

SPEECH PERCEPTION

Sven L. Mattys.

University of Bristol. UK.

(to appear in *Oxford Handbook of Cognitive Psychology*, 2011, D. Reisberg, Ed.)

Abstract

Speech perception is conventionally defined as the perceptual and cognitive processes leading to the discrimination, identification, and interpretation of speech sounds. However, to gain a broader understanding of the concept, such processes must be investigated relative to their interaction with long-term knowledge—lexical information in particular. This chapter starts by a review of some of the fundamental characteristics of the speech signal and by an evaluation of the constraints that these characteristics impose on modelling speech perception. Long-standing questions are then discussed in the context of classic and more recent theories. Recurrent themes include: (1) the involvement of articulatory knowledge in speech perception, (2) the existence of a speech-specific mode of auditory processing, (3) the multi-modal nature of speech perception, (4) the relative contribution of bottom-up and top-down flows of information to sound categorization, (5) the impact of the auditory environment on speech perception in infancy, (6) the flexibility of the speech system in the face of novel or atypical input.

Key words: coarticulation, phonemes, categorical perception, lexical access, segmentation, bottom-up, top-down, perceptual learning,

The complexity, variability and fine temporal properties of the acoustic signal of speech have puzzled psycholinguists and speech engineers for decades. How can a signal seemingly devoid of regularity be decoded and recognized almost instantly, without any formal training, and despite being often experienced in sub-optimal conditions? Without any real effort, we identify over a dozen speech sounds (phonemes) per second, recognize the words they constitute, almost immediately understand the message generated by the sentences they form, and often elaborate appropriate verbal and non-verbal responses before the utterance ends.

Unlike theories of letter perception and written-word recognition, theories of speech perception and spoken-word recognition have devoted a great deal of their investigation to a description of the signal itself, most of it carried out within the field of phonetics. In particular, the fact that speech is conveyed in the auditory modality has dramatic implications for the perceptual and cognitive operations underpinning its recognition. Research in speech perception has focused on the constraining effects of three main properties of the auditory signal: sequentiality, variability, and continuity.

NATURE OF THE SPEECH SIGNAL

Sequentiality

One of the most obvious disadvantages of the auditory system compared to its visual counterpart is that the distribution of the auditory information is time-bound, transient, and solely under the speaker's control. Moreover, the auditory signal conveys its acoustic content in a relatively serial fashion, one bit of information at a time. The extreme spreading of information over time in the speech domain has important consequences for the mechanisms involved in perceiving and interpreting the input.

In particular, given that relatively little information is conveyed per unit of time, the extraction of meaning can only be done within a window of time that far exceeds the amount of information that can be held in echoic memory (Huggins, 1975; Nooteboom, 1979). Likewise, given that there are no such things as "auditory saccades," in which listeners would be able to skip ahead of the signal or replay the words or sentences they just heard, speech perception and lexical-sentential integration must take place sequentially, in real time (Figure 1).

For a large part, listeners are extremely good at keeping up with the rapid flow of speech sounds: Marslen-Wilson (1984) showed that many words in sentences are often recognized well before their offset, sometimes as early as 200 ms after their onset, the average duration of one or two syllables. Other words, however, can only be disentangled from competitors later on, especially when they are short and phonetically reduced, e.g., "you are" pronounced as "you're" (Bard, Shillcock, & Altmann, 1988). Yet, in general, there is a consensus that speech perception and lexical access closely shadow the unfolding of the signal (e.g., the Cohort Model, Marslen-Wilson, 1987), even though "right-to-left" effects can sometimes be observed as well (Dahan, 2010).

Given the inevitable sequentiality of speech perception and the limited amount of information that humans can hold in their auditory short-term memory, an obvious question is whether fast speech, which allows more information to be packed into the same amount of time, helps listeners handle the transient nature of speech and, specifically, whether it affects the mechanisms leading to speech recognition. A problem, however, is that fast speech tends to be less clearly articulated (hypo-articulated), and hence, less intelligible. Thus, any processing gain due to denser information packing might be offset by diminished intelligibility. However, this confound can be avoided experimentally. Indeed, speech rate can be accelerated with minimal

loss of intrinsic intelligibility via computer-assisted signal compression (e.g., Foulke & Sticht, 1969; van Buuren, Festen, & Houtgast, 1999). Time compression experiments have led to mixed results. Dupoux and Mehler (1990), for instance, found no effect of speech rate on how phonemes are perceived in monosyllabic versus disyllabic words. They started from the observation that the initial consonant of a monosyllabic word is detected faster if the word is high-frequency than if it is low-frequency, whereas frequency has no effect in multi-syllabic words. This difference can be attributed to the use of a lexical route with short words and of a phonemic route with longer words. That is, short words are mapped directly onto lexical representations, whereas longer words undergo a process of decomposition into phonemes first. Critically, Dupoux and Mehler reported that a frequency effect did not appear when the duration of the disyllabic words was compressed to that of the monosyllabic words, suggesting that whether listeners use a lexical or phonemic route to identify phonemes depends on structural factors (number of phonemes or syllables) rather than time. Thus, on this account, the transient nature of speech has only a limited effect on the mechanisms underlying speech recognition.

In contrast, others have found significant effects of speech rate on lexical access. For example, both Pitt and Samuel (1995) and Radeaux, Morais, Mousty, and Bertelson (2000) observed that the uniqueness point of a word, i.e., the sequential point at which it can be uniquely specified (e.g., "spag" for "spaghetti"), could be dramatically altered when speech rate was manipulated. However, most changes were observed at slower rates, not at faster rates. Thus, changes in speech rate can have effects on recognition mechanisms, but these are observed mainly with time expansion, not with time compression. In sum, although the studies by Dupoux and Mehler (1990), Pitt and Samuel (1995), and Radeau et al. (2000) highlight different effects of time manipulation on speech processing, they all agree that packing more information per unit of time by accelerating speech rate does not compensate for the transient nature of the speech signal and for memory limitations. This is probably due to intrinsic perceptual and mnemonic limitations on how fast information can be processed by the speech system—at *any* rate.

In general, the sequential nature of speech processing is a feature that many models have struggled to implement not only because it requires taking into account echoic and short-term memory mechanisms (Mattys, 1997), but also because the sequentiality problem is compounded by a lack of clear boundaries between phonemes and between words, as described below.

Continuity

The inspection of a speech waveform does not reveal clear acoustic correlates of what the human ear perceives as phoneme and word boundaries. The lack of boundaries is due to coarticulation between phonemes (the blending of articulatory gestures between adjacent phonemes) within and across words. Even though the degree of coarticulation between phonemes is somewhat less pronounced across than within words (Fourgeron & Keating, 1997), the lack of clear and reliable gaps between words, along with the sequential nature of speech delivery, makes speech continuity one of the most challenging obstacles for both psycholinguistic theory and automatic speech recognition applications. Yet, the absence of phoneme and word boundary markers hardly seems to pose a problem for everyday listening, as the subjective experience of speech is not one of continuity but, rather, of discreteness—i.e., a string of sounds making up a string of words.

A great deal of the segmentation problem can be solved, at least in theory, based on lexical knowledge and contextual information. Key notions, here, are lexical competition and segmentation by lexical subtraction. In this view, lexical candidates are activated in multiple

locations in the speech signal—i.e., multiple alignment—and they compete for a segmentation solution that does not leave any fragments lexically unaccounted for (e.g., "great wall" is favored over "gray twall", because "twall" is not an English word). Importantly, this knowledge-driven approach does not assign a specific computational status to segmentation, other than being the mere consequence of mechanisms associated with lexical competition (e.g., McClelland & Elman, 1986; Norris, 1994).

Another source of information for word segmentation draws upon broad prosodic and segmental regularities in the signal, which listeners use as heuristics for locating word boundaries. For example, languages whose words have a predominant rhythmic pattern (e.g., word-initial stress is predominant in English; word-final lengthening is predominant in French) provide a relatively straightforward—though probabilistic—segmentation strategy to their listeners (Cutler, 1994). The heuristic for English would go as follows: *every time a strong syllable is encountered, a boundary is posited before that syllable*. For French, it would be: *every time a lengthened syllable is encountered, a boundary is posited after that syllable*. Another documented heuristic is based on phonotactic probability, that is, the likelihood that specific phonemes follow each other in the words of a language (McQueen, 1998). Specifically, phonemes that are rarely found next to each other in words (e.g., very few English words contain the /fh/ diphone) would be probabilistically interpreted as having occurred across a word boundary (e.g., "toughh hero"). Finally, a wide array of acoustic-phonetic cues can also give away the position of a word boundary (Umeda & Coker, 1974). Indeed, phonemes tend to be realized differently depending on their position relative to a word or a syllable boundary. For example, in English, word-initial vowels are frequently glottalized (brief closure of the glottis, e.g., /e/ in "isle end", compared to no closure in "I lend"), word-initial stop consonants are often aspirated (burst of air accompanying the release of a consonant, e.g., /t/ in "grey tanker" compared to no aspiration in "great anchor").

It is important to note that, in everyday speech, lexically- and sublexically-driven segmentation cues usually coincide and reinforce each other. However, in sub-optimal listening conditions (e.g., noise) or in rare cases where a conflict arises between those two sources of information, listeners have been shown to downplay sublexical discrepancies and give more heed to lexical plausibility (Mattys, White, & Melhorn, 2005; Figure 2).

Variability

Perhaps the most defining challenge for the field of speech perception is the enormous variability of the signal relative to the stored representations onto which it must be matched. Variability can be found at the word level, where there is an infinity of ways a given word can be pronounced depending on accents, voice quality, speech rate, etc., leading to a multitude of surface realizations for a unique target representation. But this many-to-one mapping problem is not different from the one encountered with written words in different hand-writings or object recognition in general. In those cases, signal normalization can be effectively achieved by defining a set of core features unique to each word or object stored in memory and by reducing the mapping process to those features only.

The real issue with speech variability happens at a lower level, namely, phoneme categorization. Unlike letters whose realizations have at least some commonality from one instance to another, phonemes can vary widely in their acoustic manifestation—even within the same speaker. For example, as shown in Figure 3A, the realization of the phoneme /d/ has no immediately apparent acoustic commonality in /di/ and /du/ (Delattre, Liberman, & Cooper,

1955). This lack of acoustic invariance is the consequence of coarticulation: The articulation of /d/ in /di/ is partly determined by the articulatory preparation for /i/, and likewise for /d/ in /du/. The power of coarticulation is easily demonstrated by observing a speaker's mouth prior to saying /di/ compared to /du/. The mode of articulation of /i/ (unrounded) vs. /u/ (rounded) is visible on the speaker's lips even before /d/ has been uttered. The resulting acoustics of /d/ preceding each vowel have therefore little in common.

The success of the search for acoustic cues, or invariants, capable of uniquely identifying phonemes or phonetic categories has been highly feature-specific. For example, as illustrated in Figure 3A, the place of articulation of phonemes (i.e., the place in the vocal tract where the airstream is most constricted, which distinguishes, e.g., /b/, /d/, /g/) has been difficult to map onto specific acoustic cues. However, the difference between voiced and unvoiced stop consonants (/b/, /d/, /g/ vs. /p/, /t/, /k/) can be traced back fairly reliably to the duration between the release of the consonant and the moment when the vocal folds start vibrating, i.e., the voice onset time, VOT (Liberman, Delattre, & Cooper, 1958). In English, the VOT of voiced stop consonants is typically around 0 ms (or at least shorter than 20 ms) whereas it is generally over 25 ms for voiceless consonants. Although this contrast has been shown to be somewhat influenced by consonant type and vocalic context (e.g., Lisker & Abramson, 1970), VOT is a fairly robust cue for the voiced-voiceless distinction.

Vowels are subject to coarticulatory influences too, but the spectral structure of their middle portion is usually relatively stable, and hence, a taxonomy of vowels based on their unique distribution of energy bands along the frequency spectrum, or formants, can be attempted. However, such distribution is influenced by speaking rate, with fast speech typically leading to the target frequency of the formants being missed or leading to an asymmetric shortening of stressed vs. unstressed vowels (Lindblom, 1963; Port, 1976). In general, speech rate variation is particularly problematic for acoustic cues involving time. Even stable cues such as VOT can lose their discriminability power when speech rate is altered. For example, at fast speech rates, the VOT difference between voiced and voiceless stop consonants decreases, making the two types of phonemes more difficult to distinguish (Summerfield, 1981). The same problem has been noted for the difference between /b/ and /w/, with /b/ having rapid formant transitions into the vowel and /w/ less rapid ones. This difference is less pronounced at fast speech rates (Miller & Liberman, 1979).

Yet, except for those conditions in which subtle differences are manipulated in the laboratory, listeners are surprisingly good at compensating for the acoustic distortions introduced by coarticulation and changes in speech rate. Thus, input variability, phonetic-context effects, and the lack of invariance do not appear to pose a serious problem for everyday speech perception. As reviewed below, however, theoretical accounts aiming to reconcile the complexity of the signal with the effortlessness of perception vary greatly.

BASIC PHENOMENA AND QUESTIONS IN SPEECH PERCEPTION

Below are some of the observations that have shaped theoretical thinking in speech perception over the past sixty years. Most of them concern, in one way or another, the extent to which speech perception is carried out by a part of the auditory system dedicated to speech and involving speech-specific mechanisms not recruited for non-speech sounds.

Categorical perception

Categorical perception is a sensory phenomenon whereby a physically continuous dimension is perceived as discrete categories, with abrupt perceptual boundaries between categories and poor discrimination within categories (e.g., perception of the visible electromagnetic radiation spectrum as discrete colors). Early on, categorical perception was found to apply to phonemes—or at least some of them. For example, Liberman, Harris, Hoffman, and Griffith (1957) showed that synthesized syllables ranging from /ba/ to /da/ to /ga/ by gradually adjusting the transition between the consonant and the vowel's formants (i.e., the formant transitions) were perceived as falling into coarse /b/, /d/, and /g/ categories, with poor discrimination between syllables belonging to a perceptual category and high discrimination between syllables straddling a perceptual boundary (Figure 4). Importantly, categorical perception was not observed for matched auditory stimuli devoid of phonemic significance (Liberman, Harris, Eimas, Lisker, & Bastian, 1961). Moreover, since categorical perception meant that easy-to-identify syllables (spectrum end-points) were also easy syllables to pronounce, whereas less-easy-to-identify syllables (spectrum mid-points) were generally less easy to pronounce, categorical perception was seen as a highly adaptive property of the speech system, and hence, evidence for a dedicated speech mode of the auditory system. This claim was later weakened by reports of categorical perception for non-speech sounds (e.g., Miller, Wier, Pastore, Kelly, & Dooling, 1976) and for speech sounds by non-human-species (e.g., Kluender, Diehl, & Killeen, 1987; Kuhl, 1976).

Effects of phonetic context

The effect of adjacent phonemes on the acoustic realization of a target phoneme (e.g., /d/ in /di/ vs. /du/) was mentioned earlier as a core element of the variability challenge. This challenge, i.e., achieving perceptual constancy despite input variability, is perhaps most directly illustrated by the converse phenomenon, namely, the varying perception of a constant acoustic input as a function of its changing phonetic environment. Mann (1980) showed that the perception of a /da/-/ga/ continuum was shifted in the direction of reporting more /ga/ when it was preceded by /al/ and more /da/ when it was preceded by /ar/. Since these shifts are in the opposite direction of coarticulation between adjacent phonemes, listeners appear to compensate for the expected consequences of coarticulation. Whether compensation for coarticulation is evidence for a highly sophisticated mechanism whereby listeners use their implicit knowledge of how phonemes are produced—i.e., coarticulated—to guide perception (e.g., Fowler, 2006) or simply a consequence of long-term association between the signal and the percept (e.g., Diehl, Lotto, & Holt, 2004; Lotto & Holt, 2006) has been a question of fundamental importance for theories of speech perception, as discussed later.

Integration of acoustic and optic cues

The chief outcome of speech production is the emission of an acoustic signal. However, visual correlates, such as facial and lip movements, are often available to the listener as well. The effect of visual information on speech perception has been extensively studied, especially in the context of the benefit provided by visual cues for listeners with hearing impairments (e.g., Lachs, Pisoni, & Kirk, 2001) and for speech perception in noise (e.g., Sumbly & Pollack, 1954). Visual-based enhancement is also observed for un-degraded speech with a semantically complicated content or for foreign-accented speech (Reisberg, McLean, & Goldfield, 1987). In the laboratory, audio-visual integration is strikingly illustrated by the well-known McGurk effect. McGurk and

McDonald (1976) showed that listeners presented with an acoustic /ba/ dubbed over a face saying /ga/ tended to report hearing /da/, a syllable whose place of articulation is intermediate between /ba/ and /ga/. The robustness and automaticity of the effect suggest that the acoustic and (visual) articulatory cues of speech are integrated at an early stage of processing. Whether early integration indicates that the primitives of speech perception are articulatory in nature or whether it simply highlights a learned association between acoustic and optic information has been a theoretically divisive debate (see Rosenblum, 2005, for a review).

Lexical and sentential effects on speech perception

Although traditional approaches to speech perception often stop where word recognition begins (in the same way that approaches to word recognition often stop where sentence comprehension begins), speech perception has been profoundly influenced by the debate on how higher-order knowledge affects the identification and categorization of phonemes and phonetic features. A key observation is that lexical knowledge and sentential context can aid phoneme identification, especially when the signal is ambiguous or degraded. For example, Warren and Obusek (1971) showed that a word can be heard as intact even when a component phoneme is missing and replaced with noise, e.g., "legi*lature," where the asterisk denotes the replaced phoneme. In this case, lexical knowledge dictates what the listener should have heard rather than what was actually there, a phenomenon referred to as phoneme restoration. Likewise, Warren and Warren (1970) showed that a word whose initial phoneme is degraded, e.g., "*eel," tends to be heard as "wheel" in "It was found that the *eel was on the axle" and as "peel" in "It was found that the *eel was on the orange." Thus, phoneme identification can be strongly influenced by lexical and sentential knowledge even when the disambiguating context appears later than the degraded phoneme.

But is this truly of interest for speech *perception*? In other words, could phoneme restoration (and other similar speech illusions) simply result from post-perceptual, strategic biases? In this case, "*eel" would be interpreted as "wheel" simply because it makes pragmatic sense to do so in a particular sentential context, not because our perceptual system is genuinely tricked by high-level expectations. If so, contextual effects are of interest to speech-perception scientists only insofar as they suggest that speech perception happens in a system that is unpenetrable by higher-order knowledge—an unfortunately convenient way of indirectly perpetuating the confinement of speech perception to the study of phoneme identification. The evidence for a post-perceptual explanation is mixed. While Norris, McQueen, and Cutler (2000), Massaro (1989), and Oden and Massaro (1978), among others, found no evidence for online top-down feedback to the perceptual system and no logical reasons why such feedback should exist, Samuel (1981; 1997), Connine and Clifton (1987), and Magnuson, McMurray, Tanenhaus, and Aslin (2003), among others, have reported lexical effects on perception that challenge feed-forward models—e.g., evidence that lexical information truly alters low-level perceptual discrimination, Samuel, 1981. This debate has fostered extreme empirical ingenuity over the past decades, but comparatively little change to theory. One exception, however, is that the debate has now spread to the *long-term* effects of higher-order knowledge on speech perception. For example, while Norris et al. (2000) argue against online top-down feedback, the same group (2003) recognizes that perceptual (re-)tuning can happen over time, in the context of repeated exposure and learning. Placing the feedforward/feedback debate in the time domain provides an opportunity to examine the speech system at the interface with cognition, and memory functions in particular. It also allows more applied considerations to be introduced, such as the role of

perceptual recalibration for second-language learning and speech perception in difficult listening conditions (Samuel & Kraljic, 2009), as discussed later.

THEORIES OF SPEECH PERCEPTION (NARROWLY AND BROADLY CONSTRUED)

Motor and Articulatory-Gesture Theories

The Motor Theory of speech perception, reported in a series of articles in the early 1950s by Liberman, Delattre, Cooper, and other researchers from the Haskins Laboratories, was the first to offer a conceptual solution to the lack-of-invariance problem. As mentioned earlier, the main stumbling block for speech-perception theories was the observation that many phonemes cannot uniquely be identified by a set of stable and reliable acoustic cues. For example, the formant transitions of /d/, especially the second formant, differ as a function of the following vowel. However, Delattre et al. (1955) found commonality between different /d/s by extrapolating the formant transitions back in time to their convergence point, or *locus* (or *hub*, Potter, Kopp, & Green, 1947), as shown in Figure 3B. Thus, what is common to the formants of all /d/s is the frequency at their origin, that is, the frequency that would best reflect the position of the articulators prior to the release of the consonant. This led to one of the key arguments in support of the motor theory, namely that a one-to-one relationship between acoustics and phonemes can be established if the speech system includes a mechanism that allows the listener to work backward through the rules of production in order to identify the speaker's intended phonemes. In other words, the lack-of-invariance problem can be solved if it can be demonstrated that listeners perceive speech by identifying the speaker's intended speech gestures rather than (or in addition to) relying solely on the acoustic manifestation of such gestures. The McGurk effect, whereby auditory perception is dramatically altered by seeing the speaker's moving lips (articulatory gestures), was an important contributor to the view that the perceptual primitives of speech are gestural in nature.

In addition to claiming that the motor system is recruited for perceiving speech (and partly because of this claim), the Motor Theory also posits that speech perception takes place in a highly-specialized and speech-specific module that is neurally isolated and is most likely a unique and innate human endowment (Liberman, 1996; Liberman & Mattingly, 1985). However, even among supporters of a motor basis for speech perception, agreeing upon an operational definition of intended speech gestures and providing empirical evidence for the contribution of such intended gestures to perception proved difficult. This led Fowler and her colleagues to propose that the objects of speech perception are not *intended* articulatory gestures but *real* gestures, that is, actual vocal tract movements that are inferable from the acoustic signal itself (e.g., Fowler, 1986, 1996). Thus, although Fowler's Direct Realism approach aligns with the Motor Theory in that it claims that perceiving speech is perceiving gestures, it asserts that the acoustic signal itself is rich enough in articulatory information to provide a stable (i.e., invariant) signal-to-phoneme mapping algorithm. In doing so, Direct Realism can do away with claims about specialized and/or innate structures for speech perception.

Although the popularity of the original tenets of the Motor Theory—and, to some extent, associated gesture theories—has waned over the years, the theory has brought forward essential questions about the specificity of speech, the specialization of speech perception, and, more recently, the neuroanatomical substrate of a possible motor component of the speech apparatus (e.g., Gow & Segawa, 2009; Pulvermüller et al., 2006; Sussman, 1989; Whalen et al., 2006), a topic that regained interest following the discovery of mirror neurons in the premotor cortex

(e.g., Rizzolatti & Craighero, 2004; but see Lotto, Hickok, & Holt, 2009). The debate has also shifted to a focus on the extent to which the involvement of articulation during speech perception might in fact be under listener's control and its manifestation partly task-specific (Yuen, Davis, Brysbaert, & Rastle, 2010, Figure 5; see comments by McGettigan, Agnew, & Scott, 2010; Rastle, Davis, & Brysbaert, 2010). The Motor Theory has also been extensively reviewed—and revisited—in an attempt to address problems highlighted by auditory-based models, as described below (e.g., Fowler, 2006, 2008; Galantucci, Fowler, & Turvey, 2006; Lotto & Holt, 2006; Massaro & Chen, 2008).

Auditory Theory(ies)

The role of articulatory gestures in perceiving speech and the special status of the speech-perception system progressively came under attack largely because of insufficient hard evidence and lack of computational parsimony. Recall that recourse to articulatory gestures was originally posited as a way to solve the lack-of-invariance problem and turn a many(acoustic traces)-to-one(phoneme) mapping problem into a one(gesture)-to-one(phoneme) mapping solution. However, the lack of invariance problem turned out to be less prevalent and, at the same time, more complicated than originally claimed. Indeed, as mentioned earlier, many phonemes were found to preserve distinctive features across contexts (e.g., Blumstein & Stevens, 1981; Stevens & Blumstein, 1981). At the same time, lack of invariance was found in domains for which a gestural explanation was only of limited use, e.g., voice quality, loudness, speech rate.

Perhaps most problematic for gesture-based accounts was the finding by Kluender, Diehl, and Killeen (1987) that phonemic categorisation, which was viewed by such accounts as necessitating access to gestural primitives, could be observed in species lacking the anatomical prerequisites for articulatory knowledge and practice (Japanese quail; Figure 6). This result was seen by many as undermining both the motor component of speech perception and its human-specific nature. Parsimony became the new driving force. As Kluender et al. put it, "A theory of human phonetic categorization may need to be no more (and no less) complex than that required to explain the behavior of these quail" (p. 1197). The gestural explanation for compensation for coarticulation effects (Mann, 1980) was challenged by a general auditory mechanism as well. In Mann's experiment, the perceptual shift on the /da/-/ga/ continuum induced by the preceding /a/ vs. /ar/ context was explained by reference to articulatory gestures. However, Lotto and Kluender (1998) found a similar shift when the preceding context consisted of non-speech sounds mimicking the spectral characteristics of the actual syllables (e.g., tone glides). Thus, the acoustic composition of the context, and in particular its spectral contrast with the following syllable, rather than an underlying reference to abstract articulatory gestures, was able to account for Mann's context effect (but see Fowler, Brown, and Mann's, 2000, subsequent multimodal challenge to the auditory account).

However, auditory theories have been criticized for lacking in theoretical content. Auditory accounts are indeed largely based on counter-arguments (and counter-evidence) to the Motor and gestural theories, rather than resting on a clear set of falsifiable principles (Diehl, Lotto, & Holt, 2004). While it is clear that a great deal of phenomena previously believed to require a gestural account can be explained within an arguably simpler auditory framework, it remains to be seen whether auditory theories can provide a satisfactory explanation for the entire class of phenomena in which the many-to-one puzzle has been observed (e.g., Pardo & Remez, 2006).

Top-down Theories

This rubric and the following one (Bottom-up Theories) review theories of speech perception *broadly construed*. They are broadly construed in that they consider phonemic categorization, the scope of the *narrowly construed* theories, in the context of its interface with lexical knowledge. Although the traditional separation between narrowly- and broadly-construed theories originates from the respective historical goals of speech perception and spoken-word recognition research (Pisoni & Luce, 1987), an understanding of speech perception cannot be complete without an analysis of the impact of long-term knowledge on early sensory processes (see useful reviews in Goldinger, Pisoni, & Luce, 1996; Jusczyk & Luce, 2002).

The hallmark of top-down approaches to speech perception is that phonetic analysis and categorization can be influenced by knowledge stored in long-term memory, lexical knowledge in particular. As mentioned earlier, phoneme restoration studies (e.g., Warren & Obusek, 1971; Warren & Warren, 1970) showed that word knowledge could affect listeners' interpretation of what they heard, but they did not provide direct evidence that phonetic categorization *per se* (i.e., *perception*, as it was referred to in that literature) was modified by lexical expectations. However, Samuel (1981) demonstrated that auditory acuity was indeed altered when lexical information was available (e.g., "pr*gress" [from "progress"], with * indicating the portion on which auditory acuity was measured) compared to when it was not (e.g., "cr*gress" [from the non-word "crogess"]).

This kind of result (see also, e.g., Ganong, 1980; Marslen-Wilson & Tyler, 1980; and, more recently, Gow, Segawa, Ahlfors, & Lin, 2008) led to conceptualizing the speech system as being deeply interactive, with information flowing not only from bottom to top but also from top down. For example, the TRACE model (more specifically TRACE II, McClelland & Elman, 1986) is an interactive-activation model made of a large number of units organized into three levels: features, phonemes, and words (Figure 7A). The model includes bottom-up excitatory connections (from features to phonemes and from phonemes to words), inhibitory lateral connections (within each level), and, critically, top-down excitatory connections (from words to phonemes and from phonemes to features). Thus, the activation levels of features, e.g., voicing, nasality, burst, are partly determined by the activation levels of phonemes, and these are partly determined by the activation levels of words. In essence, this architecture places speech perception within a system that allows a given sensory input to yield a different *perceptual experience* (as opposed to interpretive experience) when it occurs in a word vs. a non-word or next to phoneme x vs. phoneme y, etc. TRACE has been shown to simulate a large range of perceptual and psycholinguistic phenomena, e.g., categorical perception, cue trading relations, phonetic context effects, compensation for coarticulation, lexical effects on phoneme detection/categorization, segmentation of embedded words, etc. All this takes place within an architecture that is neither domain- nor species-specific. Later instantiations of TRACE have been proposed by McClelland (1991) and Movellan and McClelland (2001), but all of them preserve the core interactive architecture described in the original model.

Like TRACE, Grossberg's Adaptive Resonance Theory (ART, e.g., Grossberg, 1986; Grossberg & Myers, 1999) suggests that perception emerges from a compromise, or stable state, between sensory information and stored lexical knowledge (Figure 7B). ART includes *items* (akin to sub-phonemic features or feature clusters) and *list chunks* (combinations of items whose composition is the result of prior learning; e.g., phonemes, syllables, or words). In ART, a sensory input activates items which, in turn, activate list chunks. List chunks feed back to

component items, and items back to list chunks again in a bottom-up/top-down cyclic manner that extends over time, ultimately creating stable resonance between a set of items and a list chunk. Both TRACE and ART posit that connections between levels are only excitatory and connections within levels are only inhibitory. In ART, in typical circumstances, attention is directed to large chunks (e.g., words), and hence the content of smaller chunks is generally less readily available. Small mismatches between large chunks and small chunks do not prevent resonance, but large mismatches do. In other words, unlike TRACE, ART does not allow the speech system to "hallucinate" information that is not already there (however, for circumstances in which it could, see Grossberg, 2000a). Large mismatches lead to the establishment of new chunks, and these gain resonance via subsequent exposure. In doing so, ART provides a solution to the stability-plasticity dilemma, that is, the unwanted erasure of prior learning by more recent learning (Grossberg, 1987), also referred to as *catastrophic interference* (e.g., McCloskey & Cohen, 1989).

Thus, like TRACE, ART posits that speech perception results from an online interaction between prelexical and lexical processes. However, ART is more deeply grounded in, and motivated by biologically plausible neural dynamics, where reciprocal connectivity and resonance states have been observed (e.g., Felleman & Van Essen, 1991). Likewise, ART replaces the hierarchical structure of TRACE with a more flexible one, in which tiers self-organize over time through competitive dynamics—as opposed to being predefined. Although sometimes accused of placing too few constraints on empirical expectations (Norris et al., 2000), the functional architecture of ART is thought to be more computationally economical than that of TRACE and more amenable to modelling both real-time and long-term temporal aspects of speech processing (Grossberg, Boardman, & Cohen, 1997).

Bottom-up Theories

Bottom-up theories describe effects of lexical and sentential knowledge on phoneme categorization as a consequence of post-perceptual biases. In this conceptualization, reporting "progress" when presented with "pr*gress" simply reflects a strategic decision to do so and the functionality of a system that is geared towards meaningful communication—we generally hear words rather than nonwords. Here, phonetic analysis itself is incorruptible by lexical or sentential knowledge: It takes place within an autonomous module that receives no feedback from lexical and post-lexical layers. In Cutler and Norris' (1979) Race model, phoneme identification is the result of a time race between a sublexical route and a lexical route activated in parallel in an entirely bottom-up fashion (Figure 7C). In normal circumstances, the lexical route is faster, which means that a sensory input that has a match in the lexicon (e.g., "progress") is usually read out from that route. A non-lexical sensory input (e.g., "crogress") is read out from the pre-lexical route. In this model, "pr*gress" is reported as containing the phoneme /o/ because the lexical route receives enough evidence to activate the word "progress" and, being faster, this route determines the response. In contrast, "cr*gress" does not lead to an acceptable match in the lexicon, and hence, read-out is performed from the sublexical route, with the degraded phoneme being faithfully reported as degraded.

Massaro's Fuzzy Logical Model of Perception (FLMP, Massaro, 1987, 1996; Oden & Massaro, 1978) also exhibits a bottom-up architecture, in which various sources of sensory input—e.g., auditory, visual—contribute to speech perception without any feedback from the lexical level (Figure 7D). In FLMP, acoustic-phonetic features are activated multi-modally and each feature accumulates a certain level of activation (on a continuous 0-to-1 scale) reflecting the

degree of certainty that the feature has appeared in the signal. The profile of features' activation levels is then compared against a prototypical profile of activation for phonemes stored in memory. Phoneme identification occurs when the match between the actual and prototypical profiles reaches a level determined by goodness-of-fit algorithms. Critically, the match does not need to be perfect to lead to identification; thus, there is no need for lexical top-down feedback. Pre-lexical and lexical sources of information are then integrated into a conscious percept. Although the extent to which the integration stage can be considered a true instantiation of bottom-up processes is a matter for debate (Massaro, 1996), FLMP also predicts that auditory acuity of * is fundamentally comparable in "pr*gress" and "cr*gress"—like Race and unlike top-down theories.

From an architectural point of view, integration between sublexical and lexical information is handled differently by Norris et al.'s (2000) Merge model. In Merge, the phoneme layer is duplicated into an input layer and a decision layer (Figure 7E). The phoneme input layer feeds forward to the lexical layer (with no top-down connections) and the phoneme decision layer receives input from both the phoneme input layer and the lexical layer. The phoneme decision layer is the place where phonemic and lexical inputs are integrated and where standard lexical phenomena arise (e.g., Ganong, 1980; Samuel, 1981). While both FLMP and Merge produce a decision by integrating unaltered lexical and sublexical information, the input received from the lexical level differs in the two models. In FLMP, lexical activation is relatively independent from the degree of activation of its component phonemes, whereas, in Merge, lexical activation is directly influenced by the pattern of activation sent upward by the phoneme input layer. While Merge has been criticized for exhibiting a contorted architecture (Gow, 2000; Samuel, 2000), being ecologically improbable (e.g., Grossberg, 2000b; Montant, 2000; Stevens, 2000), and being simply a late instantiation of FLMP (Massaro, 2000; Oden, 2000), it has gathered the attention of both speech-perception and spoken-word-recognition scientists around a question that is as yet unanswered.

Bayesian Theories

Despite important differences in functional architecture between top-down and bottom-up models, both classes of models agree that speech perception involves distinct levels of representations (e.g., features, phonemes, words), multiple lexical activation, lexical competition, integration (of some sort) between actual sensory input and lexical expectations, and corrective mechanisms (of some sort) to handle incompleteness or uncertainty in the input. A radically different class of models based on optimal Bayesian inference has recently emerged as an alternative to the above—recently in psycholinguistics at least. These models eschew the concept of lexical activation altogether, sometimes doing away with the bottom-up/top-down debate itself—or at a minimum blurring the boundaries between the two mechanisms. For instance, in their Shortlist B model, Norris and McQueen (2008) have replaced activation with the concepts of likelihood and probability, which are seen as better approximations of actual (i.e., imperfect) human performance in the face of actual (i.e., complex and variable) speech input. The appeal of Bayesian computations is substantial because output (or posterior) probabilities, e.g., probability that a word will be recognized, are estimated by tabulating both confirmatory and disconfirmatory evidence accumulated over past instances, as opposed to being tied to fixed activation thresholds (Figure 8). In particular, Shortlist B has replaced discrete input categories such as features and phonemes with phoneme likelihoods calculated from actual speech data. Because they are derived from real speech, the phoneme likelihoods vary from instance to

instance and as a function of the quality of the input and the phonetic context. Thus, while noisier, these probabilities are a better reflection of the type of challenge faced by the speech system in every-day conditions. They also allow the model to provide a single account for speech phenomena that usually require distinct ad-hoc mechanisms in other models. A general criticism levelled against Bayesian models, however, concerns the legitimacy of their *priors*, that is, the set of assumptions used to determine initial probabilities before any evidence has been gathered (e.g., how expected is a word or a phoneme *a priori*). Because priors can be difficult to establish, their arbitrariness or the modeller's own biases can have substantial effects on the model's outcome. Likewise, compared to the models reviewed above, models based on Bayesian inference often lead to less straightforward hypotheses, which makes their testability somewhat limited—even though their performance level in terms of replicating known patterns of data is usually high.

TAILORING SPEECH PERCEPTION: LEARNING AND RELEARNING

Learning

The literature reviewed so far suggests that perceiving speech involves a set of highly sophisticated processing skills and structures. To what extent are these skills and structures in place at birth? Of particular interest in the context of early theories of speech perception is the way in which speech perception and speech production develop relative to each other and the degree to which perceptual capacities responsible for subtle phonetic discrimination (e.g., voicing distinction) are present in pre-linguistic infants. Eimas, Siqueland, Jusczyk, and Vigorito (1971) showed that 1-month-old infants perceive a voicing-based /ba-/pa/ continuum categorically, just as adults do. Similarly, like adults (Mattingly, Liberman, Syrdal, & Halwes, 1971), young infants show a dissociation between categorical perception with speech and continuous perception with matched non-speech (Eimas, 1974). Infants also seem to start off with an open-ended perceptual system allowing them to discriminate a wide range of subtle phonetic contrasts—far more contrasts than they will be able to discriminate in adulthood (e.g., Aslin, Werker, Morgan, 2002; Trehub, 1976). There is therefore strong evidence that fine speech-perception skills are in place early in life—at least well before the onset of speech production—and operational with minimal, if any, exposure to ambient speech. These findings have led to the idea that speech-specific mechanisms are part of the human biological endowment and have been taken as evidence for the innateness of language, or at least some of its perceptual aspects (Eimas et al., 1971). In that sense, an infant has very little to *learn* about speech perception. If anything, attuning to one's native language is rather a matter of losing sensitivity to (or *un-learning*) phonetic contrasts that have little communicative value for that particular language, e.g., the /r/-/l/ distinction for Japanese listeners.

However, the idea that infants are born with a universal discrimination device operating according to a use-it-or-lose-it principle has not been unchallenged. For instance, on closer examination, discrimination capacities at the end of the first year appear far less acute and far less universal than expected (e.g., Lacerda & Sundberg, 2001). Likewise, discrimination of irrelevant contrasts does not wane as systematically and as fully as the theory would have it (e.g., Polka, Colantonio, & Sundara, 2001). For example, Bowers, Mattys, and Gage (2009) showed that language-specific phonemes learned in early childhood but never heard or produced subsequently, as would be the case for young children of temporarily expatriate parents, can be re-learned relatively easily even decades later (Figure 9 A). Thus, discriminatory attrition is not as widespread and severe as previously believed, suggesting that the representations of

phonemes from "forgotten" languages, i.e., those we stop practicing early in life, may be more deeply engraved in our long-term memory than we think.

By and large, however, the literature on early speech perception indicates that infants possess fine language-oriented auditory skills from birth as well as impressive capacities to learn from the ambient auditory scene during the first year of life (Figure 10). Auditory deprivation during that period (e.g., otitis media; delay prior to cochlear implantation) can have severe consequences on speech perception and later language development (e.g., Clarkson, Eimas, & Marean, 1989; Mody, Schwartz, Gravel, & Ruben, 1999), possibly due to a general decrease of attention to sounds (e.g., Houston, Pisoni, Kirk, Ying, & Miyamoto, 2003). However, even in such circumstances, partial sensory information is often available through the visual channel (facial and lip information), which might explain the relative resilience of basic speech perception skills to auditory deprivation. Indeed, Kuhl and Meltzoff (1982) showed that, as early as four months of age, infants show a preference for matched audio-visual inputs (e.g., audio /a/ with visual /a/) over mismatched inputs (e.g., audio /a/ with visual /i/). Even more striking, infants around that age seem to integrate discrepant audio-visual information following the typical McGurk pattern observed in adults (Rosenblum, Schmuckler, & Johnson, 1997). These results suggest that the multi-modal (or amodal) nature of speech perception, a central tenet of Massaro's Fuzzy Logical Model of Perception (FLMP, cf. Massaro, 1987), is present early in life and operates without much prior experience with sound-gesture association. Although the *strength* of the McGurk effect is lower in infants than adults (e.g., Massaro, Thompson, Barron, & Laren, 1986; McGurk & MacDonald, 1976), early cross-modal integration is often taken as evidence for gestural theories of speech perception and as a challenge to auditory theories.

Relearning

A question of growing interest concerns the flexibility of the speech-perception system when it is faced with an unstable or changing input. Can the perceptual categories learned during early infancy be undone or retuned to reflect a new environment? The issue of perceptual (re)learning is central to research on second-language (L2) perception and speech perception in degraded conditions. Evidence for a speech-perception sensitive period during the first year of life (Trehub, 1976) suggests that attuning to new perceptual categories later on should be difficult and perhaps not as complete as it is for categories learned earlier. Late learning of L2 phonetic contrasts (e.g., /r/-/l/ distinction for Japanese L1 speakers) has indeed been shown to be slow, effortful, and imperfect (e.g., Logan, Lively, & Pisoni, 1991). However, even in those conditions, learning appears to transfer to tokens produced by new talkers (Logan et al., 1991) and, to some degree, to production (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997). Successful learning of L2 contrasts is not systematically observed, however. For example, Bowers et al. (2009) found no evidence that English L1 speakers could learn to discriminate Zulu contrasts (e.g., /b/-/b̥/) or Hindi contrasts (e.g., /t/ vs. /t̪/) even after thirty days of daily training (Figure 9 B). Thus, although possible, perceptual learning of L2 contrasts is greatly constrained by the age of L2 exposure, the nature and duration of training, and the phonetic overlap between the L1 and L2 phonetic inventories (e.g., Best, 1994; Kuhl, 2000).

Perceptual learning of accented L1 and non-canonical speech follows the same general patterns as L2 learning, but it usually leads to faster and more complete retuning (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004). A reason for this difference is that, while L2 contrast learning involves the formation of new perceptual categories, whose boundaries are sometimes in direct conflict with L1 categories, accented L1 learning "simply" involves re-tuning existing

perceptual categories, often by broadening their mapping range. This latter feature makes perceptual learning of accented speech a special instance of the more general debate on the episodic versus abstract nature of phonemic and lexical representations. At issue, here, is whether phonemic and lexical representations consist of a collection of episodic instances in which surface details are preserved (voice, accent, speech rate) or, alternatively, single, abstract representations (i.e., one for each phoneme and one for each word). That at least some surface details of words are preserved in long-term memory is undeniable (e.g., Goldinger, 1998). The current debate focuses on: (1) whether lexical representations include both indexical (e.g., voice quality) and allophonic (e.g., phonological variants) details (Luce, McLennan, & Charles-Luce, 2003), (2) whether such details are of a lexical nature (i.e., stored within the lexicon), rather than sublexical (i.e., stored at the sub-phonemic, phonemic or syllabic level; McQueen, Cutler, & Norris, 2006), (3) the online time-course of episodic trace activation (e.g., Luce et al., 2003; McLennan, Luce, & Charles-Luce, 2005), (4) the mechanisms responsible for consolidating newly learned instances or new perceptual categories (e.g., Fenn, Nusbaum, & Margoliash, 2003), (5) the possible generalization to other types of non-canonical speech, such as disordered speech (e.g., Lee, Whitehill, & Ciocca, 2009; Mattys & Liss, 2008)

According to Samuel and Kraljic (2009), the above literature should be distinguished from a more recent strand of research which focuses on the specific variables affecting perceptual learning and the mechanisms linking such variables to perception. In particular, Norris, McQueen, and Cutler (2003) found that lexical information is a powerful source of perceptual recalibration. For example, Dutch listeners repeatedly exposed to a word containing a sound half-way between two existing phonemes (e.g., *witlo**, where * is ambiguous between /f/ and /s/, with *witlof* a Dutch word—chicorey—and *witlos* a nonword) subsequently perceived a /f/-/s/ continuum as biased in the direction of the lexically-induced percept (more /f/ than /s/ in the *witlo** case). Likewise, Bertelson, Vroomen, and de Gelder (2003) found that repeated exposure to McGurk audio-visual stimuli (e.g., audio /a*a/ and visual /aba/ leading to the auditory perception of /aba/) biased the subsequent perception of an audio-only /aba/-/ada/ continuum in the direction of the visually-induced percept. Although visually-induced perceptual learning seems to be less long-lasting than its lexically-induced counterpart (Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004), the Norris et al. and Bertelson et al. studies demonstrate that even the mature perceptual system can show a certain degree of flexibility when it is faced with a changing auditory environment.

SPEECH RECOGNITION BY MACHINES

This chapter was mainly concerned with Human Speech Recognition (HSR), but technological advances in the past decades have allowed the topic of speech perception and recognition to become an economically profitable challenge for engineers and applied computer scientists. A complete review of Automatic Speech Recognition's (ASR) historical background, issues, and state of the art is beyond the scope of this chapter. However, a brief analysis of ASR in the context of the key topics in HSR reviewed above reveals interesting commonalities as well as divergences among the preoccupations and goals of the two fields.

Perhaps the most notable difference between HSR and ASR is their ultimate aim. Whereas HSR aims to provide a description of how the speech system works (processes, representations, functional architecture, biological plausibility), ASR aims to deliver speech transcriptions as error-free as possible, regardless of the biological and cognitive validity of the underlying algorithms. The success of ASR is typically measured by the percentage of words

correctly identified from speech samples varying in their acoustic and lexical complexity. While increasing computer capacity and speed have allowed ASR performance to improve substantially since the early systems of the 1970s (e.g., Jelinek, 1976; Klatt, 1977), ASR accuracy is still about an order of magnitude behind its HSR counterpart (Moore, 2007, see Figure 11).

What is the cause of the enduring performance gap between ASR and HSR? Given that the basic constraints imposed by the signal (sequentiality, continuity, variability) are the same for humans and machines, it is tempting to conclude that the gap between ASR and HSR will not be bridged until the algorithms of the former resemble those of the latter. And currently, they do not. The architecture of most ASR systems is almost entirely data-driven: Its structure is expressed in terms of a network of sequence probabilities calculated over large corpora of natural speech (and their supervised transcription). The ultimate goal of the corpora, or training data, is to provide a database of acoustic-phonetic information sufficiently large that an appropriate match can be found for any input sound sequence. The larger the corpora, the tighter the fit between the input and the acoustic model (e.g., triphones instantiated in hidden Markov models, HMM, cf. Rabiner & Juang, 1993), and the lower the ASR system's error rate (Lamel, Gauvain, & Adda, 2000). By that logic, hours of training corpora, not human-machine avatars, are the solution for increased accuracy, giving support to the controversial assertion that human models have so far hindered rather than promoted ASR progress (Jelinek, 1985). However, Moore and Cutler (2001) estimated that increasing corpus sizes from their current average capacity (one thousand hours or less, which is the equivalent of the average hearing time of a two-year-old) to 10,000 hours (average hearing time of a ten-year-old) would only drop the ASR error rate to 12%.

Thus, a data-driven approach to speech recognition is constrained by more than just the size of the training data set. For example, the lexical and syntactic content of the training data often determines the application for which the ASR system is likely to perform best. Domain-specific systems (e.g., banking transactions by phone) generally reach high recognition accuracy levels even when they are fed continuous speech produced by various speakers, whereas domain-general systems (e.g., speech-recognition packages on personal computers) often have to compromise on the number of speakers they can recognize and/or training time in order to be effective (Everman et al., 2005). Therefore, one of the current stumbling blocks of ASR systems is language modelling (as opposed to acoustic modelling), that is, the extent to which the systems include higher-order knowledge—syntax, semantics, pragmatics—from which inferences can be made to refine the mapping between the signal and the acoustic model. Existing ASR language models are fairly simple, drawing upon the distributional methods of acoustic models in that they simply provide the probability of all possible word sequences based on their occurrences in the training corpora. In that sense, an ASR system can predict that "necklace" is a possible completion of "The burglar stole the..." because of its relatively high transitional probability in the corpora, not because of the semantic knowledge that burglars tend to steal valuable items, and not because of the syntactic knowledge that a noun phrase typically follows a transitional verb. Likewise, ASR systems rarely include the kind of lexical feedback hypothesized in HSR models like TRACE (McClelland & Elman, 1986) and ART (Grossberg, 1986). Like Merge (Norris et al., 2000), ASR systems only allow lexical information and the language model to influence the relative weights of activated candidates, but not the fit between the signal and the acoustic model (Scharenborg Norris, ten Bosch, & McQueen, 2005).

While the remaining performance gap between ASR and HSR is widely recognized in the ASR literature, there seems to be no clear consensus on the direction to take in order to reduce it

(Moore, 2007). Given today's ever expanding computer power, increasing the size of training corpora is probably the easiest way of gaining a few percentage points in accuracy, at least in the short-term. More radical solutions are also being envisaged, however. For example, attempts are being made to build more linguistically plausible acoustic models by using phonemes (as opposed to di/triphone HMMs) as basic segmentation units (Ostendorf, Digilakis, & Kimball, 1996; Russell, 1993) or by preserving and exploiting fine acoustic detail in the signal instead of treating it as noise (Carlson & Hawkins, 2007; Moore & Maier, 2007).

CONCLUSION

The scientific study of speech perception started in the early 1950s under the impetus of research carried out at the Haskins Laboratories, following the development of the *Pattern Playback* device. This machine allowed Franklin S. Cooper and his colleagues to visualize speech in the form of a decomposable spectrogram and, reciprocally, to create artificial speech by *sounding out* the spectrogram. Contemporary speech perception research is both a continuation of its earlier preoccupations with the building blocks of speech perception and a departure from them. On the one hand, the quest for universal units of speech perception and attempts to crack the many-to-one mapping code are still going strong. Still alive, too, is the debate about the involvement of gestural knowledge in speech perception, re-ignited recently by neuro-imaging techniques and the discovery of mirror neurons. On the decline are the ideas that speech is special with respect to audition and that infants are born with speech- and species-specific perceptual capacities. On the other hand, questions have necessarily spread beyond the sublexical level, following the assumption that decoding the sensory input must be investigated in the context of the entirety of the language system—or, at the very least, some of its phonologically-related components. Indeed, lexical feedback, online or learning-related, has been shown to modulate the perceptual experience of an otherwise unchanged input. Likewise, what used to be treated as speech surface details (e.g., indexical variations), and commonly filtered out for the sake of modelling simplicity, are now more fully acknowledged as being preserved during encoding, embedded in long-term representations, and used during retrieval. Speech-perception research in the coming decades is likely to expand its interest not only to the rest of the language system but also to domain-general cognitive functions such as attention and memory as well as practical applications (e.g., ASR) in the field of artificial intelligence. At the same time, researchers have become increasingly concerned with the external validity of their models. Attempts to enhance the ecological contribution of speech research is manifest in a sharp increase in studies using *natural* speech (conversational, accented, disordered) as the front-end of their models.

REFERENCES

- Aslin, R. N., Werker, J. F., & Morgan, J. L. (2002). Innate phonetic boundaries revisited. *Journal of the Acoustical Society of America*, *112*, 1257-1260.
- Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592-597.
- Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In H. Nusbaum & J. Goodman (Eds.), *The transition from speech sounds to spoken words: The development of speech perception* (pp. 167-224). Cambridge, MA: MIT Press.
- Blumstein, S. E. & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, *10*, 25-32
- Bowers, J. S., Mattys, S. L., & Gage, S. H. (2009). Preserved implicit knowledge of a forgotten childhood language. *Psychological Science*, *20*, 1064-1069.
- Bradlow, A. R. & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707-729.
- Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2299-2310.
- Carlson, R. & Hawkins, S. (2007). When is fine phonetic detail a detail? *Proceedings of the 16th ICPHS Meeting* (pp. 211-214), Saarbrücken, Germany.
- Clarke, C. M. & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116*, 3647-3658.
- Clarkson, R. L., Eimas, P. D., & Marean, G. C. (1989). Speech perception in children with histories of recurrent otitis media. *Journal of the Acoustical Society of America*, *85*, 926-933.
- Connine, C. M. & Clifton, C. (1987) Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 291-299.
- Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, *92*, 81-104
- Cutler, A. & Norris, D. (1979). Monitoring sentence comprehension. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 113-134). Lawrence Erlbaum Associates: NJ.
- Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science*.
- Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*, 769-773.
- Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149-179.

- Dupoux, E., & Mehler, J. (1990). Monitoring the lexicon with normal and compressed speech: Frequency effects and the prelexical code. *Journal of Memory & Language*, 29, 316-335.
- Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception & Psychophysics*, 16, 513-521.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171, 303-306.
- Everman, G., Chan, H.Y., Gales, M.J.F, Jia, B., Mrva, D, Woodland, P.C. (2005). Training LVCSR systems on thousands of hours of data. Proceedings of the IEEE ICASSP, Philadelphia, pp. 209-212.
- Felleman, D. & Van Essen, D. (1991). Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, 1, 1-47.
- Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, 425, 614-616.
- Fougeron, C. & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, 101, 3728-3740.
- Foulke, E. & Sticht, T. G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, 72, 50-62.
- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3-28.
- Fowler, C. A. (1996). Listeners do hear sounds not tongues. *Journal of the Acoustical Society of America*, 99, 1730-1741.
- Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, 68, 161-177.
- Fowler, C. A. (2008). The FLMP STMPed. *Psychonomic Bulletin & Review*, 15, 458-462
- Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception & Performance*, 26, 877-888.
- Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13, 361-377.
- Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 110-125.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251-279.
- Goldinger, S. D., Pisoni, D. B., & Luce, P. A. (1996). Speech perception and spoken word recognition: Research and Theory. In N.J. Lass (Ed.), *Principles of Experimental Phonetics* (pp. 277-327). St. Louis: Mosby.
- Gow, D. W. (2000). One phonemic representation should suffice. *Behavioral and Brain Science*, 23, 331.

- Gow D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F. H. (2008). Lexical influences on speech perception: A Granger causality analysis of MEG and EEG source estimates. *Neuroimage*, *43*, 614-23.
- Gow D. W., & Segawa, J. A. (2009). Articulatory mediation of speech perception: A causal analysis of multi-modal imaging data. *Cognition*, *110*, 222-236.
- Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab and H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines, Vol 1: Speech perception* (pp. 187-294). Academic Press.
- Grossberg, S. (1987). Competitive learning: From interactive activations to adaptive resonance. *Cognitive Science*, *11*, 23-63
- Grossberg, S. (2000a). How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society*, *6*, 579-588.
- Grossberg, S. (2000b). Brain feedback and adaptive resonance in speech perception. *Behavioral and Brain Science*, *23*, 332-333.
- Grossberg, S. & Myers, C. (1999). The resonant dynamics of conscious speech: Interword integration and duration-dependent backward effects. *Psychological Review*, *107*, 735-767.
- Grossberg, S., Boardman, I., & Cohen, M. A. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 481-503.
- Houston, D. M., Pisoni, D. B., Kirk, K. I., Ying, E. A., & Miyamoto, R. T. (2003). Speech perception skills of infants following cochlear implantation: A first report. *International Journal of Pediatric Otorhinolaryngology*, *67*, 479-495.
- Huggins, A.W. F. (1975). Temporally segmented speech and “echoic” storage. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 209-225). New York: Springer-Verlag.
- Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*, 532-556.
- Jelinek, F. (1985). *Every time I fire a linguist, the performance of my system goes up*. Public statement at the IEEE ASSPS Workshop on Frontiers of Speech Recognition, Tampa, Florida.
- Jusczyk, P. W., & Luce, P. A. (2002). Speech Perception and Spoken Word Recognition: Past and Present. *Ear and Hearing*, *23*, 2-40.
- Klatt, D. H. (1977). Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America*, *62*, 1345-1366.
- Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese Quail can form phonetic categories. *Science*, *237*, 1195-1197.
- Kuhl, P. K. (1981). Discrimination of speech by non-human animals: Basic auditory sensitivities conducive to the perception of *speech-sound* categories, *Journal of the Acoustical Society of America*, *95*, 340-349.

- Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences USA*, 97, 11850-11857.
- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5, 831-843.
- Kuhl, P. K. & Meltzoff, A. N. (1982). The bimodal development of speech in infancy. *Science*, 218, 1138-1141.
- Lacerda, F. & Sundberg, U. (2001). Auditory and articulatory biases influence the initial stages of the language acquisition process. In F. Lacerda, C. von Hofsten, & M. Heimann (Eds.), *Emerging Cognitive Abilities in Early Infancy* (pp. 91-110). Mahwah, NJ: Lawrence Erlbaum.
- Lachs, L., Pisoni, D. B., & Kirk, K. I. (2001). Use of audio-visual information in speech perception by pre-lingually deaf children with cochlear implants: A first report. *Ear & Hearing*, 22, 236-251.
- Lee, A., Whitehall, T. L., & Coccia, V. (2009). Effect of listener training on perceptual judgement of hypernasality. *Clinical Linguistics & Phonetics*, 23, 319-334.
- Lamel, L., Gauvain, J.-L., & Adda, G. (2000). Lightly supervised acoustic model training. *Proceeding of the ISCA Workshop on Automatic Speech Recognition*, 150-154.
- Lieberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.
- Lieberman, A. M., Delattre, P.C., & Cooper, F.S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position, *Language and Speech*, 1, 153-167.
- Lieberman, A. M., Harris, K. S., Eimas, P., Lisker, L., & Bastian, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 4, 175-195.
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.
- Lieberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1-36.
- Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773-1781.
- Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication*, 22, 1-16.
- Lisker, L., & Abramson, A. S. (1970). The voicing dimensions: Some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences*. Prague: Academia.
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89, 874-886.
- Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, 68, 178-183.

- Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*, 602-619.
- Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Science*, *13*, 110-114.
- Luce, P. A., McLennan, C. T., & Charles-Luce, J. (2003). Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In J. Bowers & C. Marsolek (Eds.), *Rethinking implicit memory* (pp. 197-214). New York: Oxford University Press.
- Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmas past. *Cognitive Science*, *27*, 285-298.
- Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*, 407-412.
- Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71-102.
- Marslen-Wilson, W. D. & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1-71.
- Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Erlbaum.
- Massaro, D. W. (1989). Testing between the TRACE model and the Fuzzy Logical Model of Speech Perception. *Cognitive Psychology*, *21*, 398-421.
- Massaro, D. W. (1996). Integration of multiple sources of information in language processing. In T. Inui & J. L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 397-432). MIT Press.
- Massaro, D. W. (2000). The horse race to language understanding: FLMP was first out of the gate and has yet to be overtaken. *Behavioral and Brain Science*, *23*, 338-339.
- Massaro, D. W. & Chen, T. H. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review*, *15*, 453-457.
- Massaro, D. W., Thompson, L. A., Barron, B. , & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, *41*, 93-113.
- Mattingly, I. G., Liberman, A. M., Syrda, A. K., & Halwes T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, *2*, 131-157.
- Mattys, S. L. (1997). The use of time during lexical processing and segmentation: A review. *Psychonomic Bulletin & Review*, *4*, 310-329.
- Mattys, S.L. & Liss, J.M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as from artificial normality. *Perception & Psychophysics*, *70*, 1235-1242.
- Mattys, S. L., White, L., & Melhorn, J. F (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*, 477-500.

- McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, *23*, 1-44.
- McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.
- McCloskey, M. & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, *24*, 109-165.
- McGettigan, C. , Agnew, Z. K., & Scott, S. K. (2010). Are articulatory commands automatically and involuntarily activated during speech perception? *Proceedings of the National Academy of Sciences USA*, *107*, E42.
- McGurk, H. & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*, 306-321.
- McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*, 21-46.
- McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113-1126.
- Miller, J. L. & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*, 457-465.
- Miller, J. D. Wier, C. C., Pastore, R., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, *60*, 410-417.
- Mody, M., Schwartz, R. G., Gravel, R. S., & Ruben, R. J. (1999). Speech perception and verbal memory in children with and without histories of otitis media. *Journal of Speech, Language and Hearing Research*, *42*, 1069-1079.
- Montant, M. (2000). Feedback: A general mechanism in the brain. *Behavioral and Brain Science*, *23*, 340-341.
- Moore, R. K. (2007). Spoken language processing by machine. In G. Gaskell (Ed.), *Oxford Handbook of Psycholinguistics* (pp. 723-738). Oxford, UK: Oxford University Press.
- Moore, R. K. & Cutler, A. (2001). Constraints on theories of human vs. machine recognition of speech. Proceedings of SPRAAC Workshop on Human Speech Recognition as Pattern Classification, Max-Planck-Institute for Psycholinguistics, Nijmegen, 11-13 July, pp. 145-150.
- Moore, R. K. & Maier, V. (2007). Preserving fine phonetic detail using episodic memory: Automatic speech recognition using MINERVA2. *Proceedings of the 16th ICPHS Meeting*, Saarbrücken, Germany.
- Movellan, J. R. & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113-148.

- Nittrouer, S. (2001). Challenging the notion of innate phonetic boundaries. *Journal of the Acoustical Society of America*, *110*, 1598-1605.
- Nooteboom, S. G. (1979). The time course of speech perception. In W. J. Barry & K. J. Kohler (Eds.), *"Time" in the production and perception of speech* (Arbeitsberichte 12). University of Kiel: Institut für Phonetik.
- Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189-234.
- Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral & Brain Sciences*, *23*, 299-370.
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204-238.
- Oden, G. C. (2000). Implausibility versus misinterpretation of the FLMP. *Behavioral and Brain Science*, *23*, 344.
- Oden, G.C. & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172-191.
- Ostendorf, M., Digilakis, V., & Kimball, O. A. (1996). From HMMs to segment models: A unified view of stochastic modelling for speech recognition. *IEEE Transactions, Speech and Audio Processing*, *4*, 360-378.
- Pardo, J. S., & Remez, R. E. (2006). The perception of speech. In M. Traxler and M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics, 2nd Edition* (pp. 201-248). New York: Academic Press.
- Pisoni, D. B. & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*, 21-52.
- Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of /d/-/ð/ perception: Evidence for a new developmental pattern. *Journal of the Acoustical Society of America*, *109*, 2190-2201.
- Potter, R. K., Kopp, G. A., & Green, H. C. (1947). *Visible speech*. D. Van Nostrand Company, Inc. New York.
- Port, R. F. (1977). The influence of speaking tempo on the duration of stressed vowel and medial stop in English Trochee words. Doctoral dissertation, Indiana University. Bloomington, IN: Indiana University Linguistics Club.
- Pulvermüller, F., Huss, M., Kherif, F., Moscoso Del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences USA*, *103*, 7865-7870.
- Radeau, M., Morais, J., Mousty, P., & Bertelson, P. (2000). The effect of speaking rate on the role of the uniqueness point in spoken word recognition. *Journal of Memory and Language*, *42*, 406-422.

- Rastle, K., Davis, M. H., & Brysbaert, M., (2010). Response to McGettigan et al.: Task-based accounts are not sufficiently coherent to explain articulatory effects in speech perception. *Proceedings of the National Academy of Sciences USA*, 107, E43.
- Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In R. Campbell and B. Dodd (Eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 97-114). Hillsdale, N.J.: Erlbaum Associates.
- Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169-192,
- Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni and R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51-78). Oxford: Blackwell.
- Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, 59, 347-357.
- Russell, M. J. (1993). A segmental HMM for speech pattern modeling. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 640-643.
- Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, 110, 474-94.
- Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General*, 125, 28-51.
- Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, 32, 97-127.
- Samuel, A. G. (2000). Merge: Contorted architecture, distorted facts, and purported autonomy. *Behavioral and Brain Science*, 23, 345-346.
- Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, 12, 348-351.
- Samuel, A. G. & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71, 1207-1218.
- Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science*, 29, 867-918.
- Stevens, K. N. (2000). Recognition of continuous speech requires top-down processing. *Behavioral and Brain Science*, 23, 348.
- Stevens, K. N. & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. Eimas and J. Miller (Eds.), *Perspectives on the study of speech* (pp. 1-38). Hillsdale, NJ: Lawrence Erlbaum.
- Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.
- Sussman H. M. (1989). Neural coding of relational invariance in speech: Human language analogs to the barn owl. *Psychological Review*, 96, 631-642.

Summerfield, A. Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1074-1095.

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development*, 47, 466-472.

Umeda, N. & Coker, C. H. (1974). Allophonic variation in American English. *Journal of Phonetics*, 2, 1-5.

van Buuren RA, Festen J, Houtgast T. (1999). Compression and expansion of the temporal envelope: evaluation of speech intelligibility and sound quality. *Journal of the Acoustical Society of America*, 105, 2903-2913.

Vroomen, J., Van Linden, B., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication*, 44, 55-61.

Whalen, D. H., Benson, R. R., Richardson, M., Swainson, B., Clark, V. P., Lai, S., Mencl, W. E., Fulbright, R. K., Constable, R. T, & Liberman, A. M. (2006). Differentiation of speech and nonspeech processing within primary auditory cortex. *Journal of the Acoustical Society of America*, 119, 575-581.

Yuen, I., Davis, M. H., Brysbaert, M., Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences USA*, 107, 592-597.

Figure captions

Figure 1. Illustration of the sequential nature of speech processing. A. Waveform of a complete sentence, i.e., air pressure changes (Y axis) over time (X axis). B-C-D. Illustration of a listener's progressive processing of the sentence at three successive points in time. The visible waveform represents the portion of signal that is available for processing at time t1 (B), t2 (C) and t3 (D).

Figure 2. Sketch of Mattys, White, and Melhorn's (2005) hierarchical approach to speech segmentation. The relative weights of speech segmentation cues are illustrated by the width of the grey triangle. In optimal listening conditions, the cues in Tier I dominate. When lexical access is compromised or ambiguous, the cues in Tier II take over. Cues from Tier III are recruited when both lexical and segmental cues are compromised (e.g., background of severe noise). [reprinted from Mattys, S. L., White, L., & Melhorn, J. F (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, 134, 477-500 (Figure 7), by permission of the American Psychological Association].

Figure 3. A. Stylized spectrograms of /di/ and /du/. The dark bars, or formants, represent areas of peak energy on the frequency scale (Y axis), which correlate with zones of high resonance in the vocal tract. The curvy leads into the formants are formant transitions. They show coarticulation between the consonant and the following vowel. Note the dissimilarity between the second formant transitions in /di/ (rising) and /du/ (falling). However, as shown in B, the extrapolation back in time of the two second formants' transitions point to a common frequency locus.

Figure 4. Idealized identification pattern (solid line, left Y axis) and discrimination pattern (dashed line, right Y axis) for categorical perception. Illustration with a /ba/ to /da/ continuum. Identification shows a sharp perceptual boundary between categories. Discrimination is finer around the boundary than inside the categories.

Figure 5. Electropalatographic data showing the proportion of tongue contact on alveolar electrodes during the initial and final portions of /k/-initial (e.g., *kib*) or /s/-initial (e.g., *sib*) syllables (collapsed) while a congruent or incongruent distractor is presented (Yuen et al., 2010). The distractor was presented auditorily in conditions A and B and visually in condition C. With the target *kib* as an example, the congruent distractor in the A condition was *kib* and the incongruent distractor started with a phoneme involving a different place of articulation (e.g., *tib*). In condition B, the incongruent distractor started with a phoneme that differed from the target only by its voicing status, not by its place of articulation (e.g., *gib*). Condition C was the same as condition A, except that the distractors was presented visually. The results show “traces” of the incongruent distractors in target production when the distractor is in articulatory competition with the target, particularly in the early portion of the phoneme (condition A), but not when it involves the same place of articulation (condition B), or when it is presented visually (condition C). The results suggest a close relationship between speech perception and speech production. [reprinted from Yuen, I., Davis, M. H., Brysbaert, M., Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences USA*, 107, 592-597 (Figure 2), by permission of the National Academy of Sciences].

Figure 6. Pecking rates at test for positive stimuli (/dVs/) and negative stimuli (all others) for one of the quail in Kluender et al.'s (1987) study in eight vowel contexts. The test session was

preceded by a learning phase in which the quail learned to discriminate /dVs/ syllables (i.e., syllables starting with /d/ and ending with /s/, with a varying intervocalic vowel) from /bVs/ and /gVs/ syllables, with four different intervocalic vowels not used in the test phase. During learning, the quail was rewarded for pecking in response to /d/-initial syllables (positive trials) but not to /b/- and /g/-initial syllables (negative trials). The figure shows that, at test, the quail pecked substantially more to positive than negative syllables even though these syllables contained entirely new vowels, that is, vowels leading to different formant transitions with the initial consonant than those experienced during the learning phase. [reprinted from Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese Quail can form phonetic categories. *Science*, 237, 1195-1197 (Figure 1), by permission of the National Academy of Sciences]

Figure 7. Simplified architecture of: (A) TRACE, (B) ART, (C) Race, (D) FLMP, and (E) Merge. Layers are labeled consistently across models for comparability. Excitatory connections are denoted by arrows. Inhibitory connections are denoted by closed black circles.

Figure 8. Main Bayesian equation in Shortlist B (Norris & McQueen, 2008). $P(word_i|evidence)$ is the conditional probability of a specific word ($word_i$) having been heard given the available (intact or degraded) input ($evidence$). $P(word_i)$ represents the listener's prior belief, before any perceptual evidence has been accumulated, that $word_i$ will be present in the input. $P(word_i)$ can be approximated from lexical frequencies and contextual variables. The critical term of the equation is $P(evidence|word_i)$, which is the likelihood of the evidence given $word_i$, that is, the product of the probabilities of the sublexical units (e.g., phonemes) making up $word_i$. This term is important because it acknowledges and takes into account the variability of the input (noise, ambiguity, idiosyncratic realization, etc.) in the input-to-representation mapping process. The probability of $word_i$ so calculated is then compared to that of all other words in the lexicon (n). Thus, Bayesian inference provides an index of word recognition that considers both lexical and sublexical factors as well as the complexity of a real and variable input.

Figure 9. A. AX discrimination scores over 30 consecutive days (50% chance level; feedback provided) for Zulu contrasts (e.g., /b/-/β/) and Hindi contrasts (e.g., /t/ vs. /ʈ/) by DM, a 20-year-old male native-English speaker who was exposed to Zulu from four to eight years of age but never heard Zulu subsequently. Note DM's improvement with the Zulu contrasts over the 30 days, in sharp contrast with his inability to learn the Hindi contrasts. B. Performance on the same task by native English speakers with no prior exposure to Zulu or Hindi. [adapted from Bowers, J. S., Mattys, S. L., & Gage, S. H. (2009). Preserved implicit knowledge of a forgotten childhood language. *Psychological Science*, 20, 1064-1069 (partial Figure 1)]

Figure 10. Summary of key developmental landmarks for speech perception and speech production in the first year of life. [reprinted from Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5, 831-843 (Figure 1), by permission of the Nature Publishing Group]

Figure 11. Comparison of speech recognition error rate by machines (ASR) and humans. The logarithmic scale on the Y axis shows that ASR performance is approximately one order of magnitude behind human performance across various speech materials (ASR error rate for telephone conversation: 43%). The data were collated by Lippmann (1997). [reprinted from

Moore, R. K. (2007). Spoken language processing by machine. In G. Gaskell (Ed.), *Oxford Handbook of Psycholinguistics* (pp. 723-738). Oxford, UK: Oxford University Press (Figure 44.6), by permission of Oxford University Press]

Figure 1.

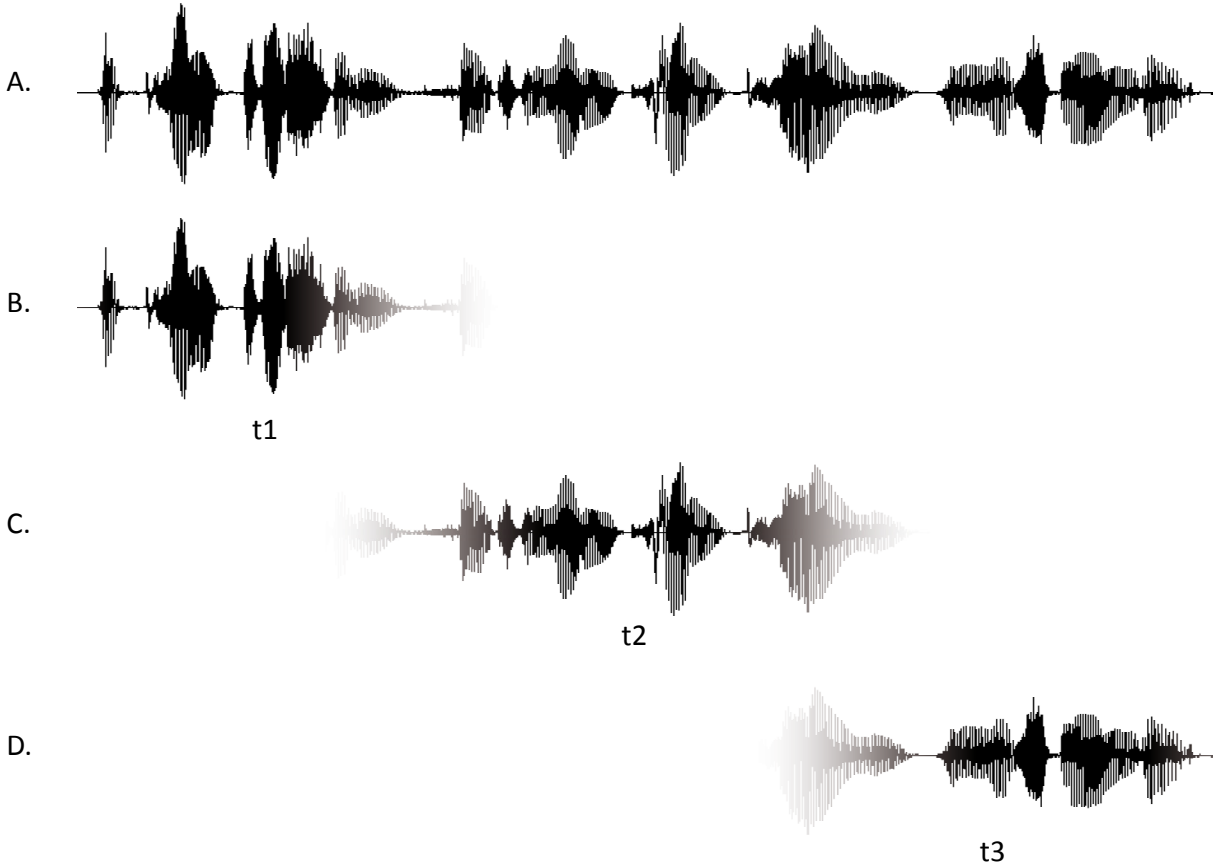


Figure 2.

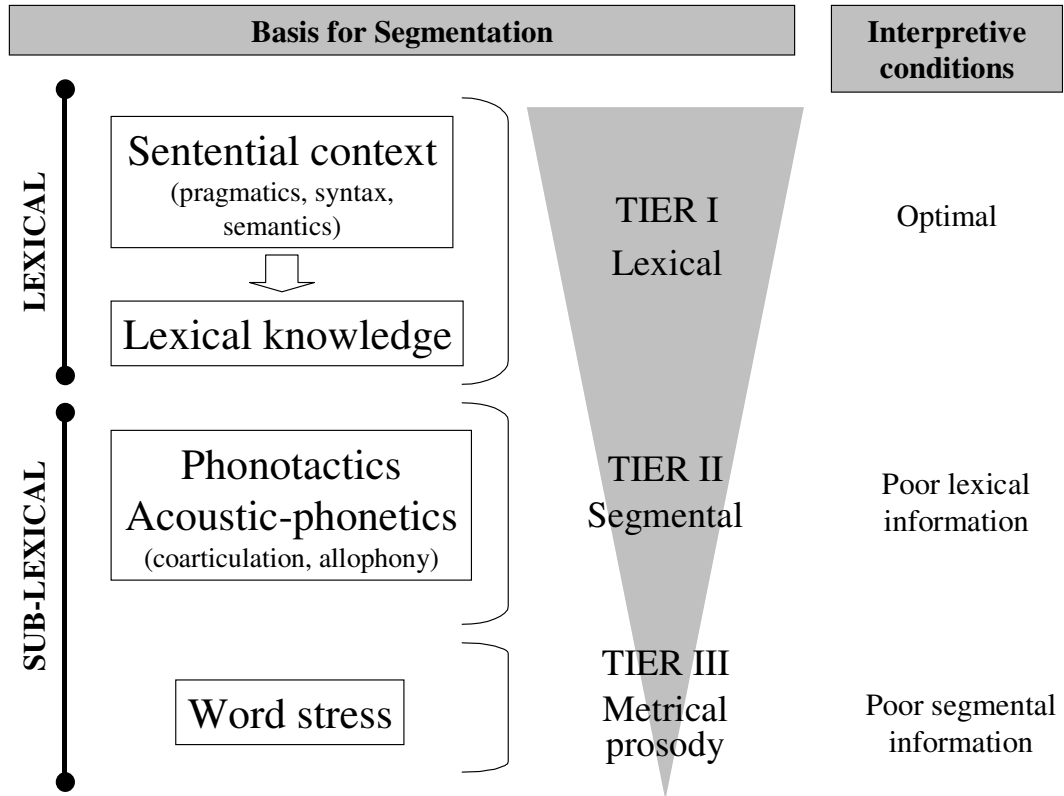


Figure 3.

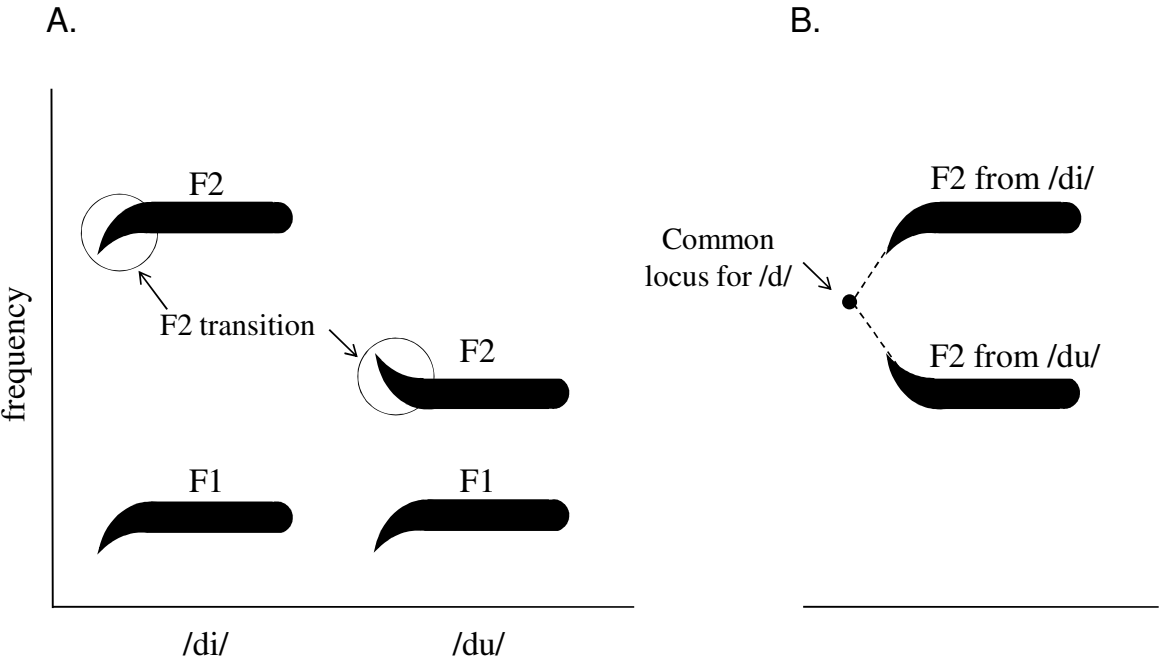


Figure 4.

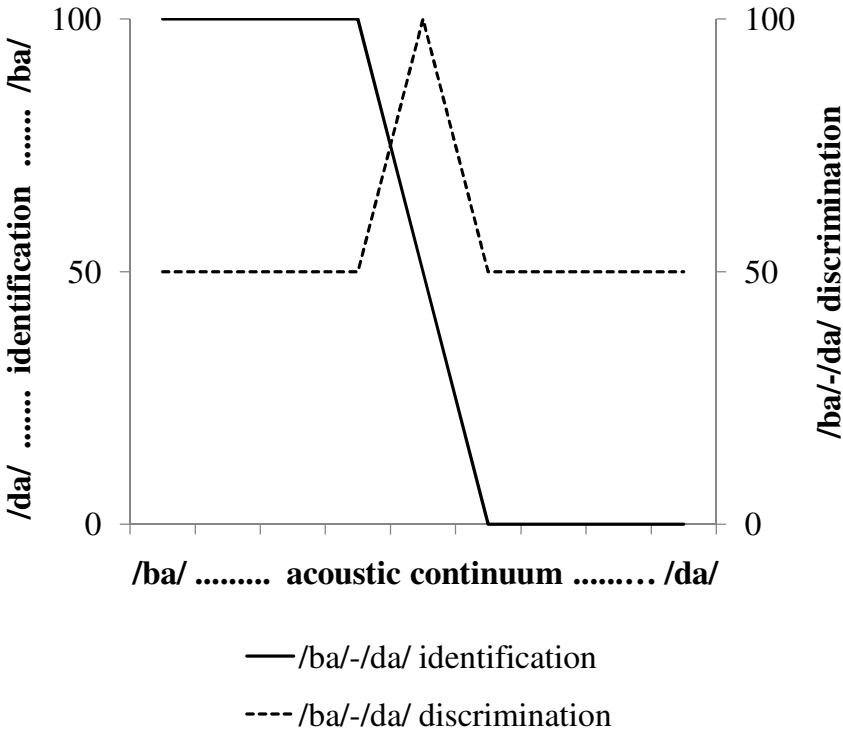


Figure 5.

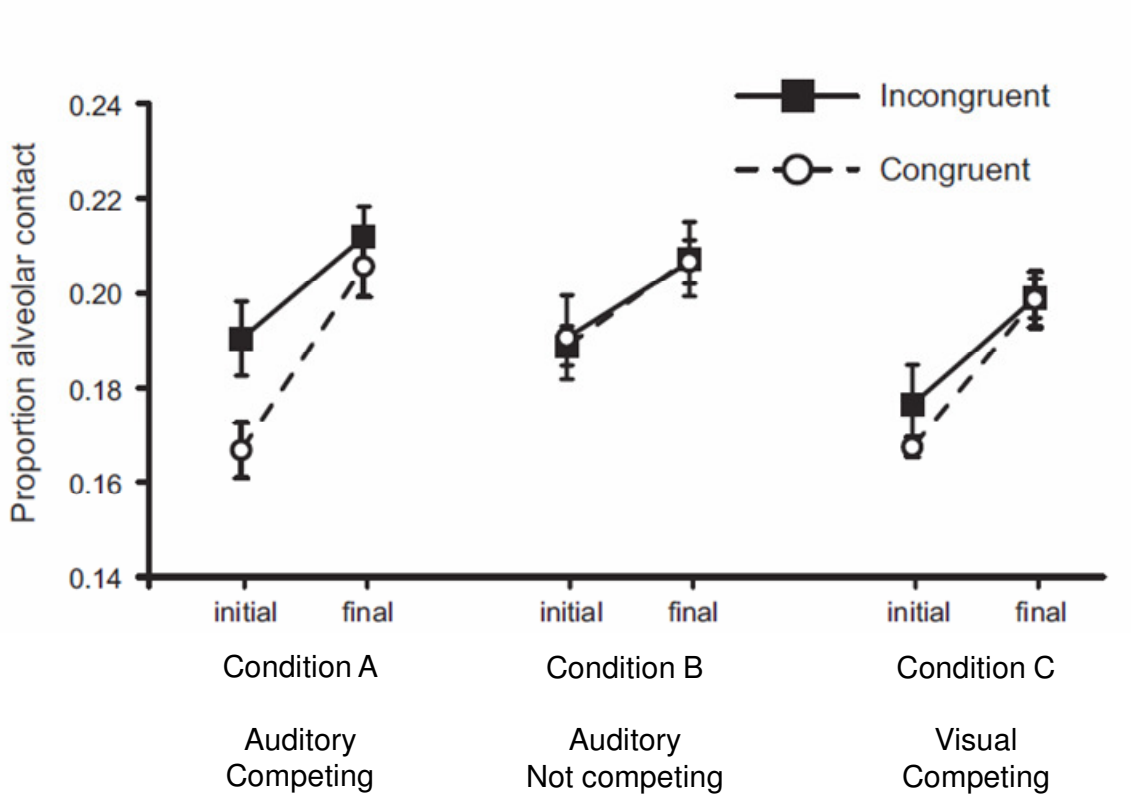


Figure 6.

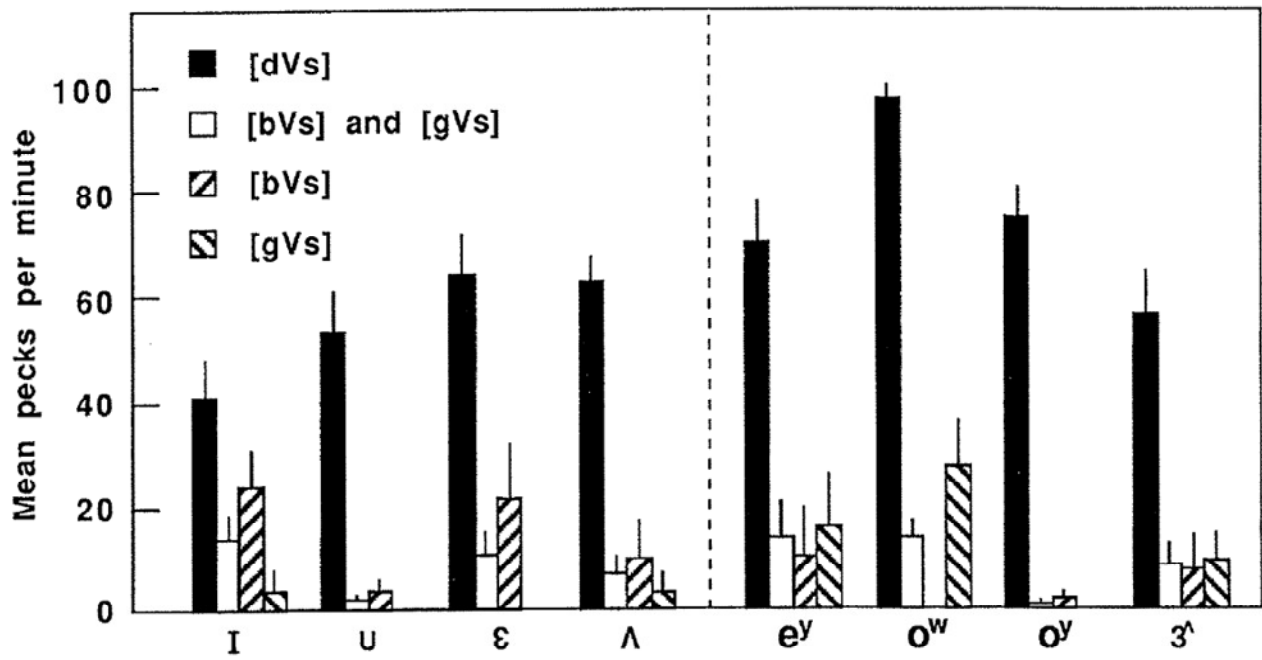


Figure 7.

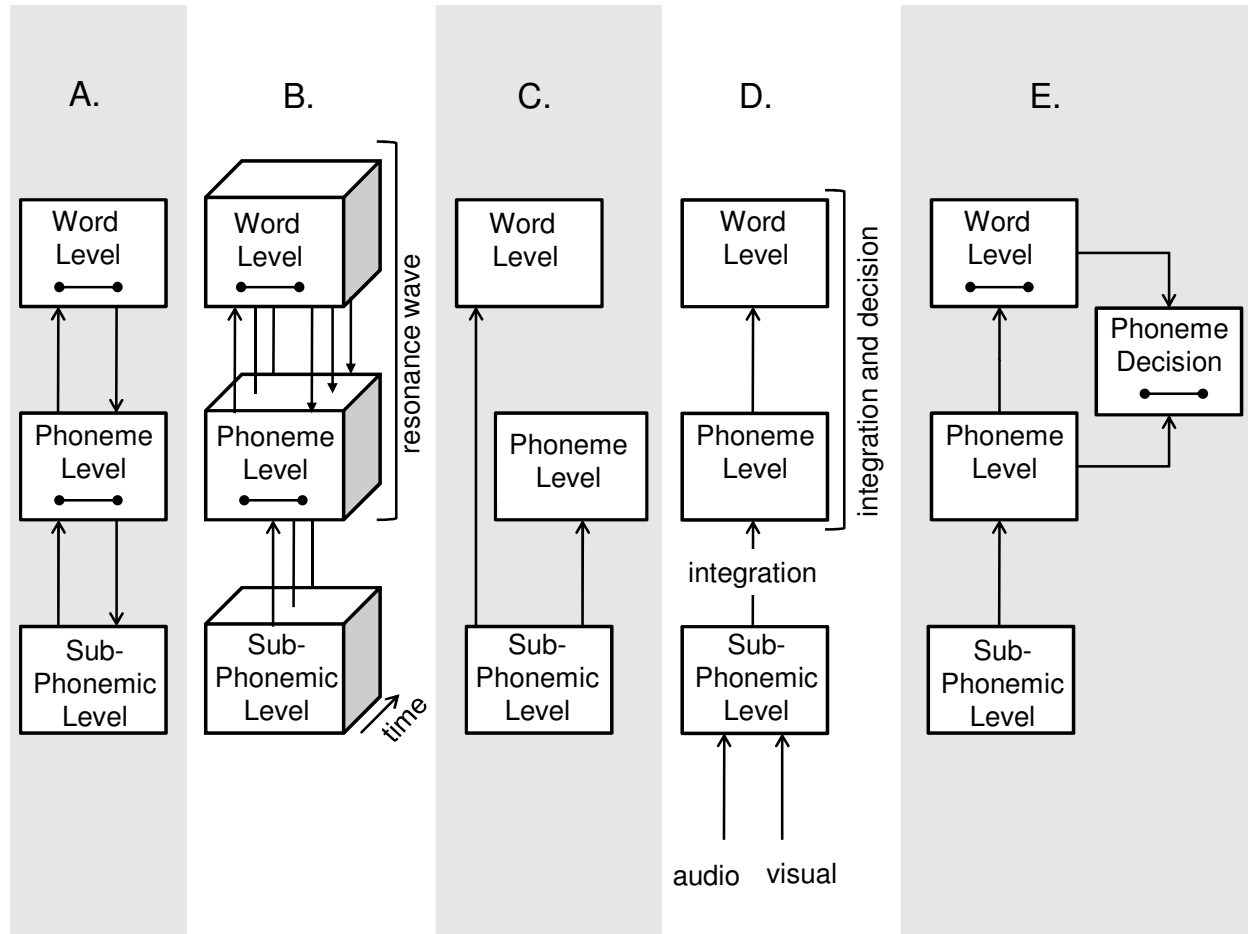


Figure 8.

$$P(\text{word}_i | \text{evidence}) = \frac{P(\text{evidence} | \text{word}_i) \times P(\text{word}_i)}{\sum_{j=1}^{j=n} P(\text{evidence} | \text{word}_j) \times P(\text{word}_j)}$$

Figure 9.

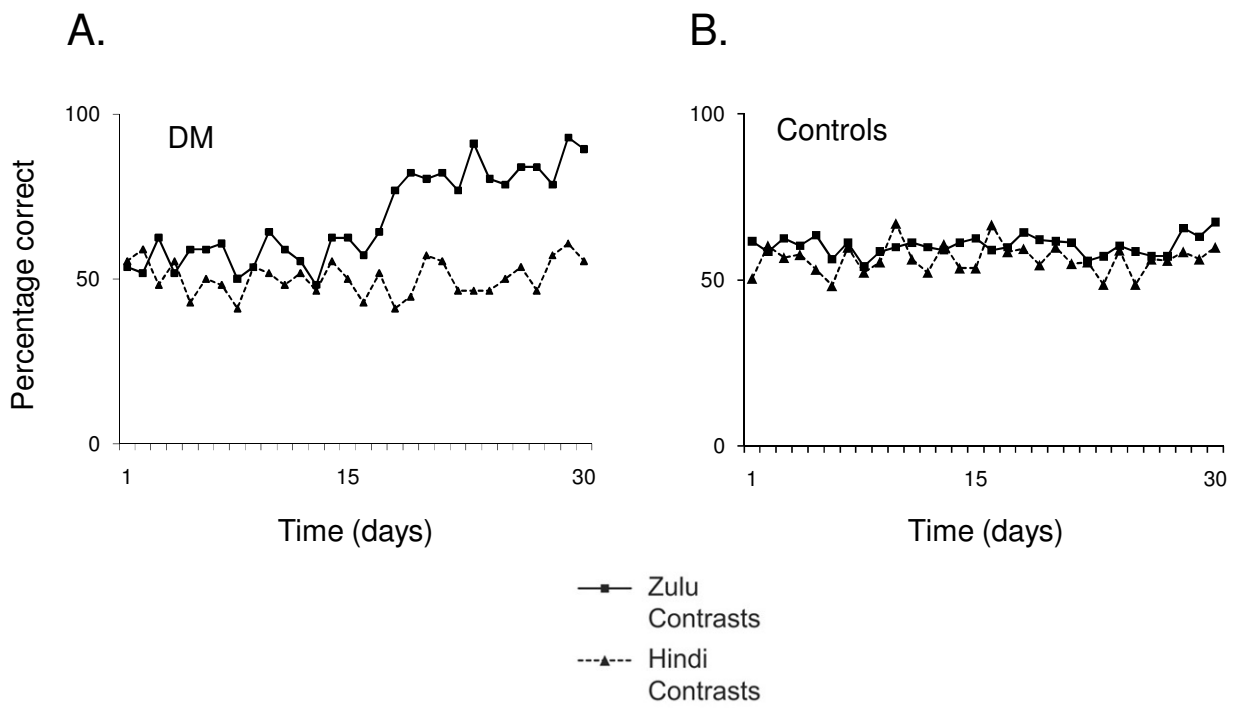


Figure 10.

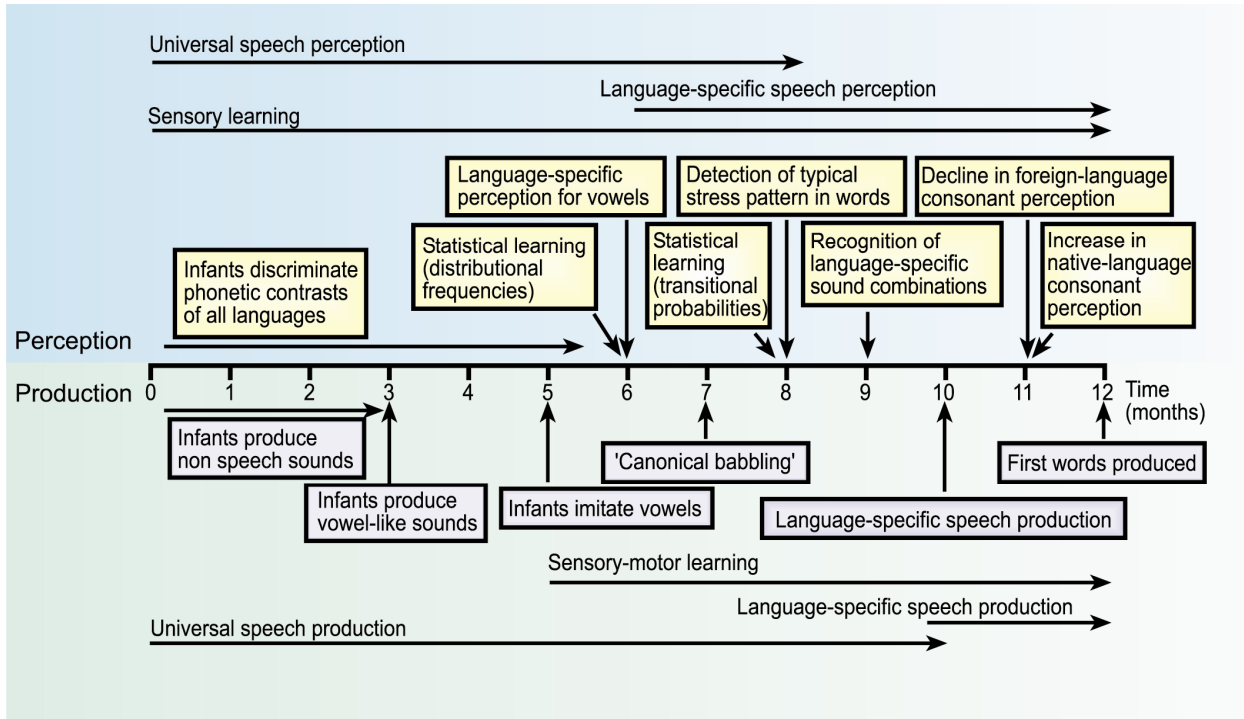


Figure 11.

