal as being the /p/ phoneme. (And so do babies! Eimas et al., 1971). Further, your ability to discriminate between different acoustic signals depends on whether two signals come from the same side of the voice onset time "border" or whether they come from opposite sides. If two stimuli come from the same side of the border (with voice onset times of, say, 10 and 17 ms), you have a lot of trouble hearing the difference. But if two stimuli having the same absolute difference in voice onset time come from opposite sides of the border (17 and 24 ms, say), you have a much greater chance of hearing the difference. Liberman argued that this categorical perception of speech sounds provided further evidence that the speech perception system was special and different from the auditory perception processes that dealt with nonspeech sounds.

The McGurk effect: Visual gestures affect speech perception

According to the motor theory of speech perception, understanding speech requires you to figure out which gestures created a given acoustic signal. Because figuring out the gestures is the primary goal of the speech perception system, you might expect that system to use any sort of information that could help identify the gestures. While acoustic stimuli offer cues to what those gestures are, other perceptual systems could possibly help out, and if they can, motor theory says that the speech perception system will take advantage of them. In fact, two non-auditory perceptual systems—vision and touch—have been shown to affect speech perception. The most famous demonstration of *multi-modal speech perception* is the McGurk effect (Kerzel and Bekkering, 2000; McGurk and MacDonald, 1976). The McGurk effect happens when people watch a video of a person talking, but the audio portion of the tape has been altered. For example, the video image might show a person saying *ga*, but the audio signal is of a person saying *ba*. What people actually perceive is someone saying *da*. If the visual information is removed (when individuals shut their eyes, for example), the auditory information is accurately perceived and the person

hears *ba*. (You can experience the McGurk effect yourself by typing "McGurk effect" into your favorite web browser and following the links to any of several demonstrations of the effect.) The McGurk effect is incredibly robust: It happens even when people are fully warned that the auditory and visual information do not match; and it happens even if you try to pay close attention to the auditory information and ignore the visual (unless you look away or close your eyes). It happens when real words are used rather than nonsense syllables (Dekle et al., 1992). It happens even if the auditory and visual information is processed only by one of the brain's two hemispheres (Baynes et al., 1994).

The McGurk effect happens because your speech perception system combines visual and auditory information when perceiving speech, rather than relying on auditory information alone. Of course, the auditory information by itself is sufficient for perception to occur (otherwise, we would not be able to communicate over the phone), but the McGurk effect shows that visual information influences speech perception when that visual information is available. The McGurk effect is an example of multi-modal perception because two *sensory modalities*, hearing and vision, contribute to the subjective experience of the stimulus (two modes of perception, therefore multi-modal perception).

The vision–hearing combination is not the only way to alter speech perception. There is a more "icky" (Carol Fowler's term; Fowler, 2008) way to create another variant of the standard McGurk effect. In this alternative method, information from touch (*haptic* perception) is combined with auditory information to change the way people perceive a spoken syllable (Fowler and Dekle, 1991). This kind of speech perception occurs outside the laboratory from time to time in a specialized mode called *tadoma*. Hearing- and vision-impaired individuals, such as Helen Keller, have learned to speak by using their sense of touch to feel the articulatory movements involved in speech. In the lab, haptic perception has been used to investigate the limits of multi-modal speech perception. According to the motor theory, information about speech gestures should be useful, regardless of the source, auditory or otherwise. That being the case, information about articulatory gestures that is gathered via the perceiver's sense of touch should affect speech perception, similar to the way visual information does. To find out whether that happens, Carol Fowler had experimental participants feel her lips while they listened to a recording of a female speaker (also Fowler) speaking a variety of syllables. Blindfolded and gloved,[19] experimental participants heard the syllable /ga/ over a speaker (or over headphones in a separate experiment) while Fowler simultaneously (silently) mouthed the syllable /ba/. As a result, the experimental participant felt the articulatory gestures appropriate to one syllable, but heard the acoustic signal appropriate to a different syllable. As in the visual version of the McGurk effect, what participants actually perceived was a compromise between the auditory signal and the haptic (touch) signal. Instead of perceiving the spoken syllable /ga/, or the felt syllable /ba/, they heard the "hybrid" syllable /da/. Just as in the visual McGurk effect, speech perception was influenced by input from two perceptual modalities.

Motor theory explains both versions of the McGurk effect, the visual one and the haptic one, as stemming from the same basic process. The goal of the speech perception system is not a spectral analysis of the auditory input. Rather, it is figuring out what set of gestures created the auditory signal in the first place. Motor theory straightforwardly handles visual and haptic effects on speech perception by arguing that both vision and touch can contribute information that helps the perceiver figure out what gesture the speaker made. Under natural conditions, the visual, touch, and auditory information will all line up perfectly, meaning that secondary sources of information (non-auditory sources, that is) will be perfectly valid cues. While speech perception does not absolutely require visual or haptic input, those sources can certainly be useful. Think about what you do in a noisy bar when the background noise makes it hard to hear your conversational partner. Odds are, you look at their mouth. Why? Because the visual information

helps to supplement the noisy and degraded auditory input. Why is that useful? According to motor theory, the visual information is useful because what you are really trying to do is figure out what speech gestures your partner is making. That's useful, because figuring out the gestures leads you back to the gestural score, figuring out the gestural score leads you back to the phonemes, and figuring out the phonemes gets you back to the message.

Mirror neurons: The motor theory enjoys a renaissance

Motor theory has been enjoying a renaissance recently sparked off by new evidence about monkey neurons (Gallese et al., 1996; Gentilucci and Corballis, 2006; Kohler et al., 2002; Rizzolatti and Arbib, 1998). More specifically, researchers working on Macaque monkeys (*Macaca nemestina*) discovered neurons in a part of the monkey's frontal lobes that responded when a monkey performed a particular action, when the monkey watched someone else perform that action, or when the monkey heard a sound associated with that action. These neurons were called *mirror neurons*. The existence of mirror neurons in monkeys was established by invasive single-cell recording techniques. Similar experiments in humans are ethically impossible, and so the existence of the human equivalent of Macaca mirror neurons remains a hypothesis, rather than an established fact.

The part of the brain where mirror neurons were found in monkeys is called *area F5*, which bears some resemblance to a part of the human brain that is important for language processing, *Broca's area* (see Chapter 13). Neuroimaging and research involving direct recording from neurons in Broca's area (part of the frontal lobes of the brain in the left hemisphere) both show that it participates in speech perception (Sahin et al., 2009; St. Heim et al., 2003). The researchers who discovered mirror neurons proposed that mirror neurons could be the neurological mechanism that the motor theory of speech perception requires. That is, mirror neurons in Broca's area could fire when an individual produces a particular set of phonemes. The same mirror neurons would fire when the same individual heard those same phonemes, providing a bridge between speaking and listening. (Keep in mind, this all presupposes that mirror neurons exist in human brains, which has not been demonstrated at the time of writing.)

Although it is not possible (yet) to record from single human neurons, other kinds of experiments have been conducted to try to find evidence for the participation of the human motor cortex in speech perception. The experimental logic is as follows: Motor theory says that accessing representations of specific speech gestures underlies speech perception. Those representations of speech gestures must be stored in the parts of the brain that control articulatory movements. The parts of the brain that control articulation are the motor cortex in the frontal lobes of the brain and the adjacent premotor cortex. Put that all together and it means that, according to motor theory, you should activate the motor cortex when you perceive speech. Proponents of mirror neurons argue that mirror neurons are the neural (brain) mechanism that establishes the link between heard speech and motor representations that underlie speech production. Mirror neurons have recently been found in the monkey equivalent of the motor cortex (they have also been found in the monkey equivalent of the human premotor cortex and in the monkey equivalent of the parietal lobes). Proponents of mirror neurons view evidence that the motor cortex responds to speech as supporting their view of speech perception. Some mirror neuron enthusiasts argue further that mirror neurons play a role in speech perception in modern humans because our speech production and perception processes evolved from an older manual gesture system (Gentilucci and Corballis, 2006).[20]

Although mirror neurons have not been found in humans, proponents of the mirror neuron hypothesis have used slightly less direct ways to find evidence for the involvement

of motor and premotor cortices in speech perception. This evidence comes in two distinct forms: neuroimaging data and *transcranial magnetic stimulation* (TMS) studies (Benson et al., 2001; Binder et al., 1997; Cappelletti et al., 2008; Fadiga et al., 2002; Gow and Segawa, 2009; McNealy et al., 2006; Meister et al., 2007; Pulvermüller et al., 2006; Sato et al., 2009; St. Heim et al., 2003; Watkins et al., 2003). In Pulvermüller et al. (2006), participants listened to syllables that resulted from bilabial stops (/pa/, /ba/) or alveolar stops (/ta/, /da/) on *listening* trials. On *silent production* trials, participants imagined themselves making those sounds. Measurements of their brains' activity were gathered using functional magnetic resonance imaging (fMRI). Listening to speech caused substantial brain activity in the superior (top) parts of the temporal lobes on both sides of the participants' brains (which correspond to primary and secondary auditory receiving areas), but it also caused a lot of brain activity in the motor cortex in the experimental participants' frontal lobes. Further, brain activity in the motor cortex depended on what kind of speech sounds the participants were listening to. If they were listening to a bilabial stop syllable, activity was observed in one part of motor cortex. If they were listening to an alveolar stop syllable, activity was observed in a different part of the motor cortex. The brain areas that responded when participants listened to speech were similar to the brain areas that responded when participants imagined saying the same syllables. That is, listening to or imagining saying the syllable /ba/ was correlated with brain activity in one part of the motor cortex. Listening to or imagining saying /ta/ was correlated with brain activity in a different part of the motor cortex. Motor theory explains these results by arguing that the same brain areas that produce speech are involved in perceiving it. Hearing or saying /ba/ activates the same part of motor cortex because listening to /ba/ activates stored representations that are involved in moving the lips. Hearing or producing /da/ activates a different part of the motor cortex from /ba/ because tongue movements (involved in producing /da/) rely on motor representations that are stored in a different part of the motor cortex. Other neuroimaging studies also show activity in the frontal lobes when people listen to speech, although some studies find frontal lobe activity only when the experimental participants have to explicitly compare different syllables or phonemes (so the frontal lobe activity may be related to the process of comparing speech sounds rather than the act of perceiving those speech sounds in the first place; Buchanan et al., 2000; Newman and Twieg, 2001; Scott et al., 2009; Zatorre et al., 1992).[21]

TMS experiments have also been used to bolster the motor theory of speech perception (Fadiga et al., 2002; Meister et al., 2007; Watkins et al., 2003). In this kind of experiment, a strong magnetic field is created right next to an experimental participant's head. The magnetic field interferes temporarily with the normal functioning of neurons in the cortex just below the magnetic coil. Magnetic stimulation can alter an individual's behavior on various cognitive tasks, and the results of stimulation can be measured by neural responses at other locations on the body. For example, magnetic stimulation of parts of the motor cortex can lead to increases in neural activity in the muscles of the hand and fingers. These enhanced responses are called *motor-evoked potentials*. When TMS was applied to participants' motor cortex in one study, participants were less able to tell the difference (*discriminate*) between two similar phonemes.[22] Further, when people listen to speech sounds that involve tongue movements, and have TMS applied to the parts of motor cortex that control the tongue, increased motor-evoked potentials are observed in the participants' tongue muscles. When TMS is applied elsewhere, or when the speech sounds do not involve tongue movements, motor-evoked potentials measured at the tongue are no different than normal. Motor-evoked potentials at the tongue are also obtained when TMS is applied and people watch videos of other people talking (Watkins et al., 2003). All of these experiments show that the motor cortex generates neural activity in response to speech, consistent with the motor theory of speech perception.

The mirror neuron theory of speech perception jumps the shark

And then it gets a little bit crazy. If you ask the average psycholinguist or neurolinguist[23] whether the parts of the motor cortex that control leg movements should be involved in speech perception, they tend to say things like "No," "No way," or "Huh?" However, the same kinds of TMS manipulations that lead to motor-evoked potentials in the tongue muscles also produce motor-evoked potentials in the leg muscles (Liuzzi et al., 2008). It makes sense, from the motor theory perspective, that TMS should lead to activity in the tongue muscles when we listen to speech because motor theory says the representations we need to figure out the speech gestures reside in the motor cortex (the mirror neuron variant of motor theory makes the same claim). But how much sense does it make to say that perceptual representations for speech perception reside in the leg-control part of the motor cortex? The authors of the leg study concluded that speech perception depends on "an extended action–language network, also including the leg motor circuits" (Liuzzi et al., 2008, p. 2825). They propose a link between non-verbal gestures and speech gestures, and a further link between leg movements (which do not play a major role in human communication, despite claims to the contrary) and manual (hand and arm) gestures (which do).

Instead of taking the leg results as strong evidence for motor theory, the disinterested observer might actually use these results to call into question the entire TMS/motor-evoked potential research enterprise. If your experimental technique produces a thoroughly anomalous result, it might just be possible that there is something wrong with that technique as a research tool. On the other hand, widespread activity in motor cortex in response to speech would make sense, if listening to speech triggers circuits that people use to prepare behavioral responses, which could include a variety of both verbal and non-verbal movements (Scott et al., 2009). Alternatively, motor neurons might respond to speech because they are involved in a monitoring and correction circuit. When we speak, we monitor it for errors (as mentioned earlier). When an individual's own speech is electronically altered as it is being produced, that individual will alter their spoken output to compensate for the electronic changes in less than 150 ms (Tourville et al., 2007; see also Okada and Hickok, 2006). Neuroimaging shows that this feedback loop involves groups of both posterior, temporal lobe neurons, and neurons in the frontal lobes. So, activity in motor cortex could involve neural circuits that normally respond to speech perception processes (that are carried out elsewhere in the brain) by dynamically adjusting speech output. Alternatively, one way to verify that you have heard a speech sound correctly would be to covertly produce your own version of the speech sound and compare the two examples. This would account for motor activation during speech perception—it would reflect self-generation of phonemes for comparison to the input.

Other problems for mirror neuron/motor theory

Motor theory has faced a number of challenges besides some odd results in the TMS research (Hickok, 2008; Lotto et al., 2009). Some challenges to motor theory are rooted in the strong connection it makes between perception and production (based on the idea that perception involves the activation of motor representations of specific speech gestures). Infants, for example, are fully capable of perceiving the differences between many different speech sounds, despite the fact that they are thoroughly incapable of producing those speech sounds (Eimas et al., 1971; see Chapter 9). To account for this result, we

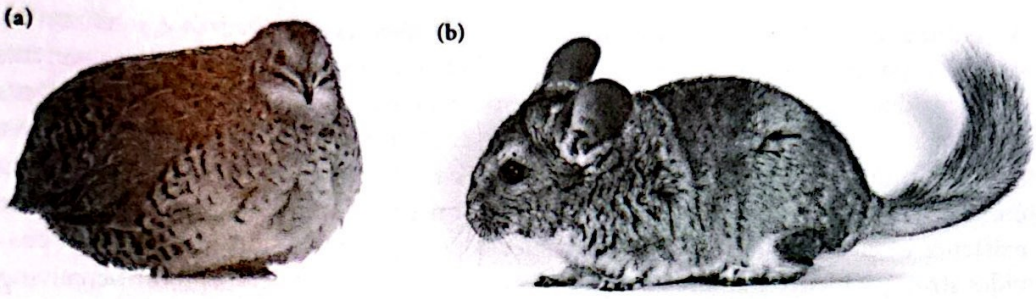


Figure 2.8 Japanese quail (left) and Chinchilla (right). They perceive differences between different phonemes, they look good, and they taste good. *Source:* (a) Eric Isselée/Adobe Stock; (b) Sergey Goruppa/Adobe Stock

either have to conclude that infants are born with an innate set of speech–motor representations (and are incapable of making the appropriate gestures only because they have not yet learned to control their articulators well enough) or that having a set of speech–motor representations is not necessary to perceive phonemes.

Additional experimental observations have also cast doubt on whether speech–motor representations are necessary for speech perception. No one would suggest, for example, that nonhuman animals have a supply of speech–motor representations, especially if those animals are incapable of producing anything that sounds like human speech. Two such animals are Japanese quail and chinchillas (see Figure 2.8). Japanese quail and chinchillas, once they are trained to respond to one class of speech sounds, and refrain from responding to another class, will demonstrate aspects of speech perception that resemble human performance. More specifically, both kinds of animal show categorical perception of speech, and both show compensation for coarticulation (Diehl et al., 2004; Kluender et al., 1987; Kluender and Kiefte, 2006; Kuhl and Miller, 1975).[24] Since these animals lack the human articulatory apparatus, they cannot have speech–motor representations. Because they respond to aspects of speech very much like humans do, motor theory's claim that speech–motor representations are necessary for speech perception is seriously threatened.

Japanese quail and chinchillas show that aspects of speech perception are not limited to human perceivers. Other research shows that duplex perception and categorical perception are not limited to speech perception. Sounds other than speech, such as slamming doors, produce duplex perception when the original signals are edited so that two parts of the acoustic signal are played to different ears. Sounds other than speech, such as the sound a violin makes when it is bowed versus plucked, show categorical shifts in perception. When a violin string is plucked, there is a burst, a brief delay, and then the onset of steady-state vibration (comparable to the burst and vocal fold vibration in speech signals that are perceived as stop consonants). When the same violin string is played with a bow, the vibration and burst begin nearly simultaneously. When violin sounds are edited to vary the lag between the burst and the onset of vibration, short lags are perceived as bowed sounds, but there is a sudden change to perceiving the sound as a string being plucked when the burst–vibration lag gets longer. Both of these effects run contrary to the motor theory's claim that these aspects of speech perception are the result of a specially tuned processing module (Kluender and Kiefte, 2006; Lotto et al., 2009). If categorical and duplex perception were the result of a special speech processing module, they would occur only for speech sounds.

Research with aphasic patients (patients who have a variety of language-related problems as the result of brain damage) casts further doubt on the motor theory.

A century and a half ago, Paul Broca and Carl Wernicke showed that some brain-damaged patients could understand speech, but not produce it, while other patients could produce fluent speech, but could not understand it (see Chapter 13). Two patients (Messrs Leborgne and Lelong), both of whom could understand speech, had extensive damage in the frontal lobes of their brains, specifically in the part of their brains that corresponds to area F5 in *Macaca* (where mirror neurons are located in monkeys). The existence of clear dissociations between speech perception and speech production provides strong evidence that intact motor representations are not necessary for perceiving speech. Although the kinds of language disorders that result from brain damage are complex, and not all cases neatly fit into the "broken perception" or "broken production" categories, numerous cases that show the selective impairment of either perception or production (but not both) have been described (see Caplan and Hildebrandt, 1988; note that subtle comprehension impairments have been shown in patients in the "broken production" category, but these involve syntax rather than phonology). If speech perception requires access to intact motor representations, then brain damage that impairs spoken language output should also impair spoken language perception, but this pattern does not appear much of the time.

A motor or mirror neuron advocate might argue that the damage in the reported cases is not extensive enough to wipe out the speech–motor representations, or that *unilateral* damage (limited to one side of the brain) does not wipe out all of the relevant motor representations. However, even with both motor cortices (the one in the left hemisphere and the one in the right hemisphere of the brain) thoroughly damaged, at least some patients can still understand speech quite well (Caltagirone, 1984). The motor theory of speech perception claims that speech is understood because listeners can use the incoming acoustic signal to activate representations of the physical motions that created it. Because the motor (muscle movement) representations are thought to be stored in the parts of the brain that control movement (i.e. the motor and premotor cortices in the frontal lobes), motor theory predicts that damage to those frontal regions should produce significant problems with speech perception. After all, if you understand speech by activating motor representations (which allows you to tell which gestures created the acoustic signal), and if those motor representations are stored in a particular part of the brain, then damaging those parts of the brain should cause problems understanding speech because you can no longer find the motor representations you need.[25]

Motor theory and mirror neuron theory have been criticized because they do not indicate how acoustic signals allow listeners to figure out which gestures produced those signals (other than "it's done with mirror neurons"). Another problem for either account is that there is a many-to-one mapping between gestures and phonemes. That is, the same speech sound can be produced by different articulatory gestures (MacNeilage 1970). More specifically, different people can produce the same phoneme by using different configurations of the vocal tract. Because the vocal tract offers a number of locations where the air flow can be restricted, and because different combinations of air-flow restriction have the same (or nearly the same) physical effect, they wind up producing acoustic signals that are indistinguishable to the perceiver. That means that there is no single gesture for syllables like /ga/. Studies involving the production of bite-block vowels also show that very different gestures can lead to the same or nearly the same acoustic signal, and perception of the same set of phonemes (Gay et al., 1981). In this kind of experiment, speakers hold an object between their teeth and attempt to say a given syllable. When they do, the way they move their articulators is different than normal, but the acoustic signal that comes out can be very close to the normal one. Motor theory can account for this set of facts in one of two ways. It could propose that more than one

speech-motor representation goes with a given phoneme. But that would complicate the representation of speech sounds, and the perceiver could wind up needing separate sets of speech-motor representations for each speaker. Alternatively, motor theory could propose that there is a single set of "ideal" or "prototype" speech-motor representations, and that an acoustic analysis of the speech signal determines which of these "ideal" gestures most closely matches the acoustic input, but that would violate the spirit and letter of the motor theory.

The general auditory approach to speech perception

The general auditory (GA) approach to speech perception starts with the assumption that speech perception is not special (Diehl and Kluender, 1989; Diehl et al., 2004, 1991; Kluender and Kiefte, 2006; Pardo and Remez, 2006). Instead, "speech sounds are perceived using the same mechanisms of audition and perceptual learning that have evolved in humans ... to handle other classes of environmental sounds" (Diehl et al., 2004, p. 154). Researchers in this tradition look for consistent patterns in the acoustic signal for speech that appear whenever particular speech properties are present. Further, they seek to explain commonalities in the way different people and even different species react to aspects of speech. For example, some studies have looked at the way people and animals respond to *voicing contrasts* (the difference between unvoiced consonants like /p/ and voiced consonants like /b/). These studies suggest that our ability to perceive voicing is related to fundamental properties of the auditory system. We can tell whether two sounds occurred simultaneously if they begin more than 20 ms apart. If two sounds are presented starting within about 20 ms of each other, we will perceive them as being simultaneous in time. If one starts more than 20 ms before the other, we perceive them as occurring in a sequence, one before the other. The voicing boundary for people and Japanese quail sits right at that same point. If vocal fold vibration starts within 20 ms of the burst, we perceive the phoneme as voiced. But if there's more than a 20 ms gap between the burst and vocal fold vibration, we perceive an unvoiced stop. Thus, this aspect of phonological perception could be based on a fundamental property of auditory perception, rather than the peculiarities of the gestures that go into voiced and unvoiced stop consonants.

Because the acoustic signals created by speech are tremendously complex, the general acoustic approach, as it stands, does not offer explanation of the full range of human (or animal) speech perception abilities. Its chief advantages lie in its ability to explain common characteristics of human and nonhuman speech perception, as well as common properties of human speech and nonspeech perception. Because the GA approach is not committed to gestures as the fundamental unit of phonological representation, it is not vulnerable to many of the criticisms leveled at the motor theory.

The *fuzzy logical model of speech perception* (FLMP), one of the better known approaches within the GA tradition, incorporates the idea that there is a single set of "ideal" or "prototype" representations of speech sounds, as determined by their acoustic characteristics (Massaro and Chen, 2008; Massaro and Oden, 1995; Oden and Massaro, 1978; see also Movellan and McClelland, 2001). According to FLMP, speech perception reflects the outcomes of two kinds of processes: *bottom up* and *top down*. Bottom-up processes are those mental operations that analyze the acoustic properties of a given speech stimulus. These bottom-up processes activate a set of potentially matching phonological representations. Stored representations of phonemes are activated to the degree that they are similar to acoustic properties in the speech stimulus; more similar

phonemes attain higher degrees of activation, less similar phonemes attain lower degrees of activation. Top-down processes are those mental operations that use information in long-term memory to try to select the best possible candidate from among the set of candidates activated by the bottom-up processes. This may be especially important when the bottom-up information is ambiguous or degraded. For example, when the /n/ phoneme precedes the /b/ sound (as in *lean bacon*), oftentimes coarticulation makes the /n/ phoneme come out sounding more like an /m/. When someone listens to *lean bacon*, bottom-up processes will activate both the prototype /n/ phoneme and the prototype /m/ phoneme, because the actual /n/ part of the signal will be intermediate between the two prototypes. According to the FLMP, our knowledge that *lean bacon* is a likely expression in English should cause us to favor the /n/ interpretation, because there is no such expression as "leam bacon." However, if the /n/ sound were in a nonword, such as *pleam bacon*, a listener would be more likely to pick the /m/ interpretation, because the competing /n/ sound would not receive any support from "top-down" processes. This effect, the tendency to perceive ambiguous speech stimuli as real words if possible, is known as the *Ganong* effect, after its discoverer, William Ganong (1980).

FLMP also offers a mechanism that can produce *phonemic restoration effects* (Bashford et al., 1992; Bashford and Warren, 1987; Bashford et al., 1996; Luthra et al., 2021; Miller and Isard, 1963; Samuel, 1981, 1996; Sivonen et al., 2006; Warren, 1970). Phonemic restoration happens when speech stimuli are edited to create gaps. For instance, you might record the word *legislators*, and delete the middle "s" sound. When you play that stimulus with the "s" deleted, people oftentimes notice that there is a gap in the word, and it sounds funny. However, if you insert a noise, like the sound of someone coughing, or even white noise, people experience *phonemic restoration*—they hear the word as if the middle "s" sound were present, as if someone had pronounced *legislators* perfectly. If you put your specially edited word in the middle of a sentence, as in *It wasn't until midnight that the legi(cough)lators finished the bill*, people again hear the word *legislators* as if it had been pronounced perfectly, with the middle "s" sound in its normal place, and they hear the cough as if it happened just before or just after the edited word. (People hear *It wasn't until midnight that the (cough) legislators finished the bill.*) These phonemic restoration effects are stronger for longer words than shorter words, and they are stronger for sentences that are grammatical and make sense than sentences that are ungrammatical or don't make sense. Further, the specific phoneme that is restored can depend on the meaning of the sentence that the edited word appears in. For example, if you hear *The wagon lost its (cough)eel*, you will most likely hear the phoneme /w/ in place of the cough. But if you hear *The circus has a trained (cough)eel*, you will most likely hear the phoneme /s/. Research involving *evoked response potentials* (ERPs) that are created when groups of neurons fire in response to a stimulus show that the nervous system does register the presence of the cough noise very soon after it appears in the stimulus (within about 200 ms).

All of these results suggest that a variety of possible sources of top-down information affect the way the acoustic signal is perceived. Further, they suggest that perception of speech involves analyzing the signal itself as well as biasing the results of this analysis based on how well different candidate phonological interpretations fit in with other aspects of the message. These other aspects could include whether the phonological interpretation results in a real word or not (as in *lean* vs. *leam*), whether the semantic interpretation of the sentence makes sense (as in *I saw them kiss* vs. *I saw them dish*), and how intact the top-down information is (a poorly constructed sentence is less likely to make up for a degraded acoustic signal).

Summary and Conclusions

Speaking requires you to have an idea and it requires you to move your articulators. Sounds simple, but there are a lot of steps you have to take after you have an idea and before it makes it into the world as a set of sound waves. You have to find the right lexicalized concepts in your language, you have to activate the lemma representations that correspond to those lexicalized concepts. Having done that, you have to find the right forms for those lemmas, which involves both morphological and syntactic processing. One you have activated the right set of morphemes and have arranged them in a series, you can start activating sounds that will express your idea. Activating sound codes entails a set of processes that lead to syllabification, where specific activated speech sounds are assigned to specific positions in specific syllables. Having accomplished that much, the syllabified representation is turned over to the motor system, which creates a gestural score that your motor control systems use to signal over 100 muscles that are involved in speech. The final outcome of that process is a set of muscle movements that drive the articulators, which perturb the flow of air coming out of your body and create the characteristic patterns that we perceive as speech.

Understanding speech requires that you register the acoustic pattern created by the movement of the articulators and use it to recover the speaker's intended meaning. Sounds simple, but there are a lot of steps you have to take after you register the presence of a speech stimulus before you can figure out what it means. Coarticulation makes the analysis of the speech signal especially challenging because there are no clear temporal breaks that signal where one phoneme ends and the next one begins, and because the gestures used to produce a phoneme are affected by the preceding and following phonemes. Because the articulators are moving simultaneously, and because the precise nature of the movements for a given phoneme depends on both the preceding and the following phonemes, there is no one-to-one relationship between acoustic signals and phonemes. Motor theory, and its mirror neuron variant, propose that we "see through" the complexity of the acoustic characteristics of speech by using the speech signal to activate representations of the movements (gestures) that created the speech signal. Motor theory advocates propose that speech perception is carried about by a specially functioning and dedicated processing module. According to motor theory, this module leads to special properties of speech perception, including duplex and categorical perception. Mirror neuron advocates point to parts of monkey brains that respond when the monkey makes a gesture (e.g. grasping an object) or sees someone else make the same gesture. Mirror neurons are seen as the vital bridge between perception and production that the motor theory requires.

Critics of motor theory, on the other hand, have shown that speech perception is not "special" as defined by motor theory. Nonhuman animals, like Japanese quail and chinchillas, perceive aspects of speech much the same way humans do; and humans experience duplex and categorical perception for nonspeech sounds. As an alternative to motor theory, some accounts propose that general-purpose auditory processing mechanisms are deployed for speech. The GA approach can explain why nonhuman animals perceive some kinds of phonemes, and why speech has some of the characteristics that it has—such as having perceptual boundaries at specific voice onset times. The FLMP falls within this tradition. It proposes that both signal analysis and stored information influence the perception of any given speech stimulus. Such interactions of bottom-up and top-down information are demonstrated by phenomena like the Ganong effect and different kinds of phonemic restoration. However, the GA approach does not yet constitute a complete theory of speech perception, and so speech perception continues to be actively and intensively researched by language scientists.

TEST YOURSELF

1. What kinds of mental processes do speakers go through prior to articulation?
2. According to the WEAVER++ model, what kinds of representations do speakers activate before they speak? What evidence supports the psychological reality of models such as WEAVER++? What observations suggest that aspects of the WEAVER++ system may not be present in human speakers?
3. Describe the difference between a concept and a lexicalized concept. What roles do each of them play in speech production?
4. What kinds of errors do people make when they speak? What do the errors tell us about the speakers' mental processes?
5. Describe similarities and differences between Gary Dell's spreading activation model and the WEAVER++ model of speech production. What evidence favors each account?
6. Describe the tip-of-the tongue (TOT) phenomenon. What kinds of words are most likely to produce a TOT and why?
7. How is speech perceived according to Liberman's motor theory? What is coarticulation and what role does it play in the theory? What is the McGurk effect and what does it tell us about speech perception? Why do some people believe that motor neurons provide the physical/neural basis for speech perception? Is there anything wrong with the mirror neuron hypothesis?
8. What are the chief theoretical alternatives to motor theory? Why might one prefer these alternatives?

THINK ABOUT IT

1. Try to induce TOT states. Design an experiment (for example, you could compare different kinds of words). Use the definitions from the earlier TOT section or come up with some of your own. Test your classmates or your friends. How often are you able to induce TOT states? Do some kinds of words work better than others? Are your results consistent with the experimental results?
2. Take some time to listen to conversations around you. When two people are conversing, are there similarities in what the two people say or how they say it? What do you think accounts for these similarities?
3. Find a quiet place to work, a partner, and a pencil. Sit so that you and your partner can hear each other but not see each other. Have your partner speak a short list of words, like *pencil, box, toaster, walnut, camera,* and *thing*. Your partner should flip a coin before saying each word. If the coin comes up "heads," your partner should speak the word while holding the pencil between their teeth. (This is kind of like doing a bite-block production experiment.) See if you can hear when your partner has the pencil in their mouth. Which kinds of speech sounds are most affected by the pencil? See if you can figure out why. See if you can determine what cues you are using to figure out when your partner is using a pencil.

Notes

1 Caber (n.): A very large wooden log that is thrown in contests of strength.
2 To be truly complete, the theory would also have to explain how the articulatory apparatus is controlled, but that is a conceptually separate topic. Most theories of speech production are satisfied to let the motor system deal with the actual movements, although some evidence suggests that articulator movement in speech is programmed dynamically each time speech is produced, rather than being controlled by an inventory of precompiled gestural plans. For example, speakers can produce acoustic signals that are within the range of normal variation even if their vocal apparatus is significantly perturbed by bite-blocks or other mechanical methods (e.g. Gay et al., 1981).
3 There is a multi-word expression in English that expresses the concept "tattoo on the lower back"; the German equivalent is *Arschgeweih*, which translates literally as "ass antlers."
4 The process is called resyllabification in some accounts of speech production, but this seems to imply an initial stage of processing in which syllables are tied to individual words and then reorganized, which may not be accurate.
5 Syllable frequency effects also suggest that they are a psychologically real representational unit that participates in production (Levelt and Wheeldon, 1994).
6 At parties, we used to play a profane version of this experiment called "fuzzy duck–ducky fuzz." Oh, the laughter we enjoyed. Good times.
7 See also Moss et al. (1997), for evidence of semantic over phonological priority in lexical access.
8 Ebenezer, sampan, ambergris, philatelist.
9 The precise mechanism that produces "competition" effects is still under investigation. Some accounts favor mutual inhibition within the conceptual and lemma levels (e.g. Dell et al., 1997), while others favor non-inhibitory processes in networks where multiple sources can feed activation to different candidates (e.g. Roelofs et al., 1996).
10 Facilitatory and inhibitory effects depend on the precise timing of the onset of the target picture and the word. Interested readers may wish to consult Griffin and Ferreira (2006) and Levelt (1989).
11 And vice versa in interactive accounts like Dell's spreading activation model (Dell, 1986; Dell et al., 1997) and related accounts (e.g. Cooper and Ferreira, 1999).
12 But see Roelofs et al. (1996) for a strictly feedforward, serial-selection model that can produce mixed errors as well as the lexical bias.
13 In another study, Motley and colleagues showed that slips of the tongue that led to sexually suggestive statements were more frequent when experimental participants were in a sexually charged frame of mind (Motley and Baars, 1979). Participants made more pain-related slips of the tongue when they expected to receive an electric shock.
14 Alvin Liberman, the founder of the modern study of speech perception, argued that we are specially adapted by evolution for just this purpose—to produce and understand coarticulated speech. Otherwise, he argues, we could only talk as fast as we can spell (i.e. really slowly), and communication would suffer.
15 Other names also appear in the literature, including sonogram, sonograph, and spectrograph. This chapter follows Liberman and uses spectrogram.
16 Fowler's direct realist perspective offers a different theory of perception within the motor theory tradition. The chief difference between the two approaches is that the most current version of Liberman's motor theory treats prototype "intended gestures" as being the fundamental units of speech perception, while Fowler believes that the fundamental units are the actual speech gestures that speakers produce (see e.g. Fowler, 2008).
17 In another version of the experiment, the two parts of the stimulus were both played to both ears, but the relative loudness of each component was manipulated. The elided transition began to affect perception at intensities below the detection threshold for the transition when presented in isolation, and produced duplex perception when the intensity of the transition was about 20 dB greater than the base. Liberman and colleagues view both of these effects—sub-threshold effects on phonological perception and duplex perception with large intensity differences between transition and base—as evidence for modular speech processing.
18 Other examples of perceptual variables that can vary continuously are hue in vision (related to color perception) or saturation in gustation (which can lead to gradual changes in taste perception).
19 It would have been much ickier without the gloves.
20 Of course, there is no direct evidence for this hypothesis and these authors have not ruled out the equally likely possibility that modern speech evolved from more primitive systems of vocal signals (e.g. alarm calls rather than manual gestures).
21 But note that while fMRI and other imaging studies also find frontal activity correlated with phoneme comparison and judgment tasks, they often do not find frontal activity for speech perception tasks that do not involve comparison and judgment. Thus, frontal activations may reflect perceptual processes, but they might also reflect working memory processes, executive function, attention, or other subcomponent processes involved in phonological comparisons.

22 See Hickok (2008) for a wide-ranging critique of the mirror neuron theory of action understanding.

23 And I have ...

24 Pinker (1994) objects to these findings because, he argues, the animals require thousands of training trials, while human infants require few or none. But this criticism is really misplaced. While the animals may require many trials to learn the experimental procedure (that they get rewarded for particular behaviors under particular contingencies), they do not in fact need thousands of trials to respond appropriately to a given stimulus after this basic training. While the animals are trained on a specific set of training stimuli, their ability to discriminate phonemes and to compensate for coarticulation generalizes to novel stimuli (that they were not exposed to during the basic training period; see e.g. Kluender and Kiefte, 2006).

25 Of course, it is always possible that there are multiple sets of motor representations for speech gestures, stored in multiple parts of the brain (just as monkeys have multiple somatotopic maps), but motor theory clearly associates speech perception with motor representations stored in the motor strip and adjacent premotor areas.

References

Arieh, Y., and Algom, D. (2002). Processing picture-word stimuli: The contingent nature of picture and of word superiority. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 221–232.

Baars, B.J., and Motley, M.T. (1974). Spoonerisms: Experimental elicitation of human speech errors. *JSAS Catalog of Selected Documents in Psychology, 4*, 118.

Baars, B.J., Motley, M.T., and MacKay, D. (1975). Output editing for lexical status from artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 14*, 382–391.

Bashford, J.A. Jr., Riener, K.R., and Warren, R.M. (1992). Increasing the intelligibility of speech through multiple phonemic restorations. *Perception and Psychophysics, 51*, 211–217.

Bashford, J.A. Jr., and Warren, R.M. (1987). Multiple phonemic restorations follow the rules for auditory induction. *Perception and Psychophysics, 42*, 114–121.

Bashford, J.A. Jr., Warren, R.M., and Brown, C.A. (1996). Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. *Perception and Psychophysics, 58*, 342–350.

Baynes, K., Funnell, M.G., and Fowler, C.A. (1994). Hemispheric contributions to the integration of visual and auditory information in speech perception. *Perception and Psychophysics, 55*, 633–641.

Benson, R.R., Whalen, D.H., Richardson, M., et al. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain & Language, 78*, 364–396.

Binder, J.R., Frost, J.A., Hammeke, T.A., et al. (1997). Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience, 17*, 353–362.

Blackmer, E.R., and Mitton, J.L. (1991). Theories of monitoring and timing of repairs in spontaneous speech. *Cognition, 39*, 173–194.

Blumstein, S.E., Alexander, M.P., Ryalls, J.H., et al. (1987). On the nature of foreign accent syndrome: A case study. *Brain & Language, 31*, 215–244.

Browman, C.P., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201–251.

Browman, C.P., and Goldstein, L. (1990). Representation and reality: Physical systems and phonological structure. *Journal of Phonetics, 18*, 411–424.

Browman, C.P., and Goldstein, L. (1991). Gestural structures: Distinctiveness, phonological processes and historical change. In I.G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception: Proceedings of a Conference to Honor Alvin M. Liberman* (pp. 313–338). Erlbaum.

Brown, A.S. (1991). A review of tip-of-the-tongue experience. *Psychological Bulletin, 109*, 204–223.

Brown, A.S., and McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5*, 325–337.

Buchanan, T.W., Lutz, K., Mirzazade, S., et al. (2000). Recognition of emotional prosody and verbal components of spoken language: An fMRI study. *Cognitive Brain Research, 9*, 227–238.

Bürki, A., Elbuy, S., Madec, S., et al. (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language, 114*, 104–125.

Caltagirone, V.G. (1984). Speech suppression without aphasia after bilateral perisylvian softenings (bilateral rolandic operculum damage). *Italian Journal of Neurological Science, 5*, 77–83.

Caplan, D., and Hildebrandt, N. (1988). *Disorders of Syntactic Comprehension*. MIT Press.

Cappelletti, M., Fregni, F., Shapiro, K., et al. (2008). Processing nouns and verbs in the left frontal cortex: A transcranial magnetic stimulation study. *Journal of Cognitive Neuroscience, 20*, 707–720.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14*, 177–208.

Cooper, C.J., and Ferreira, V.S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 318–344.

Cooper, F.S., Delattre, P.C., Liberman, A.M., et al. (1952). Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America, 24*, 597–606.

Cutting, J.C., and Ferreira, V.S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 318–344.

Dankovičová, J., Gurd, J.M., Marshall, J.C., et (2001). Aspects of non-native pronunciation in a case of altered accent following stroke (Foreign Accent Syndrome). *Clinical Linguistics and Phonetics, 15*, 195–218.

Dekle, D.J., Fowler, C.A., and Funnell, M.G. (1992). Audiovisual integration in perception of real words. *Perception and Psychophysics, 51*, 355–362.

Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Dell, G.S., and Reich, P.A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20*, 611–629.

Dell, G.S., Schwartz, M.F., Martin, N., et al. (1997). Lexical access in normal and aphasic speakers. *Psychological Review, 104*, 801–838.

Diehl, R.L., and Kluender, K.R. (1989). On the objects of speech perception. *Ecological Psychology, 1*, 121–144.

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

Diehl, R.L., Walsh, M.A., and Kluender, K.R. (1991). On the interpretability of speech/nonspeech comparisons: A reply to Fowler. *Journal of the Acoustical Society of America, 89*, 2905–2909.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., et al. (1971). Speech perception in infants. *Science, 171*, 303–306.

El-Zawawy, A.M. (2021). On-air slips of the tongue: a psycholinguistic-acoustic analysis. *Journal of Psycholinguistic Research, 50*, 463–505.

Fadiga, L., Craighero, L., Buccino, G., et al. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience, 15*, 399–402.

Fodor, J. (1983). *Modularity of Mind*. MIT Press.

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3–28.

Fowler, C.A. (2008). The FLMP STMPed. *Psychonomic Bulletin & Review, 15*, 458–462.

Fowler, C.A., and Dekle, D.J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 816–828.

Galantucci, B., Fowler, C.A., and Turvey, M.T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*, 361–377.

Gallese, V., Fadiga, L., Fogassi, L., et al. (1996). Action recognition in the premotor cortex. *Brain, 119*, 593–609.

Ganong, W.F., III. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110–125.

Garrett, M.F. (1975). The analysis of sentence production. In G. Bower (Ed.), *Psychology of Learning and Motivation* (pp. 133–177). Academic Press.

Garrett, M.F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language Production, Vol. 1* (pp. 177–220). Academic Press.

Gaskell, M.G., and Marslen-Wilson, W.D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613–656.

Gauvin, H.S., and Hartsuiker, R.J. (2020). Towards a new model of verbal monitoring. *Journal of Cognition, 3*(1).

Gauvin, H.S., Jonen, M.K., Choi, J., et al. (2018). No lexical competition without priming: Evidence from the picture–word interference paradigm. *Quarterly Journal of Experimental Psychology, 71*, 2562–2570.

Gay, T., Lindblom, B., and Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America, 69*, 802–810.

Gentilucci, M., and Corballis, M.C. (2006). From manual gesture to speech: A gradual transition. *Neuroscience and Biobehavioral Reviews, 30*, 949–960.

Goldstein, E.B. (2006). *Sensation and Perception*. Wadsworth.

Goldstein, E.B. (2007). *Cognitive Psychology*. Wadsworth.

Gow, D.W., Jr., and Segawa, J.A. (2009). Articulatory mediation of speech perception: A causal analysis of multi-modal imaging data. *Cognition, 110*, 222–236.

Griffin, Z.M., and Ferreira, V.S. (2006). Properties of spoken language production. In M.J. Traxler and M.A. Gernsbacher (Eds.), *The Handbook of Psycholinguistics* (2nd ed., pp. 21–59). Elsevier.

Hartsuiker, R.J., and Kolk, H.H.J. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology, 42*, 113–157.

Hickok, G. (2008). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience, 21*, 1229–1243.

Jescheniak, J.D., and Levelt, W.J.M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 824–843.

Jescheniak, J.D., Meyer, A.S., and Levelt, W.J.M. (2003). Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 432–438.

Kempen, G., and Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition, 14*, 185–209.

Kerzel, D., and Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 634–647.

Kluender, K.R., Diehl, R.L., and Killeen, P.R. (1987). Japanese quail can learn phonetic categories. *Science, 237*, 1195–1197.

Kluender, K.R., and Kiefte, M. (2006). Speech perception within a biologically realistic information theoretic framework. In M.J. Traxler and M.A. Gernsbacher (Eds.), *The Handbook of Psycholinguistics* (2nd ed., pp. 153–200). Elsevier.

Kohler, E., Keysers, C., Umilta, A., et al. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297*, 846–848.

Kuhl, P.K., and Miller, J.D. (1975). Speech perception by the chinchilla: Voiced–voiceless distinctions in alveolar plosive consonants. *Science, 190*, 69–72.

Kurowski, K., Blumstein, S.E., and Alexander, M. (1996). The foreign accent syndrome: A reconsideration. *Brain & Language, 54*, 1–25.

Levelt, W.J.M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.

Levelt, W.J.M. (1989). *Speaking: From Intention to Articulation*. MIT Press.

Levelt, W.J.M. (1994). Do speakers have access to a mental syllabary? *Cognition, 50*, 239–269.

Levelt, W.J.M., Roelofs, A., and Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–75.

Levelt, W.J.M., Schriefers, H., Vorberg, D., et al. (1991). The time course of lexical access in speech production: A study in picture naming. *Psychological Review, 98*, 122–142.

Levelt, W.J.M., and Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50*, 239–269.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., et al. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.

Liberman, A.M., Delattre, P., and Cooper, F.S. (1952). The role of selected stimulus-variables in the perception of unvoiced stop consonants. *The American Journal of Psychology, 65*, 497–516.

Liberman, A.M., Delattre, P.C., Cooper, F.S., et al. (1954). The role of consonant–vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs, 68*, 1–13.

Liberman, A.M., Delattre, P.C., Gerstman, L.J., et al. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology, 52*, 127–137.

Liberman, A.M., Harris, K.S., Hoffman, H.S., et al. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*, 358–368.

Liberman, A.M., and Mattingly, I.G. (1985). The motor theory of speech perception revisited. *Cognition, 21*, 1–36.

Liberman, A.M., and Mattingly, I.G. (1989). A specialization for speech perception. *Science, 243*, 489–494.

Liberman, A.M., and Whalen, D.H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4*, 187–196.

Liuzzi, G., Ellger, T., Flöel, A., et al. (2008). Walking the talk: Speech activates the leg motor cortex. *Neuropsychologia, 46*, 2824–2830.

Lohman, A. (2018). Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun-verb conversion pairs. *Journal of Linguistics, 54*, 753–777.

Lotto, A.J., Hickok, G.S., and Holt, L.L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences, 13*, 110–114.

Lovelace, E. (1987). Attributes that come to mind in the TOT state. *Bulletin of the Psychonomic Society, 25*, 370–372.

Luthra, S., Peraza-Santiago, G., Beeson, K.N., et al. (2021). Robust lexically mediated compensation for coarticulation: Christmash time is here again. *Cognitive Science, 45*(4), e12962.

Mackay, D. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology, 86*, 210–227.

MacNeilage, P.F. (1970). Motor control of serial ordering of speech. *Psychological Review, 62*, 615–625.

Mariën, P., Keulen, S., and Verhoeven, J. (2019). Neurological aspects of foreign accent syndrome in stroke patients. *Journal of Communication Disorders, 77*, 94–113.

Mariën, P., Verhoeven, J., Wackenier, P., et al. (2009). Foreign accent syndrome as a developmental motor speech disorder. *Cortex, 45*, 870–878.

Marslen-Wilson, W.D., and Warren, P. (1994). Levels of representation and process in lexical access. *Psychological Review, 101*, 653–675.

Martin, J.G., and Brunell, H.T. (1982). Perception of anticipatory coarticulation effects in vowel–stop consonant–vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576–585.

Massaro, D.W., and Chen, T.H. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review, 15*, 453–457.

Massaro, D.W., and Oden, G.C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1053–1064.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., et al. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology, 2*, 131–157.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–747.

McNealy, K., Mazziotta, J.C., and Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *The Journal of Neuroscience, 26*, 7629–7639.

Meister, I.G., Wilson, S.M., Deblieck, C., et al. (2007). The essential role of premotor cortex in speech perception. *Current Biology, 17*, 1692–1696.

Miller, G.A., and Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2*, 217–228.

Moen, I. (2000). Foreign accent syndrome: A review of contemporary explanations. *Aphasiology, 14*, 5–15.

Moss, H.E., McCormick, S.F., and Tyler, L.K. (1997). The time course of activation of semantic information during spoken word recognition. *Language and Cognitive Processes, 12*, 695–731.

Motley, M.T., and Baars, B.J. (1979). Effects of cognitive set upon laboratory induced verbal (Freudian) slips. *Journal of Speech & Hearing Research, 22*, 421–432.

Motley, M.T., Baars, B.J., and Camden, C.T. (1983). Formulation hypotheses revisited: A reply to Stemberger. *Journal of Psycholinguistic Research, 12*, 561–566.

Motley, M.T., Camden, C.T., and Baars, B.J. (1982). Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 21*, 578–594.

Movellan, J.R., and McClelland, J.L. (2001). The Morton–Massaro law of information integration: Implications for models of perception. *Psychological Review, 108*, 113–148.

Nelson, T.O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

Newman, S.D., and Twieg, D. (2001). Differences in auditory processing of words and pseudowords: An fMRI study. *Human Brain Mapping, 14*, 39–47.

Nooteboom, S. (1973). The tongue slips into patterns. In V. Fromkin (Ed.), *Speech Errors as Linguistic Evidence* (pp. 144–156). Mouton.

Nooteboom, S.G. (1969). The tongue slips into patterns. In A.G. Schiarone, A.J. van Essen, and A.A. van Raad (Eds.), *Leyden Studies in Linguistics and Phonetics* (pp. 114–132). Mouton.

Nozari, N., and Pinet, S. (2020). A critical review of the behavioral, neuroimaging, and electrophysiological studies of co-activation of representations during word production. *Journal of Neurolinguistics, 53*, 100875.

Oden, G.C., and Massaro, D.W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*, 172–191.

Okada, K., and Hickok, G. (2006). Left posterior auditory-related cortices participate in both speech perception and speech production: Neural overlap revealed by fMRI. *Brain & Language, 98*, 112–117.

Oldfield, R.C., and Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology, 17*, 273–281.

Pardo, J.S., and Remez, R.E. (2006). The perception of speech. In M.J. Traxler and M.A. Gernsbacher (Eds.), *The Handbook of Psycholinguistics* (2nd ed., pp. 201–248). Elsevier.

Pinker, S. (1994). *The Language Instinct*. Harper.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition, 77*, 97–131.

Pulvermüller, F., Huss, M., Kheriff, F., et al. (2006). *Motor Cortex Maps Articulatory Features of Speech Sounds. Proceedings of the National Academy of Sciences, 103*, 7865–7870.

Remez, R.E., Rubin, P.E., Berns, S.M., et al. (1994). On the perceptual organization of speech. *Psychological Review, 101*, 129–156.

Rizzolatti, G., and Arbib, M.A. (1998). Language within our grasp. *Trends in Neurosciences, 21*, 188–194.

Roelofs, A., Meyer, A.S., and Levelt, W.J.M. (1996). Interaction between semantic and orthographic factors in conceptually driven naming: Comment on Starreveld and La Heij (1995). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 246–251.

Roelofs, A., Özdemir, R., and Levelt, W.J.M. (2007). Influences of spoken word planning on speech recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 900–913.

Rubin, D.C. (1975). Within word structure in the tip-of-the-tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior, 14*, 392–397.

Sahin, N.T., Pinker, S., Cash, S.S., et al. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science, 326*, 445–449.

Samuel, A.G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474–494.

Samuel, A.G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*, 28–51.

Sato, M., Tremblay, P., and Gracco, V.L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain & Language, 111*, 1–7.

Schwartz, B.L., and Pournaghdali, A. (2020). Tip-of-the-tongue states: Past and future. In A.M. Cleary and B.L. Schwartz (Eds.), *Memory Quirks* (pp. 207–223). Routledge.

Scott, S.K., McGettigan, C., and Eisner, F. (2009). A little more conversation, a little less action: Candidate roles for the motor cortex in speech perception. *Nature Reviews: Neuroscience, 10*, 295–302.

Sivonen, P., Maess, B., Lattner, S., et al. (2006). Phonemic restoration in a sentence context: Evidence from early and late ERP effects. *Brain Research, 1121*, 177–189.

St. Heim, B., Opitz, K., and Friederici, A.D. (2003). Phonological processing during language production: FMRI evidence for a shared production–comprehension network. *Cognitive Brain Research, 16*, 285–296.

Stokes, R.C., Venezia, J.H., and Hickok, G. (2019). The motor system's [modest] contribution to speech perception. *Psychonomic Bulletin & Review, 26*(4), 1354–1366.

Streeter, L.A., and Nigro, G.N. (1979). The role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America, 65*, 1533–1541.

Tourville, J.A., Reilly, K.J., and Guenther, F.H. (2007). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage, 39*, 1429–1443.

Warren, P., and Marslen-Wilson, W.D. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics, 41*, 262–275.

Warren, P., and Marslen-Wilson, W.D. (1988). Cues to lexical choice: Discriminating place and voice. *Perception and Psychophysics, 43*, 21–30.

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392–393.

Watkins, K.E., Strafella, A.P., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*, 989–994.

Whalen, D.H., and Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. *Science, 237*, 169–171.

Wheeldon, L.R., and Levelt, W.J.M. (1995). Monitoring and the time course of phonological encoding. *Journal of Memory and Language, 34*, 311–334.

Wingfield, A. (1968). Effects of frequency on identification and naming of objects. *American Journal of Psychology, 81*, 226–234.

Zatorre, R.J., Evans, A.C., Meyer, E., et al. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science, 256*, 846–849.