Phonology and Phonetics

Editor
Aditi Lahiri

Lexical Representation
A Multidisciplinary Approach

edited by
Gareth Gaskell
Pienie Zwitserlood

De Gruyter Mouton
Recognizing words from speech:
The perception-action-memory loop

David Poeppel and William Idsardi

1. Conceptual preliminaries

1.1. Terminological

The failure to be sufficiently careful about terminological distinctions has resulted in some unnecessary confusion, especially when considering the neurobiological literature. For example, the term *speech perception* has unfortunately often been used interchangeably with *language comprehension*. We reserve the term *language comprehension* for the computational subroutines that occur subsequent to the initial perceptual analyses. In particular, *language comprehension* can be mediated by ear, eye, or touch. The linguistic system can be engaged by auditory input (speech), visual input (text or sign), and tactile input (Braille). In other words, the processes that underlie language comprehension build on sensorimotor input processes that appear to be, at least in part, independent. While this point may seem pedantic, the literature contains numerous reports that do not respect these distinctions and that conflate operations responsible for distinct aspects of perception and comprehension.

We focus here on *speech perception* proper, the perceptual analysis of auditory input. Importantly, further distinctions must be considered. There are at least three experimental approaches grouped under the rubric ‘speech perception,’ and because they are different in the structure of the input, the perceptual subroutines under investigation, and the putative endpoint of the computations, it is important to be cognizant of these distinctions, too.

(a) Most research on speech perception refers to experimentation on specific contrasts across *individual speech sounds*, i.e., sub-lexical/pre-lexical units of speech. Subjects may be presented with single vowels or single syllables and asked to execute particular tasks, such as discrimination or identification. In a typical study, subjects listen to consonant-vowel (CV) syllables drawn from an acoustic continuum – for example series exemplifying the /ra/-/la/ tongue-shape contrast or the /bi/-/pi/ voicing contrast – and are asked upon presentation of a single token to identify the stimulus category. This research strategy focuses on *sublexical properties...*
of speech and typically examines questions concerning the nature of categorical perception in speech (e.g., Liberman 1996), the phonemic inventory of speakers/listeners of different languages (e.g., Harnsberger 2000), perceptual magnet effects (e.g., Kuhl et al. 2007), the changes associated with first (e.g., Eimas et al. 1971) and second language learning (e.g., Flege and Hillenbrand 1986), phonotactic constraints (e.g., Dupoux et al. 1999; Kawasaki and Idsardi 2007), the role of distinctive features (e.g., Kingston 2003), and other issues productively addressed at the pre-lexical level of analysis.

This work has been immensely productive in the behavioral literature and is now prominent in the cognitive neurosciences. For example, using fMRI, several teams have examined regionally specific hemodynamic effects when subjects execute judgments on categorically varying stimuli (Blumstein, Myers, and Rissman 2005; Liebenthal et al. 2005; Raizada and Poldrack 2007). These studies aim to show that there are regions responding differentially to signals that belong to different categories, or that are speech versus non-speech. Interestingly, no simple answer has resulted from even rather similar studies, with temporal, parietal and frontal areas all implicated. Similarly, electrophysiological methods (EEG, MEG) have been used to probe the phonemic inventories of speakers of different languages. For example, Näätänen et al. (1997) were able to show subtle neurophysiological distinctions that characterize the vowel inventories of Finnish versus Estonian speakers. Kazanina, Phillips, and Idsardi (2006), discussed further below, used MEG data to illustrate how language-specific contrasts (Russian versus Korean), including allophonic distinctions, can be quantified neurophysiologically.

Despite its considerable influence, it must be acknowledged that this research program has noteworthy limitations. For example, a disproportionately large number of studies examine categorical perception as well as the notion of 'rapid temporal processing', all typically based on plosive contrasts (especially voice-onset time, VOT). While syllables with plosive onsets are admittedly fascinating in their acoustic complexity (and VOT is easily manipulated) a rich variety of other phenomena at the pre-lexical level have not been well explored. Moreover, these types of studies are 'maximally ecologically invalid': experimenters present single, sub-lexical pieces of speech in the context of experimental settings that require 'atypical' attention to particular features – and by and large engage no further linguistic processing, even designing studies with non-words so as to preclude as much as possible any interference from other linguistic levels of analysis. The results obtained are therefore in danger of masking or distorting the processes responsible for ecologically natural speech perception. Speakers/listeners do not consciously attend to sub-lexical material, and therefore the interpretation of these results, especially in the context of neurobiological findings, requires a great deal of caution, especially since task effects are known to modulate normal reactivity in dramatic ways.

b) A second line of research investigates speech perception through the lens of spoken word recognition. These studies have motivated a range of lexical access models (for instance lexical access from spectra, Klatt 1979, 1989; instantiations of the cohort model, e.g., Gaskell and Marslen-Wilson 2002; neighborhood activation model, Luce and Pisoni 1998; continuous mapping models, Allopenna et al. 1998, and others) and have yielded critical information regarding the structure of mental/neural representations of lexical material. Behavioral research has made a lot of significant contributions to our understanding and has been extensively reviewed prior to the advent of cognitive neuroscience techniques (see, for example, influential edited volumes by Marslen-Wilson 1989 and Altmann 1990). Typical experimental manipulations include lexical decision, naming, gating, and priming. Recognizing single spoken words is considerably more natural than performing unusual task demands on sub-lexical material. Some models, such as the influential TRACE model (McClelland and Elman 1986), view featural and lexical access as fully integrated, others argue for more cascaded operations.

Some important cognitive neuroscience contributions in this domain have been made by Blumstein and colleagues who have examined aspects of spoken word recognition using lesion and imaging data (e.g., Misiruki et al. 2005; Prabhakaran et al. 2006; Uzman, Blumstein, and Sullivan 2001). The data support a model in which superior temporal areas mediate acoustic-phonetic analyses, temporoparietal areas perform the mapping to phonological-lexical representations, and frontal areas (specifically the inferior frontal gyrus) play a role in resolving competition (i.e., deciding) between alternatives when listeners are confronted with noisy or underspecified input. The effect of lexical status on speech-sound categorization has been investigated extensively in the behavioral literature (typically in the context of evaluating top-down effects), and Blumstein and colleagues, using voicing continua with word or non-word endpoints, have recently extended this work using fMRI (Myers and Blumstein 2008). They demonstrate that fMRI data show dissociations between functionally 'earlier' effects in the temporal lobes (related to perceptual analyses) and putatively 'later,' downstream decision processes implicating frontal lobe structures. A behavioral task that has been used productively in studies of lexical representation is repetition priming, and Gagnepain et al. (2007) used word and non-word repetition priming to elucidate which cortical structures are specifically sensitive to the activation of lexical entries. Bilateral superior temporal
sulcus and superior temporal gyrus (STS, STG) are particularly prominent, suggesting that the mapping to lexical information occurs in cortical regions slightly more ventral than perceptual computations (and bilaterally; cf. Hickok and Poeppel 2000, 2004, 2007). Finally, subtle theoretical proposals about lexical representation have recently been tested in electrophysiological studies. Eulitz and colleagues (Friedrich, Eulitz, and Lahiri 2006), for example, have used lexical decision designs to support underspecification models of lexical representation.

c) A third way in which speech perception is examined is in the context of recognizing spoken sentences and assessing their intelligibility. In these studies, participants are presented with sentences (sometimes containing acoustic manipulations) and are asked to provide an index of intelligibility, for example by reporting key words or providing other metrics that reflect performance. Understanding spoken sentences is, naturally, a critical goal because it is the perceptual task we most want to explain — but there is a big price to pay for using this type of ecologically natural material. In using sentential stimuli, it becomes exceedingly difficult to isolate input-related perceptual processes per se (imagine teasing out effects of activation, competition, and selection à la Marslen-Wilson), because presentation of sentences necessarily entails lexical processes, syntactic processes, both lexical semantic and compositional semantic processes — and therefore engages numerous ‘top-down’ factors that demonstrably play a critical role in the overall analysis of spoken input.

Cognitive neuroscience methodologies have been used to test intelligibility at the sentence level as well. In a series of PET and fMRI studies, for example, Scott and colleagues have shown that anterior temporal lobe structures, especially anterior STS, play a privileged role in mediating intelligibility (e.g., Scott et al. 2000). Electrophysiological techniques have also been used to study sentence-level speech intelligibility, and Luo and Poeppel (2007) have argued that phase information in the cortical signal of a particular frequency, the theta band, is particularly closely related to and modulated by the acoustics of sentences.

In summary, the locution ‘speech perception’ has been used in at least three differing ways. Important attributes of the neurocognitive system underlying speech and language have been discovered using all three approaches discussed. This brief outline serves to remind the reader that it is challenging to isolate the relevant perceptual computations. Undoubtedly, we need to turn to all types of experimental approaches to obtain a full characterization. For example, to understand the nature of distinctive features for perception and representation, experimentation at the level of subphonemic, phonemic, or syllabic levels will be critical; to elucidate how words are represented and accessed, research on spoken-word recognition is essential; and it goes without saying that we cannot do without an understanding of the comprehension of spoken sentences. Here, we take speech perception to refer to a specific set of computational subroutines (discussed in more detail in section 1.3 below): speech perception comprises the set of operations that take as input continuously varying acoustic waveforms made available at the auditory periphery and that generate as output those representations (morphemic, lexical) that serve as the data structures for subsequent operations mediating comprehension. More colloquially, our view can be caricatured as the collection of operations that lead from vibrations in the periphery to abstractions in cortex (see Figure 1).

1.2. Methodological

Brain science needs gadgets, and practically every gadget usable on humans has been applied to speech and lexical access. There are two types of approaches that the consumer of the literature should know: ‘direct’ techniques using electrical or magnetic measurement devices and ‘indirect’ recording using hemodynamically based measurements as proxies for brain activity. The different methods are suited to address different kinds of questions about speech and language, and the careful alignment of research question with technique should be transparent.

The electrical and electromagnetic techniques directly measure different aspects of neuronal activity. Electrophysiological approaches applied to spoken-language recognition range from, on the one hand, very invasive studies with high spatial resolving power — single-unit recording in animals investigating the building blocks underpinning phonemic representation (Enginere et al. 2008; Rauschecker, Tian, and Hauser 1995; Schroeder et al. 2008; Steinschneider et al. 1994; Young 2008) and pre-surgical subdural grid recording in epilepsy patients (e.g., Boatman 2004; Crane et al. 2001) — to, on the other hand, noninvasive recording using electroencephalography (EEG/ERP) and magnetoencephalography (MEG). These methods share the high temporal resolution (on the order of milliseconds) appropriate for assessing perceptual processes as they unfold in real time, but the methods differ greatly in the extent to which one can identify localized processes. Insofar as one has mechanistic processing models/hypotheses that address how speech is represented and processed in neuronal tissue, electrophysiological techniques are critical. Spoken language unfolds quickly, with acoustic signal changes in the millisecond range having specific consequences for perceptual classification. Accordingly, these tech-
niques are necessary to zoom (in time) into such granular temporal changes. Moreover, although many aspects of speech cannot be addressed in animal models (for example lexical representation), the single-unit and local-field-potential (LFP) animal work informs us about how single neurons and neuronal ensembles encode complex auditory signals. Thus, even though the perceptual endgame is not the same for ferrets and Francophones, some of the subroutines that constitute perception can be probed effectively using animal models.

The hemodynamic techniques, principally fMRI and PET, and more recently NIRS (near infra-red spectroscopy) have been used extensively since the late 1980s to study speech perception (Binder et al. 2000; Blumstein et al. 2005; Burton, Small, and Blumstein 2000; Meyer et al. 2005; Olesen et al. 2007; Raettig and Kotz 2008; Scott and Wise 2004). The major advantages—especially of fMRI—are its spatial resolution, and now, ubiquitous availability of the machines. It is now possible to detect activity differentially at a spatial scale of a millimeter and better, and therefore these noninvasive recordings are approaching a scale that is familiar from animal studies (roughly the scale of cortical columns) (Bandettini 2003, Logothetis 2008). However, the temporal resolution is limited, roughly to changes occurring over hundreds of milliseconds (i.e., about a word or so). The main contribution of these approaches is to our understanding of the functional anatomy (see Section 3). Note, also, that these techniques provide a ‘spatial answer’—requiring as a hypothesis a ‘spatial question.’ While the contribution of hemodynamic imaging to anatomy is considerable, questions about representation—and especially online processing—are difficult to address using such methods. Recent reviews of fMRI, in particular, emphasize the need to complement such data with electrophysiological recordings (Logothetis 2008). As a leading neuroscientist and imaging expert, Nikos Logothetis, points out, ‘fMRI is a measure of mass action. You almost have to be a professional moron to think you’re saying something profound about the neural mechanisms. You’re nowhere close to explaining what’s happening, but you have a nice framework, an excellent starting point’ (http://www.sciencenews.org/view/feature/id/50295/title/Trawling_the_brain).

Historically, neuropsychological data have been the most widely available, consequently deficit-lesion correlation research forms the basis for the functional anatomy of speech sound processing as we conceive it here (see Section 3). In recent years, the reversible (in)activation of neuronal tissue using transcranial magnetic stimulation, TMS, has received much attention, although as yet few studies have investigated speech—and those that have yielded very dodgy results, e.g., D’Ausilio et al. 2009). The careful dissection of deficits and associated lesions has played a big role in establishing some of the key insights of current models, including that speech perception is more bilaterally mediated than common textbook wisdom holds to be true, and that frontal areas contribute to perceptual abilities under certain task configurations (see, e.g., work by Blumstein for elaboration). Neuropsychological data establish both (a) that speech processing clearly dissociates from language processing as well as from other parts of auditory cognition (Poeppel 2001) and (b) that the classical view that the left temporal lobe subserves speech and language comprehension is dramatically underspecified.

While these school-marmish reminders regarding the benefits and limitations of techniques may seem irritating and perhaps even obvious, it is remarkable how often research is insensitive to crucial methodological limitations, thereby furthering interpretations that are untenable given the origin of the data. Insofar as we seek a theoretically sound, formally explicit, and neuronally realistic model of spoken language processing and brain, a thoughtful consideration of which techniques answer which questions is essential.

1.3. ‘Function-o-logical’

The perspective summarized here has been developed in recent pieces (Poeppel and Hackl 2008; Poeppel, Ildsardi, and Wassenhove 2008; Poeppel and Monahan 2008). What we hope to provide is a serviceable definition for the cognitive neuroscience of speech perception that links various interrelated questions from acoustics to phonology to lexical access. Figure 1, from Poeppel et al. (2008), summarizes what we take to be the problem.

The starting point for the perceptual-computational system is the acoustic signal, a continuously varying waveform that encodes information on different timescales (Fig 1a). For example, the amplitude envelope of the signal correlates well with properties of the syllabic structure of an utterance; the fine structure of the signal, in contrast, carries information over shorter timescales (including features and segments). This input array must ultimately be transformed into a series of discrete segments that constitute a morpheme/word. Because we believe the key goal to be the identification of words, specifying the format of lexical representation is necessary. Moreover, the morphemes/words must be stored in a format that permits them to enter into subsequent linguistic computation (including, e.g., combinatoric operations that underlie language comprehension); identifying a word is not nearly enough—the listener must be able to connect it formally
The input waveform (representation R1) is analyzed by the auditory periphery and is presumably represented in auditory cortex by neurons with sophisticated spectro-temporal receptive field properties (STRFs). One can think of this as a neural version of a spectrogram, albeit one composed of numerous mini-spectrograms with specializations for certain spectro-temporal patterns (Fig 1b), such as the characteristic convergent second and third formant trajectories near velars (Stevens 1998). This type of representation (R2) is most likely a property of neurons in the auditory cortex, and it does not differentiate between speech and non-speech signals. Moreover, given the highly conserved nature of mammalian auditory cortex, these representations are very likely shared with other species, and consequently these representations can be investigated using animal models and single-cell recording approaches. Based on this initial (high resolution) auditory cortical pattern, multiple representations on different scales are constructed, in parallel. In this next step, “auditory primitives” are built out of early auditory cortical elements, with one key feature being the time scale of the new representations. This third type of representation (R3) must be of a granularity that permits mappings (linking operations) from the encoding of simple acoustic properties in early auditory cortical areas to speech primitives in more downstream areas (arguably including STG andSTS). We conjecture that these intermediate representations encompass at least two subtypes (temporal primitives) commensurate with syllabic and segmental durations (Boemio et al. 2005; Giraud et al. 2007; Poeppel 2001, 2003; Poeppel et al. 2008). The initial cortical representation is fractionated into (at least) two streams, and concurrent multi-time resolution analysis then lies at the basis of subsequent processing. The specific nature of R3 is a critical research question, and we have characterized the question as arriving at a ‘primal sketch’ for speech perception (Fig 1c), akin to Marr’s famous hypothesis about intermediate representations for object recognition; one possibility for the primal sketch is the PFNA coarse coding (plosive-fricative-nasal-approximant), discussed below. The final, featurally specified representation (R4) constitutes the format that is both the endpoint of perception but which is also the set of instructions for articulation. As discussed further below, the loop between perception, memory, and action is enabled because the representational format used for words in memory, distinctive features, allows the mapping both from the input to words (identify features) and from words to action (features are in motoric coordinates).

Obviously, a major goal now must be to look for a Hegelian synthesis for these various antitheses, i.e., levels of representation with competing structures and affordances. In particular, how is it that we have so much solid evidence for both cohorts and neighborhoods, whose guiding assump-
tions seem irreconcilable? What kind of system is this that illustrates both phonetic specificity (a surface property of speech sounds) and phonological underspecification (a generalization following from a highly abstract code)? Again, we believe that in order to understand this panoply of confusing results we need to draw further distinctions, and we offer up a modest proposal in order to have our exemplar cake and eat it too. Stealing a line from Cutler and Fay (1982) we agree that there is “one mental lexicon, phonologically arranged.” But the operative word here is “arranged”. We envision a 3-step process that offers a place for each of the kinds of findings. The first step is a coarse-coding of the signal into universal speech categories (akin if not identical to Stevens’ (1998) landmarks). For concreteness let us say that this code is just the speech stream coded into four categories (PFNA: plosives, fricatives, nasals and approximants). Preliminary modeling of English-like lexicons suggests that this coding yields pools of words of approximately the same size as the usual lexical neighborhoods and with a fair overlap between various pools and neighborhoods. Within these pools we can now conduct a directed left-to-right search using contextually defined featural definitions (i.e., the cues for [labial] within [nasal] are different than those within [plosive], and differ language to language). Moreover, this search can be guided by the differences amongst the words in the active pool using analysis-by-synthesis and Bayesian inference (see below). Finally, once the best-available word-form has been selected, the contents of that lexical item are examined, compared to the memory trace of the incoming signal, and verified to in fact be the word we’re looking for. Since the lexical entry contains a great deal of information (morphology, syntax, semantics, pragmatics, usage) there is little harm or cost (and much benefit) in storing a detailed phonetic summary of the form’s pronunciation (though we would prefer a model-based statistical summary to an exemplar cloud). In sum, we get to the entry via a coarse-coded search with subsequent directed refinement, but the choice needs to be verified to be accepted. Thus we expect (eventually) to see in the time-course of word-recognition early effects of coarse-coding followed later by exemplar-like effects of lexical item phonetic specificity, even if our current methods are perhaps too crude to pick up this distinction.

One way to think about the challenge is to consider the analogy to visual object recognition. Research there has attempted to identify which intermediate representations can link the early cortical analyses over small spatial receptive fields (edges or Gabor patches, or other early visual primitives) with the representation of objects. There have been different approaches to intermediate representations, but every computational theory, either explicitly or implicitly, acknowledges the need for them. The more traditional hypothesis – a mapping from acoustic to phonetic to phonological representations – is no longer central to the problem as we define it (although the mapping from R1/R2 to R3 to R4 is reminiscent of similar challenges). The multiple levels of representation we envision are simultaneous representations on different time-scales corresponding to different linguistic ‘views’ of the speech material.

2. Linguistic bases of speech perception

2.1. Features

Because most modern societies are literate and often familiar with a language with an alphabetic script, there is a tendency to identify speech perception with the perception of whole, single speech segments (phones or phonemes) – the amount of speech generally captured by a single letter in an alphabetic script. However, segmental phonemes are not the smallest units of representation, but are composed of distinctive features which connect articulatory goals with auditory patterns, and provide a discrete, modality – and task-neutral representation suitable for storage in long-term memory (see Jakobson, Fant and Halle 1952, for the original proposals, and Halle 2002, for a spirited defense of this position; see Mielke 2007, for a contrasting view). For example the feature [+round] encodes a speech sound component that in articulation involves rounding the lips through the enervation of the orbicularis oris muscle, and on the auditory side a region of speech with a downward sweep of all of the formants (when formant transitions are available), or diffuse spectra (in stop bursts and fricatives). The features thus are the basis of the translation (coordinate transformation) between acoustic-space and articulator-space, and moreover are part of the long-term memory representations for the phonological content of morphemes, forming the first memory-action-perception loop.

Phonetic features come in two kinds: articulator-bound and articulator-free. The articulator-bound features (such as [+round]) can only be executed by a particular muscle group. In contrast, the articulator-free, or “manner” features, which (simplifying somewhat) specify the degree of constriction at the narrowest point in the vocal tract, can be executed by any of several muscles along the vocal tract. Specifying the degree of constriction defines the sonority scale, and thus the major classes of segments: plosives (with complete constriction), fricatives (with constrictions sufficiently narrow to generate turbulent noise), sonorants (including nasals,
with little constriction), glides and vowels (i.e., approximants, with virtually no constriction). Moreover, as noted above, this division suggests a computational technique for calculating R2 and R3: build a set of major-class detectors from R1 representations (Stevens 2002; Juneja and Espy-Wilson 2008). To a crude first approximation, this consists of detectors for quasi-silent intervals (plosives), regions with significant amounts of non-periodicity (fricatives), regions with only one significant resonance (nasals) and regions with general formant structure (approximants, which then must be sub-classified). These definitions are plausibly universal, and all of these detectors are also plausibly ecologically useful for non-speech tasks (such as predator or prey detection), and thus should be amenable to investigation with animal models, and are good candidates for hard-wired circuits. Once the major class is detected, dedicated sub-routines particular to the recovered class are invoked to subsequently identify the contemporaneous articulator-bound features. In this way, features such as [+round] may have context-sensitive acoustic definitions, such as diffuse falling spectra in stop bursts, a relatively low spectral zero in nasals, and lowered formants in vowels.

2.2. Groupings

Even though the individual features are each tracked as a separate stream (like instruments in an orchestra), identification of the streams of phonetic features by themselves is not sufficient to adequately capture the linguistically structured representations. The features must be temporally coordinated, akin to the control exerted by the conductor. Speech-time is quantized into differently-sized chunks of time. There are two critically important chunk-sizes that seem universally instantiated in spoken languages: segments and syllables. Temporal co-ordination of distinctive features overlapping for relatively brief amounts of time (10-80 ms) comprise segments; longer coordinated movements (100-500 ms) constitute syllabic prosodies. For instance “we” and “you” differ in the synchronization of [+round]: in “we,” rounding coincides with the initial glide, in “you,” the rounding is on the vowel, and in “wu” rounding covers both segments. This first aggregation of features must somehow ignore various coarticulation and imprecise articulation effects which can lead to phantom (excrecent) segments, as can be seen in pronunciations of “else” which rhyme with “welts” (familiar to Tom Lehrer fans). At the syllable level, English displays alternating patterns of weak and strong syllables, a distinction which affects the pronunciation of the segments within the syllables, with weak syllables having reduced articulations along several dimensions. It is possible that groupings of other sizes (morae, feet) are also relevant; certainly linguistic theories postulate manageries of such chunks. We believe that the syllable level may begin to be calculated from the major-class detectors outlined in the previous section; typologically, language syllable structure seems to be almost exclusively characterized by sonority, with the articulator-bound features playing little role in the definition of the constitution of syllables. We hypothesize that the parallel sketch of major-class based syllables and the elaboration of segments via the identification of articulator-bound features offers a potential model for the synthesis of the so-far irreconcilable findings for cohort and neighborhood models of lexical access.

2.3. Predictable changes in pronunciation: phonological process

Speech is highly variable. One of the goals of distinctive feature theory is to try to identify higher-order invariants in the speech signal that correlate with the presence of particular features like the [+round] example above (Perkell and Klatt 1986). However, even if we had a perfect theory of phonetic distinctive features, there is variation in the pronunciation of features and segments due to surrounding context, starting with co-articulation effects inherent in the inertial movements of the articulators in the mouth. The net result of these patterned variations in pronunciation is that we are willing to consider disparate pronunciations to be instances of “the same speech sound” because we can attribute the differences in pronunciation to the surrounding context of speech material. A particularly easy way to observe this phenomenon is to consider different forms of the same word which arise through morphological operations like prefixation and suffixation. The ‘t’ in "atom" and "atomic" are not pronounced the same way - "atom" is homophonous with "Adam" for many speakers of American English, whereas "atomic" has a portion homophonous with the name "Tom". In technical parlance, the ‘t’ in "atom" is flapped, whereas the ‘t’ in "atomic" is aspirated. This is by no means an unusual case. Every known language has such contextually determined pronunciations (allophonic variation) that do not affect the meanings of the words, and which, for the purpose of recovering the words, appear to be extra noise for the listener. Even worse, languages pick and choose which features they employ for storing forms in memory. English, for example, considers the difference between [l] and [r],
[±lateral], to be contrastive, so that "rip" and "lip" are different words, as are "more" and "mole". Korean, on the other hand, treats this difference as allophonic, a predictable aspect of the position of the segment in the word; the word for water is "mul" but the term for freshwater is "muri chōn". For Koreans [I] and [r] are contextual pronunciations of the same sound – they use [r] before vowels and [I] before consonants and at the ends of words.

Recent MEG studies (Kazanina et al. 2006) have confirmed that listeners do systematically ignore allophonic differences (Sapir 1933). Using a mismatch design, Kazanina and colleagues compared the behavioral and neural responses of Russian and Korean speakers to items containing “ta” or “da”. The difference in the feature [voice] between “t” and “d” is significant (contrastive) in Russian, as it serves to distinguish pairs of words such as “dom” (house) and “tom” (volume). In Korean, however, this difference is predictable, with “d” occurring only between sonorants, as can be seen in the word “totuk” 도둑 meaning ‘thief’, pronounced “toduk” (and spelled that way in the McCune-Reischauer Romanization system). In this word, the second “t” is pronounced as a “d” because it is flanked by vowels (similar to the English flapping rule). Subjects listened to sequences of items in which one of the two types (for instance “da”) was much more frequent (the “standard”); the other item (the “deviant”, here “ta”) occurred 13% of the time. Neural responses to the items in different were compared (i.e., “ta” as standard was compared with “ta” as deviant). Russian speakers showed a reliable difference in their responses to standards and deviants, indicating that they detected the deviant items in a stream of standards. Conversely, Korean speakers showed no differences, suggesting that they form a perceptual equivalence class for “t” and “d”, mapping these two sounds onto the same abstract representation.

Similar phonological processes can also change the pronunciation of otherwise contrastive speech sounds. For instance, in English “s” and “z” are contrastive, differing in the feature [voice], as can be seen by the minimal pair “seal” and “zeal”. However, the plural ending is pronounced either “s” or “z” depending on the adjacent sound: “cats” and “dogz”. English listeners are sensitive to this sequential patterning rule, showing longer reaction times and differences in the neural responses when “s” and “z” are cross-spliced into incongruent positions. *“utz”, *“uds” (Hwang et al. submitted). Thus, in later morphological computations, contrastive sounds are organized into higher-level equivalence classes displaying functional identity (such as the plural ending).

Phonological perception thus requires the identification of major class features and articulator-bound features, the coordination of these features into segment-sized units and larger chunks, and the identification of equivalence classes of features and segments at various levels of abstraction, with this stage of processing culminating in the identification of a stored word form, which can then be mapped back out by the motor system in pronunciation.

3. Cortical basis of speech perception: fractionating information flow across a distributed functional anatomy

Historically, the reception of speech is most closely associated with the discoveries by the German neurologist Wernicke. Based on his work, popularized in virtually every textbook since the early 20th century, it was hypothesized that posterior aspects of the left hemisphere, in particular the left superior temporal gyrus (STG), were responsible for the analysis of the input. Perception (modality-specific) and comprehension (modality-independent) were not distinguished, and so the left temporal lobe has become the canonical speech perception region. Because speech perception dissociates clearly from the auditory perception of non-speech, as well as from more central comprehension processes (see, for example, data from pure word deafness, reviewed in Poeppel 2001; Stefanatos, Gershkoff, and Madigan 2005), the search for a ‘speech perception area’ is, in principle, reasonable (and the search for such a specialized region has, of course, yielded a rich and controversial research program in the case of face recognition and the fusiform face area). However, data deriving from lesion studies, brain imaging, and intracranial studies have converged on a model in which a distributed functional anatomy is more likely.

The major ingredients in this model are two concurrent processing streams. Early auditory areas, bilaterally, are responsible for creating a high-resolution spectro-temporal representation. In the terminology developed in Section 1, the afferent auditory pathway from the periphery to cortex executes the set of transformations from R1 to R2. In superior temporal cortex, two parallel pathways originate. One pathway, the ventral stream, is primarily involved in the mapping from sound to lexical meaning. We hypothesize that lexical representations per se (the mappings from concepts to lexical-phonological entries) ‘reside’ in middle temporal gyrus (MTG) (for a recent analysis of lexical processing and MTG based on neurophysiology, see Lau, Phillips, and Poeppel 2008), and the cortical regions that are part of the ventral stream perform the operations that transform acoustic signals into a format that can make contact with these long-term representations.
One crucial cortical region involved in the mapping from sound structure to lexical representation is the superior temporal sulcus (STS). Neurons in STS appear to be executing some of the essential computations that generate the speech primitives. Our examination of both lesion and imaging data suggests that the ventral pathway is bilateral. Importantly, the left and right contributions are overlapping but not identical; for example, the fractionation of the auditory signal into temporal primitives of different granularity (i.e., different time-scales) occurs differentially on the two sides (Boemio et al. 2005; Giraud et al. 2007). In short the ventral pathway itself can be subdivided into concurrent processing streams that deal preferentially with different ‘linguistic views’ of the input signal, ideally mapping directly onto the parallel linguistically motivated representations for segments and syllables.

The second pathway is the dorsal stream. The areas comprising the dorsal stream perform the mapping from sensory (or perhaps phonological) representations to articulatory and motor representations (R4). Various parts of the dorsal stream lie in the frontal lobe, including in premotor areas as well as in the inferior frontal gyrus. One critical new region that is motivated – and now identified – by research on this model is area SPT (Sylvian parietal-temporal; see Hickok et al. 2003; Pa and Hickok 2008). Acoustic information is represented in a different coordinate system than articulatory information, and thus the mapping from acoustic to motor requires a coordinate transformation. Moreover, the SPT “sensorimotor interface” provides the substrate for dealing with working memory demands as well. Because the distributed functional anatomy has been described at length elsewhere (Hickok and Poeppel 2000, 2004, 2007) we aim here simply to emphasize two core features of the model that have stimulated the formulation of new hypotheses about the organization of the speech perception system: there are two segregated processing streams, each of which has functionally relevant subdivisions; second, the organization of the ventral stream is bilateral, unlike the striking lateralization often observed in language processing. On our view, it is the dorsal stream, principally involved in production, which exhibits cerebral dominance. The ventral stream, on the other hand is asymmetric but has important contributions from both hemispheres. The model highlights the distributed nature of the cortical fields underlying speech processing. Moreover, the model illustrates the perception-memory-action loop that we have described. The loop at the basis of speech processing ‘works’ because of the nature of the shared currency that forms the basis of the representation. We contend that this distinctive featural representation is one that permits natural mappings from input to memory to output.

Figure 2. Dual stream model (Hickok and Poeppel 2007). A functional anatomic model of speech perception that hypothesizes speech perception to be executed by two concurrent processing streams. The green box is the starting point, the auditory input. The ventral stream (pink) mediates the mapping from sounds to lexical-semantic representations. The dorsal stream (blue) provides the neuronal infrastructure for the mapping from sound analysis to articulation. For discussion, see Hickok and Poeppel (2000, 2004, 2007). For colour version see separate plate.

4. Analysis by synthesis: an old algorithm, resuscitated

The recurrent theme has been the memory-action-perception (MAP) loop. We pointed out that the necessity of being able to both produce and perceive speech leads to the necessity of coordinate transformations between
acoustic-space and articulatory-space, and economy considerations dictate that we look for a memory architecture that would enable and facilitate this. Distinctive features, as developed by Jakobson, Fant and Halle in the 1950s, have such desirable characteristics, with both acoustic and articulatory definitions, and provide a basis for long-term memory representations. These (static) representational considerations are mirrored in the computational algorithms for (dynamic) speech production and perception; the algorithms are similarly intertwined. Such a system was proposed at the very beginning of modern speech perception research, in the analysis-by-synthesis approach of MacKay (1951) and Halle and Stevens (1962). Bever and Poeppel (2010) and Poeppel and Monahan (2010) review this idea. Here we have the MAP loop with both large and small. At each level, speech perception takes the form of a guess at an analysis, the subsequent generation of a predicted output form, and a correction to that guess based on an error signal generated by the comparison of the predicted output form against the incoming signal. The initial hypotheses ('guesses') are generated based on the current state and the smallest bit of input processed (say a 30 ms sample of waveform), and whatever additional information may have been used to predict the signal (the prior, in a Bayesian framework). The initial set of guesses that trigger synthesis will be large, but at each subsequent processing step, the set of supported guesses gets smaller, therefore

the set of synthesized representations gets smaller, and therefore verification or rejection based on subsequent input gets quicker. Although little empirical data exist to date that test these (old) ideas, recent studies on audiovisual speech perception support analysis by synthesis and illustrate how the amount of prior information modulates (in time) cortical responses to speech sounds (van Wassenhove, Grant, and Poeppel 2005).

The appeal of the analysis-by-synthesis view is threefold. First, it provides a link to motor theories of perception, albeit at a more abstract level. Motor theories in their most direct form are not well supported empirically (for a recent discussion, see Hickok, 2010), but the hypothesis that a perceptual role for some of the computations underlying motoric action is worth exploring. There is an intriguing connection between perception and production, repeatedly observed in many areas of perception, and narrowly perceptual or motoric theories seem not to be successful at capturing the observed phenomena. Second, analysis-by-synthesis for speech links to interesting new work in other domains of perception. For example, new research on visual object recognition supports the idea (Yuille and Kersten 2006); and the concept has been examined in depth in work on sentence comprehension by Bever (Townsend and Bever 2001). There is a close connection between the analysis-by-synthesis program and Bayesian models of perception (a point also made by Hinton and Nair 2005), and thereby also to link up with more tenable accounts of what mirror neurons are (Kilner, Friston, and Frith 2007). Third, this approach provides a natural bridge to concepts gaining influence in systems neuroscience. The view that the massive top-down architectural connectivity in cortex forms the basis for generating and testing expectations (at every level of analysis) is gaining credibility, and predictive coding is widely observed using different techniques. In our view, the type of computational infrastructure afforded by the analysis-by-synthesis research program provides a way to investigate speech perception in the service of lexical access in a way that naturally links the computational, algorithmic, and implementational levels advocated by Marr.

Research on the cognitive science and cognitive neuroscience of speech perception seems to us a productive approach to investigate questions about cognition and its neural basis more generally. The psychological models are increasingly detailed and well articulated and facilitate a principled investigation of how the brain computes with complex representations. For the development of these important models, we owe a debt of gratitude.
References


Bever, Thomas and David Poeppel in press Analysis by synthesis: A (Re-)emerging program of research for language and vision. *Biolinguistics*.


Halle, Morris 2002 *From Memory to Speech and Back*. Berlin: Mouton de Gruyter.


Hickok, Gregory 2010 The role of mirror neurons in speech perception and action word semantics. *Language and Cognitive Processes*. 
Kuhl, Patricia, Barbara T. Conboy, Sharon Coffey-Corina, Denise Padden, Maritza Rivera-Gaxiola, and Tobey Nelson

Lau, Ellen, Colin Phillips, and David Poeppel

Liberman, Alvin M.

Liberman, Alvin M.

Luce, Paul A and David B Pisoni

Luo, Huan and David Poeppel

MacKay, Donald M.
1951 Mindlike behavior in artefacts. British Journal for Philosophy of Science 2: 105-121.

Marslen-Wilson, William

McClelland, James L. and Jeff Elman

Meyer, Martin, Stefan Zysset, Yves D. von Cramon, and Kai Alter

Mielke, Jeff

Misiurski, Cara, Sheila E. Blumstein, Jesse Rissman, and Daniel Berman
Raettig, Tim and Sonja A. Kotz

Raizada, Rajiv D. and Russell A. Poldrack

Rauschecker, Joseph P., Biao Tian, and Marc Hauser

Sapir, Edward

Schroeder, Charles E., Peter Lakatos, Joshinao Kajikawa, Sarah Partan, and Aina Puce


Scott, Sophie K. and Richard J. S. Wise

Stefanatos, Gerry A., Arthur Gershkoff, and Sean Madigan

Steinschneider, Mitchell, Charles E. Schroeder, Joseph C. Arezzo, and Herbert G. Vaughan Jr.

Stevens, Kenneth N.


Townsend, David J., and Thomas G. Bever
Brain structures underlying lexical processing of speech: Evidence from brain imaging

Matthew H. Davis and Jennifer M. Rodd

1. The neural foundations of lexical processing of speech

A mental lexicon that links the form of spoken words to their associated meanings and syntactic functions has long been seen as a central to speech comprehension. As should be apparent from other papers in this volume, William Marslen-Wilson's experimental and theoretical contributions have been supremely influential in guiding research on the psychological and computational properties of the lexicon. With recent developments in functional brain imaging, methods now exist to map these processes onto neuroanatomical pathways in the human brain.

In the current paper we will argue that theoretical proposals made in various iterations of the Cohort account remain just as relevant to this neuroscientific endeavour as they were for previous generations of researchers in the psychological and cognitive sciences. Here we will review recent brain imaging work on lexical processing in the light of these theoretical principles.

Accounts of the neural processes underlying spoken word recognition have converged on the proposal that brain regions centred on the superior temporal gyrus (STG) are critical for pre-lexical processing of spoken words. That is, this region is engaged in the acoustic-phonetic processes that provide the input for later lexical and semantic analysis of the speech signal. In the neuropsychological literature, cases of pure word deafness (an isolated impairment of speech perception in the presence of intact linguistic skills in other modalities and perception of non-speech sounds) is typically observed following bilateral lesions to these superior temporal regions (Saffran, Marin, and Komshian 1976; Stefanatos, Gershkoff, and Madigan 2005). Functional imaging studies that compare neural responses to speech sounds and acoustically matched sounds that do not evoke a speech percept evoke differential responses in superior temporal gyrus regions that surround but do not include primary auditory cortex (Davis and Johnsrude 2003; Scott et al. 2000; Uppenkamp et al. 2006; Vouloumanos et al. 2001). Although activation for speech greater than non-speech is often seen bilaterally, those studies that focus on phonological aspects of speech processing...