

FROM SOUND TO SENSE AND BACK AGAIN: THE INTEGRATION OF LEXICAL AND SPEECH PROCESSES

David W. Gow Jr.^{1,2} & Bob McMurray³

¹ Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

² Department of Psychology, Salem State College, Salem, MA, USA

³ Department of Brain and Cognitive Sciences, University of Rochester,
Rochester, NY, USA

gow@helix.mgh.harvard.edu

ABSTRACT

The path from sound to sense crosses several disciplinary boundaries. Unfortunately, compelling early demonstrations of categorical perception helped to create a historical wedge between speech and word recognition. This talk highlights some advantages of the re-integration of these fields. First, task-dependencies found in studies of speech perception suggest lexical processes as a more appropriate locus of study. Examination of sensitivity to within-category VOT variation reveals that while such variation is perceived categorically in explicitly metalinguistic tasks, it leads to continuous sensitivity in an implicit measure of lexical activation. Similarly, work on place assimilation shows that the same segment may receive different interpretations in metalinguistic offline tasks and online tasks implicitly reflecting automatic word recognition. To the extent that metalinguistic and lexical tasks produce different interpretations of stimuli, tasks stressing lexical activation are closer to core processing phenomena. Second, we argue that work on word recognition minimizes the importance of phonetic detail at its own peril. A series of studies using paradigms including head-mounted eyetracking techniques show that listeners rely on subphonemic detail to recognize words that have undergone lawful assimilation. Moreover, the same detail facilitates the recognition of neighboring items that drive the assimilation. Thus, subphonemic variability may serve as a processing asset, rather than an obstacle to processing. We suggest a consolidation of research paradigms. Speech processes responding to subsegmental acoustic detail both facilitate word recognition and resolve ambiguity. At the same time lexical dynamics are exquisitely sensitive to such detail and provide a more accurate window on speech processing.

INTRODUCTION

There is a curious, and potentially damaging, distinction between the study of speech and spoken word recognition. Students of speech and spoken word recognition typically inhabit different departments, publish in different journals, and attend different conferences. Were this purely a matter of organization, little would be lost. However, this division has fundamentally shaped views of what the important computational problems are, what mechanisms are available to solve these problems, and what methodologies are appropriate to understand them. We reexamine the assumptions that promoted this distinction and outline an approach based on new findings about the relationship between speech and spoken word recognition.

The speech/spoken word recognition distinction, as well as many of the assumptions, methods, and research goals of the two areas, is largely attributable to early work suggesting the categorical perception (CP) of speech sounds. CP was demonstrated using metalinguistic tasks

that showed that listeners appear to impose relatively sharp phonetic boundaries over continuous acoustic transformations, and that within-category discrimination is much poorer than between-category discrimination (Liberman et al, 1957; see Repp, 1984 for a review). This finding cemented the intuition of users of alphabetic orthographies that the phoneme, or at least the features that define phonemes, exist as discrete, abstract units, insulated from the notorious variability of the speech signal. Moreover, they provided organizing principles for the disciplines. Speech became the problem of mapping from signal to features or phonemes, and word recognition became an independent process of mapping from these idealized representations to lexical representations. This division sanctioned the independence of the two fields, and also the independence of phonetics and phonology. Moreover this division (and its assumptions) exists despite a long-standing literature that demonstrates the limitations of CP and undermines the fundamental assumptions that strict CP is an obligatory, necessary, or even desirable component of spoken language processing.

Surveying the field and drawing on our own work, we see the emergence of a new way of thinking about the perception of within-category variance and its roles in the perception of spoken language. This approach stresses the integration of work across traditional disciplinary boundaries, and promises to consolidate and extend the great progress in speech acoustics, phonetics, phonology, and spoken word recognition that the last 50 years have seen. Our case is outlined in four parts. First we argue that CP has fundamentally defined the implicit dominant paradigms in acoustic phonetics, spoken word recognition and phonology. Second, we present evidence that challenges the relevance of CP to natural processes of speech and lexical processing. Given this background we review some of our work to illustrate the uses of gradient phonetic and lexical activation in spoken language perception. Finally we propose a framework for new thinking on the problems of speech perception and spoken word recognition.

CATEGORICAL PERCEPTION AND THE STANDARD PARADIGM

CP suggested a discrete processing step in the speech chain that allowed speech, spoken word recognition and phonology to be treated as largely independent of one another. Over time, this organization crystallized into what we will refer to as the Standard Paradigm. Within disciplines, the Standard Paradigm is constrained by research methods and what Marr (1982) referred to as the computational problem at the heart of each field—the mapping between input and output. CP enabled a recasting of the broader question of this mapping from sound to sense, into a series of smaller, independent discipline-specific mappings.

In *phonetics and speech perception*, the mapping of interest became the one between the highly variable speech signal and abstract, discrete phonetic categories and segments. This framework invited a focus on sources of invariance in the signal that might facilitate this mapping, and how context might modify it. The success of CP studies also established explicit metalinguistic tasks like phoneme identification and discrimination as dominant experimental paradigms for evaluating listeners' interpretation of speech. This methodological framework implicitly minimized the role of phonological and lexical factors in speech perception. Over time though, this view has become less tenable given evidence that both of these factors affect speech perception (c.f. Ganong, 1980; Warren, 1970; Samuel & Pitt, 2003; Magnuson, et al., 2003; Massaro & Cohen, 1983; Dupoux et al., 1999; Halle et al., 1998).

For *phonology*, the central question became one of understanding the processes governing the distribution and dynamic ecology of these idealized featural representations. CP provided empirical justification for the examination of the behavior of these abstract units. It also suggested that phonology could proceed without significant contact with phonetics or the study of spoken word recognition. However, the advent of laboratory phonology changed this situation, by introducing a research strategy for integrating phonetic/perceptual constraints into phonological theory (c.f. Hayes, 1999; Steriade, 2000). In addition, Optimality Theory (Prince & Smolensky, 1993) suggested that phonology could work on the basis of multiple graded constraint satisfaction, an approach that does not require a symbolic substrate in the way that generative approaches did.

Finally, for psychologists studying *spoken word recognition*, CP framed the word recognition problem as one of mapping between idealized featural or segmental representations and equally abstract lexical representations. This approach placed word recognition in a cognitive domain removed from general perceptual processes and the acoustic signal. Within this framework, subphonemic variation became largely a problem of concern only to phoneticians, with models of spoken word recognition paying scant attention to the signal. The exceptions to this generalization are cases in which a particular stimulus property were shown to affect specific “higher level” processes. Lexical segmentation provides an example of this approach. When non-segmental signal properties (like systematic allophonic and duration variation) was discovered to play a role in segmentation, CP-constrained models could not handle these continuous properties. The result was that these cues were viewed as extra-segmental cues that trigger discrete processes and placed in simple models in which the detection of such a cue triggered the activation of a discrete segmentation process (c.f.; Nakatani & Dukes, 1977; Cutler & Norris, 1988). Models of this sort retain primacy of segmental input to lexical processing. However, as analogous results are found for other speech phenomenon, this approach will ultimately cast word recognition as a hodge-podge of extra-segmental processes layered on an increasingly obsolete segmental lexicon.

From a historical standpoint, CP has not fared well as a theory of speech perception. A case in point is Fry et al.’s finding (1962) that vowel perception is less categorical than consonant perception (also see Healy & Repp, 1982). This finding calls into question the notion that categorical perception is precondition for stable, robust word recognition. Moreover, the central claim, that listeners are insensitive to within category variance, is not supported when experimental paradigms either use an implicit performance measure such as reaction time or speeded identification (Pisoni & Tash, 1974; Carney et al., 1977); assess discrimination at auditory rather than linguistic (or metalinguistic) modes of processing (Pisoni & Lazarus, 1974); or explicitly address how well individual stimuli exemplify a category (Allen & Miller, 2001; Miller, 2001; Massaro & Cohen, 1983b). As a result, few researchers today acknowledge strict CP as necessary, or even involved in speech perception. However, the standard paradigm instigated by CP has not been significantly altered. A large proportion of empirical work is still based on the meta-linguistic tasks that were inspired by CP (and which may be its locus). Speech perception, phonology, and word recognition are still studied as modular independent units.

CP continues to shape the Standard Paradigm because studies refuting CP have placed a narrow focus on *perception*, to the exclusion of the broader question of how within-category

variability might influence language *understanding*. This narrow focus is understandable given the historical separation between the speech and spoken word recognition.

VARIABILITY FROM SIGNAL TO WORD

Given the role that CP has played as the implicit cornerstone of the Standard Paradigm, we now turn to limitations of CP and the theoretical perspective it inspires. CP has only been demonstrated with metalinguistic tasks that assume the necessity or accessibility of the phoneme. But consider the minimal computational problem presented by speech communication as a whole: how do listeners derive meaning from sound? Since listeners seem to perform this task in a manner that is rapid, spontaneous, and automatic, the most meaningful and defensible task would be one that asked listeners to spontaneously and automatically interpret a meaningful utterance in real time. This rules out metalinguistic tasks such as phoneme identification or discrimination. However, meaning-based tasks could reveal or rule out CP by demonstrating systematic effects of within-category variation on *interpretative* processes and conclusively answer the question of whether CP is fundamental to recognition.

Experiments using the visual world paradigm (Cooper, 1974; Tanenhaus et al., 1995) provide one example of such a task by combining a referential visual context and task with exquisite sensitivity to probabilistic interpretation of utterances as they unfold. In this paradigm, subjects' eye movements are monitored while they receive verbal instructions to select or manipulate one of the four objects on a computer screen or table. With careful control of the verbal instructions and the set of visual competitor objects, eye-movements can reveal unfolding representations at many levels of linguistic processing. For example, Allopenna, et al., (1998), presented subjects with screens containing a target (e.g. beaker), a cohort competitor (e.g. beetle), a rhyme competitor (e.g. speaker) and an unrelated item (e.g. carrot). While selecting the target, subjects made significantly more eye-movements to cohort and rhyme competitors than unrelated items, replicating lexical competitor effects in a referential, natural task. Importantly, fixation probabilities over time reflected the temporal similarity of competitors to the target, and were well fit by activation functions from the TRACE model of word recognition (McClelland & Elman, 1986) when coupled with a simple linking hypothesis. Subsequent work has shown that this measure replicates a range of classic findings in spoken word recognition and yields a picture of lexical activation unfolding over time. This work has shown sensitivity to lexical frequency (Dahan et al., 2001), lexical neighborhood (Magnuson et al., 2003b) and mismatching coarticulatory information (Dahan et al., 2001b). Thus, these methods meet our standards of providing a rapid, spontaneous and automatic measure of subjects' understanding of simple verbal instructions, while providing a detailed picture of lexical processing dynamics

McMurray et al. (2002) used this technique to determine whether within category phonetic variation affects higher-level language processes, specifically, lexical access. In this experiment, each trial contained a visual display with a target/competitor pair from a b/p (Voice Onset Time) continuum (e.g. beach/peach) and two unrelated objects (e.g. lamp and ship). Subjects were instructed to select the items on the screen that matched a single auditory target item. Similar to previous studies of categorical perception (e.g. Liberman et al., 1961; Pisoni & Tash, 1974) the auditory targets were synthesized from a nine-step a VOT continuum, with VOTs ranging from 0-40 ms.

Results indicated that even when subjects' ultimate responses (mouse-clicks) were factored out, subjects were increasingly more likely to fixate the competitor object (e.g. the peach after hearing *beach*) as the VOT approached the category boundary. This same linear trend was found when tokens adjacent to the category boundary were excluded. These results provide strong evidence that within-category VOT variation affects the online interpretation of language, specifically, the activation among lexical competitors during word recognition.

In a subsequent experiment, McMurray et al. (in preparation) used the same stimuli in a purely metalinguistic task in which subjects were asked to identify the initial consonant of the tokens by clicking on buttons containing "b", "p", "l" or "sh". Again, eye-movements were monitored during this task, with subtly different results. While there were effects of subphonemic detail on competitor fixations when the stimuli that abutted the category boundaries were included, these effects largely disappeared when these partially ambiguous tokens were excluded.

These experiments support two conclusions. First, they show that demonstrations of CP are task dependent. CP was found in the metalinguistic task, but not in eye movement data from the picture selection task. The fact that CP is shown in eye movement data for the metalinguistic task demonstrates that eye movements can in principle reveal CP, and so the findings from the picture selection task cannot be attributed to limitations of the measure. Instead, they suggest that CP is a task artifact, rather than an obligatory component of spoken word recognition.

The other conclusion to be drawn from this work is that phonetic variability produces variability in lexical activation. This conclusion is consistent with findings from other paradigms including lexical priming (Andruski et al., 1994; Utman et al., 2000). While this could be modeled using the "extra-segmental" approach applied to word segmentation, the fact the VOT and voicing are universally considered core segmental phenomena rules out this approach. Considered with respect to previously mentioned evidence for lexical and phonological effects on the speech categorization, these results suggest significant, direct, two-way connections between raw acoustic/phonetic variability and lexical activation.

THE STANDARD PARADIGM REDUX: USES OF SUBPHONEMIC VARIABILITY

In this section we illustrate the beginnings of an alternative approach by considering the possible implications of the relationship between phonetic variation and lexical activation using several examples drawn from our own work. Our premise is that by discarding the requirements of a discrete output to speech perception and input to spoken word recognition three theoretical advantages arise. First, subphonemic variation can be harnessed to improve word recognition by taking advantage of systematic, subsegmental covariation in signal. Second, higher-level, lexical processes are now available to solve problems in speech perception. Finally, lower-level perceptual processes (e.g. perceptual grouping principles) are now available to solve problems in spoken word recognition, such as the resolution of lexical ambiguity.

These advantages of our framework will now be illustrated in three domains. We will discuss problems of temporal integration and rate normalization and show how lexical activation coupled with sensitivity to subphonemic detail provides a novel approach to the problem. Work on word segmentation reveals that if standard lexical processes are augmented by subphonemic detail and perceptual enhancement at word boundaries, a unitary account can explain segmentation.

Finally we will discuss the role of perceptual grouping processes in coping with lawful variability and lexical ambiguity created by English place assimilation.

Temporal Integration and Rate Normalization

Consider the McMurray et al. (2002) results showing that listeners display within-category sensitivity to VOT variation. Within the Standard Paradigm such variation is treated as noise, or an obstacle to language understanding. We suggest, however, that continuous variation and graded activation may be meaningful and useful. Listeners exploit systematic variation or more appropriately, **covariation**, to guide and refine word recognition (McMurray et al, 2003).

This hypothesis is based on a large number of studies in laboratory phonology that demonstrate that acoustic variation (such as VOT) is not random. Rather, continuous cues like VOT *covary* with other aspects of its phonetic environment that could be exploited to facilitate language perception. For example, evidence from Fougeron and Keating (1997) suggests that VOT and other articulatory factors covary with prosodic strength. VOT is also strongly correlated with speaking rate and vowel length (e.g. Kessinger & Blumstein, 1998) as well as upcoming phonetic context in languages such as Hungarian, in which voicing assimilates across word or syllable boundaries (Gow & Im, in press). Significantly, in each case information in one part of the signal provides information that may facilitate the interpretation of other parts. While VOT remains an important cue for research in speech perception (both methodologically and theoretically), this broad approach speech extends to many (if not all) continuous acoustic cues. That is, covariation exists in many phonetic domains. Examples include anticipatory nasalization and r-coloration in vowels (Lahiri & Marslen-Wilson, 1991; Clark & Hillenbrand, 2003), and subphonemic remnants of vowel deletion (Manuel, 1991).

While currently limited by the failure to instantiate many relevant characteristics of the speech signal, activation-based computational models of word recognition provide a valuable framework for understanding how phonetic detail might influence the dynamics of lexical activation. The systematic sensitivity of lexical activation functions to within-category detail, coupled with their broader temporal domain, suggests the outline of a potential mechanism for integrating covarying cues into stable percepts. Activation-based models of spoken word-recognition all share the assumption that activation at all levels of representation decays over time. How quickly activation of a candidate falls below threshold then is partially determined by peak activation prior to decay and continuing match to the input. At a featural level, for example, activation functions related to VOT variability could guide the integration and interpretation of speech cues. Thus, higher-level cognitive processes (the growth and decay of activation of lexical items) may in fact provide the necessary representation for temporal integration.

This hypothesis is beginning to be tested empirically. McMurray et al. (2004) used an eye-tracking task to examine rate normalization in the context of word initial /b/-/w/ (formant transition slope) continua with varying vowel lengths. Like VOT, formant transition slope is a temporal cue sensitive to vowel length or rate effects. They examined the temporal pattern of eye-movements do determine when formant transition slope and vowel length effects could be seen. If activation is dependent on a purely sublexical integration mechanism that modularly outputs its results to lexical processing, effects of vowel length should be simultaneous with formant transition slope. An activation-based account, on the other hand would be supported by immediate integration of cues *as they arrive*.

Results support activation as an integration mechanism. Lexical activation was affected by formant transition slope *before* vowel length. That is, lexical activation reflected very early effects of formant transition slope, representing this information for later integration with vowel length. Lexical activation provides both a useful mechanism for understanding temporal integration, and in some contexts may play an important computational role as a meaningful unit of abstraction over large time windows. Moreover, this sort of integration could not occur if lexical dynamics were not sensitive to within-category detail. Thus, moving beyond CP and the standard paradigm may significantly extend our understanding of these processes.

Segmentation: Activation, Competition and Perceptual Enhancement

Gow and Gordon's (1995) work on lexical segmentation provides another example of how phonetic variation might be exploited to facilitate language perception. This work probed the relationship between lexical and prelexical factors in word segmentation in connected speech. The TRACE model (McClelland & Elman, 1986) demonstrated the utility of treating lexical segmentation as an emergent process arising from word recognition (rather than as a separate, extrasegmental mechanism). This is an appealing mechanism because it replaces two processes, lexical access and segmentation, with a single process. Unfortunately, the high frequency of lexical embedding (e.g. *ham* in *hamster*) (Cutler et al., 1994) means that this approach would lead to an overgeneration of lexical candidates by embedded words. This suggests that segmentation by access is only viable if an additional mechanism rejects spurious access of unintended words TRACE limits overgeneration through a competition mechanism that is sensitive to the strength of bottom-up activation, which is implicitly modulated by word length. Unfortunately, this mechanism is of limited use, as speakers may intend to produce short (e.g. *car go*) and longer words (*cargo*) at different times. Moreover, there is evidence that segmentation may be influenced by a variety of prosodic and subphonemic factors including metrical pattern, the duration of onset segments, the degree of aspiration of voiceless segments, glottalization of initial vowels and systematic allophonic variation (see Gow & Gordon, 1995 for a review). Significantly, these cues are correlated with word boundaries, suggesting a complement to recognition-driven segmentation.

In a series of cross-modal priming experiments, Gow and Gordon examined the online interpretation of lexically ambiguous items such as *cargo/car go* in sentential contexts in which either interpretation was contextually viable. This work showed that listeners who hear the sentence *She saw the car go around the corner* access both CARGO and GO at the offset of the ambiguity. This pattern of parallel activation is consistent with the notion that segmentation is a byproduct of word recognition. If segmentation were a prerequisite for word recognition it would follow that listeners would only access an interpretation consistent with one segmentation of the string. Listeners showed a different pattern of segmentation when *cargo* was pronounced as a single word rather than two. Listeners who hear the same sentence with *cargo* replacing *car go* access CARGO, but show a strong trend towards inhibiting GO. The only difference between the two conditions was the pronunciation of the prime. Post-hoc analyses revealed that onset segment were longer than comparable non-onset segments. These within-category phonetic effects, along with those associated with other putative segmentation cues, are part of a larger pattern of covariation between phonetic variables and position within a word.

Gow and Gordon note that all known acoustic segmentation cues have the common effect of enhancing the perception of word-initial features. Gow and Gordon's Good Start Model (1995)

proposes that the perceptual enhancement of word onsets leads to strong early activation of intended lexical candidates. This activation makes them stronger competitors, which enables them to inhibit unintended lexical candidates and limit the overgeneration of lexical candidates in a recognition-driven segmentation scheme. Moreover, embedded words (*go* in *cargo*) lack these enhancing cues (by virtue of being non word-initial) further limiting their ability to compete.

This interpretation is supported by recent eye-tracking results (Salverda et al, 2003) showing that competition between embedded and embedding words is modulated by subphonemic variation related to perceptual strengthening and weakening. This approach to thinking about lexical access and segmentation suggests another way in which phonetic variability and ensuing variability in the strength of lexical activation may facilitate language perception. Moreover, it was the combination of cognitive processes (lexical activation and competition) with perceptual ones (sensitivity to fine-grained detail) that provided the solution.

Place Assimilation: Perceptual Grouping and Lexical Activation

Our last example relates to the perception of lawfully modified speech. In English coronal place assimilation, coronal segments approximate the place of articulation of a subsequent non-coronal. For example, after labial assimilation, *right berries* may resemble *ripe berries*, and *green boat* may contain the nonword *greem*. Within the standard paradigm, many treatments of the phenomenon view this process as one of discrete feature substitution. Thus, the coronal /t/ in *right* may be viewed as becoming a labial [p] when followed by the labial /b/ in the phrase *right berries*. Thus, such neutralization would have few implications for speech processing, but significant implications for the problem of word recognition. Within this framework Gaskell and Marslen-Wilson (1998; 2001) suggested that listeners might infer the underlying form of such segments. Knowledge that a coronal becomes a labial before another labial may drive an inference that in any labial-labial sequence, the first labial is actually a coronal. This is at best a limited inference, because in some cases it would lead a listener to misperceive a labial. Such an inference would lead to the access of *right* when *ripe* was intended, and may transform valid words such as *up* into nonwords (*ut*). However, evidence from several studies suggests that listeners do not rely on this type of phonological inference to recover the underlying place of assimilated segments (Gaskell & Marslen-Wilson, 2001; Gow, 2003).

Gow (2001) has now shown evidence that these modifications are, in fact, subphonemic. Acoustic measures, including measures of F2 immediately prior to closure, show that assimilated segments have physical characteristics intermediate between coronal and noncoronal segments—a subphonemic modification. Gow (2001; 2002; 2003; Gow & Im, in press) has demonstrated that listeners rely on a combination of subphonemic acoustic variation in assimilated segments, and post-assimilation context to resolve lexical ambiguity. Perceptual studies reinforce this view (Gow, 2003; Gow & Hussami, 1999). Behavioral results also show that assimilation actually enhances the perception of post-assimilation context (Gow, 2001; 2003)—subphonemic variation in the assimilated segment improves recognition of the context. These results contradict the intuition rooted in the Standard Paradigm, that within-category variation potentially undermines recognition.

In a series of ongoing experiments using the head-mounted eye-tracking paradigm described above, we are exploring the role of progressive and regressive context effects in the processing of assimilated speech. In one experiment, subjects viewed four pictures depicting a maroon

goose, a maroon duck, a patriotic duck (a duck with a flag), and a patriotic goose. In a given trial, eye movements were tracked while subjects were instructed to use a computer mouse to *select the maroon goose*. Listeners heard one of two recorded and cross-spliced versions of the instruction. In one version the word *maroon* was initially pronounced in a coronal context (*maroon duck*) leaving the final coronal unmodified. In the other version *maroon* was initially pronounced in a velar context to produce a natural velar assimilation of the coronal.

Listeners showed earlier looks to the maroon goose after the assimilated instruction than after the unassimilated instruction. This effect becomes apparent roughly 100 milliseconds after the offset of the assimilated item. Given that it takes approximately 200 ms to plan and launch an eye-movement, the effect seems to arise sometime during the end of the assimilated item. This result is consistent with prior evidence that subphonemic detail is useful when considered with regards to the regular covariation created by assimilation (Gow, 2001; 2003).

A second experiment uses a similar paradigm to examine regressive context effects. In this experiment assimilatory contexts produced lexical ambiguity. For example, coronal to labial assimilation may cause the coronal /t/ in *eight* to approximate a /p/ in the phrase *eight babies*. This created potential ambiguity between the *eight* and *ape* in the phrase *eight_p (assim) babies*. Subjects' eye movements were tracked while they were instructed to use a mouse click to select an image from a matrix showing pictures depicting eight babies, ape babies, eight dolls and ape dolls. Using a similar combination of assimilated and unassimilated tokens of *eight* and *ape*, we were able to examine evidence for concurrent progressive and regressive context effects.

The results of this study show a pattern of same-trial regressive and progressive context effects that is consistent with earlier studies that examined these effects separately (Gow, 2001; 2002; 2003). Subjects who heard labialized items in labial contexts (*eight_p babies*) showed more looks to pictures depicting *eight* than *ape*. Conversely, subjects who heard the same labialized segments in coronal contexts (*eight_p dolls*) showed more looks to pictures showing *ape* than *eight*. These regressive context effects accompanied progressive effects analogous to those found in the previous experiment. Subjects showed earlier looks to pictures of babies after hearing the labialized tokens (*eight_p babies*) than they did after hearing unmodified coronal tokens (*eight babies*). These effects only appeared when assimilation was present but incomplete. Thus, these perceptual processes rely on subphonemic information.

Gow (2003) proposes a novel interpretation of these results, arguing that they are the result of a fundamental *perceptual* process in spoken language perception, feature cue parsing. In general, individual phonetic features are encoded in multiple characteristics of the speech signal that are distributed over time. Studies of cue trading show that in such cases, multiple distributed cues are integrated to make phonetic distinctions (c.f. Summerfield & Haggard, 1977). The feature cue parsing account of these is based on the inference that there must be a grouping process that determines *which* cues to integrate. One such basic perceptual principle is that perceptually salient features attract weaker similar features (Bregman, 1990).

Consider the interpretation of the labialized /t/ ([^t_p]) in the phrase *eight babies*. Assimilation produces a pattern of feature cues that might be viewed as simultaneous weak evidence of labial *and* coronal place. Thus, strong evidence of word-initial labiality from *babies* should attract weaker evidence of labiality from the assimilated item. This would have two effects. By

pulling away evidence of assimilated labiality it would disambiguate the modified *eight*, since only evidence of coronality (consistent with *eight*) would be left word-finally. At the same time, associating assimilated labiality from the offset of *eight* with the onset of *babies* would increase bottom-up support for *babies*. Thus, one act of grouping would produce both progressive and regressive effects. The same mechanism also explains effects of coronal context when the labialized *eight* is followed by the coronal onset of *dolls*. In this case strong evidence of initial coronality from *dolls* attracts the residual underlying evidence of offset coronality in *eight*, leaving only evidence of assimilated offset labiality. Thus subjects hear *eight* as *ape*.

Feature cue parsing thus accounts for a complicated set of context effects related to assimilation through a fundamental grouping mechanism that is guided by the patterning of gradient cue information consistent with multiple features. Moreover, it explains these effects in the context of a process that plays a role in all spoken language processing.

IMPLICATIONS

This work underlines the limits of the Standard Paradigm by showing that listeners are sensitive to within-category phonetic variation on the way to the computation of meaning, and that this variation may often facilitate word recognition. It also suggests that the traditional boundaries between disciplines limit progress in all fields. While the argument against CP is not new, until recently, its implications for the value of integration between the disciplines has not been formally explored. The work cited here illustrates the value of understanding phonetic variation as potential signal and not noise, and of overlooking traditional disciplinary boundaries.

Just as CP has helped frame research questions and methods over the last half century, we believe that evidence for signal-dependent gradient lexical activation has implications for how we study and understand the path from sound to sense. One is methodological. Following the observations of Massaro and Cohen (1983b) and others, we believe that CP reflects task-specific decision processes that are not part of normal spoken language perception. Spoken language perception is best studied through paradigms that tap real language understanding rather than metalinguistic awareness. By tapping the minimal computational problem presented to language perceivers (deriving meaning from the speech signal), such tasks avoid the trap of imposing experimenters' preconceived notions about the existence and accessibility of intermediate representations on subject performance.

An important caveat is that this strategy is only appropriate if experiments can be formulated and results predicted that could in principle confirm or disconfirm the existence of hypothesized intermediate processes or representations. Fifty years ago there were no experimental paradigms available that met these standards. Now there are several. In this paper we have emphasized the usefulness of variants of the visual world eyetracking technique developed by Tanenhaus et al. (1995). This paradigm has a number of advantages including its sensitivity to small differences in lexical activation and its ability to track the evolution of activation of multiple lexical candidates over time with precise temporal resolution. Additionally, the ability to control context (in this case visual context) couples the experimental task to the minimal computational problem of speech. This effectively creates a situation in which the listener's interpretation of the spoken language *is approximately equal to* the processing unit being studied (in this case lexical activation). The role of this paradigm in studies of word frequency, categorical perception, lexical competition, and a variety of effects of fine phonetic detail (c.f. Dahan et al.,

2001a; Dahan et al., 2001b; McMurray et al., 2002; Magnuson et al., 2003;) clearly demonstrate the utility and potential of this paradigm. Several physiological paradigms may also meet this criterion. For example, ERP and MEG techniques may be used in the context of passive listening tasks using meaningful linguistic stimuli (see Gow and Holcomb, 2002, for an example of such techniques applied to assimilation phenomena). This is not to say that traditional tasks including phoneme classification, discrimination, and monitoring, or priming paradigms that require lexical decision or naming should be abandoned or ignored. Research using these methods continues to produce valuable data. However, these data can be understood more thoroughly given an awareness of the role of task-specific processes, and the context of data from paradigms that do not invoke such processes.

A second implication of signal-dependent gradient lexical activation is the importance of within-category phonetic variation to the process of spoken word recognition. There is a recent trend towards examining the role of acoustic fine detail in studies of word recognition (c.f. Andruski et al., 1994; Gow & Gordon, 1995; Utman et al., 2000; Spinelli et al., 2003; Dahan et al., 2001a, 2001b). This development suggests that a paradigm shift has already begun in the field. While this trend is encouraging to us, we see a lag in the development of explicit models to address it. Phonetic detail is minimal in several current prominent models of spoken word recognition including Shortlist (Norris, 1994) and MERGE (Norris et al, 2000). Nevertheless, there are signs of an evolutionary move towards the use of input representation that capture aspects of fine phonetic detail. Early TRACE simulations (McClelland & Elman, 1986) addressed aspects of phonetic fine detail, and later simulations of word level phenomenon use graded acoustic features. While Gaskell et al.'s (1995) model employed only discrete input feature representations, this was superseded by the Gaskell (2003) model, which accommodated input representations of graded, overlapping pace features. Ultimately, however, the formulation of the computational models will require not only raw sensitivity to graded features (or better yet acoustics), but some representation of the covariation between such continuous features.

Explicitly representing this continuous covariation is the domain of phonetics, phonology and laboratory phonology. While the Standard Paradigm implied discrete categories as the input to phonology and abstract rules as the goal, emerging approaches should (and often do) embrace the continuous nature of phonetic cues. Combined with ongoing efforts to ground phonology in (or develop phonologies from) perceptual and articulatory constraints (c.f. Gow & Zoll, 2002), this promises to further integrate work in speech perception, word recognition and phonology.

Integration is the hallmark of the emerging paradigm. This is characterized by increased interest in measures of lexical activation by speech researchers, interest in low-level phonetic properties by students of lexical processes, and attention paid to the role perceptual and phonetic constraints in phonological theory. In addition to producing integration across traditional disciplinary boundaries, it also has the potential to increase the integration of theoretical accounts of the relationship between speech and word recognition. We have argued that within the CP-constrained paradigm phonetic fine detail can only influence lexical processes indirectly through discrete, extra-segmental processing mechanisms. However, as the models of Gow and Gordon (1995) or Grossberg et al. (1997) suggest, activation dynamics may provide a means for directly integrating fine phonetic factors into accounts of lexical processes without positing additional processes. This approach to conceptualizing spoken language perception

may lead to more parsimonious explanations of a variety of phenomena including rate normalization, and the use of prosody to facilitate syntactic parsing and indicate semantic focus.

In conclusion, research in spoken language perception is in the midst of a major paradigm shift. Evidence for the limits of categorical perception and the role of within-category phonetic variation in lexical processes is prompting a reconsideration of the way spoken language perception is studied and understood. The emerging paradigm is characterized by a blurring of disciplinary boundaries, the ascendancy of experimental paradigms that limit the role of metalinguistic processes, renewed focus on the role of phonetic factors in word recognition, and the development of models that address disparate processing phenomena through the unified perspective of lexical activation dynamics.

ACKNOWLEDGEMENTS

This work was supported by NIH grants R29DC03108 and 2R01DC3108 to the Massachusetts General Hospital (David Gow, PI), grant F31DC006537 to the University of Rochester (Bob McMurray, PI) and grant DC-005071 (Michael Tanenhaus & Richard Aslin, PI's). We would like to thank Stefanie-Shattuck-Hufnagel, Michael Tanenhaus, Michael Spivey, and Joyce McDonough for their encouragement and constructive comments on earlier drafts of this paper.

REFERENCES

- Allen, J.S. and Miller, J.L. (2001) Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate, *Perception & Psychophysics*, 63(5), 798-810.
- Allopenna, P.D., Magnusen, J.S., & Tanenhaus, M.K. (1998). Tracking the timecourse of spoken word recognition: Evidence for continuous mapping models, *Journal of Memory and Language*, 38, 798-810.
- Andruski, J.E., Blumstein, S.E. and Burton, M.W. (1994) The effect of subphonetic differences on lexical access, *Cognition*, 52, 163-187.
- Bregman, A. S. (1990) *Auditory Scene Analysis*, Cambridge, MA: MIT Press.
- Carney, A.E., Widin, G.P. and Viemeister, N.F. (1977) Non categorical perception of stop consonants differing in VOT, *Journal of the Acoustical Society of America*, 62, 961-970.
- Clark, M.J. and Hillenbrand, J.M. (2003) Quality Of American English Front Vowels Before /R/, *Journal of the International Phonetic Association*, 33(1), 1-16.
- Cooper, R. (1974). The control of eye fixation by the meaning of spoken language, *Cognitive Psychology*, 6, 84-107.
- Cutler, A. and Norris, D. (1988) The role of strong syllables for segmentation in lexical access, *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 113-121.
- Cutler, A., McQueen, J., Baayen, H. and Drexler, H. (1994). Words within words in a real-speech corpus. *Proceedings of the 5th Australian International Conference on Speech Science and Technology, Vol. 1*, pp. 362--367. Perth: Australian Speech Science and Technology Association.
- Dahan, D., Magnuson, J. and Tanenhaus, M. (2001) Time course of frequency effects in spoken-word recognition: Evidence from eye movements, *Cognitive Psychology*, 42, 317-367.

- Dahan, D., Magnuson, J., Tanenhaus, M., & Hogan, E. (2001b). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition, *Language and Cognitive Processes*, 16, 507-534.
- Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. (1999) Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception & Performance*, 25(6), 1568-1578.
- Fougeron, C. and Keating, P. (1997) Articulatory strengthening at edges of prosodic domains, *Journal of the Acoustical Society of America*, 101, 36728-3740.
- Fry, D.B., Abramson, A.S., Eimas, P.D. and Liberman, A.M. (1962) The identification and discrimination of synthetic vowels, *Language and Speech*, 5, 171-189.
- Ganong, W.F. (1980) Phonetic Categorization in Auditory Word Recognition, *Journal of Experimental Psychology: Human Perception and Performance*, 6(1), 110-125.
- Gaskell, M. (2003). Modelling regressive and progressive effects of assimilation in speech perception, *Journal of Phonetics*, 31, 447-463.
- Gaskell, M., Hare, M., and Marslen-Wilson, W. (1995) A connectionist model of phonological representation in speech perception, *Cognitive Science*, 19, 407-439.
- Gaskell, M., & Marslen-Wilson, W. (1997) Integrating form and meaning: A distributed model of speech perception, *Language and Cognitive Processes*, 12, 613-656.
- Gaskell, M., & Marslen-Wilson, W. (1998) Phonological variation and inference in lexical access, *Journal of Experimental Psychology: Human Perception and Performance*, 22, 144-158.
- Gaskell, M., & Marslen-Wilson, W. (2001) Lexical ambiguity resolution and spoken word recognition: Bridging the gap, *Journal of Memory and Language*, 44, 325-349.
- Gow, D. (2001) Assimilation and Anticipation in continuous spoken word recognition, *Journal of Memory and Language*, 45, 133-139.
- Gow, D. (2002) Does English Coronal Place Assimilation Create Lexical Ambiguity? *Journal of Experimental Psychology: Human Perception and Performance*, 28(1), 163-179.
- Gow, D. (2003) Feature parsing: Feature cue mapping in spoken word recognition, *Perception & Psychophysics*, 65(4), 575-590.
- Gow, D. and Gordon, P. (1995) Lexical and prelexical influences on word segmentation: Evidence from priming, *Journal of Experimental Psychology: Human Perception and Performance*, 21(2), 344-359.
- Gow, D.W. and Holcomb, P. (November, 2002). Electrophysiological correlates of place assimilation context effects. Paper presented at the meeting of the Psychonomic Society, Society, Kansas City, MO.
- Gow, D.W., and Hussami, P. (November, 1999). Acoustic modification in English place assimilation. Paper presented at the meeting of the Acoustical Society of America, Columbus, OH.
- Gow, D.W., & Im, A. (in press). A cross-linguistic examination of assimilation context effects. *Journal of Memory and Language*.
- Gow, D.W., and Zoll, C. (2002) The role of feature parsing in speech processing and phonology, *MIT Working Papers in Linguistics*, 42, 55-68.

- Grossberg, S., Boardman, I., and Cohen, M. (1997) Neural dynamics of variable-rate speech categorization, *Journal of Experimental Psychology: Human Perception and Performance*, 23, 481-503.
- Halle, P., Segui, J., Frauenfelder, U., and Meunier, C. (1998) Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception & Performance*, 24(2), 592-608.
- Hayes, B. (1999) Phonetically-driven phonology: The role of optimality theory and inductive reasoning. In *Functionalism and Formalism in Linguistics, vol 1: General Papers* M. Darnell et al., Eds, John Benjamins, Amsterdam 243-285.
- Healy, A. and Repp, B. (1982) Context independence and phonetic mediation in categorical perception, *Journal of Experimental Psychology: Human Perception and Performance*, 8(1), 68-80.
- Kessinger, R. and Blumstein, S. (1998) Effects of speaking rate on voice onset time and vowel production: some implications for perception studies, *Journal of Phonetics*, 26, 117-128.
- Lahiri, A. and Marslen-Wilson, W. (1991) The mental representation of lexical form: A phonological approach to the recognition lexicon, *Cognition*, 38(3), 245-294.
- Lieberman, A., Harris, K., & Hoffman, H. and Griffith, B. (1957) The discrimination of speech sounds within and across phoneme categories, *Journal of Experimental Psychology*, 54, 358-368.
- Lieberman, A.M., Harris, K.S., Kinney, J. and Lane, H. (1961) The discrimination of relative onset-time of the components of certain speech and non-speech patterns, *Journal of Experimental Psychology*, 61, 379.
- Magnuson, J., McMurray, B., Tanenhaus, M., and Aslin, R. (2003) Lexical effects on compensation for coarticulation: the ghost of Christmash past, *Cognitive Science*, 27, 285-298.
- Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (2003b). The microstructure of spoken word recognition: Studies with artificial lexicons, *Journal of Experimental Psychology: General*, 132(2), 202-227.
- Manuel, S. Y. (1991) Recovery of "deleted" schwa, *Perilius: Papers from the Symposium on Current Phonetic Research Paradigms for Speech Motor Control*, 115-118.
- Marr, D. (1982) *Vision*, San Francisco: Freeman.
- Massaro, D.W. and Cohen, M.M. (1983) Phonological context in speech perception, *Perception & Psychophysics*, 34, 338-348.
- Massaro, D.W. and Cohen, M.M. (1983b) Categorical or continuous speech perception: a new test, *Speech Communication*, 2, 15-35.
- McClelland, J. and Elman, J. (1986) The TRACE model of speech perception, *Cognitive Psychology*, 18(1), 1-86.
- McMurray, B., Aslin, R., Tanenhaus, M., Spivey, M., and Subik, D. (in preparation). Two "B" or not 2 /b/: categorical perception in lexical and nonlexical tasks.
- McMurray, B., Clayards, M., Aslin, R., and Tanenhaus, M. (May, 2004) Gradient sensitivity to acoustic detail and temporal integration of phonetic cues, Poster presented at the Acoustical Society of America, New York, NY.

- McMurray, B., Tanenhaus, M., and Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access, *Cognition*, 86(2), B33-B42.
- McMurray, B., Tanenhaus, M., Aslin, R. and Spivey, M. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access, *Journal of Psycholinguistic Research*, 32(1), 77-97.
- Miller, J.L. (2001) Mapping from acoustic signal to phonetic category: Internal category structure, context effects, and speeded categorization, *Languages and Cognitive Processes*, 16, 683-690.
- Nakatani, L.H., and Dukes, K.D. (1977) Locus of segmental cues for word juncture, *Journal of the Acoustical Society of America*, 62, 714-719.
- Norris, D. (1994) Shortlist: A connectionist model of continuous speech recognition, *Cognition*, 52(3), 189-234.
- Norris, D., McQueen, J. and Cutler, A. (2000) Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Science*, 23(3), 299-370.
- Pisoni, D. and Lazarus, J. (1974) Categorical and noncategorical modes of speech perception along the voicing continuum, *Journal of the Acoustical Society of America*, 55(2), 328-333.
- Pisoni, D.B. and Tash, J. (1974) Reaction times to comparisons within and across phonetic categories, *Perception & Psychophysics*, 15(2), 285-290.
- Prince, A. & Smolensky (1993) *Optimality Theory: Constraint interaction in generative grammar*, Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ. (234 pages). ROA 537.
- Repp, B.H. (1984) Categorical perception: Issues, methods and findings, In *Speech and Language (vol. 10): Advances in Basic Research and Practice* (edited by N. Lass), Orlando: Academic, 244-335.
- Salverda, A.P., Dahan, D. and McQueen, J. (2003) The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension, *Cognition*, 90(1), 51-89.
- Samuel, A. and Pitt, M. (2003) Lexical Activation (and other factors) can mediate compensation for coarticulation, *Journal of Memory and Language*, 48(2), 416-434.
- Spinelli, E., McQueen, J.M. and Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language*, 48, 233-254.
- Steriade, D. (2000) Directional asymmetries in place assimilation: A perceptual account, *Perception in Phonology* (E. Hume and K. Johnson, Eds.), New York: Academic.
- Summerfield, Q., and Haggard, M. (1977) On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants, *Journal of the Acoustical Society of America*, 62, 435-448.
- Tanenhaus, M.K., Spivey, M.J., Eberhard, K.M. and Sedivy, J.C. (1995) Integration of visual and linguistic information in spoken language comprehension, *Science*, 268, 1632-1634.
- Utman, J.A., Blumstein, S.E. and Burton, M.W. (2000) Effects of subphonetic and syllable structure variation on word recognition, *Perception & Psychophysics*, 62(6), 1297-1311.
- Warren, R. (1970). Perceptual restoration of missing speech sounds, *Science*, 167, 392-393.