Magnuson, J. S. & Crinnion, A. M. (2022). Spoken word recognition. In A. Papafragou, J. C. Tru

SPOKEN WORD RECOGNITION

Chapter in preparation for the Oxford University Press ***Handbook of the Mental Lexicon***

James S. Magnuson

Anne Marie Crinnion

Department of Psychological Sciences & CT Institute for the Brain and Cognitive Sciences
University of Connecticut

CORRESPONDING AUTHOR

James S. Magnuson
Psychological Sciences
University of Connecticut
Storrs, CT 06269-1020
james.magnuson@uconn.edu

The problem: scope and division of labor

Recognizing words seems easy. We just hear them. Naïve listeners report that word recognition is normally effortless, and at least as easy as reading words on a printed page. You might share the intuition of a prominent auditory neurophysiologist who said that he was surprised that speech scientists struggled so much to explain speech perception, since there are clear breaks between words in fluent speech. A quick look at some spectrograms disabused him of this opinion (there are no reliable acoustic cues to boundaries between words, let alone phonemes [consonants and vowels]). In fact, acoustic breaks are more likely within words than between words (the one clear break in "where were you a year ago" occurs as part of the /g/ in *ago*).

It is easy to underestimate the challenge of explaining human speech recognition, where the buzzes, hisses, pops and whistles emanating from a speaker's vocal tract result in a listener typically recovering the message the speaker wishes to transmit. It is in fact so challenging that multiple subfields of cognitive science and neuroscience are devoted to small pieces of the puzzle. We have largely broken the problem into four pieces: speech perception (roughly, mapping from acoustics to phonological categories like phonemes), spoken word recognition (mapping from phonemes or something similar to sound forms of words), sentence processing (mapping series of word forms to syntactic structures, constrained by semantics), and pragmatics and discourse (situating sentence processing within a larger set of utterances, such as a narrative or conversation). You may notice some gaps. For example, why should spoken word recognition be conceived as ending at form recognition? There is indeed a rich literature on accessing semantic representations in spoken word recognition (see Rodd, this volume, Piñango, this volume, and Magnuson, 2017). Of course, one should also question what might be lost by not considering the lowest levels of language processing in the context of the highest, and vice-versa, and we will return to this at the end of the chapter. For now, let's consider why the subfield of SWR typically starts with abstract phonological inputs rather than the speech signal and why it typically ends with form and not meaning -- a state of affairs we will call the *Simplified Mapping Perspective* (SMP).

The SMP stems from very practical **simplifying assumptions** that emerged in the 1980s when theories of spoken word recognition came into their own. Although there were some attempts to develop models that could operate directly on speech (Elman & McClelland, 1986; Klatt, 1979), a set of longstanding unsolved problems in theories of speech perception (mainly unsolved to this day) and lack of computing power motivated this simplification. For at least 120 years (Bagley, 1900-1901), psychologists have recognized that spoken word recognition offers plenty of mysteries of its own, even if we step away from speech perception and constrain the problem to mapping from something like a string of phonemes to lexical forms. If we simplify the scope of our problem in this way (i.e., adopt the SMP), and leave the acoustic-phonetic challenges to specialists in speech perception, we can posit a fundamental process that seems to be necessary for recognizing spoken words: mapping an uninterrupted stream of phonemes onto a series of word forms stored in memory (indices to the so-called *mental lexicon*). This is spoken word recognition construed as *form* recognition. For now, we will not consider meanings of individual words, complications of morphological processes[1], lexical semantics of sets of words, nor potential constraints of syntax or discourse. And again, the SMP starts with something like phonemes as inputs, so for now we will not concern ourselves with lower-level questions that we relegate to the domain of speech perception.[2]

---

[1] Typically, the SMP focus is restricted to *lemma* forms: 'citation', uninflected forms, such as BARK but not BARKS or BARKED.

[2] The SMP also harks back to Marslen-Wilson's (1987) proposal that SWR emerges from three distinct functions that operate in parallel: access [form activation], selection (discrimination among forms, essentially recognition via determining best fit), and integration (with higher levels of processing).

Defining the scope like this requires us to adopt at least three simplifying assumptions that we know are wrong: that we can isolate a process of spoken word recognition from the rest of language and cognition, that the input can be conceived of as something like phonemes, and that word forms are a plausible stopping point (i.e., that there are not rich interactions with processing of morphology, meaning and syntax). But if we reject these simplifying assumptions, we must adopt others, or face an endless regress where at the "low" end we go deeper and deeper, from phonemes to phonetic cues to acoustic energy impinging on hair cells to the dynamics of networks of cells to individual cells, *ad infinitum*. We can do the same thing at the "high" end, progressing from word meaning to lexical semantics to sentence level constraints to discourse constraints to a listener's personal experience with each word and phrase and any memories or emotions they trigger, to social interactions with an interlocutor, to cultural concerns, etc. It is not possible to consider all these levels at once. So in any domain, developing theories and understanding requires a reductionist approach to break problems into manageable pieces. A degree of fundamental understanding of the pieces is a prerequisite for developing integrative theories. As understanding of smaller pieces of the puzzle emerges, the scope of the problem can be expanded. Of course, developing micro-theories poses risks (cf. the parable of blind men examining an elephant, coming to radically different conclusions about its nature depending on whether they inspect the trunk, ear, leg, tusk, etc.). Later, we will consider the potential perils of segmenting the system incorrectly, or of losing sight of simplifying assumptions. For now, we will adopt the conventional phonemes-to-form definition of spoken word recognition, and focus on the great strides psycholinguists have made within this scope.

We will divide the rest of this chapter into two major sections: core challenges that current SWR theories and models face within the SMP (a mainly historical overview of SWR theories and models and the data that motivate them), and future challenges, where we review phenomena that are clearly essential for true understanding of human SWR, but are mostly outside the scope of current theories and models. In a nutshell, by walling off SWR from the unsolved challenges of speech perception and sentence processing (not to mention neural-level findings, which we will touch upon briefly later), the SMP has allowed tremendous progress towards understanding crucial components of word-level processing and representation that is largely prerequisite to integrative theories spanning these levels. It is possible that solutions to currently unresolved challenges and debates may emerge as we develop models that allow lower and higher levels of language and cognition to constrain SWR.[3]

Core challenges for the simplified mapping perspective on SWR

The primary challenge is mapping the stream of phonemes onto words. One possibility would be to buffer some amount of speech into short-term memory and analyze it in chunks. But what would the chunks be? This idea runs right into what is known as the **segmentation problem**: there are few cues to word boundaries. Because the input must be processed before word boundaries can be discovered, the chunks could not be words. Perhaps the chunks could be longer utterances, such as phrases or sentences. However, phrase and even sentence boundaries can also be uncertain -- another segmentation problem.[4] Also, in many contexts, speech proceeds quickly with few pauses (e.g., interlocuters taking immediate or even overlapping turns), so this idea might pose impractical computational demands.

[3] One might argue that the SMP is an oversimplification; as we shall see, several SWR studies since the 1980s have strayed beyond its boundaries. This is a fair point, and -- spoiler alert -- we will reject the SMP ourselves. However, we find it a useful *approximation* of the modal approach in SWR over the last few decades (most models of SWR conform to it, for example), as well as a foil that highlights exceptions to it and potential alternatives.

[4] A third segmentation problem is found in the domain of speech perception, where the information specifying adjacent (or sometimes more distant) consonants and vowels overlaps in time, precluding clear phonemic segmentation.

Another intuitive approach might be to consider the lexicon as a series of branching paths, as in Figure 1. There would be one tree for every possible onset phoneme, which could potentially branch to every other possible phoneme, and so on (though in fact only a subset of possible branches occurs in English [for example], and word length is finite). Recognizing a word would simply be a matter of mapping the phoneme stream onto the correct branch. As we can see in Figure 1, a challenge crops up immediately: the **embedding problem** (McQueen, Cutler, Briscoe, & Norris, 1995). Many words are embedded within longer words. We need a mapping process that won't prematurely "recognize" BASS or BASK when hearing BASKET. If we only consider words spoken in isolation, this is not much of an issue; the system can wait for silence to mark the end of the word. But how will the system manage embeddings in fluent speech, without clear word boundaries?

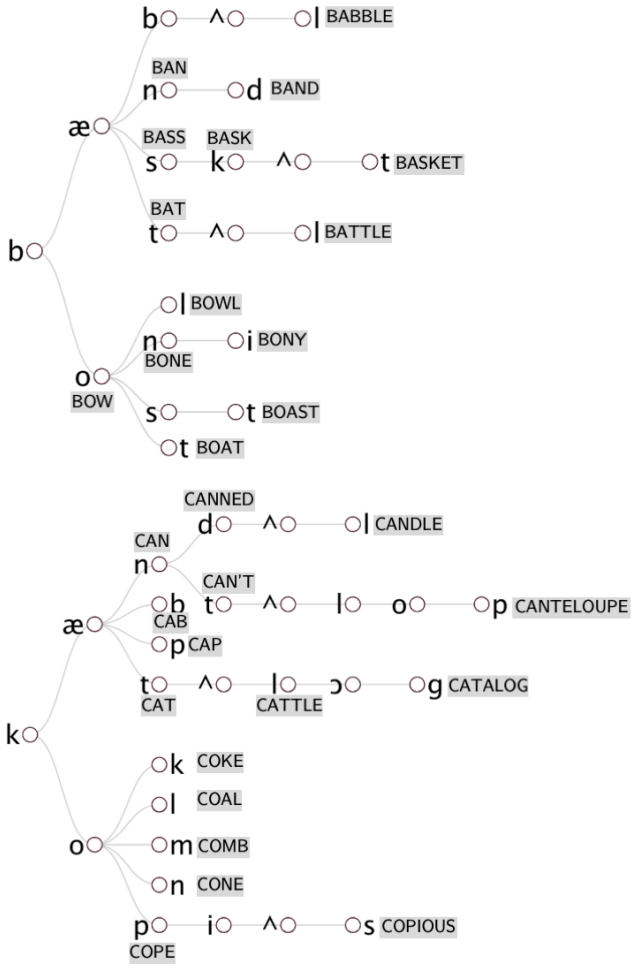

Figure 1: Examples of lexical trees beginning with /b/ and /k/, showing only a small subset of branches (some of which are truncated). Note that in some cases the path to a terminal node (which cannot be continued to form a word without considering morphological inflections) passes through multiple words that are onset-embedded in one or more longer words (e.g., the branch for BASKET passes through BASS and BASK). Source: Magnuson (2020a).

In fact, the foundational modern theory of spoken word recognition -- the **Cohort Model** (Marslen-Wilson & Welsh, 1978) -- posited that a process operating over something like this tree structure would provide a basis for mapping phonemes to words, as well as an emergent solution to word segmentation.

The basic idea is depicted in Figure 2. When the first phoneme in the stream is encountered, it is added to a buffer. Each subsequent phoneme is added to the string in the buffer *if* doing so results in a continuation of a word in the lexicon. If adding the phoneme does not yield a match to a word in the lexicon, this likely indicates a word boundary (e.g., /kæt/ is a word, but if the next phoneme is /r/ [as in "CAT RUNS"], it does not match a word, and so a boundary is posited after /kæt/ and the process begins again with /r/ as the first phoneme). However, if the phoneme cannot be added to the prior string *and* the prior string does not correspond to a word, this means an error has occurred, and reanalysis is needed. For example, if the input were /kʌlæps/ (COLLAPSE) but the /l/ was not clearly articulated and the listener mapped it to /r/, the string /kʌræ-/ could be mapped to CARAFE but at the /p/ at the next position, neither the prior string (/kʌræ/) nor the new string (/kʌræp/) would match a lexical item, so reanalysis would be required.

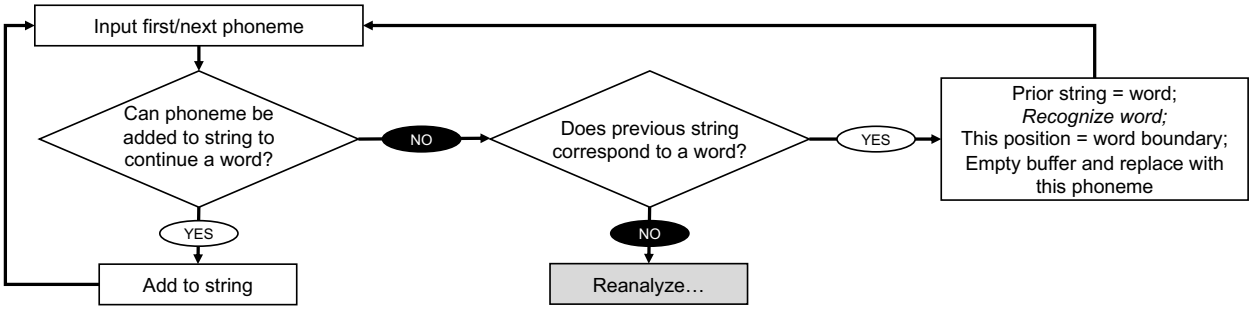

Figure 2: Flow chart illustrating the basic algorithmic principles of the Cohort Model (Marslen-Wilson & Welsh, 1978). Source: Magnuson (2020b).

The reason this was called the *Cohort Model* is that at the beginning of a string, all words that match the first phoneme are 'activated' and form the potential recognition *cohort*.[5] Thus, a more typical way of describing the process in Figure 2, rather than as traversing a tree, is generating a list of all matching items at the first position, and then winnowing that list as each new phoneme comes in. On this view, word onsets hold a privileged place. Even if two words have great global similarity -- e.g., BATTLE and CATTLE overlap in four of five phonemes -- hearing one is not predicted to activate the other if they mismatch at onset. The /b/ at BATTLE's onset constitutes positive evidence for any word beginning with /b/ and evidence *against* any word that does not begin with /b/. On the other hand, hearing one of two words with relatively little global overlap but with the same onset (BATTLE and BAG, which match in only two of five possible positions) would be predicted to strongly activate both.

Although the primacy of onsets had been long recognized (Bagley, 1900-1901), Marslen-Wilson and his colleagues transformed this intuitive idea into a formal model that generated clear, testable predictions. Indeed, they conducted comprehensive tests of the model over several years. Their methods included *gating* (playing a listener progressively longer snippets of a word and asking them to guess what word it would turn out to be; Grosjean, 1980) and a variety of *pairwise approaches* (cf. Magnuson, 2017). In pairwise approaches, a specific word is presented and the impact of hearing (or seeing) that word on the activation of one specific word (i.e., a specific target and potential competitor pair) is assessed. For example, one can test whether hearing BATTLE *primes* cohorts (BAND, BAG, etc.) or rhymes (CATTLE).

[5] Note, however, that *cohort* is often used in SWR to refer to an onset competitor, e.g., CAB is a cohort *of* CAT. Note also that the *competition cohort*, that is, the set of items predicted to compete strongly with a give target word, is not typically defined based on overlap in only the first phoneme. Here, we adopt a common standard of defining the competition cohort as words overlapping in the first two phonemes (other variants include 'overlap in the first 200 msecs' or 'overlap through the first vowel').

When Marslen-Wilson and his colleagues used the *cross-modal semantic priming* paradigm (Swinney, 1979) to assess spread of activation based on phonological similarity (e.g., Marslen-Wilson & Zwitserlood, 1989), an asymmetry emerged. In cross-modal semantic priming, the participant hears a series of words but is focused on a visual lexical decision task. The participant decides whether each written string they see on a screen is a word or not (and of course, the domains can be flipped such that the lexical decision task is focused on spoken words). Although the spoken words are superfluous, they impact the visual lexical decision via semantic relations. For example, participants are faster to identify LOG (associate of CABIN) as a word after hearing CATTLE than after a phonologically-unrelated word, such as BRAIN. This suggests that CABIN was so strongly activated when CATTLE was heard that a detectable degree of activation spread to its semantic associates.

Work by Marslen-Wilson and colleagues and others motivated by the Cohort Model led to multiple discoveries regarding the incremental process of spoken word recognition. For example, the simple view of the Cohort Model depicted in Figures 1 and 2 predicts that recognition should happen at the **uniqueness point**. Given a word like COCONUT, the /n/ represents the uniqueness point; once the /n/ is encountered there is only one lexical possibility (or two, taking into account morphological marking for number, i.e., COCONUT vs. COCONUTS). The fact that the uniqueness point is at or after word offset for many words (e.g., CAT can continue as CATTLE, CATALOG, etc.) also motivates the parsing strategy depicted in Figure 2. Given the string, /ðʌkætəloʊ/ (we can approximate this phonetic transcription roughly as *tha-kat-a-low*, with no pauses or breaks), a word boundary would be discovered between the second and third phonemes (THE and then something beginning with /k/). A potential boundary exists after the /t/ (THE CAT…), but the schwa and /l/ afterward could be added to /kæt/ to form CATTLE. The final vowel still leaves an ambiguity; a possible parse is THE CATTLE LOW (the final word being a low-frequency, archaic verb, meaning *to moo*), and if the utterance ended there, it would be the parse. However, if the utterance continued as /ðʌkætəloʊpt/, the system would have to discover the parse THE CAT ELOPED.

The low-frequency example of the verb TO LOW leads us to two other critical considerations in the time course SWR that emerged in Cohort-motivated studies: the **isolation point** and the **recognition point** (Grosjean, 1980). The *recognition point* is where we can establish definitively that a listener has *decided* upon the identity of a spoken word (not necessarily correctly). This could be the point when the listener presses the YES button in a lexical decision (deciding that the stimulus is a word), although there is some controversy about whether a correct YES response in lexical decision necessarily requires actual recognition (vs., e.g., high but not definitive activation, depending on the nature of nonwords used, among other factors; see Balota, 1990, on the slippery notion of a "magical moment" of recognition). Grosjean (1980) instead asked participants (in a gating task) to listen to increasingly longer portions of words (from word onset), provide their best estimate of the identity of the word, and to rate their confidence about that estimate. He defined the *isolation point* as the time a participant settled upon the correct response for the target word (that is, they correctly identified the actual word and did not change their response on subsequent gates).

The isolation point necessarily precedes (or coincides with) the recognition point, but both may occur prior to or later than the uniqueness point. Grosjean found that for single words presented in without context, the isolation point was relatively late. A relatively neutral context (constraining form class and possibly concreteness, e.g., *my son asked for a...*) led to an earlier isolation point for appropriate words (e.g., PARROT), and semantically constraining contexts shifted the isolation point earlier (*because he loves pets, my son asked for a ...*). This raises an interesting question that we will return to later: does top-down context directly affect recognition processes and therefore contribute causally to perception (e.g.,

via feedback), *or* does context *bias* word recognition *after* bottom-up information drives activation (e.g., in a decision-stage process)? For now, we will focus on "context-free" SWR (i.e., isolated words), and will shortly consider how the lexical context of isolated words may mediate or moderate sublexical processing.

However, despite the terrific amount of data that had accumulated supporting the Cohort Model, Luce (1986; Luce & Pisoni, 1998) took a contrarian perspective with the *Neighborhood Activation Model (NAM)*. They proposed that similarity defined in a more global manner might capture aspects of competition that were not apparent in studies using gating or pairwise approaches, which strongly supported the Cohort Model's emphasis on temporally 'left-to-right' sequential processing. In contrast, they took what we call a *lexical dimensions* approach to investigating phonological similarity (cf. Magnuson, 2017). On a lexical dimensions approach, predictors are selected for factorial and/or regression analyses examining performance measures for large numbers of items (rather than specific pairs). Luce and Pisoni focused on two lexical dimensions: frequency of occurrence, and *neighborhoods*.

They adopted a specific and simple neighborhood definition drawing on prior work (e.g., Greenberg & Jenkins, 1964; Landauer & Streeter, 1973; and Sankoff & Kruskall, 1983): two words are *neighbors* if they differ by no more than a single phonemic deletion, addition, or substitution (the so-called *DAS rule*). For example, CAT has the deletion neighbor AT, addition neighbors SCAT and CAST, and many substitution neighbors (e.g., BAT, COT, CAN). Luce and Pisoni acknowledged that the DAS rule makes very strong assumptions, in that any it ignores potentially gradient phonetic similarity (e.g., CAD is more phonetically similar to CAT than is CALF) or potentially differential effects of position of similarity (e.g., do BAT and CAT compete as strongly as CAB and CAT?). However, they argued that the computational simplicity of the DAS rule provides an excellent starting point for considering effects of phonological neighborhoods (and they also explored gradient measures of similarity, as we discuss shortly).

They proposed that frequency and neighborhood could be integrated into a simple choice model for spoken word recognition, paraphrased in Equation 1. A word's *frequency-weighted neighborhood probability (FWNP)* is calculated as the ratio of a target word *t*'s frequency (*f*) to the summed frequencies of all *n* of its neighbors (including itself). The simple, elegant idea here is that words differing by a single phoneme are similar enough to activate each other if one of the words is heard, that strength of activation will be proportional to word frequency (i.e., prior probability), and that activated words will compete for recognition. Thus, the larger the proportion of the frequency-weighted neighborhood that the target contributes, the more easily it should be recognized. So if two target words had the same frequency-weighted neighborhood, the target with higher frequency (i.e., the one with the larger numerator) would be (predicted to be) recognized more easily. If two target words had the same frequency, the one with the 'sparser' neighborhood (i.e., the one with the smaller denominator), should recognized more easily.

$$FWNP_t = \frac{f_t}{\sum_{j=1}^{n} f_j} \qquad (1)$$

Luce and Pisoni (1998) conducted factorial studies using categorical definitions of high and low word (log) frequency, neighborhood density (count of neighbors) and neighborhood frequency (summed log frequencies of neighbors). In general, higher frequency predicted better lexical decision performance (higher accuracy and faster reaction times), while high neighborhood density (count) and high neighborhood frequency predicted worse performance. They also used a graded form of the FWNP rule, as in Equation 2 (note that we use a simplified notation, but Equation 2 is equivalent to Equation 6 in Luce & Pisoni, 1998).

$$FWNP_t = \frac{p(t|i)f_t}{\sum_{j=1}^{l} p(j|i)f_j} \qquad (2)$$

In Equation 2, *p(t|i)* is the probability that the word is target *t* given the input *i.* Given that the input does indeed correspond to word *t*, you might expect this to be 1.0. However, this may not be 1.0 given the possibility of external or internal noise. In the denominator, *p(j|i)* is the probability that the input word is actually word *j*. To put this slightly differently, the *p(t|i)* and *p(j|i)* are the calculated *similarity* between *i* and target word *t* or word *j*. Note that *n* (all neighbors) in Equation 1 is replaced here with *l*, because now every word in the entire lexicon is considered, not just the neighbors that conform to the DAS rule. As in Equation 1, the denominator includes the target, *t*. Note also the relation between Equations 1 and 2; in Equation 1, the similarity calculation is dropped because words are defined categorically as neighbors (those conforming to the DAS rule) or not. Luce and Pisoni (1998) reported that for an identification-in-noise task, the FWNP accounted for approximately four times as much variance as word frequency at high and moderate signal-to-noise ratios (SNRs): 16% vs. 4% with a +15 dB SNR, and 22% vs. 5% at +5 dB SNR. The FWNP and frequency accounted for similar variance at a low SNR (5% vs. 6% at a -5 dB SNR).

The graded similarity metric makes some surprising predictions. For example, it correctly predicts that VEER should prime BULL (which differ only by a single phonetic feature at each position; Luce, Goldinger, Auer, & Vitevitch, 2000).[6] However, the DAS variant of the FWNP (Equation 1) has come into common usage, with little exploration of the graded version (Equation 2). An obstacle to using the graded version is the need for a precise basis for calculating similarity. Luce and his colleagues derived confusion probabilities for the specific parameters of the experiments in which they have applied the graded metric (e.g., phoneme or diphone confusions for specific SNRs, talkers, etc.). This is a gap that could be filled by the use of a generic similarity metric based on acoustic-phonetic features, but to our knowledge, this approach has not yet been taken.

Let's consider how the Neighborhood Activation Model and Cohort Model relate to one another. In Figure 3, we present subsets of the neighbors and cohorts of the word CAT (/kæt/). Neighborhoods and cohorts overlap in substitution neighbors that overlap at all but the final position, and addition neighbors that preserve the first two phoneme positions. But neighbors include non-cohorts with substitutions or deletions at the first or second positions, and cohorts include items that mismatch by more than one phoneme, so long as they overlap in the first two positions. Which should we prefer? Informally, it would appear that the NAM has outstripped Cohort in SWR. Anecdotally, reviewers routinely insist upon neighborhood being controlled in word recognition studies, but rarely comment upon cohort size or density. The NAM approach has also proved useful in designing clinical assessments and interventions for language impairments (e.g., Kirk et al., 1995; Morrisette & Geirut, 2002; Sommers & Danielson, 1999; Storkel et al., 2010, 2013).

[6] An important extension of the framework is the concept of *probabilistic phonotactics* (Vitevitch & Luce, 1999), which considers positional likelihoods of each phoneme and diphone in a word. While these measures tend to increase with frequency-weighted neighborhood, probabilistic phonotactics adds important information, and many studies now control probabilistic phonotactics as well as neighborhood.

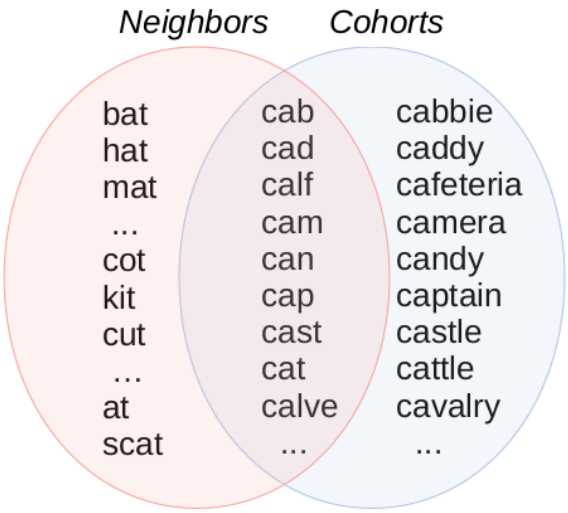

Figure 3: Illustration of the differences and overlap in competitor sets predicted for the target word CAT by the Neighborhood Activation and Cohort Models (Magnuson, 2020c).

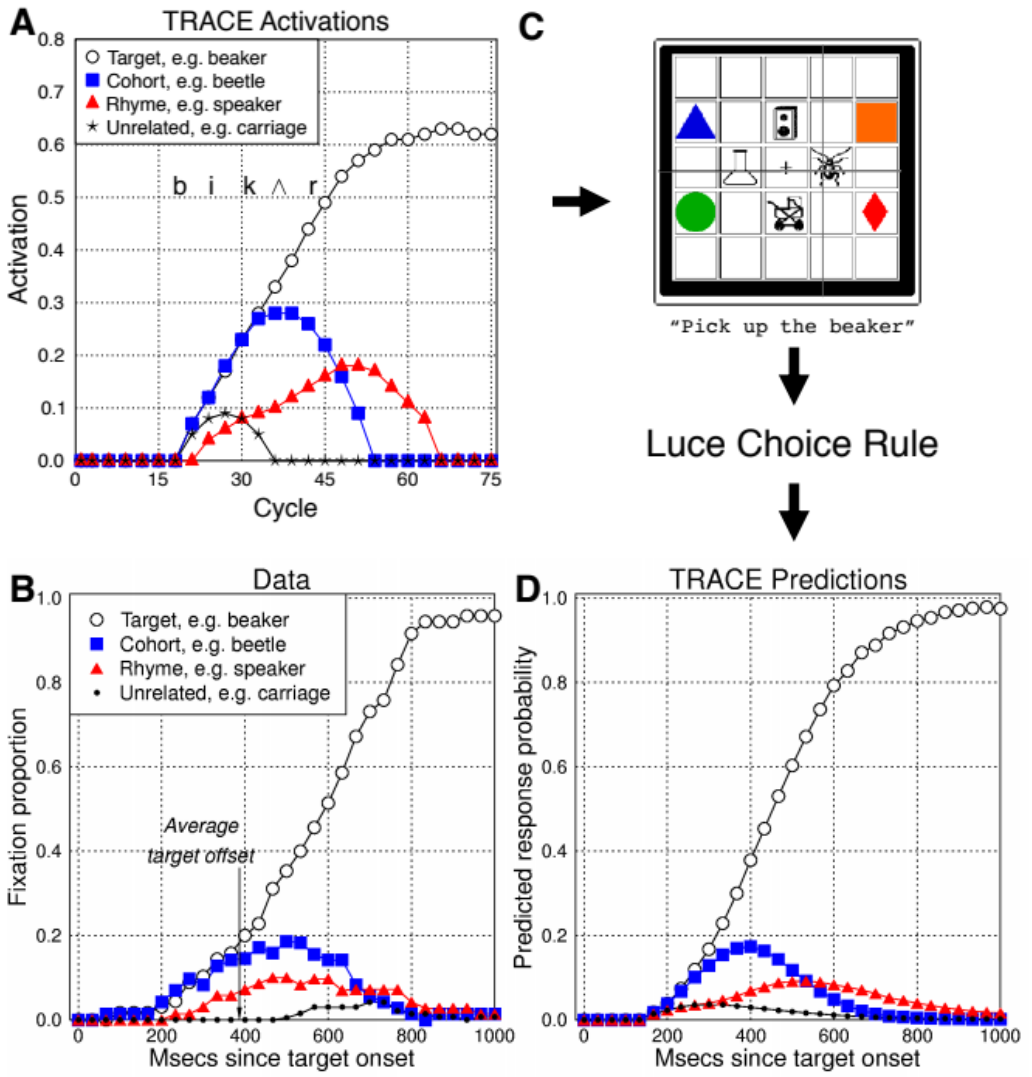

Figure 4: How the TRACE model is linked to the visual world paradigm task, adapted from Allopenna et al. (1998). Participants saw displays like the one in C. As participants followed spoken instructions like "pick up the beaker", the proportion of fixations to different object types mapped onto phonetic similarity over time with an approximately 200-ms lag (panel B). When TRACE simulations were conducted with analogous items, similar trends were observed (A). When TRACE activations were constrained to the same four-alternative choices presented to participants (C), and thus rescaled as *predicted response probabilities* (D), a very close fit was observed.

However, it's not clear that the apparent dominance of the NAM approach is fully warranted, nor has there been a clear reconciliation of the two approaches. A study by Allopenna, Magnuson, and Tanenhaus (1998; see Figure 4) suggested a middle ground that is consistent with one of the best-known computational models of human SWR, TRACE (McClelland & Elman, 1986). In an early "visual world paradigm" (Tanenhaus, Spivey-Knowlton, Sedivy, & Eberhard, 1995) study, Allopenna et al. presented participants with displays containing four shapes and four images of objects. Participants followed spoken instructions to interact with one of the images (e.g., "pick up the beaker; now put it below the diamond"). On critical trials, the target image (e.g., BEAKER) was accompanied by a potential cohort competitor (BEETLE) and/or a rhyme competitor (SPEAKER) along with images of one or two phonologically unrelated items (Figure 4C). Allopenna et al. tracked the proportion of fixations to each object from the onset of the target name in the "pick up the…" instruction. The time course is shown in Figure 4B. There was early, strong competition between cohorts (which makes intuitive sense, since the input initially matches the target and its cohort equally well), but also later, weaker competition between rhymes. Changes in fixation proportions over time mapped onto phonetic similarity with an approximately 200-ms lag, which is a typical latency for saccades in cognitive tasks (Carpenter, 1988; Viviani, 1990), and nearly as fast as could be expected, given that saccade latencies to a point of light in a darkened room are approximately 150 ms (Fischer, 1992; Matin, Shao, & Boff, 1993; Saslow, 1967).

Notably, the cohort competition they observed is not predicted by NAM, since the cohorts differed by more than one phoneme. Similarly, the rhyme competition is not predicted by the Cohort Model.[7] Thus, the time course suggests a possible reconciliation between Cohort and NAM predictions. While cohort competition is indeed strong, there is also weaker competition from at least one non-cohort type of neighbor. The fact that competition between rhymes is relatively weak suggests that the failure to observe rhyme effects in pairwise approaches was likely due to the use of low-sensitivity paradigms rather than the actual dynamics of lexical access and competition in online SWR. For example, consider again cross-modal semantic priming. To detect effects in this paradigm, activations must be strong enough to cross modalities (auditory-visual) and spread form-based activation via semantic relations. If form-based rhyme effects are relatively weaker than cohort effects (Figure 4B), it is unsurprising that rhyme effects might be too weak to drive cross-modal semantic priming.

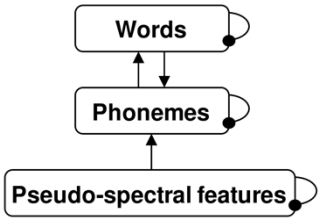

Figure 5: Schematic of the architecture of the TRACE model. Arrows indicate positive, selective connections (features connect to phonemes for which they are relevant, phonemes connect to words that contain them, and words connect back to constituent phonemes). Knobs indicate lateral inhibition. (Source: Magnuson, 2020d.)

Let us consider how the human-like competition timecourse emerges in the TRACE model. The basic architecture of TRACE is presented schematically in Figure 5. In TRACE, "pseudo-spectral feature" inputs (elements representing features such as *diffuseness* or *burstiness* that ramp on and off over time, with adjacent phoneme patterns overlapping as a coarse analog to coarticulation) activate corresponding phonemes. Activated phonemes stimulate word nodes that contain them. As word nodes become

[7] In addition, the rhymes in the example are not neighbors, as BEAKER and SPEAKER differ by two phonemes. However, all other rhyme pairs differed by just one phoneme (e.g., SANDAL/CANDLE).

activated, they send feedback to constituent phonemes. Crucially, activations are governed by competition via *lateral inhibition.*[8]

As a word is input incrementally to TRACE, items overlapping at onset (cohorts) naturally become activated since they match the beginning of the input. Rhymes can be activated as the input begins to activate their constituent phonemes. Importantly, lexical feedback will even drive modest activation of the initial phoneme of a rhyme (though not sufficiently for that phoneme to be as strongly activated as the phonemes in the bottom-up input). One might expect that because BEAKER and SPEAKER overlap in four phonemes but BEAKER and BEETLE only overlap in two, BEAKER should activate SPEAKER *more* strongly than BEETLE. However, this does not happen due to lateral inhibition. By the time the input begins to activate the rhyme, the target word and *all of its cohorts* are moderately activated (e.g., around cycle 20 in Figure 4A). The inhibition they jointly send to the rhyme prevents the rhyme from becoming as strongly activated as cohort items despite greater global similarity.

Thus, the incremental nature of the input promotes a strong advantage for cohort competition (note that McClelland & Elman [1986] cite the Cohort Model as their inspiration.). Since there is no explicit tracking of word onsets or removal of items with mismatches (lateral inhibition *suppresses* rather than categorically *removes* items from the competition set) items that mismatch at one or more positions can become activated, with degree of activation depending on many complex factors (position[s] of mismatch[es], number of other items similar to the target and/or potential competitor, etc.). One way that TRACE could be made more consistent with the Cohort Model would be to use *bottom-up inhibition:* that is, phonemes would send inhibition to words they do *not* occur in (in a position-specific fashion; see Footnote 4; to our knowledge, this has not been attempted with TRACE). Note that lateral inhibition has important advantages over categorical removal based on mismatches; in particular, it allows rhyme effects to emerge, and it also makes TRACE quite robust against noise (see, e.g., Magnuson, Mirman, Luthra, Strauss, & Harris, 2018), and replaces the need for the (unspecified) reanalysis mechanism in the Cohort Model with a of more tolerant form of constraint satisfaction.

So, does TRACE reconcile differences between NAM and Cohort? Yes and no. On the one hand, it provides a mechanistic explanation for how the timecourse of phonological competition might emerge in human listeners, with surprisingly fine-grained, nearly millisecond-scale simulations (Figure 4A and 4D). However, while some general principles that govern TRACE's behavior can be articulated (as we have just attempted), it does not provide a similarity metric, let alone a categorical (or even graded) definition of competitors. Its central tendencies (e.g., the averages presented in Figure 4A and 4D) closely correspond to those of human behavior (e.g., Figure 4B). But TRACE does not generate convenient, precise numbers corresponding to its word-specific predictions like NAM does. One could attempt to do this by operationalizing recognition time in TRACE, for example. However, TRACE has a modest phoneme inventory (14 rather than the ~40 required for English) and a fairly small lexicon (212 words originally, which has been extended to ~1000 in two reports [Frauenfelder & Peeters, 1998; Magnuson et al., 2018]).

Another challenge for models of SWR is the fact that phonemes may not be processed in strict temporal

[8] The schematic in Figure 5 belies a considerably more complex architecture, where many copies of each feature, phoneme, and word node are tiled in the TRACE memory over time (essentially creating a temporotopic map, which allows interaction not just vertically, from words to phonemes, but horizontally between units that are aligned with different stretches of time in TRACE's memory). Lateral inhibition is relatively 'local', with connections constrained to units with similar temporal alignments. For detailed discussions of the TRACE architecture, see McClelland and Elman (1986), Magnuson, Mirman and Myers (2013), and Magnuson, Mirman, and Harris (2012).

order. Toscano, Anderson, and McMurray (2013) used a visual world paradigm to show that listeners look more to phonemic anadromes, or words that share the same phonemes but in a different order (e.g., *gut* for target word *tug*), than to another word that shares the same vowel and one consonant (i.e., has word initial mismatch; *gun* for target word *tug*) or to an unrelated word. Notably, TRACE was unable to simulate these effects (although it might if one were to extend the spread of coarcticulation further left and right in TRACE). Similar effects have also been shown in a short-term priming task (i.e., a phonemic anadrome primes a target better than an unrelated prime or a prime that has word initial mismatch but shares a vowel and consonant; Dufour & Grainger, 2019). Furthermore, Grainger and Dufour (2020) found that these priming effects occur only when the target has a higher frequency than the prime, suggesting that lexical representations underlie these effects. While these findings suggest that spoken word recognition does not occur in a strictly sequential way, findings from Gregg, Inhoff, and Connine (2019) suggest that overlap in vowel position is important for these effects to be observed. In other words, while Gregg and colleagues found more looks to an anagram than an unrelated word, for pairs like leaf (target) and flea (competitor), where the same consonants are present in a different order, but the position of the vowels does not overlap in the two words, they observed no more looks to the competitor than to a completely unrelated word. These effects suggest that models of spoken word recognition need to reconsider strict temporal ordering of constituent phonemes[9], consider frequency as a core concern in models, and also that models need to consider different saliency of vowels vs. consonants -- none of these are considered by default in TRACE, although jTRACE (Strauss, Harris & Magnuson, 2007) includes three different ways to enable frequency.

Despite these limitations, TRACE has proved to account for a broad and deep set of phenomena in human SWR and speech perception (see Table 1). There are also a few reported failures (e.g., Chan & Vitevtich [2009] reported that simulations with jTRACE [Strauss et al., 2007] failed to simulate differences they observed with a factorial manipulation of *clustering coefficient* [proportion of a word's neighbors that are also neighbors of each other] that we review briefly below). The fact that TRACE simulates so many results with its default parameter settings is important to note, as some other models require parameter changes for simulating different phenomena (see Magnuson et al., 2012, for a review).

However, it is crucial to note that successful simulations do not establish that the mechanism proposed within a model is correct. Indeed, much of the SWR literature since the mid-1980s has revolved around disagreements regarding the algorithms proposed in competing models. We do not have space in this chapter to discuss most models of SWR (e.g., the Distributed Cohort Model [Gaskell & Marslen-Wilson, 1997, 1999], PARSYN [Luce et al., 2000], TISK [Hannagan et al., 2013] or various models within the Adaptive Resonance Theory [ART] framework [Grossberg, Boardman, & Cohen, 1997; Grossberg, Govindarajan, Wyse, & Cohen, 2004; Grossberg & Myers, 2000]). Instead, we will focus briefly on the Shortlist and Merge ‘A’ and 'B' models proposed by Norris (1994) and Norris and McQueen (2008), respectively, as they represent a theoretical position (*autonomy*) fundamentally opposed to TRACE’s interactive activation.

[9] The Time Invariant String Kernel (TISK) model of spoken word recognition (Hannagan, Magnuson, & Grainger, 2013) may account for these results. TISK is an interactive activation model like TRACE, but instead or reduplicated, strictly ordered phoneme templates (see Footnote 7), it uses a form of "open diphone" coding, where diphone units are activated by every ordered pair of phonemes in a word. For example, when *cat* is presented, nodes for /kæ/ and /æt/ are activated, but so is a node for /kt/ (hence, "open", or not necessarily adjacent). This allows for a much more compact model compared to TRACE. It would also seem likely to tolerate phoneme transpositions better than TRACE, but to our knowledge, it has not been applied to such results yet.

Table 1: Effects accounted for by the TRACE model. As we discuss, there are also a few reported failures with the TRACE model, and successful simulations *do not* establish that algorithms implemented in a model are correct – they simply confirm that the model's algorithm is *plausible and not-yet-falsified*. This list provides a sense of the breadth and depth that has been achieved with TRACE, and a set of candidate benchmarks for other models.

	Effect / phenomenon	Reference	Notes
1.	Lexical context mediates phoneme perception ("Ganong effect" [Ganong, 1980])	McClelland & Elman (1986), *p. 24*	Original parameters
2.	Elimination of lexical effects under time pressure	*ibid, p. 26, Figs. 5 & 6*	Original parameters
3.	Phoneme position impacts lexical effects (weaker at word onset than offset)	*ibid, pp. 27, 29 and 30, Figs. 8-11*	Original parameters
4.	Dependence of lexical effects on phonological ambiguity (demonstrating bottom-up priority)	*ibid, p. 28*	Original parameters
5.	Lexical basis for phonotactic influences	*ibid, pp. 33-35, Fig. 12*	Original parameters
6.	Trading relations among phonetic cues	*ibid, p. 38*	Original parameters
7.	Categorical perception (identification and discrimination, including reaction time and timecourse effects)	*ibid, p. 42*	Original parameters
8.	Recovery from noisy input, and influences of lexical neighborhood	*ibid, p. 55*	Original parameters
9.	Time course of word recognition	*ibid, p. 57*	Original parameters
10.	Frequency and context effects	*ibid, p. 60*	Original parameters
11.	Lexical segmentation, including multiword sequences, and impact of right context	*ibid, pp. 61-69*	Original parameters
12.	Compensation for coarticulation	Elman & McClelland (1988), *Fig. 3A*	Original parameters
13.	Lexically-mediated compensation for coarticulation	*ibid, Fig. 3B*	Original parameters
14.	Stochastic noise allows correct simulation of classical context effects that the basic TRACE model fails to simulate correctly	McClelland (1991)	Modified parameters and noise
15.	Timecourse of cohort and rhyme competition observed in human eyetracking study (eye tracking data)	Allopenna et al. (1998), *Expt. 1 (Fig. 4, this chapter)*	Original parameters + Luce choice rule
16.	Elimination of rhyme effects in a gating paradigm (eye tracking data)	*ibid, Experiment 2*	Original parameters
17.	Time course of word frequency effects (eye tracking data)	Dahan, Magnuson, & Tanenhaus (2001)	Original parameters + 3 alternative implementations of frequency
18.	Subcategorical mismatch (eye tracking data)	Dahan, Magnuson, Tanenhaus, & Hogan (2001)	Original parameters
19.	Subcategorical mismatch (lexical decision data, words only -- nonwords not included)	Magnuson, Dahan, & Tanenhaus (2001)	Original parameters
20.	Lexically-induced delays in phoneme recognition	Mirman, McClelland & Holt (2005)	Original parameters
21.	Lexically-guided tuning of speech perception with Hebb-TRACE	Mirman, McClelland, & Holt (2006)	Extensions for learning
22.	Attentional modulation of lexical effects	Mirman, McClelland, Holt, & Magnuson (2008)	Original parameters + attentional scaling parameter
23.	Individual differences relate to lexical decay	McMurray et al. (2010)	Variety of parameters explored
24.	Changes in timecourse of phonological competition in Broca's and Wernicke's aphasia (eye tracking data)	Mirman, Yee, Blumstein, & Magnuson (2011)	Original parameters + modified Luce choice rule (other parameters explored)
25.	Variation in phoneme inhibition relates to reduced rhyme + enhanced subcategorical mismatch in individuals with weaker language abilities	Magnuson et al. (2011)	Variety of parameters explored
26.	Duality of length effects (early short-word advantage, later long-word advantage)	Magnuson et al. (2013)	Original parameters; see Pitt & Samuel (2006) for human data
27.	Suppression of embedded words	Magnuson et al. (2013)	Original parameters
28.	Phoneme restoration as a function of word length and position	Magnuson (2015)	Original parameters
29.	Flexible modulation of inhibition may explain experience-based changes in lexical competition	Kapnoula & McMurray (2016)	Original parameters + changes in phonetic features tailored to materials + inhibition variation
30.	Feedback promotes resistance to noise (model-specific results [feedback turned on vs. feedback off])	Magnuson et al. (2018)	Original parameters + parameters for large lexicon (Frauenfelder & Peeters, 1998)

Currently, the primary alternative to interactive models like TRACE is an *autonomous* framework without feedback. The fundamental premise is that any phenomenon that can be simulated with feedback could be simulated by an autonomous model. Consider, for example, one of the most familiar effects that appears to support the notion of lexical feedback: the *Ganong effect* (Ganong, 1980). In the Ganong paradigm, a continuum is created from a word to a nonword, e.g., SHAPE-*SAPE, and participants identify the phoneme that is changing ('sh' vs. 's' in this case). Compared to a nonword-nonword continuum or a word-word continuum (e.g., SHIP-SIP), where one would observe a classic categorical perception pattern with an apparent boundary in the middle of the continuum (first few steps identified as 'sh', last few as 's', with intermediate response rates for items in the middle of the continuum, reflecting their ambiguity), the pattern shifts for a word-nonword continuum; the boundary shifts towards the nonword (such that the previously ambiguous items are identified as consistent with the word, and previously unambiguous steps nearer to the nonword continuum become ambiguous).

This pattern is naturally explained under a feedback account. Phonemes receive two sources of input in a model like TRACE: bottom-up input and lexical feedback. When there is a lexical item consistent with one endpoint of a continuum but not the other, the phonemes consistent with the lexical endpoint receive more total activation, because lexical knowledge directly mediates sublexical activation. This proposed mechanism also readily accounts for other top-down effects; we will discuss two more examples. The first is the *word superiority effect*, where phonemes can be detected more quickly in word than nonword contexts (Rubin, Turvey, & van Gelder, 1976) -- on the interaction account, this arises due to the boost of top-down feedback for phonemes in words. The second is *phoneme restoration* (e.g., Warren, 1970). If a phoneme in a word is replaced by noise (e.g., the /t/ in RESTORE), participants report hearing noise, but also report hearing the missing phoneme, and have difficulty identifying where the noise was positioned within the word; however, if the phoneme is replaced by silence, participants precisely identify the position of the gap and do not report hearing the missing phoneme (see, e.g., Samuel, 1981, 1996, 1997). This phenomenon demonstrates both the potential for top-down feedback to restore an ambiguous or noise-masked (or replaced) portion of a stimulus and the need for models to exhibit *bottom-up priority* (under the assumption that noise allows partial activation of all phonemes, which is enough to allow lexical feedback to boost the missing phoneme sufficiently).[10]

Norris et al. (2000) propose that such effects arise post-perceptually -- that is, that lexical knowledge moderates a decision process rather than mediating sublexical activation. In their MERGE model, phonemes feed to words, and then both levels – phonemes and words – feed to a post-perceptual bank of phoneme decision nodes. Norris et al. argue that any result appearing to support lexical feedback could be simulated by an autonomous model like MERGE with post-perceptual integration.

Norris et al. presented two simulations with MERGE demonstrating its abilities to simulate apparent lexical effects. The first was *subcategorical mismatch,* where words are cross-spliced to introduce misleading coarticulatory cues consistent with a word or nonword. Marslen-Wilson and Warren (1994) used a lexical decision paradigm and found that words cross-spliced with a word (e.g., NET with coarticulation on the vowel consistent with NECK) or a nonword (NET with misleading coarticulation consistent with NEP on the vowel) were both recognized more slowly than a version of NET with consistent coarticulation (created by splicing together two instances of NET to ensure that any differences were not due to the cross-splicing operation). The found reported that TRACE predicted a different pattern, with misleading coarticulation from a cross-spliced word leading to slower target activation than from misleading coarticulation from a cross-spliced nonword. Norris et al. found that MERGE predicted

[10] For a debate as to whether TRACE appropriately accounts for phoneme restoration, see Grossberg and Kazerounian (2011, 2016) vs. Magnuson (2015).

the pattern observed by Marslen-Wilson and Warren (equivalently slowed responses for word or nonword cross splicing). Subsequently, Dahan, Magnuson, Tanenhaus, and Hogan (2001) used the visual world paradigm to revisit this finding. They found that activations in TRACE mapped tightly onto participants' fixation proportions. They attributed the difference between eye tracking and lexical decision data to participants responding 'yes' in response to the activation of the lexical competitor when the misleading coarticulation was consistent with a word (Magnuson, Dahan, & Tanenhaus [2001] simulated 'yes' responses from TRACE activations and the Dahan et al. fixation proportions; both predict the Marslen-Wilson & Warren data pattern for fairly fast responses). We will return to this result when we discuss Shortlist B.

Their second simulation focused on a result reported by Connine et al. (1997), where final phonemes in nonwords were detected more quickly when the nonword was more similar to a real word (e.g., participants detected the final /t/ in *GABINET more quickly than in *MABINET; the former differs from CABINET by only a single phonetic feature). MERGE readily simulated this patter with post-perceptual lexical integration in its phoneme decision nodes. Based on these results, and detailed critical examinations of other purportedly top-down effects, Norris et al. argued that there was no evidence that definitively supported an interactive architecture over an autonomous one.

Before turning to Shortlist B, we need to go on a brief tangent about another paradigm that proponents of both views agreed had strong potential to provide definitive support for interaction: *lexically-mediated compensation for coarticulation (LCfC).* Elman and McClelland (1988) devised this paradigm by combining the Ganong effect (discussed above) with the *compensation for coarticulation* paradigm (Mann & Repp, 1981). In the latter paradigm, participants perform an identification task on a front-back place of articulation (POA) continuum (e.g., from TAPES [front] to CAPES [back]) without context, and with a context word ending with a sound with a front place of articulation (e.g., MUSS) or a back place of articulation (e.g., MUSH). Perception of the continuum shifts *away* from the place of articulation of the context sound. Mann and Repp proposed that listeners learn that in real speech, if a talker has to produce two segments with distant places of articulation, they are unlikely to reach the canonical place of articulation for the second sound. Thus, we learn to compensate for coarticulation, and accept noncanonical place of articulation based on context.

Elman and McClelland proposed that if compensation for coarticulation reflected sublexical interactions, one could test for lexical mediation of sublexical processing by replacing clear context sounds with ambiguous sounds midway between front and back places of articulation, but then use lexical context to restore one or the other. They used items like CHRISTMAS and SPANISH, with the final segment replaced with the same ambiguous fricative. Consistent with their prediction, these restored phonemes shifted the perception of the target continuum (e.g., TAPES-CAPES) in the same direction as clear tokens of 's' or 'sh'. However, Pitt and McQueen (1998) reported that it seemed that a sublexical explanation was possible: they were able to drive compensation for coarticulation with ambiguous phonemes in nonword contexts based on the likelihood of the preceding phoneme. Subsequently, Magnuson et al. (2003a) reported LCfC with words where transitional probabilities were pitted against lexical context, and lexical context won. Magnuson et al. (2003b) also assessed all items that had been used in previous LCfC studies and found that the diphone transitional probabilities were at odds with lexical context in several items, and that larger *n*-phones could not explain all extant results (on average, the necessary *n*-phone was approximately the same as word length). Samuel and Pitt (2003) also reported additional positive LCfC results (along with analyses of factors that appear to affect the strength of effects). However, McQueen, Jesse and Norris (2009), using materials supplied by Magnuson et al., were unable to replicate the earlier results. Although more positive than negative results have been reported (possibly reflecting a "file drawer" phenomenon due to the difficulty of publishing null results), the apparent fragility of the LCfC

paradigm has led to an impasse.

With this context, let us consider Shortlist B. The 'B' in Shortlist B stands for *Bayesian*, as Shortlist B is a radical revision of the original Shortlist model. Norris and McQueen (2008) redubbed the original model 'Shortlist A', with 'A' standing for *activation*. Shortlist A was a neural network model, and thus its currency was node activation. It differed from the TRACE model by building a competition network on-the-fly based on the top lexical matches (intended to be generated by simple recurrent networks [Elman, 1990; 1991], but generated via a simple lexical lookup) and, crucially, eschewing feedback. (For an extensive review of the case for rejecting lexical feedback, see Norris, McQueen, & Cutler, 2000.)

Norris and McQueen (2008) reject the notion of linking activation in network models to human behavior and cognition as convoluted (additional parameters and assumptions are frequently required to make such links). They propose instead that probabilities from a low-parameter Bayesian model can be linked directly to human behavior, while also providing a principled and optimal model of human SWR. Like Shortlist A, Shortlist B rejects top-down information flow. Shortlist B also uses a different form of input (diphone confusion probabilities measured from human participants, sampled in 3 'gates' per phoneme). At each input step, up to 50 words aligned with the current phoneme can be added to a 'lattice' (a set of paths corresponding to possible lexical segmentations of the input) consisting of a maximum 500 paths -- the 50-word and 500-path limits are the 'shortlist' parameters. Words and paths are ranked according to their probabilities (with words evaluated relative to their frequency and fit with the bottom-up input, and paths evaluated based on conditional probabilities). Probabilities over all paths within the lattice are ranked, and the model works in some ways very much like the Cohort Model logic we presented in Figures 1 and 2, such that for many word sequences, only one complete path is possible. When more than one path is possible (e.g., perhaps RECOGNIZE SPEECH and WRECK A NICE BEACH), prior probabilities of words or conditional probabilities of multiple words identify the *most likely* parse. One way in which Shortlist B compares very favorably with TRACE is that it includes the full Dutch phoneme inventory (37 phonemes) and a lexicon of 20,250 Dutch words.

Norris and McQueen reported seven simulations (see Table 2). They also reported a new Merge model, Merge B and used it to simulate Marslen-Wilson and Warren's (1994) subcategorical mismatch paradigm. They compared Merge B to data from Marslen-Wilson and Warren as well as a study by McQueen, Norris and Cutler (1999). Merge B provided quite good quantitative fits for both word and nonword cross-spliced items, with some minor discrepancies. (However, they neither mention the Dahan et al. [2001] eye tracking data and TRACE simulations, nor provide simulations of the time course.)

Thus, Shortlist B accounts for 7 phenomena, including a more detailed assessment of neighborhood and frequency interactions than has been conducted with TRACE, and simulations of the stress-based *Possible Word Constraint* (Norris, McQueen, Cutler & Butterfield, 1997) that would not be possible in TRACE without adding representations for syllabic stress. This is far fewer than the list in Table 1, and the lack of overlap (with a few exceptions) impedes direct model comparisons. The full phoneme and large lexical inventories in Shortlist B is a considerable advance compared to other models. One potential weakness of Shortlist B is that the model is told explicitly where each phoneme begins, whereas in TRACE, phonemes overlap and their onsets are not directly encoded. As we already mentioned, there are not robust cues for phoneme segmentation, so this is a simplifying assumption that may need to be reconsidered in the future in Shortlist B.

Table 2: Effects accounted for by the Shortlist B model. All are reported by Norris and McQueen (2008). Note that an eighth simulation of subcategorical mismatch is not included because it was conducted with a different model ("Merge B").

	Effect / phenomenon	Reference	Notes
1.	Parsing a multiword sequence and overcoming embedded words via right context	*Fig. 3*	TRACE: #11 in Table 1
2.	Word frequency x neighborhood density x neighborhood frequency	*Fig. 4*	TRACE has only been applied to word frequency (Table 1, #17) and selective neighborhood characteristics (Table 1, #8)
3.	Word frequency x neighborhood characteristics x stimulus quality	*Fig. 5*	Not directly simulated with TRACE, though #27 in Table 1 could be comparable for neighborhood and stimulus quality)
4.	Time course of word frequency effects	*Fig. 7*	TRACE: #17 in Table 1
5.	"Word spotting" on the basis of the stress-based Possible Word Constraint	*Fig. 8*	Not possible in TRACE without adding stress representation
6.	Identity priming based on Possible Word Constraint	*Fig. 9*	Not possible in TRACE without adding stress representation
7.	Recovery from onset mispronunciation	*Fig. 10*	Conceptually related in TRACE: Table 1 #4, #8

What of the feedback debate? It remains unresolved. Norris, McQueen and Cutler (2016) present an extended case for their position that feedback is never *necessary*, but have extended their treatment of this issue to accept some forms of potential feedback, while continuing to reject (only) feedback as implemented in TRACE. They suggest that feedback in service of Bayesian processing is compatible with their view, because it is qualitatively different from what they label "activation feedback." Magnuson et al. (2018) report simulations showing that feedback helps make TRACE robust against noise and makes TRACE "Bayes approximant". They also provide a critique of Norris et al. (2016)'s arguments. In addition, McClelland (2013) and McClelland et al. (2014) provide detailed explanations of how interactive activation models relate to Bayesian models, and under what conditions they are Bayes-equivalent. Norris, McQueen, and Cutler (2018) published a commentary laying out their disagreement with Magnuson et al. (2018). On our view, this debate remains at an impasse. Until a paradigm is devised that can definitively falsify one model or the other -- or until the field of SWR simply moves on to more realistic models -- no resolution is in sight.

Indeed, one might take the view that the field is moving on, with great excitement about the concept of *predictive coding (PC)*. Rao and Ballard (1999) originally proposed PC as an extension of theoretical work in vision. In their work, a visual stimulus (image patch) was encoded by a hierarchy of 'modules' with progressively larger spatial scale (e.g., first-level modules take input from small, overlapping grids of pixels tiled over part of an image, and second-level modules take input from multiple, overlapping first-level modules, and so on). Higher levels send predicted states to their immediately inferior level, and those predictions are compared to the states of the inferior-level nodes. Inferior levels, rather than passing forward their actual state, pass forward the discrepancy between the top-down prediction and their state. This provides a potentially compact and robust code. Rao and Ballard's PC model developed receptive fields at the lowest levels that resembled wavelets based on the statistics of natural images. Higher levels developed progressively larger and more complex receptive fields for visual features (and note that that is where the model stops; it does not perform image or object recognition, for example). A possible prediction from such a mechanism is that when an input conforms to top-down expectations, but bottom-up signal should be *weaker* since error (which is what is passed forward) is smaller.

Several neuroimaging studies have reported results consistent with this hallmark of PC. In an example from SWR, Gagnepain, Henson, and Davis (2012) presented listeners with (lexical) expectation-

confirming inputs (i.e., words, e.g., *formula*) or (lexical) expectation-violating inputs (e.g., *formubo*). Gagnepain et al. found relatively weaker activation in superior temporal gyrus at the *la* of *formula* compared to the *bo* of *formubo* (i.e., for items like these; this is just an example). We note that such results are consistent with longstanding evidence that unexpected inputs can drive event-related potentials like the phonological mismatch negativity or N400 (e.g., Kutas, Van Petten, & Kluender, 2004). But do such results demonstrate *predictive coding* or simply *predictive processing*? Davis and colleagues have done some proof-of-concept modeling with simple Bayesian prediction (Gagnepain et al., 2012) and variants of interactive activation models (Blank & Davis, 2016) with multiplicative feedback from words to phonemes (somewhat like TRACE) or subtractive feedback (intended to be like predictive coding). A full discussion of these studies is beyond the scope of this chapter, but in a nutshell, these models are neither faithful implementations of PC, nor have they been validated on basic aspects of SWR. Luthra, Li, You, and Magnuson (under review) provide a fuller review of this literature, and also present simulations using the Gagenpain et al. simple Baysesian model (which we call "predictive cohort"), TRACE, and a simple recurrent network (SRN; Elman, 1990) trained on TRACE-like inputs (it predicts the current word from acoustic-phonetic features over time). All three models display predictive processing, but surprisingly, the only model that shows the putative hallmark of PC (a model-internal reduction in signal for expectation-confirming stimuli akin to *formula* relative to expectation-violating inputs akin to *formubo*) is TRACE. Luthra et al. argue that true understanding of the nature of neural signal changes in response to expectations, and whether they actually reflect PC, will require the development of a recurrent PC model that can be applied to SWR.

New developments, unresolved challenges, and the limits of current theories and models

The breadth and depth of models like TRACE and Shortlist B are substantial and remarkable. These models have helped guide theories of SWR, and inspired a wide variety of empirical and computational investigations that have enhanced our understanding of human SWR. They complement rather than fully supplant rule-based (mathematical and verbal) models like the Neighborhood Activation Model and the Cohort Model. In particular, the Neighborhood Activation Model's DAS rule for defining neighbors continues to contribute to most studies of SWR, as controlling neighborhood characteristics has become the convention.

In recent years, Vitevitch and his colleagues have added new force and scope to the NAM approach by applying the tools of *network science* (e.g., Menczer, Fortunato, & Davis, 2020; Newman, 2010) to graphs created by connecting DAS neighbors (pioneered by Vitevitch, 2008). Network science provides a toolkit for characterizing interconnected systems at gradient levels of analysis ranging from local ("microscale", i.e., individual nodes, corresponding to words in this case) to subcomponents ("mesoscale") to global ("macroscale", characterizing the entire network based on statistics aggregated over all nodes[11]).

[11] Readers may be familiar with the notion of "small world networks" (Watts & Strogatz, 1998), where most nodes have fairly few connections, but enough nodes have many connections that the number of "hops" from node-to-node it takes to get from any node to any other node is small. This was first observed informally in networks of human acquaintances (Milgram [1967] famously asked people in the midwest to pass along a letter [without an address] to a named individual on the east coast by sending it to someone they knew who might know someone [who might know someone else] more likely to know the addressee; on average, letters that arrived had passed through 6 individuals – the basis for the possibly familiar concept that there are only '6 degrees of separation' between randomly selected individuals). A surprising number of social, biological, and artificial systems have small world structure, though the tools of network science can classify many kinds of networks, and network type has significant implications (e.g., for a system where information is transmitted, how efficiently information can be transmitted and how robust the system may be to noise or damage; see Strogatz [2003] for an accessible overview).

Vitevitch and colleagues have shown that a graph-theoretic network of lexical forms based on the DAS rule immediately increases NAM's scope. For example, when sets of words are selected to be matched on neighborhood but differ in *clustering coefficient* (the proportion of a word's neighbors that are also neighbors of each other), sets with higher clustering coefficient are processed more slowly (Chan & Vitevitch, 2009). Similarly, Siew (2017) examined sets of words matched on DAS neighborhood ("1-hop" neighbors in a graph, since they are directly connected) but varying in number of 2-hop neighbors (words that are 2 links apart in the DAS network); words with more 2-hop neighbors were processed more slowly. It is still early days with respect to applying this unconventional, innovative approach,[12] but it has the potential to provide novel insights into SWR. For example, one possibility that could be explored is comparing similarity metrics by using them as the connecting rules for such networks.

Note that there are many aspects of human speech that are undoubtedly crucial for human SWR that are outside the scope of current models and theories (for the most part, in the sense that we do not have theories that provide an integrated account of the phenomena reviewed above and those reviewed below, even if theoretical accounts have been offered for individual phenomena). Let's consider a subset of these unresolved challenges.

One challenge for theories of spoken word recognition is how to account for the fact that listeners process speech under a variety of contexts, often in environments with more variation than we typically allow in a laboratory (see Purse, Tamminga, & White, this volume), which also poses significant challenges for understanding the development of lexical knowledge (see Creel, this volume). Variation can occur at the level of the ambient acoustic environment (which may be noisy or echoey), details of the speech input (such as speaking rate, or number of talkers, and differences between them in size, age, sex, or dialect/accent). Mullennix, Pisoni, and Martin (1989) reported foundational studies illustrating the impact of **talker variation**. In a series of tasks, Mullennix and colleagues found that both the speed and accuracy of SWR are reduced under conditions of talker changes. Across perceptual identification and naming tasks, they found that the impact of talker variation was more consistent than the impact of word frequency or neighborhood density. Magnuson and Nusbaum (2007) provide a review of talker variation effects, and suggest ways theories of speech perception and SWR might accommodate talker variation; a key point is that *features* of various sorts are *stipulated* (given to models) rather than discovered. For example, phoneme boundaries are stipulated in Shortlist B, and phoneme templates are stipulated in TRACE. Many additional features would have to be stipulated in such models to account for natural variation in the speech signal.

Another challenge is that **speaking rate varies** dynamically in everyday speech, and this alters the mapping from acoustics to perceptual categories. A classic example is that formant transition durations that correspond to /w/ at a relatively fast speaking rate map to /b/ at a relatively slower rate (e.g., Miller & Baer, 1983), suggesting listeners must have to accommodate this variation somehow. The impact of variation can be seen in SWR tasks. Francis and Nusbaum (1994) found an interaction between cognitive load and variation in speaking rate, and Nusbaum and Morin (1992) found a similar interaction between load and talker variation. McLennan and Luce (2005) reported similar interactions, where talker and rate variation interacted with task difficulty. Such results suggest that accommodating variation is an effort- and attention-demanding process.

However, recall that one of the core simplifying assumptions of the simplified mapping perspective is that

[12] For an application of graph theoretic networks to speech perception (phonological categories and talker variability), see Crinnion, Malmskog and Toscano (2020).

a speech perception module provides something like a stream of phonemes as input to SWR. So perhaps we can simply propose that prelexical normalization mechanisms of some sort take care of these sources of variation. An impediment to this view is that lexical and sentential information appear to provide crucial bases for accommodating variation. For example, a lexical mismatch might be a better indicator of a rate change than, e.g., vowel durations (hearing what seems to be a /w/ in a lexical context that calls for a /b/, e.g., /wol/ [nonword *WOLE] might suggest the actual production is /bol/ [BOWL]). Similarly, context may indicate that a change in talker requires a change in acoustic-phonetic mapping (e.g., hearing THE CAP WAS ON HER HID might suggest a change in the /ɛ/-/ɪ/ boundary).[13]

Indeed, over the last two decades, a growing literature has shown that listeners learn to adjust acoustic-phonetic mappings for specific talkers based on lexical context – a finding known as **lexically mediated perceptual learning**. Norris, McQueen, and Cutler (2003) presented listeners with an ambiguous fricative (midway between /f/ and /s/). Critically, one group of listeners heard the ambiguous token on /f/-final words and heard unambiguous /s/-final words, while another group of listeners heard the ambiguous token on /s/-final words and heard unambiguous /f/-final words. Afterwards, when listeners categorized tokens from an /f/-/s/ continuum, the listeners who heard the ambiguous token on /f/-final words categorized more tokens along the continuum as /f/ than those who heard the ambiguous token on /s/-final words. Importantly, this shift does not occur with ambiguous tokens at the ends of non-words. This foundational study has been extended by many groups, and in particular Kraljic and Samuel (2005, 2006, 2011; Samuel & Kraljic, 2009), who have examined how such perceptual learning varies across different classes of phonemes, whether it generalizes between talkers, and contexts that can block such learning (e.g., an image of a talker with a pen in her mouth suggests abnormal patterns result from that motor difficulty rather than reflecting talker-specific patterns). We do not have space to review this literature in detail, but note that it suggests a tight linkage between signal-level details and lexical and sentential contexts.

Another challenge is that in casual, fluent speech, there are rampant phonological processes that lead to **significant deviations from canonical phonemic forms** of spoken words. For example, Gow and McMurray (2007; see also Gaskell & Marslen-Wilson, 1996) studied coronal place **assimilation**, in which a coronal segment assimilates the place of a following non-coronal segment (i.e., *clean bars* might be pronounced as *cleam bars*). They used a visual world paradigm study to examine whether context can signal lexical form *using* coronal assimilation. Because coronals assimilate in English, hearing an assimilated segment followed by a non-coronal is more likely than an assimilated segment followed by a coronal (i.e., when *bite guard* is pronounced more like *bike guard,* it is more likely to be understood as *bite guard* in context because of assimilation; when *bite damage* is pronounced more like *bike damage*, it will be understood as *bike damage*, since if the true underlying form were *bite*, there would be no assimilation due to the following coronal). Gow and McMurray found that assimilated coronals do indeed facilitate activation for non-coronals (where assimilation occurs) and inhibit activation for coronals (where assimilation does not occur). They also found regressive effects; when there is lexical ambiguity in the assimilated segment (i.e., *bike* and *bite*, as compared to *clean* and *cleam*), the class of the following segment (i.e., whether it is a coronal or not) influences processing of that initial segment (*bike* or *bite*). Complicating the situation further, assimilations tend to be partial and graded (Gow, 2001, 2002, 2003). Clearly then, these phonological processes influence the lexical processing in ways that may be difficult to reconcile within current models of SWR.

There are also **reductions** that lead to more extreme deviations from canonical forms. Consider

[13] This impediment also applies to episodic or exemplar accounts that try to avoid normalization (e.g., Goldinger, 1998), or at least suggests a limitation for them (see Magnuson & Nusbaum, 2007, for a more detailed critique of episodic accounts).

progressively more casual productions of "I am going to be there" in Table 3, which gives a sense of the challenge. Johnson (2004) found that more than 60% of words in a corpus of casual speech deviate from canonical forms, and that one or more segments are missing in nearly 30% of casually-produced words. This presents an enormous challenge to models and theories of SWR under typical listening conditions (clear speech, low noise, low cognitive effort). Janse and Ernestus (2011; see also White, Mattys, & Wiget, 2012) report experiments that suggest continuous use of syntactic and semantic context is required to overcome the rampant ambiguity that results; indeed, transcription accuracy is very poor for words extracted from fluent speech and presented in isolation. This bumps up against another core simplifying assumption of the SMP, that form recognition can proceed modularly without constraints from higher-levels of processing.

Table 3 Progressively more casual productions of "I am going to be there."

Gloss	IPA	Number of phonemes
I am going to be there	/ɑⁱæmgointubiðeⁱɹ/	14
I am going to be there	/ɑⁱmgointubiðeⁱɹ/	13
I'm gonna be there	/ɑⁱmgʌnʌbiðeⁱɹ/	11
I'munna be there	/ɑⁱmʌnʌbiðeⁱɹ/	10

Another consideration is the rich prosodic information in the speech signal. This includes the 'melody' of speech (variation in pitch over time), but also timing and stress, among other factors (see Dahan, 2015, for a review). Most aspects of prosody are absent from current models, with one salient exception: the *metrical segmentation strategy (MSS)* proposed by Cutler and Norris (1988), and integrated into the Shortlist A model by Norris, McQueen, and Cutler (1995). The MSS proposes that language-specific, probabilistic patterns of strong and weak syllable stress can constrain segmentation of fluent speech. Human behavior in several studies has been consistent with the MSS (e.g., detection of a word [e.g., MINT] embedded within a nonword is better when the word occurs at the onset of a nonword with strong-weak stress [MINTEF] rather than a strong-strong nonword [e.g., MINTAFE]). The MSS was implemented in Shortlist A simply by giving a boost to the match score of items that began with a strong syllable, and penalizing other items – in essence, creating features that code for primary stress on the first syllable of a word. While this did not make Shortlist A sensitive to stress in general, it substantially boosted the model's ability to simulate human sensitivity to the MSS. Other aspects of prosody have not been incorporated into current models, as we will discuss in the next section.

The challenges of variation, prosody, assimilation, and reduction are exacerbated by a variety of **adverse listening conditions** that are common outside the laboratory (see Mattys, Davis, Bradlow, & Scott, 2012, for a review). Noise, cognitive load, anxiety (Mattys, Seymour, Atwood, & Munafo, 2013), and chronic or age-related reductions in perceptual acuity (along with perceptual learning) are, for the most part, outside the scope of current models (noise is easy to apply, though more work is required to link noise in models to human speech recognition under noise, and Mirman et al. [2008] address attentional effects and Mirman et al. [2006] address perceptual learning with the TRACE model). Models must be extended to real-world conditions, learning over the lifespan, and age-related changes.

We have avoided grappling with **neural-level** responses to SWR in this chapter so far. One might justify this on the common interpretation of Marr (1982) that *algorithmic* theories can be developed independently of implementational (neurobiological) details. Marr, however, also argued that complete understanding requires eventual linkage of computational, algorithmic, and implementational theories. In fact, as we consider how interactions between speech perception and SWR, cognitive neuroscience findings may be crucial. For example, using an effective connectivity analysis, Gow and Olson (2016)

found influences of brain regions associated with lexical processing areas on regions associated with lower-level acoustic information, suggesting that the neural representations of ambiguous speech sounds may be influenced by sentential and lexical context (i.e., lexical information may *mediate* perception, even at a neural level). Furthermore, Gwilliams, Linzen, Poeppel, and Marantz (2018) used MEG to demonstrate that acoustic-phonetic information can be maintained so that later information in the processing stream can influence perception of earlier sounds. Recent findings using EEG to measure the time course of speech perception have demonstrated that both semantic information and lexical content influence processing at lower levels (i.e., sublexical; Getz & Toscano, 2019; Noe & Fischer-Baum, 2020). Noe and Fischer-Baum (2020) showed that for an ambiguous sound between a voiced and unvoiced sound (i.e., a sound in between /t/ and /d/), participants' neural responses (as measured by the N100 ERP component, a signal known to indicate the presence of voicing; see Toscano, McMurray, Dennhardt, & Luck, 2010) demonstrated lexical mediation: when participants heard a sound ambiguous between /t/ and /d/ at the beginning of a continuum where only one sound made up a word (i.e., tape-*dape, vs. *tate-date), participants' N100 responses were more like responses to less ambiguous /t/ sounds when hearing a tape/*dape continuum and more like /d/ when hearing a *tate/date continuum. These lines of work represent an important frontier between speech perception, SWR, and neural mechanisms that operate during speech processing.

Moving forward

How might understanding of SWR move forward? It is possible that we are reaching the limits of current modeling frameworks. One of the major hurdles is relating inputs in models (ranging from phonemes in TISK [Hannagan et al., 2013] to 'pseudo-spectral' features over time in TRACE to human diphone confusion probabilities in Shortlist B) to real speech. As we note above, adding additional aspects of the speech signal to such inputs mainly involves reducing those details to features of some sort -- *stress* as present or not, for example. One could try building in multiple forms of each word in a lexicon to account for reductions. One might build in conditional probabilities to account for phenomena like assimilation or compensation for coarticulation. However, such approaches are unlikely to extend current modeling approaches much further, as we are unlikely to be able to fully anticipate the continuous, graded nature of variation in speech.

Consider a study by Salverda, McQueen, and Dahan (2003; see also Davis, Marslen-Wilson, & Gaskell, 2002). They reasoned from longstanding findings that there are systematic relationships between prosodic features and word length (Lehiste, 1970; e.g., vowel durations in stressed syllables tend to be longer in monosyllabic than bisyllabic words), and that listeners might be sufficiently sensitive to these probabilistic contingencies to constrain SWR. They recorded multiple talkers saying words like HAM, HAMMER, and HAMSTER. The initial vowel, on average, was about 20 msecs longer in HAM than in bisyllabic words, although duration distributions overlapped. They used a visual world paradigm and found that listeners fixated referents of words significantly faster when vowel duration relative to speaking rate was consistent with the number of syllables in the word (that is, durations consistent with HAM led to longer latencies to reach HAMSTER than durations consistent with a bisyllabic word).

This illustrates how the simplifying assumptions underlying the SMP may actually *complicate* aspects of SWR, as this finding suggests that words contain probabilistically-constraining information about word length that could, for example, mitigate the embedding problem we mentioned earlier. If we assume a phonemic grain of input to SWR, we hide from ourselves subphonemic constraints that could actually *simplify* a core challenge of SWR (see Magnuson, 2008, for extended discussion). And, as we discussed already, a similar situation holds in downstream direction: recognition of many words in casual, fluent speech depends crucially on morphological, semantic and syntactic context for disambiguation. The

simplifying assumption in the SMP that the goal of SWR is recognition of lexical sound forms initially improves the tractability of SWR (by giving us a more manageable scope) but at the cost of walling off constraints that may greatly simplify the problem.

These two problems – the proliferation of features that can approximate but not fully capture details of real speech, and the potential for simplifying assumptions to hide constraints – are daunting. Moving beyond them may require modeling frameworks that both use real speech as input and connect to higher levels of language comprehension. As we noted early in this chapter, the modal SMP view of SWR adopts simplifying hypotheses that we expect virtually all researchers in this field would agree are incorrect, but were crucial to pioneering work in SWR. Their utility is reductionist; they constrain theoretical problems to manageable scope. But as we also discussed, such simplifications have to be provisional; once enough smaller component problems are fairly well understood, it is time to develop larger-scope, integrative theories. We think we have reached this point in SWR.

One potential way to do this would be to embrace developments in deep learning that have allowed complex networks that have evolved from the same origin as models like TRACE (i.e., the parallel distributed processing revolution; Rumelhart, McClelland et al., 1986; McClelland, Rumelhart, et al., 1986) that currently power (fairly) robust automatic speech recognition on literally billions of smart phones worldwide every day (e.g., Hinton et al., 2012). A formidable challenge in exploring implications of deep learning approaches for cognitive theories is that the best-performing systems have many layers of multiple kinds and require complex, (arguably) biologically-implausible training regimes. This means understanding how and why the fully trained systems function as they do may present theoretical and technical challenges similar to the ones we face in trying to understand human SWR -- that is, determining how those systems work may require substantial experimentation (via simulation with the systems themselves) or even the development of simpler models to isolate and identify key algorithmic details. The prospects for progress may thus appear rather dismal.

Kietzmann, McClure, and Kriegeskorte (2019) make a compelling case that in fact, deep networks are not inscrutable black boxes, but can be understood at various levels of detail by careful analysis and by relating them to human behavior and neurobiology. Furthermore, it is not necessary to begin with the most complex models available for a domain. Magnuson, You, Luthra et al. (2020) set out to borrow the minimum possible from automatic speech recognition approaches in order to develop a neural network model of human speech recognition that would operate on real speech. They developed a two-layer recurrent network that maps spectral slice inputs from speech files to semantic output vectors via a hidden layer of *long short-term memory (LSTM)* nodes (Hochreiter & Schmidhuber, 1997). LSTMs have memory cells and gates that determine the relative weight of new and old information from long sequences of input.

This network, dubbed EARSHOT (for *Emulation of Auditory Recognition of Speech by Humans Over Time*), achieves high accuracy on 1000 words for 9 training talkers, as well as moderate generalization to subsets of words excluded from training for each talker, and for a tenth talker completely excluded from training. It also simulates the time course of lexical activation and phonological competition over time, with results that resemble those observed in TRACE (see Figure 4). Intriguingly, although the model is not trained on phonetic targets, a structured phonetic code emerges in hidden unit responses that strongly resembles phonetically-structured responses in human superior temporal cortex (Mesgarani, Cheung, Johnson, & Chang, 2014). Magnuson et al. point out that this similarity does not necessarily indicate similar functions or mechanisms in the model and cortex -- simply that EARSHOT and human cortex are both sensitive to key information in the speech signal (i.e., phonetic patterns). However, EARSHOT's hidden units also display complex spectrotemporal response patterns that could generate novel predictions

for human cortical responses to speech.

EARSHOT represents an initial step toward creating models complex and powerful enough to operate on real speech while remaining simple enough to guide theoretical understanding. It currently has significant limitations. While EARSHOT is exposed to surface variation of various sorts, whether and how it accommodates that variation has not yet been addressed. EARSHOT is also currently restricted to single words, but has the potential to take structured, continuous inputs, which could open the way to models that integrate levels from speech to sentence processing. As a learning model, it also has the potential to provide a new framework for studying the development of spoken language comprehension. It is possible that by incrementally increasing the scope, realism, and complexity of such models, the gap between current cognitive models of SWR (e.g., TRACE, Shortlist B) and deep learning networks used for robust, real-world automatic speech recognition can be bridged -- and along the way, provide bases for deeper theoretical understanding of human spoken language comprehension.

However, neural networks like EARSHOT may miss important aspects of human speech processing. While the success of deep neural networks for robust automatic speech recognition used by billions of smartphone users daily (e.g., Hinton et al., 2012) is reason for optimism, it could be that such networks are to human speech recognition as airplanes are to avian flight -- a good solution, but not homologous to biology. An alternative may come from neural networks that are able to *oscillate* in response to the temporally-varying speech signal (Giraud & Poeppel, 2012). Peele and Davis (2012) propose that if oscillating neural systems encoding speech entrain to dynamically-changing rhythms of speech, this may provide natural solutions to some of the challenges discussed above (rate variation may fall away if the entire system entrains to amplitude modulations of speech). Future progress will likely require deeper understanding of the neurobiological foundations of speech processing guided by innovative, neurally-realistic models.

References

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38, 419-439.

Bagley, W. C. (1900-1901). The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology, 12*, 80-130.

Balota, D. A. (1990). The role of meaning in word recognition. In D. A Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9–32). Hillsdale, NJ: Erlbaum.

Blank, H. & Davis, M.H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology, 14*(11): e1002577. https://doi.org/10.1371/journal.pbio.1002577

Carpenter, R. H. S. (1988). *Movements of the eyes* (2nd edition). London: Pion.

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1934–1949. https://doi.org/10.1037/a0016902

Connine, C. M., Titone, D., Deelman, T., & Blasko, D. G. (1997). Similarity mapping in spoken word recognition. *Journal of Memory and Language, 37(4)*, 463– 480.

Crinnion, A. M., Malmskog, B., & Toscano, J. C. (2020). A graph-theoretic approach to identifying acoustic cues for speech sound categorization. *Psychonomic Bulletin & Review*, 1-22.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14(1)*, 113–121.

Dahan, D. (2015). Prosody and language comprehension. *Wily Interdisciplinary Reviews: Cognitive Science, 6*, 441–452. doi: 10.1002/wcs.1355

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*, 317-367.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (2001). Tracking the time course of subcategorical mismatches: Evidence for lexical competition. *Language and Cognitive Processes, 16 (5/6)*, 507-534.

Davis, M. H., Marslen-Wilson, W. D. & Gaskell. M. G. (2002). Leading up the lexical garden-path: segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 218-244.

Dufour, S., & Grainger, J. (2019). Phoneme-Order Encoding During Spoken Word Recognition: A Priming Investigation. *Cognitive Science*, *43*(10), e12785.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195–225.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In: J. S. Perkell, & D. H. Klatt (Eds), *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory & Language, 27*, 143-165.

Fischer, B. (1992). Saccadic reaction time: Implications for reading, dyslexia and visual cognition. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 31–45). New York: Springer-Verlag.

Francis, A. L., & Nusbaum, H. C. (1996). Paying attention to speaking rate. *Proceedings: International Conference on Speech and Language Processing* (ICSLP) '96, vol. 3, SaA2L2. DOI: 10.1109/ICSLP.1996.607911

Frauenfelder, U. H., & Peeters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE. In J. Grainger & A. M. Jacobs (Eds.), *Localist Connectionist Approaches to Human Cognition,* pp. 101-146. Mahwah, NJ: Erlbaum.

Gagnepain, P., Henson, R.N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22(7)*, 615–21.

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

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 144–158

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*(5), 613-656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science, 23*, 439-462.

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science*, *30*(6), 830-841.

Giraud, A., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*, 511-517.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

Gow, D. W. Jr., & Olson, B. B. (2016). Sentential influences on acoustic-phonetic processing: A Granger causality analysis of multimodal imaging data. *Language, Cognition and Neuroscience*, *31*(7), 841-855.

Gow, D. W., Jr. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory & Language, 45*, 133–159.

Gow, D. W., Jr. (2002). Does english coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception & Performance, 28*, 163–179.

Gow, D. W., Jr. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics, 65*, 575–590.

Gow, D.,W., and McMurray, B. (2007) Word recognition and phonology: The case of English coronal place assimilation. J.S. Cole & J. Hualdo (Eds.) *Papers in Laboratory Phonology 9* (pp 173-200). New York: Mouton de Gruyter.

Grainger, J., & Dufour, S. (2020). The influence of word frequency on the transposed-phoneme priming effect. *Attention, Perception, and Psychophysics*.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word, 20*, 157-177.

Gregg, J., Inhoff, A. W., & Connine, C. M. (2019). Re-reconsidering the role of temporal order in spoken word recognition. *Quarterly Journal of Experimental Psychology*, *72*(11), 2574-2583.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*, 267-283.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review, 107*(4), 735-767.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23*(2), 481-503.

Grossberg, S., Govindarajan, K. K., Wyse, L. L., & Cohen, M. A. (2004). ARTSTREAM: a neural network model of auditory scene analysis and source segregation. *Neural Networks, 17*(4), 511-536.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585-7599.

Hannagan, T., Magnuson, J. S. & Grainger, J. (2013). Spoken word recognition without a TRACE.

Frontiers in Psychology, 4:563. doi:10.3389/fpsyg.2013.00563.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*, 82–97.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9, 1735–1780

Janse, E. & Ernestus, M. (2011). The roles of bottom-up and top-down information in the recognition of reduced speech: Evidence from listeners with normal and impaired hearing. *Journal of Phonetics, 39,* 330-343.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama, & K. Maekawa (Eds.), Spontaneous speech: Data and analysis. *Proceedings of the 1st session of the 10th international symposium* (pp. 29–54). Tokyo, Japan: The National International Institute for Japanese Language.

Kapnoula, E. C., & McMurray, B. (2016). Training alters the resolution of lexical interference: Evidence for plasticity of competition and inhibition. *Journal of Experimental Psychology: General, 145*, 8-30.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. In S. Murray Sherman (Ed.), *Oxford research encyclopedia of neuroscience*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kirk, K.I., Pisoni, D.B., & Osberger, M.J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing, 16*, 470-481.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics, 7(3),* 279–312.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141-178.

Kraljic, T., & Samuel, A.G. (2006). How general is perceptual learning for speech? *Psychonomic Bulletin and Review, 13*, 262-268.

Kraljic, T., & Samuel, A.G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition, 121*, 459-465.

Kutas, M., & Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics Electrified II (1994–2005). In 10.1016/B978-012369374-7/50018-3.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12*, 119-131.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. (Research on Speech Perception, Technical Report No. 6). Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1-36.

Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception and Psychophysics, 62*, 615-625.

Luthra, S., Li, M., You, H., & Magnuson, J. S. (under review). Does predictive processing imply predictive coding in models of spoken word recognition?

Magnuson, J. (2020a). Simple lexical tree, version 2. figshare. doi:10.6084/m9.figshare.12016962.v1

Magnuson, J. (2020b). Cohort Model Flowchart. figshare. doi:10.6084/m9.figshare.12016965.v1

Magnuson, J. (2020c). Neighbors and Cohorts Venn Diagram. figshare. Figure. https://figshare.com/articles/Neighbors_and_Cohorts_Venn_Diagram/12287396

Magnuson, J. (2020d). Very simple TRACE schematic REVISED (Version 1). figshare.

https://doi.org/10.6084/m9.figshare.12410420.v1

Magnuson, J. S. (2008). Nondeterminism, pleiotropy, and single word reading: Theoretical and practical concerns. In E. Grigorenko & A. Naples (Eds.), *Single Word Reading,* pp. 377-404. Mahweh, NJ: Erlbaum.

Magnuson, J. S. (2015). Phoneme restoration and empirical coverage of interactive activation and adaptive resonance models of human speech processing. *Journal of the Acoustical Society of America, 137(3)*, 1481-1492. http://dx.doi.org/10.1121/1.4904543.

Magnuson, J. S. (2017). Mapping spoken words to meaning. In G. Gaskell & J. Mirkovic (Eds.), *Speech Perception and Spoken Word Recognition* (pp. 76-96). New York: Routledge.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. J*ournal of Experimental Psychology: Human Perception and Performance, 33*, 391-409.

Magnuson, J. S., Dahan, D., & Tanenhaus, M. K. (2001). On the interpretation of computational models: The case of TRACE. In J. S. Magnuson and K.M. Crosswhite (Eds.), *University of Rochester Working Papers in the Language Sciences, 2 (1)*, 71 – 91.

Magnuson, J. S., Kukona, A., Braze, B., Johns. C.L., Van Dyke, J., Tabor, W., Mencl, E., Pugh, K.R., & Shankweiler, D. (2011). Phonological instability in young adult poor readers: Time course measures and computational modeling. In P. McCardle, B. Miller, J.R. Lee, & O. Tseng, *Dyslexia Across Languages: Orthography and the Brain-Gene-Behavior Link*, pp. 184-201. Baltimore: Paul Brookes Publishing.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2003a). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science, 27,* 285-298.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2003b). Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science, 27*, 795-799.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, K. McRae, & M. Joanisse (Eds.), T*he Cambridge Handbook of Psycholinguistics*, pp. 76-103. Cambridge University Press.

Magnuson, J. S., Mirman, D., & Myers, E. (2013). Spoken word recognition. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 412-441). New York, USA: Oxford University Press.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, 9:369. doi:10.3389/fpsyg.2018.00369

Magnuson, J.S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P.D., Theodore, R.M., Monto, N., & Rueckl, J.G. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. C*ognitive Science, 44*, e12823. http://dx.doi.org/10.1111/cogs.12823

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69*, 548-558.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71-102.

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576-585.

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

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.

Matin, E., Shao, K., & Boff, K. (1993). Saccadic overhead: Information processing time with and without saccades. *Perception & Psychophysics, 53*, 372-380.

Mattys, S.L., Davis, M.H., Bradlow, A.R., & Scott, S.K. (2012). Speech recognition in adverse

conditions: A review. *Language & Cognitive processes, 27*, 953-978.

Mattys, S.L., Seymour, F., Attwood, A.S., & Munafò, R.R. (2013). Effects of acute anxiety induction on speech perception: Are anxious listeners distracted listeners? *Psychological Science, 24*, 1606-1608.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1–44.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. *Frontiers in Psychology*, 4:503. doi: 10.3389/fpsyg.2013.00503

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

McClelland, J. L., Mirman, D., Bolger, D. J., and Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science, 38*, 1139–1189. doi: 10.1111/cogs.12146

McClelland, J. L., Rumelhart, D. E., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* , : MIT Press.

McLennan, C. T. & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 306-321.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology, 60*, 1-39.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes, 10*, 309–331.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance, 25*(5), 1363–1389. https://doi.org/10.1037/0096-1523.25.5.1363

McQueen, J.M., Jesse, A. &Norris, D. (2009) *No lexical-prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? Journal of Memory and Language, 61(1)*, 1-18.

Menczer, F., Fortunato, S., & Davis, C. A. (2020). *A First Course in Network Science*. Cambridge, UK: Cambridge University Press.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343, 1006–1010.

Milgram, S. (1967). The small world problem. *Psychology Today, 2*, 60–67.

Miller, J. L., & Baer, T. Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America, 73*, 1751-1755.

Mirman, D., McClelland, J. L., & Holt, L. L. (2005). Computational and behavioral investigations of lexically induced delays in phoneme recognition. *Journal of Memory & Language, 52*(3), 424-443.

Mirman, D., McClelland, J.L., & Holt, L.L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review*, 13(6), 958-965.

Mirman, D., McClelland, J.L., Holt, L.L., & Magnuson, J.S. (2008). Effects of attention on the strength of lexical influences on speech perception: Behavioral experiments and computational mechanisms. *Cognitive Science, 32*, 398-417.

Mirman, D., Yee, E., Blumstein, S., & Magnuson, J.S. (2011). Theories of spoken word recognition deficits in aphasia: Evidence from eye-tracking and computational modeling. *Brain & Language, 117*, 53-68.

Morrisette, M.L. & Gierut, J.A. (2002). Lexical organization and phonological change in treatment. *Journal of Speech, Language and Hearing Research, 45,* 143-159.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*, 365-378.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Noe, C., & Fischer-Baum, S. (2020). Early lexical influences on sublexical processing in speech perception: Evidence from electrophysiology. *Cognition*, *197*, 104162.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*, 189-234.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Norris, D., McQueen, J, Cutler, A. (2018) *Commentary on "Interaction in spoken word recognition models: feedback helps*, *Frontiers in Psychology – Cognitive Science*, 9:1568.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1209-1228.

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.

Norris, D., McQueen, J. M., and Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language & Cognitive Neuroscience, 31*, 4–18. doi: 10.1080/23273798.2015.1081703

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34(3)*, 191–243.

Nusbaum, H.C., & Morin,T.M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, production, and linguistic structure* (pp. 113-134). Tokyo: Ohmsha Publishing.

Peelle, J.E. & Davis, M.H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology, 3*:320. doi:10.3389/fpsyg.2012.00320

Pitt, M. A. & Samuel, A. G. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception & Performance, 32*, 1120-1135.

Pitt, M. A. and McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39,* 347-370.

Rubin, P., Turvey, M. T., & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics, 19*, 384-398.

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1.* MIT Press.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51-89.

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

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*(1), 28-51.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*(2), 97-127.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory & Language, 48*(2), 416-434.

Samuel, A.G., & Kraljic, T. (2009). Perceptual learning in speech perception. Attention, *Perception & Psychophysics, 71*, 1207-1218.

Sankoff, D. & Kruskall, J.B. (1983). *Time warps, sting edits, and macromolecules: The theory and practice of sequence comparison*. Addison-Wesley; London.

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America, 57*, 1030–1033.

Siew, C. S. Q. (2017). The influence of 2-hop density on spoken word recognition. *Psychonomic*

Bulletin and Review, 24, 496-502.

Sommers, M.S. & Danielson, S. E. (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging, 14*, 458-472.

Storkel, H.L., Bontempo, D.E., Aschenbrenner, A.J., Maekawa, J., & Lee, S.Y. (2013). The effect of incremental changes in phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Speech, Language, and Hearing Research, 56*, 1689-1700.

Storkel, H.L., Maekawa, J., & Hoover, J.R. (2010). Differentiating the effects of phonotactic probability and neighborhood density on vocabulary comprehension and production: A comparison of preschool children with versus without phonological delays. *Journal of Speech, Language, and Hearing Research, 53*, 933-949.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE : A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. B*ehavior Research Methods, 39,* 19-30.

Strogatz, S. (2003). *Sync: How Order Emerges from Chaos in the Universe, Nature, and Daily Life*. Hachette: New York, New York.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of con- text effects. *Journal of Verbal Learning & Verbal Behavior, 18*, 645–659.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268(5217)*, 632-634.

Toscano, J. C., Anderson, N. D., & McMurray, B. (2013). Reconsidering the role of temporal order in spoken word recognition. *Psychonomic Bulletin & Review*, *20*(5), 981-987.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*(10), 1532-1540.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech Language Hearing Research*, *51*, 408–422.

Vitevitch, M.S. and Luce, P.A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language, 40*, 374-408.

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), *Eye Movements and Their Role in Visual and Cognitive Processes. Reviews of Oculomotor Research V4*. Amsterdam: Elsevier.

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

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

White, L., Mattys, S.L., & Wiget, L. (2012). Segmentation cues in conversational speech: Robust semantics and fragile phonotactics. *Frontiers, 3*, 1-9.