Speech Perception

Arthur G. Samuel[1,2,3]

[1]Department of Psychology, Stony Brook University, Stony Brook, New York 11794-2500;
[2]Basque Center on Cognition Brain and Language, Donostia-San Sebastian 20009 Spain;
[3]IKERBASQUE, Basque Foundation for Science, Bilbao 48011 Spain;
email: asamuel@ms.cc.sunysb.edu

Annu. Rev. Psychol. 2011. 62:16.1–16.24

The *Annual Review of Psychology* is online at psych.annualreviews.org

This article's doi:
10.1146/annurev.psych.121208.131643

0066-4308/11/0110-0001$20.00

Key Words

spoken word recognition, auditory language

Abstract

Speech perception has been studied for over a half century. During this time, one subfield has examined perception of phonetic information independent of its contribution to word recognition. Theories in this subfield include ones that are based on auditory properties of speech, the motor commands involved in speech production, and a Direct Realist approach that emphasizes the structure of the information reaching the perceiver. A second subfield has been less concerned with the acoustic-phonetic properties of speech and more concerned with how words are segmented and recognized. In this subfield, there has been a focus on the nature of communication among different levels of analysis (e.g., phonetic and lexical). In recent years, there has been a growing appreciation of the need to understand how the perceptual system dynamically changes in order to allow listeners to successfully process the variable input and new words that they constantly encounter.

Contents

INTRODUCTION 16.2
SPEECH PERCEPTION:
A MULTILAYERED SYSTEM . . . 16.2
Phonetic Features and Phonemes:
Below the Lexicon 16.2
Lexical and Higher Levels 16.6
Nonsegmental Factors 16.11
SPEECH PERCEPTION:
A DYNAMIC SYSTEM 16.13
Statistical Learning of Speech 16.13
Perceptual Learning of Speech 16.15
New Word Learning by Adults 16.18
SPEECH PERCEPTION:
CONCLUSION 16.20

INTRODUCTION

Speech is the primary means by which humans communicate with each other. For more than a half-century, scientists have been studying how listeners understand spoken language. Although we have learned a great deal about how the system works, a great deal is yet to be discovered. This review summarizes some of the major findings in this research area, but it is by no means exhaustive. There are two major sections: The first section describes the "steady-state" system that takes speech input and maps it onto various levels of representation (e.g., phonetic features, phonemes, words). The second section is more concerned with the dynamic nature of this system—how does it get built, and how does it change as a function of the input that it receives? Historically, there have been somewhat separate research domains for speech perception and for spoken word recognition, with the former focusing on processes that operate to decode speech sounds regardless of whether those sounds comprise words. In the most recent article on speech perception in the *Annual Review of Psychology* series (Diehl et al. 2004), the focus was exclusively on the former; readers should consult that review for an excellent discussion of these topics. The current review intentionally blurs the distinction between speech perception and spoken word recognition on the assumption that the purpose of speech perception is to allow the listener to recognize the words produced by a speaker.

Speech perception: the process that transforms speech input into a phonological representation of that input

Spoken word recognition: the process of identifying the word(s) in a stream of speech input

SPEECH PERCEPTION: A MULTILAYERED SYSTEM

Phonetic Features and Phonemes: Below the Lexicon

It seems appropriate to begin at the beginning: When scientists first began to study speech using relatively modern techniques, they observed two apparently related phenomena—categorical perception and the right ear advantage. In these early studies, researchers created sets of syllables in which a particular acoustic parameter was varied in such a way that the syllable at one end of the continuum was heard in one way (e.g., /ba/), and the syllable at the other end in a different way (e.g., /pa/). For simple nonspeech stimuli, varying a parameter this way leads to relatively continuous changes in perception. For example, if one end of the continuum is a 100 Hz tone, and the other end is a 200 Hz tone, with the intermediate items changing in frequency in a systematic way (e.g., 120 Hz, 140 Hz, 160 Hz, 180 Hz), listeners typically hear a gradual change across the continuum; each tone is a bit higher pitch than the one before it. For many speech continua, in contrast, perception seemed categorical: Listeners heard a few items as one category (e.g., /ba/) and then things abruptly changed, with the remaining items heard as the other category (e.g., /pa/) (Liberman et al. 1967). This categorical tendency in perception was strongest for stop consonants, somewhat weaker for other consonants (e.g., fricatives), and weaker still for vowels. Repp (1984) provides a thoughtful and thorough assessment of the literature on categorical perception.

The same patterning across phoneme types was found in dichotic listening experiments, studies in which headphones were used to play one speech sound to the right ear and a

different speech sound to the left ear. Listeners showed a reporting advantage for speech played to the right ear. As noted, the strength of this advantage mirrored the ordering in categorical perception studies, with a strong asymmetry for stop consonants but not for vowels (Shankweiler & Studdert-Kennedy 1967). Since the right ear has stronger connections to the left hemisphere of the brain, and language is generally processed on the left side, the right ear advantage was taken as an index of specialized language processing. Building from these two phenomena and from the fact that there did not seem to be any invariant acoustic cues that corresponded to each phoneme, Liberman et al. (1967) proposed that there was a special processor for speech sounds, different from the neural mechanisms that were involved in perceiving other sounds. This special processor was designed to extract the speaker's intended gestures that were used to produce the speech. This approach was called a Motor Theory of speech perception, as it focused on the motor commands in production rather than on the acoustic cues that were present in the auditory stream.

Proponents of the Motor Theory reported additional phenomena that were taken to support the theory, including (*a*) trading relations, (*b*) compensation for coarticulation, and (*c*) duplex perception. A number of studies reported "trading relations" among multiple acoustic cues that could be accounted for as deriving from the same motor commands. For example, prior research had shown that place information can be provided by both the spectral distribution of the "burst" at the onset of a consonant and by the patterns of the formant transitions of the consonant into the following vowel. Dorman and colleagues (1977) demonstrated that these two cues can trade off against each other—a more extreme version of one cue can make up for a weaker form of the other in signaling the consonant. They argued that this perceptual equivalence is difficult to account for from an acoustic perspective but is just what one would expect if both of these cues reflect the same motor plan for the consonant.

Work on compensation for coarticulation began with studies by Mann and her collaborators (e.g., Mann & Repp 1981). Early speech perception research had revealed that consonants and vowels are not produced independently of the sounds around them. Instead, segments are coarticulated, and this coarticulation is a major source of acoustic-phonetic variation. For example, the place of articulation for a /d/ will be somewhat different if the segment that precedes it is articulated near the front of the mouth (such as /l/), or further back (such as /r/)—the point of tongue contact on the roof of the mouth gets pulled a bit forward or back by the preceding sound's position. Because the major difference between /d/ and /g/ is that /g/ is articulated further back than /d/, the coarticulatory influences of a preceding sound like /r/ can make a /d/ more /g/-like than it is in the context of a preceding /l/. Listeners are sensitive to these effects; the phonetic boundary between /d/ and /g/ is shifted in a way that "compensates" for the articulatory drift. The apparent articulatory basis for this effect was taken to support the view that listeners are sensitive to the gestural properties underlying speech.

Another phenomenon that was taken to support the Motor Theory was "duplex perception," first reported by Rand (1974). Duplex perception occurs when a synthetic syllable is broken into two pieces, and each piece is presented to one ear. For example, Rand presented the second- and third-formant transitions of a stop-consonant syllable to one ear and the remainder of the syllable to the other ear. Under these conditions, listeners have two simultaneous percepts. They hear the isolated transitions as what they are—chirping (nonspeech) sounds—but they also hear the speech syllable with the place of articulation specified by those transitions. In other words, the transitions are simultaneously producing a nonspeech percept and a speech percept, consistent with the notion that two different processors are using this input—one processor for speech and one for other sounds.

Of course, the Motor Theory did not go unchallenged. At least two alternative views were

Motor Theory: a theory that asserts that speech perception relies on the same representations and processes used to produce speech

developed, and the debate has been ongoing for over two decades. One alternative was also a type of motor theory, but it differed from the traditional Motor Theory in at least two critical ways. Fowler (e.g., 1986, 1991) has been the leading proponent of this alternative, and she has drawn important distinctions between her approach and the classic Motor Theory. Most importantly, Fowler's overarching theoretical perspective is the Direct Realist view that is most closely associated with Gibson (1966). In Gibson's analysis, objects and events in the environment structure the media that impinge on a perceiver's perceptual system, and the structuring of the medium is rich enough to allow the perceiver to directly perceive the objects and events that caused the structuring. In the case of speech, the articulatory movements structure the air pressure variations that reach the perceiver's ear; in a Direct Realist view, the perceiver thus perceives the gestures that structured the medium. This view differs from classic Motor Theory in terms of the objects of perception: Here they are the actual gestures that structure the air pressure variations, whereas proponents of standard Motor Theory found that they needed to retreat back to the intended gestures rather than to the gestures themselves. Perhaps a deeper difference between the two views is that Direct Realism is a very general theory of perception (Gibson developed it primarily with respect to how people navigate by using patterns in the optic array). As such, no special speech processor is invoked—the same principles apply to all sounds that structure the medium.

Fowler & Rosenblum (1990) provided an elegant demonstration of their position by creating a nonspeech version of the duplex phenomenon. Proponents of the Motor Theory had taken duplex perception as evidence that listeners have two separate sound-processing subsystems, with one of them dedicated to processing speech sounds. Fowler and Rosenblum recorded the sound of a metal door slamming shut and then filtered this sound into its higher and lower frequencies. With the high frequencies removed, the sound was generally still heard as a door closing (or more generally, as the sound of an impact of some sort), but this duller sound was more like what a wooden door being closed sounds like. The high-frequency part by itself was heard as the kind of sound that comes from shaking small metal pellets in a cup. Fowler and Rosenblum used the "wooden door" and "shaking sound" in a dichotic paradigm modeled on what had been done with the speech syllable base and the formant chirps and produced essentially the same pattern of results: If the high frequency (shaking) sound was presented to one ear and the lower frequencies (wooden door) to the other, listeners often reported hearing both the shaking sound and a metal door. Note that this nicely matches the speech case in which listeners report both the chirp (shaking sound) and the full syllable (metal door). Clearly, no one would suggest that listeners have a special sound processor for slamming doors. As such, Fowler and Rosenblum argued that both the original demonstration and their own results should be understood in terms of a perceptual system that used the structured pattern of sound to directly perceive the events (articulators moving or metal doors slamming) that had imposed the structure.

Although the nonspeech duplex demonstration was clearly problematic for the Motor Theory, it was not specifically supportive of the Direct Realist view. That is, although the results were predicted by that theory, they were also consistent with a more general perspective in which perception is driven by the acoustic input rather than by either a special speech processor or by a mechanism that directly perceives the precipitating events. Diehl and colleagues (2004) describe the "General Auditory" approach as one in which speech sounds are perceived with the same mechanisms that are used in the perception of sounds more generally. In this view, there is no role for gestures in speech perception—speech is just another (important) environmental sound that maps onto whatever the representations are for sounds in general.

The General Auditory view was advocated in many papers that appeared as a response to the

"speech is special" perspective of the original Motor Theory. Like the Fowler & Rosenblum (1990) demonstration of a nonspeech case of duplex perception, much of this work was designed to show that there was no special status for speech sounds. For example, a number of studies undercut the idea that categorical perception implicated a special speech processor. These demonstrations took several forms. Pisoni & Tash (1974) showed that perception of stop consonants was not as categorical as had been claimed. These authors measured reaction times taken to identify members of a speech continuum and found that reaction times increased as tokens got closer to the phoneme boundary. This result is at odds with the notion that all tokens within a category are perceived equivalently, the fundamental idea of categorical perception. Similarly, Samuel (1977) showed that if listeners were given extensive training with feedback, they could discriminate within-category members of a test continuum varying in stop consonant voice-onset time.

These demonstrations of less-than-categorical perception of stop consonants were complemented by demonstrations that nonspeech continua can produce discrimination results that looked exactly like the ones for stop consonants. Miller and coworkers (1976) provided one such case. These authors constructed a set of nonspeech stimuli that were modeled on the voice-onset time continua that had been used to show categorical perception of voicing for stop consonants. The nonspeech stimuli varied in the relative onset time of a noise burst versus a buzzing sound. Both of these pieces were clearly not speech, but they were chosen to be somewhat analogous to the noise burst of a stop consonant onset and the voicing of a following vowel. Miller et al. (1976) ran identification and discrimination tests using their set of "noise-buzz" stimuli that mirrored the tests used to show categorical perception of speech and obtained a pattern of results that matched that for speech. From a General Auditory perspective, these results indicate that categorical perception is tied to certain complex acoustic patterns rather than to the perception of intended or actual speech gestures.

A central tenet of the Motor Theory was the existence of a special speech processor that listeners use to decode speech sounds. Clearly, it would make no sense to postulate such a special speech processor for animals other than humans. However, there have been a number of demonstrations that chinchillas, monkeys, and quail can all produce perceptual results that seem similar to those that have been attributed to a special speech processor in humans. For example, Kuhl & Miller (1978) trained chinchillas to differentiate the endpoints of a voice-onset time continuum, and when they tested them on contrasts across the continuum, they found discrimination functions that were quite similar to those found for human listeners. Another interesting result comes from Kluender et al. (1987). Those favoring the Motor Theory had taken support for this position from the fact that human listeners treat the first sound in "dee," "dih," "doo," "dah," etc., as all sounding the same – as /d/. This is actually somewhat surprising from an acoustic perspective because there is very little acoustic overlap among some of these sounds. In contrast, the motor commands to produce a "d" in these different contexts have much in common, as they all involve positioning the tongue tip on the alveolar ridge. Kluender et al. (1987) found that when Japanese quail were trained to respond to certain syllables that began with /d/ (and to refrain from responding to syllables beginning with /b/ or /g/), the quail showed the same sort of generalization to /d/ in other vowel contexts that humans show. As noted, quail should not have a special speech processor. If they produce this kind of generalization without one, presumably we need not invoke one for such generalization by humans.

The debate among those favoring each of the three perspectives (Motor Theory, Direct Realism, and General Auditory) continues. Galantucci and colleagues (2006) have offered a widely read discussion of the competing views in which they ultimately favor motor involvement in speech perception, but not a

special system; this is most consistent with the Direct Realist position. The advocates of a motor component in speech perception have in recent years drawn heavily on the discovery of mirror neurons." Di Pelligrino et al. (1992) had been investigating an area of motor cortex in monkeys that was involved with controlling hand movements. They noticed that certain cells fired not only when the monkey was going to make a particular hand movement but also when the monkey observed a human hand making that movement. These neurons were dubbed "mirror neurons" because they fire both when the monkey produces an action and when the monkey observes that action. Mirror neurons have been the source of intense study and controversy, as they appear to provide evidence for the kind of direct link between perception and production that is hypothesized by both classic Motor Theory (which posits activation of intended gestures during perception) and by Direct Realism (which posits perception of the events—the gestures—themselves) (but see Lotto et al. 2009 for a critique of the potential role of motor neurons in speech perception).

There have also been continuing research and debate that focus on the compensation for coarticulation phenomenon. Advocates of both the Direct Realism and General Auditory views have pursued this effect. The effect fits very naturally into the Direct Realist idea that listeners are perceiving speech gestures, and Fowler et al. (2000) conducted a study to emphasize this view. They had subjects identify members of a /d/–/g/ test continuum, in the context of a preceding sound that was designed to be midway between /l/ and /r/. Their new manipulation was to accompany the ambiguous context sound by a video that was unambiguously /l/ or /r/. The idea is that the video provides information about the gestures for the ambiguous sound and that if people are sensitive to gestures then the video should drive the identification of the following /d/–/g/ sounds. This is what was found. However, Holt et al. (2005) provided evidence that this effect was actually due to differences in the video accompanying the /d/–/g/ sounds, not the context /l/ or /r/. In addition, Holt and her colleagues (e.g., Holt 2006) have shown that shifts in the identification of /d/–/g/ test items can be generated by simply preceding them with pure tones that vary in frequency. Such tones reflect frequency differences in /l/ and /r/ but clearly have no articulatory basis. These findings have been taken by Holt and her colleagues to be evidence for the General Auditory view of speech perception. As this review should make clear, there is continuing debate among those who advocate for Motor, Direct Realist, and General Auditory theories. Interested readers should consult papers by Diehl et al. (2004) and by Galantucci et al. (2006).

Lexical and Higher Levels

The preceding section describes five phenomena that have played a critical role in the development and testing of theories of how the acoustic signal gets mapped onto some kind of phonetic code: categorical perception, the right ear advantage, trading relations, duplex perception, and compensation for coarticulation. The first two phenomena were foundational in the development of the Motor Theory, and the last three have provided a test-bed to choose among the Motor Theory, the Direct Realist view, and the General Auditory account. At about the same time that these last three phenomena were being established, three other phenomena were reported that focused attention on the importance of higher-level information in speech perception and spoken word recognition: phonemic restoration (Warren 1970), the McGurk effect (McGurk & MacDonald 1976), and the Ganong effect (Ganong 1980). As with the five phenomena discussed above, these three effects have played an important role in attempts to distinguish between competing theories.

Warren (1970) introduced and explored the phonemic restoration effect. To produce this effect, a small piece of speech (typically, one phoneme and its transitions to adjacent phonemes) was cut out of a word, and a sound such as a cough or white noise replaced the missing speech. Warren played a sentence with a cough replacing one phoneme in a word

and asked listeners to report the location of the replacement. Performance was quite poor on this task, indicating that people failed to notice that the speech was missing. Warren called this "phonemic restoration" because listeners seemed to perceptually restore the missing speech. Samuel (1981, 1996) created stimulus sets in which half of the words had phonemes replaced by noise, while half were intact with noise simply superimposed on the corresponding phoneme. This procedure allowed signal detection measures to be used as an index of perceptual restoration, and the results were consistent with Warren's suggestion that listeners perceptually restored the missing speech. The restoration was stronger when the signal manipulation was done in real words than in pseudowords, demonstrating that the perceptual system can use higher-order (lexical) information to help repair degraded speech input. This ability is a valuable adaptation given that speech is usually heard under less-than-optimal listening conditions.

Ganong (1980) demonstrated a similar tendency for word recognition processes to use lexical information to guide perception when the speech signal is underspecified. He created sets of stimuli that were based on speech segments that were constructed to be ambiguous. Consider, for example, a sound that has been designed to be acoustically intermediate between /d/ and /t/. Ganong created a pair of test items that began with this ambiguous stop consonant, with one member of the pair having "ask" appended, and with the other member of the pair ending in "ash." He found that listeners generally reported hearing "task" rather than "dask" and "dash" rather than "dask." In each case, perception of the ambiguous segment is biased toward a sound that produces a word rather than a nonword. The Ganong effect has been used to compare lexical influences on phonemic perception to those produced by sentence context (Connine & Clifton 1987) and to examine how lexical activation builds from the beginning of a word to its end (Pitt & Samuel 1993).

Phonemic restoration and the Ganong effect demonstrate that acoustic-phonetic encoding cannot be fully understood independent of lexical context. There is also a substantial literature showing that acoustic-phonetic processing is significantly affected by visual information when the perceiver can see the speaker's face. At a general level, it had been known very early on that seeing the speaker's mouth can improve speech recognition scores (Sumby & Pollack 1954). A particularly striking effect of this visual information was provided by McGurk & MacDonald (1976). Their procedure involved showing a headshot of someone producing simple syllables. The audio track was dubbed to create a mismatch between what the video showed and the sound that was presented. For example, a video of the face producing /ga/ was paired with an audio recording of /ba/. Under these circumstances, listeners often reported hearing /da/, a kind of compromise between the visual and auditory input streams. Since this initial demonstration of the McGurk effect, many studies have explored how these two sources of speech information get combined (e.g., Massaro 1987). Phonemic restoration, the Ganong effect, and the McGurk effect all show that speech perception cannot be understood solely in terms of the mapping between an acoustic (or gestural) event and a phonetic percept; speech perception is also guided by additional information sources available to the perceiver.

Models of spoken word recognition vary in the way that they incorporate the need to use both the acoustic signal and the other relevant sources of information. The most influential early approach was Marslen-Wilson's (1975, 1987; Marslen-Wilson & Welsh 1978) Cohort model. Marslen-Wilson (1975) presented listeners with recorded passages and had them repeat (shadow) the input as close in time as they could. Some of the subjects could stay within a few hundred milliseconds of the input, which implied extremely rapid recognition. Marlson-Wilson suggested that the first 150–200 msec of a word could be used to access lexical representations that were consistent with the input. He introduced the notion of lexical activation of these representations, and

the cohort was the set of active representations. For accurate recognition it was then necessary to winnow this cohort down to the correct item, and Marslen-Wilson suggested that words dropped out of the cohort for two reasons: Items were deactivated if subsequent acoustic information was inconsistent with them, and they were deactivated if contextual information was inconsistent.

The core ideas of the Cohort model have received a great deal of empirical support, especially the notion that multiple lexical representations become activated as a function of the initial information in the word. Zwitserlood (1989), for example, demonstrated the activation of multiple lexical representations by the first few hundred milliseconds of a word in a sentence. She had listeners make lexical decisions to printed words that appeared at certain times in the speech stream and found semantic priming of these words. Critically, the priming was found not only for words related to the actual word in the sentence but also to words related to what the listener had heard by the time the probe word was shown. For example, a listener hearing the word "bulletin" could get a visual probe word like "gun" (related to "bullet") at the end of the first syllable, and the response to such a probe would be facilitated relative to an unrelated control word. Such priming effects indicate that the initial three phonemes activated not only the actual word being said but also others in its cohort.

Allopenna and colleagues (1998) developed an eyetracking methodology that provides converging evidence for the activation of multiple lexical candidates. In these experiments, participants saw displays of four pictured objects and heard instructions that concerned one of them (e.g., "move the beaker"). On critical trials, the display included both the target picture and an item with the same speech onset (e.g., a beetle). The eyetracking results indicated that as listeners heard the unfolding speech, they tended to examine both pictures that were consistent with the input ("bee…"), eventually focusing on the one that matched the full word. Thus, as in other paradigms, it appears that the speech input activates multiple consistent lexical candidates.

An interesting prediction of the Cohort model is that the recognition of a word should depend not only on the word itself but also on its relationship to other words in the lexicon. This prediction follows from the fact that the cohort of words competing for recognition is determined by how many words share an onset pattern; some words have many such onset-matched competitors (e.g., "extinguish"), whereas others are subject to much less initial competition (e.g., "establish"). Marslen-Wilson & Welsh (1978) argued that these two cases differ in their "uniqueness points"—the moment when only one real word is consistent with all of the input received up to that point. A number of studies have shown that word recognition takes place sooner for words with early lexical uniqueness points than for those with late uniqueness points. For example, Gaskell & Marslen-Wilson (2002) found significantly stronger repetition priming and semantic priming for "early unique" words compared to "late unique" words, suggesting that the former produce faster and stronger lexical activation, yielding the stronger priming effects.

Pitt & Samuel (2006) used the Ganong effect to examine the dynamics of lexical activation, including the effects of differences in the location of uniqueness points. Listeners heard stimuli that all ended in either "s" or "sh" and responded to each stimulus by identifying the final sound as "s" or "sh." The stimuli were either monosyllables (e.g., "miss," "wish") or trisyllabic words (e.g., "arthritis," "abolish"), and each word was used to make an eight-step continuum that ranged between final "s" and final "sh" (e.g., "miss-mish," "wiss-wish"). Recall that the Ganong effect is the tendency for listeners to report ambiguous items with a lexical bias, so that with equivalent acoustics, the final sound after "arthriti_" should generate more "s" report than the final sound after "aboli_". Pitt and Samuel found that not only was this the case, but the long words also generated a much stronger Ganong effect than the short words. This is consistent with the fact that the long

words provide both more acoustic evidence and more time for lexical activation to build before the final segment is heard. In addition, long words with early uniqueness points produced stronger lexical shifts than did long words with later uniqueness points, again consistent with the idea that an early uniqueness point provides faster lexical activation that can build without further competition from similar words.

After Marlsen-Wilson's influential presentation of the Cohort model, two types of models became the focus of a great deal of research on spoken word recognition. Both types had much in common with the Cohort model and its ideas of activation of lexical and sublexical representations. Most models share the view that speech is encoded at multiple levels of analysis, and in most cases these levels include some kind of phonetic features, some kind of sublexical units (most often phonemes), and some form of lexical representations. A major division among models has been whether the flow of processing is seen as entirely bottom-up (features → phonemes → words) or is instead viewed as more interactive, with activation at a "higher" level (e.g., lexical) allowed to influence activation at a "lower" level (e.g., phonemic). Norris (1994) and his colleagues (Cutler & Norris 1979, Norris et al. 2000) have developed a number of models (including Race, Shortlist, and Merge) that were designed to account for spoken word recognition without allowing any perceptual top-down influences (Massaro's Fuzzy Logical Model of Perception, e.g., Massaro 1989, also rejects such top-down influences). They have argued that such influences cannot help perception and that they have the potential to hurt performance. In their presentation of the Merge model, Norris et al. (2000) reviewed most of the literature that seemed to implicate top-down effects and argued that all such effects can be seen as coming from postperceptual decision processes. In their view, there are two separate streams of perceptual analysis, one that produces a set of phonemic codes and one that produces a lexical output. They argue that after these perceptual processes produce their results, the outputs can be merged for decision-making purposes but that there is no perceptual effect of one on the other.

The opposing view has been presented in two prominent models, the TRACE model (McClelland & Elman 1986) and Grossberg's Adaptive Resonance Theory (e.g., Grossberg 1980). These models feature interactive architectures in which activation of a unit at any level of the system increases the activation of other units at other levels that are consistent with it. Thus, in these models, partial information (e.g., "exting. . .") can activate a lexical representation (e.g., "extinguish"), and the activated lexical representation can in turn increase the activation of sublexical representations (e.g., ". . .sh") that are consistent with the activated lexical representation. In the TRACE model, there are phonetic features, phonemes, and words—the levels are fixed. An attractive property of Adaptive Resonance Theory is that it makes no a priori commitment to units of any particular grain. Instead, it assumes that "chunks" get represented to the extent that a listener is exposed to them and that these chunks are whatever size they happen to be. This approach avoids a number of problems that come with assuming the existence of particular units, such as phonemes.

Empirically, it has proven to be extraordinarily difficult to distinguish between purely bottom-up models and ones that posit interactivity. As Norris et al. (2000) pointed out, if there is a postperceptual merging of lexical and phonemic results, most apparent top-down perceptual effects cannot be unambiguously established. There is, however, one class of effects that may be very difficult to account for without interactivity, even allowing postperceptual merging of information. This class includes studies in which an opportunity is provided for top-down lexical influences on phonetic processing, but with the listeners making no judgments about the resulting percept. Instead, the test is for a consequential effect: If there had been a top-down effect that influenced a phonetic code, is there some effect of that phonetic code on the perception of

something else? Because there is no (potentially postperceptual) judgment made for the direct phonetic percept, this procedure avoids the interpretational ambiguity that Norris et al. (2000) identified in most procedures.

A number of studies meet this criterion. These studies have built upon three of the phenomena that were discussed above: the Ganong effect, phonemic restoration, and compensation for coarticulation. In a pair of studies, Samuel used an additional phenomenon—selective adaptation—in combination with phonemic restoration (Samuel 1997) and the Ganong effect (Samuel 2001) to produce consequential tests. Eimas & Corbit (1973) first applied the selective adaptation procedure to speech stimuli. The procedure is conceptually similar to the familiar phenomenon of color aftereffects (after looking at a red patch for about 30 seconds, a person will see a green aftereffect when looking at a white background). For speech, Eimas and Corbit first had listeners identify members of speech continua. They then played the listeners a series of adaptation sequences alternating with identification of members of the speech continuum. The adaptation sequences contained many repetitions of a continuum endpoint. For example, if people first identified members of a /ba/–/pa/ continuum, they would hear either /ba/ or /pa/ as adaptors. They found the same sort of contrastive effect seen with colors: After hearing /ba/ many times, listeners were more likely to hear /pa/ than /ba/; if /pa/ was the adaptor, the reverse occurred. In many subsequent studies, researchers demonstrated the general principle that hearing a particular phoneme repeatedly reduces later report of that phoneme and ones similar to it. Samuel (1997) tested whether the phonemes that people seem to hear via phonemic restoration are really perceived or are instead just the result of merging lexical information with phonetic information postperceptually. The test was to use adaptors that consisted of words with phonemes that had been replaced by white noise. If the restored phonemes were really perceived, then they should produce adaptation—they should affect identification of test syllables that contained the same phoneme that would have been restored. In one condition, the white noise replaced /b/, and in a second condition, the white noise replaced /d/. The two conditions produced opposite adaptation effects on the identification of a test continuum that ranged between /b/ and /d/, as predicted by interactive models. Samuel (2001) conducted a similar study, but rather than using phonemic restoration of /b/ or /d/, the Ganong effect was used to influence the perception of an ambiguous fricative sound. Again, lexical context was used to influence which phoneme should be heard, and again differential adaptation effects were found. In both studies, it is difficult to see how these effects can be accounted for in a model that does not allow lexical information to affect the perception of a phoneme because in both cases the listeners made no judgments at all about the adaptors themselves.

The other type of consequential test that supports the interactive view is based on the compensation for coarticulation phenomenon. Recall that compensation effects occur when one sound affects the perception of a following sound (e.g., a sound on a /d/–/g/ continuum will be perceived differently if it is preceded by /l/ than if it is preceded by /r/). Elman & McClelland (1988) used this effect, combined with the Ganong effect, to test for interactivity. They had listeners identify members of /t/–/k/ or /d/–/g/ test series, both of which are subject to compensation for coarticulation. The sound that immediately preceded these items was an ambiguous mixture of "s" and "sh", and this ambiguous mixture either occurred as the final sound of a word that ends in "s" (e.g., "Christmas"), or in "sh" (e.g., "foolish"). Due to the Ganong effect, the ambiguous sound would be reported as "s" in the first case, but as "sh" in the second case. The question, however, is whether this report is based on perception or on a postperceptual merging of phonetic and lexical codes. If it is perceptual, then it should generate compensation for coarticulation, affecting the identification of the following sound. Elman and McClelland found such a

shift and concluded that the lexical bias on perception of the ambiguous fricative was perceptual. This interpretation was challenged by Pitt & McQueen (1998), who argued that the effect was not actually based on lexical influences but was instead due to confounded phonotactic probabilities (i.e., the probability of particular sequences of phonemes, regardless of lexicality). They demonstrated that the compensation effect does in fact vary with phonotactic probability. Subsequently, however, two studies that controlled for phonotactic probability (Magnuson et al. 2002, Samuel & Pitt 2003) demonstrated a lexical influence on compensation for coarticulation, providing further evidence from a consequential paradigm for interactive processing in speech perception.

Nonsegmental Factors

Most of the studies discussed above focused on how listeners map the acoustic input onto phonetic features, phonemes, and words because most of the literature has assumed units of these types. However, we have already seen that there are additional factors, such as coarticulation and phonotactic probabilities, that clearly affect speech perception. These can be considered as nonsegmental factors because they do not apply directly to a given segment. Additional nonsegmental factors have also been the subject of substantial research, including prosodic influences, such as lexical stress and lexical tone, and indexical factors (which refer to aspects of a word's pronunciation that are tied to the way that a particular speaker produced it). In addition, word recognition is of course better when the word is presented in a predictable sentential context than when it does not have such support.

Prosodic factors in word recognition. Substantial literatures exist on the role of lexical stress in word recognition and speech segmentation and on the role of lexical tones in word recognition in tone languages such as Mandarin. The segmentation problem was recognized in very early research on speech perception: The words in a spoken sentence do not come with the neat "white space" separation of text, and in fact they typically are not separated at all—the ending of one word blends into the beginning of the next word. Thus, "a nice bag" and "an ice bag" could have the same acoustic realization, with the /n/ run together with the preceding and following vowels in both cases. Cutler & Norris (1988) suggested that in some cases listeners could use stress cues to aid segmentation, because in languages like English and Dutch, most polysyllabic words have stress on their first syllable; in some languages (e.g., Finnish and Hungarian), the predictability is virtually perfect. This distributional information could in theory allow the perceptual system to place a word break just before the stressed syllable, and Cutler and Norris demonstrated that listeners do in fact seem to use this statistical information in segmentation. This idea, called the Metrical Segmentation Strategy (MSS), has been validated in a number of studies (e.g., Norris et al. 1995).

In a series of experiments, Mattys and coworkers (2005) have examined various cues for segmentation, including stress. In each experiment, they produced stimuli in which two possible cues would work against each other, to see which was more important. For example, they constructed stimuli in which lexical stress cues would favor one interpretation of a stimulus, whereas phonotactic cues would favor an alternative. Importantly, they tested the different cues under both clear and noisy listening conditions. The experiments led them to propose that there is a hierarchy of cues, with different cues dominating under different conditions. In particular, when listening conditions are good, sentential and lexical cues dominate word recognition; phonotactics are of moderate value, and stress is least useful. However, when listening conditions are poor, the relative value of the different types of cues reverses. They note that in many of the studies that have provided support for the MSS, stimuli were presented under noisy or otherwise difficult conditions, exactly the cases in which stress cues are most valuable.

Speech segmentation: the process that imposes boundaries between successive linguistic units (most commonly, words)

As noted above, many languages employ another important nonsegmental aspect of spoken word recognition—lexical tone. There are hundreds of tone languages, including Mandarin, one of the most widely used languages in the world. In tone languages, the identity of a word is not specified by segmental information alone but is instead determined by the combination of segmental information with a particular prosodic pattern. For example, in Mandarin there are four basic tones, and a given segmental sequence can correspond to four different words, depending on which of the four tones it is produced with. Some interesting studies have examined the relationships among segmental cues, lexical stress, and lexical tones. For example, based on tests of speech production, Levelt (1993) has suggested that speakers access the stress pattern of a word before the segmental information. Chen's (2000) examination of speech errors in a corpus of calls to a radio program supported essentially the same conclusion for a word's tonal pattern—it appears to be accessed first and to serve as a kind of frame for the vowels and consonants. Chen found that when a segmental error was produced within a word, the tone pattern was generally intact, whereas when a whole word was erroneously produced, the tonal pattern was also a mismatch to the intended word's pattern. Zhou & Marslen-Wilson (1995) provide a model of spoken word recognition in Mandarin that incorporates both segmental and tonal layers.

Note that Levelt's (1993) and Chen's (2000) conclusion that prosodic processing precedes segmental was based on errors in speech production. Soto-Faraco et al. (2001) examined whether the precedence of prosodic information over segmental information appears on the perceptual side as well. Their test was done in Spanish, which employs lexical stress. Participants heard sentences that ended in truncated words and made lexical decisions about visually presented targets that were presented at the offset of the truncated word. The sentences were designed to provide a normal prosodic structure (which was needed to test the stress manipulation), but not to provide any semantic cues to the target words. The key manipulation was the intentional mispronunciation of some aspect of the truncated word in terms of either its segmental pattern (a vowel or consonant) or its stress pattern. Lexical decision times for visually presented words showed benefits for prime matching and costs for prime mismatching for both the segmental and stress cases. Soto-Faraco et al. (2001) therefore concluded that in Spanish, lexical stress information behaved identically to segmental information in spoken word recognition. They suggested that all information (segmental and prosodic) that is relevant to lexical access is used when it is available. Given the variable results in the literature, they argued that the observed pattern will depend on the nature of the task used in a given study.

Indexical factors in word recognition. As noted above, although there has been considerable controversy about the nature of information flow during spoken word recognition, there has been relative unanimity about the existence of phonetic features, a phoneme-like level, and lexical representations. An implicit assumption of this view is that there is substantial abstraction in processing—the considerable variation across different productions of the same utterance is "cleaned up" in a way that permits a comparison between relatively normalized phonemes and those stored as the representation of a given word. Thus, if the lexical representation of "pet" includes something like /p/ + /E/ + /t/, the word produced by a given speaker must be mapped onto such a string, abstracting away details of pitch, rate, etc.

In this context, work by Goldinger (1996) and by Nygaard et al. (1994) was rather surprising. These studies included experiments in which some of the supposedly discarded details of a particular speaker's productions turned out to affect performance over a time scale that was not consistent with prior views. For example, Goldinger presented listeners with a set of monosyllables that had been produced by a number of different speakers, with a given

listener receiving words from between two and ten different speakers. The initial task was to type in the word that was heard. Six different groups were tested in a second session that was either conducted five minutes, one day, or one week after the initial session, with the second session either involving a recognition test (explicit memory) or a perceptual identification test (reporting words that were presented in a heavy masking noise, a measure of implicit memory). In all cases, the words tested in the second session were either ones that were presented in the same voice as they had been in the first session, or in a different voice. Goldinger (1996) found significantly better word recognition for words presented in the same voice, both in the immediate test and after one day, but not after one week. The same-voice advantage remained significant at all three testing intervals on the implicit memory task.

These results, along with converging evidence from other studies (e.g., Nygaard et al. 1994), led Goldinger (1998) to propose an episodic model of lexical representation. In his model, each time a word is heard a memory trace is stored that includes not only segmental information but also some aspects of the particular realization, the latter being the "indexical" properties of a token that reflect the speaker's voice, mood, speaking rate, etc. Recently, Cutler (2008) and colleagues (McQueen et al. 2006) have argued that episodic models are inherently incapable of accounting for many speech phenomena (see below); abstraction is necessary. In fact, Goldinger has conducted a good deal of previous work that was in the abstractionist tradition (e.g., Goldinger et al. 1989), and as such, it is not surprising that he has recently (Goldinger 2007) suggested that models must include both abstraction and episodic components. This position is shared by Cutler and her colleagues (e.g., Cutler & Weber 2007), suggesting that a new consensus is forming. Taken together, recent research thus indicates that the representations underlying spoken word recognition include (abstract) segmental codes along with prosodic and indexical information.

SPEECH PERCEPTION: A DYNAMIC SYSTEM

In the preceding discussion, the system that supports speech perception in the service of recognizing spoken language has been portrayed as though it is a finished product, optimized to deal with the speech a listener may encounter. The problem with this characterization is that people do not have the luxury of operating in such a static, homogeneous speech environment. Instead, they are constantly being exposed to new speakers with varying dialects and accents, and they encounter a surprisingly large number of new words all the time. Thus, it is essential that speech perception processes can operate dynamically, learning and adjusting as a function of the input that they receive. A number of literatures indicate that this is exactly what the system does. Three of these literatures are considered here: (*a*) studies looking at how the system extracts information about the structure of speech (e.g., units and their ordering) on the basis of the statistical properties of the input; (*b*) studies examining how the system adjusts in the face of phonetic variation; and (*c*) studies exploring how adults add new words to their mental lexicon.

Mental lexicon: the set of representations for all words that an individual knows

Statistical Learning of Speech

Recall that in the discussion of interactive models of spoken word recognition, one model (Adaptive Resonance Theory; Grossberg 1980) had the virtue of not having to commit a priori to units like phonemes or syllables. Instead, units in this model (chunks) develop as a function of exposure to the language: If a particular stretch of speech is encountered very often, a chunk will develop, whether that stretch of speech is very short (e.g., a particular vowel) or longer (e.g., a particular syllable or word). This feature of the model is attractive because it does not entail the potentially arbitrary assumption of a given unit. However, the approach is only viable if there is reason to believe that the system can actually develop chunks based on the statistical properties of the input. Recent

research, often called statistical learning, has provided evidence that this can in fact occur.

Much of the work in this domain comes from studies of infant speech perception that have been motivated by the classic "nature-nurture" question: If infants can extract essential speech properties from the pattern of the input, then one need not assume that all of the information must be specified innately; nurture is supported. Infant speech perception is a very large and active research domain, and a review of it is beyond the scope of the current discussion. For a recent review of the field, readers can consult Curtin & Hufnagel (2009); see Werker & Curtin (2005) for a well-developed theoretical perspective. For the current purposes, two pairs of studies in the statistical learning literature are particularly relevant. In each pair, one study examined statistical learning effects in infants, whereas the second extended the approach to adults.

The first pair of studies tested whether infants (Saffran et al. 1996a) and adults (Saffran et al. 1996b) can use the statistical properties of syllable sequences to extract stretches of speech that could be words. These experiments are based on the notion that the units (phonemes, syllables, chunks of whatever size) that make up a word remain in fixed positions relative to each other whenever the word occurs. For example, the two syllables of "baby" occur in the order "ba" + "by" each time, whether the word is in the phrase "baby blanket" "pretty baby," or "little baby." Thus, the transitional probability relating those two syllables is relatively high compared to transitions between syllables that cross word boundaries (e.g., the second syllable in "pretty" and the first syllable in "baby"). If humans are sensitive to patterns of transitional probabilities, then they can extract words (or any other chunks) simply by sufficient exposure to the statistical patterns that follow from a unit being a unit.

Saffran et al. (1996a) tested this by presenting infants with long strings of syllables that were constructed to have the necessary statistical properties. The long strings were constructed by concatenating four different three-syllable sequences ("words") many times, in a random order. For example, one "word" was "bidaku," and its three syllables would always occur together in this order among the many syllables made up by this "word" and three other such "words". Critically, the entire sequence was generated on a speech synthesizer that was not given any information about word boundaries. The synthesizer produced a two-minute sequence of syllables that corresponded to 180 "words", but it is important to understand that from a listener's point of view, the sequence was just a long string of syllables, with no acoustic cues to word boundaries. Nonetheless, the sequence provided the statistical regularity that, for example, "da" consistently followed "bi." After the infants had listened to the two-minute sequence, they were put in a situation in which they would either hear one of the four "words" or a different sequence of three syllables, repeatedly. The presentation of sounds in this phase was controlled by whether the infant maintained visual fixation on a blinking light. This procedure provides a measure of discrimination: If there is any difference between how long infants listen to the "words" versus how long they listen to three-syllable sequences that were not present in the exposure phase, that indicates that some learning of the "words" had occurred. Saffran et al. (1996a) observed such a difference, both for a comparison to syllables that had never been heard during the two-minute exposure and for a comparison of syllables that had been in the exposure sequence but that had not maintained a fixed relative position to each other as the "words" had. Thus, after only two minutes of exposure to speech that was stripped of many natural additional cues, leaving only the statistical information, infants showed evidence of learning "words."

The fact that infants can learn from statistical regularities in speech does not necessarily mean that the same would be true of adults—there is a long history in language research that suggests that adults are not as able to absorb language as children are (whether or not one accepts the notion of a critical period for

language acquisition). Thus, it was not clear in advance whether the results for infants would hold for adults. For the adults, Saffran et al. (1996b) constructed stimuli that were similar to those that they had used with the infants, but with a number of differences. The two most important differences were in the duration of exposure and in the nature of the test for word learning. The adults heard 21 minutes (broken into three seven-minute blocks) of exposure to the new "words" rather than two minutes. And, the test for learning was more direct than what could be done with infants: Pairs of three-syllable utterances were presented, with one member of each pair being a "word" from the exposure sequence, and the participant made a forced choice to identity the item from the exposure sequence. Saffran et al. (1996b) confirmed that adults could choose the trained items at better than chance (50%), whether they were tested against items made up of unheard combinations of syllables from the exposure phase (76% correct choice) or of "part-words" (65% correct choice) that combined the last two syllables of a "word" with the first syllable of another "word," or the last syllable of a "word" plus the first two syllables of another "word."

These studies by Saffran and her colleagues demonstrate that both infants and adults can use the statistical properties of a stream of syllables to extract sequences that follow the transitional probability patterns of words. What remains to be seen is whether such statistically determined stretches of the speech signal can become words in a fuller sense of the term. This is the issue addressed by Graf Estes et al. (2007) for infants and by Mirman et al. (2008) for adults. The first phase of the Graf Estes et al. study was similar to the infant study of Saffran et al. (1996a), with infants hearing a long stream of syllables that provided statistical evidence for new "words." After this exposure, the infants were given a task that involved learning to associate names with abstract three-dimensional shapes. The critical manipulation was whether the names were the "words" that had been statistically represented in the exposure phase or comparable items without such prior statistical properties. Graf Estes et al. found that a critical property of learning a new word—associating the sound with an object—was significantly improved if the sound had previously been learned as a result of its statistical properties. Mirman et al. (2008) conducted a comparable study with adults and obtained similar results. After a seven-minute exposure to a stream of syllables with the desired statistical properties, participants were asked to learn to associate "words" with abstract shapes; as in the infant study, the "words" were either those that could have been determined through statistical learning or they were comparable items lacking this statistical advantage. The associations were learned more quickly for the former than for the latter. Thus, for both infants and adults, there is evidence that not only can statistical properties help to segment stretches of speech, but the statistically implicated regions also are enhanced as candidates for word formation.

Perceptual learning for speech: a process that modifies speech representations in order to optimize their match to the prevailing speech environment

Perceptual Learning of Speech

The results of statistical learning studies show that people can use distributional patterns in the input to identify potential units—chunks (Grossberg 1980) of various sizes. Several additional clusters of studies demonstrate an ability to modify the operation of the existing units, again as a function of the input pattern that the system receives. Work in this area is broadly construed as perceptual learning for speech (see Samuel & Kraljic 2009 for a recent review of this topic). Conceptually, it is useful to separate this literature into at least two categories. One category includes cases in which the modifications made to the system as a result of the speech exposure conditions lead to measurably improved speech recognition. The second category includes cases in which the modifications lead to shifts in phonetic category boundaries (e.g., what voice-onset time marks the change from a voiced sound like /d/ to a voiceless one like /t/).

Perceptual learning: Improved perception. A wide range of situations has been examined within the first category. In general, listeners

are given experience with some kind of unfamiliar speech stimuli, and the exposure leads to improvement in their ability to identify or discriminate speech stimuli of that type. The unfamiliar speech that has been studied includes phonetic contrasts in a nonnative language, accented or dialectal speech, and degraded speech (e.g., through compression or noise); we consider an example of each of these situations. In all of these studies, perceptual learning is inferred when exposure to the challenging speech leads to a better ability to understand what is being said.

The most thoroughly studied case of perceptual learning of a nonnative phonetic contrast involves the discrimination of English /r/ and /l/ by native speakers of Japanese. Native Japanese speakers have difficulty with this distinction because only a single sound exists in Japanese in this region of phonetic space, and this sound does not match English /l/ or /r/ very well. A number of studies have been conducted in which native Japanese speakers received extensive training in discriminating /l/ and /r/ (e.g., Lively et al. 1993, Logan et al. 1991). Over the course of several weeks, the listeners heard a series of English words and were asked to decide whether a given stimulus included /r/ or /l/; they received feedback on each trial. If the training included words produced by several different speakers, there were moderate but significant improvements in how well the listeners could identify /r/ versus /l/. Critically, the learning generalized to new tokens, from different talkers. However, if a single talker's voice was used for training, the learning only generalized to new tokens, not to other talkers.

The importance of stimulus variability in training also holds for perceptual learning of accented speech. Bradlow & Bent (2008) gave American listeners exposure to English sentences that had been produced by native Chinese speakers with strong Chinese accents. Half of the listeners were trained under high-variability conditions (sentences were produced by five different Chinese speakers with accents of varying strength), and half were trained on sentences that came from a single speaker. During training the listeners transcribed the sentences they heard. Following training, they completed a test phase that also involved transcribing sentences. Listeners who had trained with multiple speakers showed about a 10% improvement over baseline, the same improvement shown by subjects who were trained on a single speaker and tested on that same speaker. Subjects who trained on a single speaker and were tested with sentences from a different speaker were no better than baseline. Thus, just as with perceptual learning of nonnative contrasts, exposure to a high-variability training set seems to be important for learning to be general.

A third domain in which training produces improved speech recognition includes various types of degraded speech, including speech compression (e.g., Dupoux & Green 1997), vocoded speech (e.g., Davis et al. 2005), and synthetic speech (e.g., Fenn et al. 2003). For all of these stimuli, listeners typically have substantial difficulty understanding what is said (depending, of course, on the degree of degradation). The general procedure here, as in the other studies in this group, is to give listeners experience with the materials and then to test them on new materials with the same kind of degradation. For example, Dupoux & Green (1997) strongly compressed sentences, making them less than half of their normal duration. During training, subjects transcribed the speech. Dupoux and Green then gave their listeners 15 to 20 training sentences and observed improvements of approximately 10% to 15% in keyword report. Experience with this small number of sentences, over the course of about one minute, was sufficient to produce significant improvement. This rapid improvement is quite different from what is typically seen for perceptual learning of nonnative contrasts, or accented speech (though see Clarke & Garrett 2004 for a case of faster adjustment to accents), suggesting that different mechanisms may be at work.

Perceptual learning: Recalibration of phonetic boundaries. The second type of perceptual learning research is a rather recent development. In these studies, the experimenters present listeners with phonetically ambiguous stimuli, with some type of contextual information that disambiguates the stimuli. Perceptual learning is defined as a shift in phonetic categorization toward the contextually defined speech environment: After exposure to acoustically ambiguous speech sounds that are contextually disambiguated, listeners increase their report of sounds consistent with the context. Presumably such shifts should help the listener understand speech better in the prevailing input environment. Studies in this domain provide a more precise indication of exactly what is being learned than do studies in the previous section because the perceptual shifts are on a particular phonetic continuum.

The two seminal papers in this field were by Bertelson et al. (2003) and Norris et al. (2003). The two studies used somewhat different procedures but shared the general approach of presenting ambiguous speech sounds together with disambiguating contextual information; both used changes in subsequent identification of a continuum of speech sounds as the index of the effect. Bertelson et al. used visual information as the context to drive learning. Listeners heard blocks in which an acoustic item was presented that was ambiguous between /aba/ and /ada/ for that listener. In each such exposure block, the ambiguous token was dubbed onto a video of a speaker articulating either /aba/ or /ada/. The visual context produces a very strong immediate bias: Listeners heard the ambiguous token as whatever the face they saw was articulating. The critical result was that subsequent auditory-only test tokens that were formerly ambiguous were now heard as /b/ if the subject had previously seen such tokens with a face articulating /b/, but were heard as /d/ if the visual exposure had been a /d/. Bertelson et al. (2003) called this phenomenon "perceptual recalibration" because the listeners had used the visual information to recalibrate their perceptual boundaries for the speech sounds.

The study by Norris et al. (2003) was conceptually similar but used a different type of contextual guidance and a slightly different procedure. In their experiment, the exposure phase was a lexical decision task—participants identified 200 items as either words or nonwords. Among the 100 real words in this phase were 20 critical items for each listener that had been experimentally manipulated. For half of the subjects, the talker's speech was manipulated so that she seemed to produce word-final instances of /s/ in an ambiguous way (i.e., as a sound midway between [f] and [s]; hereafter, [?]). For the other half of the listeners, it was word-final /f/ sounds that were replaced by the ambiguous [?]. Based on the Ganong effect, Norris et al. expected listeners to use lexical knowledge to guide their interpretation of the ambiguous fricative. The new question was whether listeners who heard [?] in [f]-final words would subsequently categorize more sounds on an [Es]-[Ef] continuum as [f], while those who heard the same sound in [s]-final words would subsequently categorize more items as [s]. This is what Norris et al. found, indicating that listeners use lexical knowledge not only to guide their interpretation of acoustic-phonetic information but also to recalibrate phonetic boundaries so that future tokens are perceived in accord with the prior contextual guidance. Together, the seminal papers by Bertelson et al. (2003) and by Norris et al. (2003) provide clear evidence that phonetic categories are not fixed—there is continuous updating of the categorization process in order to take into account new information in the linguistic environment.

Since the publication of these two papers, a substantial number of studies have begun to delineate the properties of phonetic recalibration. For example, Vroomen et al. (2007) replicated the effects of Bertelson et al. (2003) and also examined the build-up of the recalibration effect. They had listeners identify the audio-only test items after 1, 2, 4, 8, 32, 64, 128, or 256 audiovisual exposure tokens and found that recalibration occurs very rapidly: Listeners demonstrated recalibration after a single exposure token. This learning increased

through about eight exposures, after which it reached a plateau and then began to decrease. There has not been such a systematic test of the build-up of lexically guided effects, although Kraljic et al. (2008b) have shown that as few as ten exposure items are sufficient to produce the effect.

Kraljic & Samuel (2005) examined the question of how phonetic category boundaries return to their "normal" positions after recalibration has occurred. They demonstrated that recalibration remained robust after a 25-minute period with no speech input. Moreover, it remained robust even after listeners heard many canonical pronunciations of the sound that had been oddly pronounced during the exposure phase. Eisner & McQueen (2006) subsequently showed that learning remains stable over a much longer delay—12 hours—regardless of whether subjects slept in the intervening 12 hours (see the discussion of sleep-based "consolidation" effects below).

Two recent studies using the lexical context approach have provided evidence that the recalibration process is actually applied conservatively (see Samuel & Kraljic 2009)—the system only allows recalibration if the evidence indicates that the unusual pronunciations being encountered are likely to be enduring ones in the speech environment. Kraljic et al. (2008b) found that perceptual learning is subject to a primacy bias: Pronunciations that are heard upon initial exposure to a speaker are learned, whereas those same pronunciations are not learned if they do not form part of the initial listening experience. They also found that listeners did not learn a pronunciation if it could be attributed to some transient alternative (speaker-external) factor, such as a pen in the speaker's mouth. Recalibration was also not found when a pronunciation might be attributable to a known dialect or a low-level acoustic-phonetic process such as coarticulation or assimilation (Kraljic et al. 2008a). Phonetic recalibration effects remain a very active area of research, both as a function of audiovisual context and lexical context (see **Figure 1**, see color insert).

New Word Learning by Adults

Not surprisingly, much of the research investigating new word learning has focused on children because children learn many new words as they acquire a language. For example, Storkel (2001) examined how children (ages three to six) acquired new lexical entries as a function of repetition and semantic context. Children heard new words in the context of a paragraph (one type of semantic context) along with a drawing depicting the story (a second type of semantic context). The new words occurred from one to seven times (degree of repetition). Storkel asked the children to identify each new word by selecting from three recorded nonwords and by naming the word when a picture of it was presented. Performance on both measures increased through such training. Gathercole (e.g., 2006) has suggested that learning new words is essentially a verbal short-term memory task and has shown that the ability of children to learn new words correlates with their ability to repeat back nonwords of varying complexity.

Gupta (2003) has demonstrated the same correlation in adults and has recently (Gupta & Tisdale 2009) offered a computational model that illustrates the relationship between phonological short-term memory function and the ability to add a new word to the mental lexicon. These results, and those reviewed below, reflect a somewhat surprising fact: New word learning is much more frequent in adults than one might suspect, and the system must therefore be well designed for such learning. Nation & Waring (1997) estimate that people add about 1,000 word forms (a word and its close morphological relatives) per year, up to an asymptote of about 20,000 word forms. This translates to about three word forms per day, every day of the year. Thus, just as recalibration at the phonetic category level is needed to deal with the constantly changing phonetic environment, the lexicon must also constantly develop to accommodate new entries. Models of word recognition must account for how this development affects processing.

However, there are both theoretical and empirical reasons to believe that such learning has costs. McCloskey & Cohen (1989) have discussed the possibility of creating "catastrophic interference" by adding new entries to an existing memory system if the information is fed into the existing network too rapidly. Their idea is that if a memory system consists of a pattern of connections among many units, with the strength of these connections established by prior experience, inserting new patterns into the system rapidly can cause changes in connections to propagate through the system in a way that will undermine the previously established equilibrium. McClelland and colleagues (1995) have suggested that this problem could be prevented if the new information is buffered so that its rate of introduction into the network is kept at a safe level. They have argued that information is initially represented in the hippocampus and that it is then slowly fed into the neocortex over time, particularly during sleep. The hippocampal representations are not in a system with the long-term structure that is subject to catastrophic interference because less information is kept there, and for a comparatively short time. Recall that Goldinger (2007) proposed that there are both episodic and abstract speech codes; he has identified the episodic codes with the initial hippocampal representations and the abstract ones with the neocortex. The dual-code hypothesis raises the possibility that probing the memory system before the words have been transferred to the neocortex may yield worse performance than after they have been connected to the rest of the lexicon; if McClelland et al. (1995) are correct about the role of sleep, the information must be consolidated overnight to allow the words to function as other lexical items do.

There is a fairly substantial literature on memory consolidation effects, particularly for procedural knowledge. Gaskell & Dumay (2003, Dumay & Gaskell 2007) have conducted a research program that provides clear support for such effects in adding words to the mental lexicon. Gaskell & Dumay (2003) tested whether the acquisition of new words would affect processing of similar existing words. For example, for the word "cathedral," Gaskell and Dumay created the new word "cathedruke." Each day, for five days, participants were repeatedly exposed to such nonwords in the context of a phoneme-monitoring task. The participants also completed a lexical decision task each day that included real words (e.g., "cathedral") that were similar to the newly learned nonwords (e.g., "cathedruke"). If and when a functional lexical entry for "cathedruke" developed, it should compete with the entry for "cathedral" in a lexical decision task, slowing responses to such similar words (compared to controls without new competitors). By the third day of training, Gaskell and Dumay found exactly this pattern, providing evidence for the emergence of lexical competition.

Memory consolidation: a process that unfolds over time (typically, hours or days) and that is used to produce a relatively long-lasting representation for newly learned information

Dumay & Gaskell (2007) directly focused on the potential role of sleep in consolidating newly learned lexical items. They taught one group of subjects new words in the morning and them tested 12 hours later (at night) before they had slept; a second group learned the words at night and slept before being tested 12 hours later (in the morning). Dumay and Gaskell found that the subjects who learned the words at night produced significant lexical competition effects when they were tested 12 hours later, after sleep; the subjects who learned the words in the morning did not produce such lexical competition effects 12 hours later (without sleep). Given the various control conditions in the study, the results provide good evidence that sleep is important for lexical consolidation.

Leach & Samuel (2007) have introduced a distinction between "lexical configuration" and "lexical engagement" when a new word is added to the mental lexicon. Lexical configuration is the collection of information that the person has acquired about a word (e.g., its meaning, spelling, and sound). In contrast to this relatively static set of facts associated with a word, Leach and Samuel suggested that lexical engagement is a dynamic property that functional members of the lexicon have: Such items activate or compete with other lexical representations, and they can support the

perception of their components, producing phenomena such as the Ganong effect and phonemic restoration.

Leach & Samuel (2007) examined the development of lexical configuration and lexical engagement, focusing on the conditions under which words are learned. They compared a word-learning regime similar to that used by Gaskell and Dumay (exposure to words in a phoneme-monitoring task) to one in which each new word was associated with a picture of an unusual object. Lexical configuration and engagement were assessed as a function of each training regime. Configuration was assessed by how well the word could be perceived against a background of noise; the measure of lexical engagement was the ability of a newly learned word to support perceptual learning. Both training regimes promoted lexical configuration, but they were quite different in their ability to promote lexical engagement. With phoneme-monitoring training, there were small and statistically weak perceptual learning effects, with no increase in these effects over the course of training. In contrast, learning words by associating them with pictures produced lexical representations that were fully capable of engaging sublexical codes, generating large and growing perceptual learning effects over training. This contrast suggests that lexical engagement develops when there are semantic associations available for the new words, with no such necessity for lexical configuration to develop. As with perceptual learning at the phonetic level, the dynamic adjustments needed at the lexical level remain a topic of active research.

SPEECH PERCEPTION: CONCLUSION

Researchers have been studying speech perception for over a half century. For much of this time, research on speech perception per se has often been investigated separately from speech perception in the service of spoken word recognition. As Cutler (2008) has suggested, such a separation impedes progress. The field is now at a point when it is both possible and desirable to bring these two subfields together. This is in fact happening, and this development promises to bring improved insights into how humans accomplish the extraordinary feat of understanding spoken language.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author is pleased to acknowledge support from NICHD grant R01-059787.

LITERATURE CITED

Allopenna PD, Magnuson JS, Tanenhaus MK. 1998. Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38:419–39

Bertelson P, Vroomen J, DeGelder B. 2003. Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychol. Sci.* 14(6):592–97

Bradlow AR, Bent T. 2008. Perceptual adaptation to non-native speech. *Cognition* 106:707–29

Chen JY. 2000. Syllable errors from naturalistic slips of the tongue in Mandarin Chinese. *Psychologia* 43:15–26

Clarke CM, Garrett MF. 2004. Rapid adaptation to foreign-accented English. *J. Acoust. Soc. Am.* 116(6):3647–58

Connine CM, Clifton C. 1987. Interactive use of lexical information in speech perception. *J. Exp. Psychol.: Hum. Percept. Perform.* 13:291–99

Curtin S, Hufnagle D. 2009. Speech perception: development. In *Encyclopedia of Neuroscience*, ed. LR Squire, 9:233–38. Oxford, UK: Academic

Cutler A. 2008. The abstract representations in speech processing. *Q. J. Exp. Psychol.* 61:1601–19

Cutler A, Norris D. 1979. Monitoring sentence comprehension. In *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*, ed. WE Cooper, ECT Walker, pp. 113–34. Hillsdale, NJ: Erlbaum

Cutler A, Norris D. 1988. The role of strong syllables in segmentation for lexical access. *J. Exp. Psychol.: Hum. Percept. Perform.* 14:113–21

Cutler A, Weber A. 2007. Listening experience and phonetic-to-lexical mapping in L2. In *Proceedings 16th Int. Congr. Phonet. Sci. 2007*, ed. J Trouvain, WJ Barry, pp. 43–48. Dudweiler, Germany: Pirrot

Davis MH, Johnsrude IS, Hervais-Ademan A, Taylor K, McGettigan C. 2005. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J. Exp. Psychol.: Gen.* 134(2):222–41

Diehl RL, Lotto AJ, Holt LL. 2004. Speech perception. *Annu. Rev. Psychol.* 55:149–79

Di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G. 1992. Understanding motor events: a neurophysiological study. *Exp. Brain. Res.* 91:176–80

Dorman MF, Studdert-Kennedy M, Raphael LJ. 1977. Stop-consonant recognition: release bursts and formant transitions as functionally equivalent, context-dependent cues. Percept. Psychophys. 22:109–22

Dumay N, Gaskell MG. 2007. Sleep-associated changes in the mental representation of spoken words. *Psychol. Sci.* 18:35–39

Dupoux E, Green K. 1997. Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J. Exp. Psychol.: Hum. Percept. Perform.* 23:914–27

Eimas P, Corbit J. 1973. Selective adaptation of linguistic feature detectors. *Cogn. Psychol.* 4:99–109

Eisner F, McQueen JM. 2006. Perceptual learning in speech: stability over time. *J. Acoust. Soc. Am.* 119(4):1950–53

Elman JL, McClelland JL. 1988. Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. *J. Mem. Lang.* 27:143–65

Fenn KM, Nusbaum HC, Margoliash D. 2003. Consolidation during sleep of perceptual learning of spoken language. *Nature* 425:614–16

Fowler CA. 1986. An event approach to the study of speech perception from a direct realist perspective. *J. Phon.* 14:3–28

Fowler CA. 1991. Auditory perception is not special: We see the world, we feel the world, we hear the world. *J. Acoust. Soc. Am.* 89:2910–15

Fowler CA, Brown JM, Mann VA. 2000. Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *J. Exp. Psychol.: Hum. Percept. Perform.* 26:877–88

Fowler CA, Rosenblum LD. 1990. Duplex perception: a comparison of monosyllables and slamming doors. *J. Exp. Psychol.: Hum. Percept. Perform.* 16:742–54

Galantucci B, Fowler CA, Turvey MT. 2006. The motor theory of speech perception reviewed. *Psychon. Bull. Rev.* 13:361–77

Ganong WF. 1980. Phonetic categorization in auditory perception. *J. Exp. Psychol.: Hum. Percept. Perform.* 6:110–25

Gaskell MG, Dumay N. 2003. Lexical competition and the acquisition of novel words. *Cognition* 89:105–32

Gaskell MG, Marslen-Wilson WD. 2002. Representation and competition in the perception of spoken words. *Cogn. Psychol.* 45:220–66

Gathercole SE. 2006. Nonword repetition and word learning: the nature of the relationship. *Appl. Linguist.* 27:513–43

Gibson JJ. 1966. *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin

Goldinger SD. 1996. Words and voices: episodic traces in spoken word identification and recognition memory. *J. Exp. Psychol.: Learn. Mem. Cogn.* 22:1166–83

Goldinger SD. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 105:251–79

Goldinger SD. 2007. A complementary-systems approach to abstract and episodic speech perception. In *Proceedings 16th Int. Congr. Phonet. Sci. 2007*, ed. J Trouvain, WJ Barry, pp. 49–54. Dudweiler, Germany: Pirrot

Goldinger SD, Luce PA, Pisoni DB. 1989. Priming lexical neighbors of spoken words: effects of competition and inhibition. *J. Mem. Lang.* 28:501–18

Graf Estes K, Evans JL, Alibali MW, Saffran JR. 2007. Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychol. Sci.* 18(3):254–60

Grossberg S. 1980. How does a brain build a cognitive code? *Psychol. Rev.* 87:1–51

Gupta P. 2003. Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *Q. J. Exp. Psychol. A* 56(7):1213–36

Gupta P, Tisdale J. 2009. Does phonological short-term memory causally determine vocabulary learning? Toward a computational resolution of the debate. *J. Mem. Lang.* 61:481–502

Holt LL. 2006. The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorization. *J. Acoust. Soc. Am.* 120:2801–17

Holt LL, Stephens JDW, Lotto AJ. 2005. A critical evaluation of visually moderated phonetic context effects. *Percept. Psychophys.* 67:1102–12

Kluender KR, Diehl RL, Killeen PR. 1987. Japanese quail can learn phonetic categories. *Science* 237:1195–97

Kraljic T, Brennan SE, Samuel AG. 2008a. Accommodating variation: dialects, idiolects, and speech processing. *Cognition* 107:54–81

Kraljic T, Samuel AG. 2005. Perceptual learning for speech: Is there a return to normal? *Cogn. Psychol.* 51(2):141–78

Kraljic T, Samuel AG, Brennan SE. 2008b. First impressions and last resorts: how listeners adjust to speaker variability. *Psychol. Sci.* 19:332–38

Kuhl PK, Miller JD. 1978. Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. *J. Acoust. Soc. Am.* 63:905–17

Leach L, Samuel AG. 2007. Lexical configuration and lexical engagement: when adults learn new words. *Cogn. Psychol.* 55:306–53

Levelt WJM. 1993. Accessing words in speech production: stages, processes and representations. In *Lexical Access in Speech Production*, ed. WJM Levelt, pp. 1–22. Cambridge, UK: Blackwell

Liberman A, Cooper F, Shankweiler D, Studdert-Kennedy M. 1967. Perception of the speech code. *Psychol. Rev.* 74:431–61

Lively SE, Logan JS, Pisoni DB. 1993. Training Japanese listeners to identify English /r/ and /l/ II: the role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.* 94(3):1242–55

Logan JS, Lively SE, Pisoni DB. 1991. Training Japanese listeners to identify English /r/ and /l/: a first report. *J. Acoust. Soc. Am.* 89(2):874–86

Lotto AJ, Hickok GS, Holt LL. 2009. Reflections on mirror neurons and speech perception. *Trends Cogn. Sci.* 13(3):110–14

Magnuson J, McMurray B, Tanenhaus M, Aslin R. 2002. Lexical effects on compensation for coarticulation: the ghost of Christmash past. *Cogn. Sci.* 27:285–98

Mann VA, Repp BH. 1981. Influence of preceding fricative on stop consonant perception. *J. Acoust. Soc. Am.* 69:548–58

Marslen-Wilson WD. 1975. Sentence perception as an interactive parallel process. *Science* 189:226–28

Marslen-Wilson WD. 1987. Functional parallelism in spoken word recognition. *Cognition* 25:71–102

Marslen-Wilson WD, Welsh A. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* 10:29–63

Massaro DW. 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Erlbaum

Massaro DW. 1989. Testing between the TRACE model and the fuzzy logical model of speech perception. *Cogn. Psychol.* 21:398–421

Mattys SL, White L, Melhorn JF. 2005. Integration of multiple speech segmentation cues: a hierarchical framework. *J. Exp. Psychol. Gen.* 134:477–500

McClelland JL, Elman JL. 1986. The TRACE model of speech perception. *Cogn. Psychol.* 18:1–86

McClelland JL, McNaughton BL, O'Reilly RC. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102:419–57

McCloskey M, Cohen NJ. 1989. Catastrophic interference in connectionist networks: the sequential learning problem. In *The Psychology of Learning and Motivation*, ed. GH Bower, 24:109–165. New York: Academic

McGurk H, MacDonald J. 1976. Hearing lips and seeing voices. *Nature* 264:746–48

McQueen JM, Cutler A, Norris D. 2006. Phonological abstraction in the mental lexicon. *Cogn. Sci.* 30:1113–26

Miller JD, Wier CC, Pastore RE, Kelly WJ, Dooling RJ. 1976. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: an example of categorical perception. *J. Acoust. Soc. Am.* 60:410–17

Mirman D, Magnuson JS, Graf Estes K, Dixon JA. 2008. The link between statistical segmentation and word learning in adults. *Cognition* 108:271–80

Nation P, Waring R. 1997. Vocabulary size, text coverage, and word lists. In *Vocabulary: Description, Acquisition, Pedagogy*, ed. N Schmitt, M McCarthy, pp. 6–19. New York: Cambridge Univ. Press

Norris D. 1994. SHORTLIST: a connectionist model of continuous speech recognition. *Cognition* 52:189–234

Norris D, McQueen JM, Cutler A. 1995. Competition and segmentation in spoken-word recognition. *J. Exp. Psychol.: Learn. Mem. Cogn.* 21:1209–28

Norris D, McQueen JM, Cutler A. 2000. Merging information in speech recognition: Feedback is never necessary. *Behav. Brain Sci.* 23:299–370

Norris D, McQueen JM, Cutler A. 2003. Perceptual learning in speech. *Cogn. Psychol.* 47:204–38

Nygaard LC, Sommers MS, Pisoni DB. 1994. Speech perception as a talker-contingent process. *Psychol. Sci.* 5:42–45

Pisoni DB, Tash J. 1974. Reaction times to comparisons within and across phonetic categories. *Percept. Psychophys.* 15:285–90

Pitt MA, McQueen JM. 1998. Is compensation for coarticulation mediated by the lexicon? *J. Mem. Lang.* 39:347–70

Pitt MA, Samuel AG. 1993. An empirical and meta-analytic evaluation of the phoneme identification task. *J. Exp. Psychol.: Hum. Percept. Perform.* 19:1–27

Pitt MA, Samuel AG. 2006. Word length and lexical activation: Longer is better. *J. Exp. Psychol.: Hum. Percept. Perform.* 32:1120–35

Rand T. 1974. Dichotic release from masking for speech. *J. Acoust. Soc. Am.* 55:678–80

Repp BH. 1984. Categorical perception: issues, methods, and findings. In *Speech and Language. Advances in Basic Research and Practice*, ed. N Lass, 10:243–335. Orlando, FL: Academic

Saffran JR, Aslin RN, Newport EL. 1996a. Statistical learning by 8-month-old infants. *Science* 274(5294):1926–28

Saffran JR, Newport EL, Aslin RN. 1996b. Word segmentation: the role of distributional cues. *J. Mem. Lang.* 35(4):606–21

Samuel AG. 1977. The effect of discrimination training on speech perception: noncategorical perception. *Percept. Psychophys.* 22:321–30

Samuel AG. 1981. Phonemic restoration: insights from a new methodology. *J. Exp. Psychol.: Gen.* 110:474–94

Samuel AG. 1996. Does lexical information influence the perceptual restoration of phonemes? *J. Exp. Psychol.: Gen.* 125:28–51

Samuel AG. 1997. Lexical activation produces potent phonemic percepts. *Cogn. Psychol.* 32:97–127

Samuel AG. 2001. Knowing a word affects the fundamental perception of the sounds within it. *Psychol. Sci.* 12:348–51

Samuel AG, Kraljic T. 2009. Perceptual learning in speech perception. *Atten. Percept. Psychophys.* 71:1207–18

Samuel AG, Pitt MA. 2003. Lexical activation (and other factors) can mediate compensation for coarticulation. *J. Mem. Lang.* 48:416–34

Shankweiler D, Studdert-Kennedy M. 1967. Identification of consonants and vowels presented to left and right ears. *Q. J. Exp. Psychol.* 19:59–63

Soto-Faraco S, Sebastian-Galles N, Cutler A. 2001. Segmental and suprasegmental mismatch in lexical access. *J. Mem. Lang.* 45:412–32

Storkel HL. 2001. Learning new words: phonotactic probability in language development. *J. Speech Lang. Hear. Res.* 44:1321–37

Sumby WH, Pollack I. 1954. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26:212–15

Vroomen J, van Linden S, De Gelder B, Bertelson P. 2007. Visual recalibration and selective adaptation in auditory-visual speech perception: contrasting build-up courses. *Neuropsychologia* 45(3):572–77

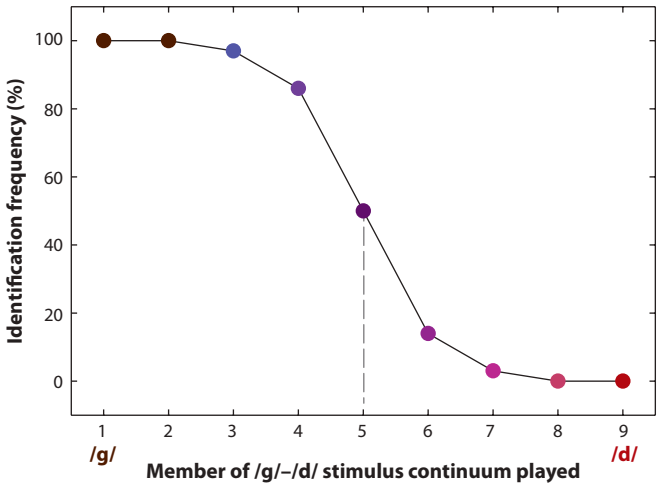
Warren RM. 1970. Perceptual restoration of missing speech sounds. *Science* 167:392–93

Werker JF, Curtin S. 2005. PRIMIR: a developmental framework of speech processing. *Lang. Learn. Dev.* 1:197–234

Zhou X, Marslen-Wilson W. 1995. Words, morphemes, and syllables in the Chinese mental lexicon. *Lang. Cogn. Proc.* 9:393–423

Zwitserlood P. 1989. The locus of effects of sentential-semantic context in spoken-word processing. *Cognition* 32:25–64

a



b

Phenomenon	Early investigation	MORE /d/ report	LESS /d/ report	Time course
Ganong effect	Ganong (1980)	Follow sound with "uck"	Follow sound with "ulf"	Immediate
Audiovisual speech perception	Sumby & Pollack (1954)	Accompany sound by face clearly articulating "d"	Accompany sound by face clearly articulating "g"	Immediate
Compensation for coarticulation	Mann & Repp (1981)	Precede sound with "esh"	Precede sound with "ess"	Immediate
Auditory context effect	Holt (2006)	Precede sound by low-frequency tones	Precede sound by high-frequency tones	Minutes
Selective adaptation	Eimas & Corbit (1973)	Precede by "guh, guh, guh..."	Precede by "duh, duh, duh..."	Minutes
Audiovisually driven recalibration	Bertelson et al. (2003)	Earlier exposure to ambiguous "d/g" sound paired with a clear visual /d/	Earlier exposure to ambiguous "d/g" sound paired with a clear visual /g/	Minutes
Lexically driven recalibration	Norris et al. (2003)	Earlier exposure to ambiguous "d/g" in a lexical context requiring /d/	Earlier exposure to ambiguous "d/g" in a lexical context requiring /g/	Days?

Figure 1

(*a*) A typical identification function for a speech experiment in which listeners identify syllables as either beginning with "d" or with "g," given many presentations of stimuli from a continuum of speech sounds. (*b*) Summary of seven different speech context effects discussed in the review. Each such context effect can shift the labeling function shown here, either toward increased or toward decreased report of "d."