The Graphical Basis of Phones and Phonemes[1]

Robert F. Port
Professor of Linguistics and Cognitive Science
330 Memorial Hall
Indiana University
Bloomington, IN 47401
port@indiana.edu

Words, main text: 7,995
Words, including references: 9,106
[July 27, 2005]

[1] To appear in Murray Munro and Ocke-Schwen Bohn (eds.) *Second-language speech learning: The role of language experience in speech perception and production*. A volume dedicated to Prof. James Flege.

> The notion of phonetic segment, phone and phoneme are closely related and all are intuitively appealing. At least one of them seems the like the right description for speech. But all those who report these intuitions happen to be people who learned to write using a phonetic alphabet in early childhood. Speech is difficult to attend to because of its rapidity, its variability and the invisibility of the most important body movements, so some cognitive scaffolding for attending to speech accurately is required. The technology of alphabetic writing was modified for this purpose about a hundred years ago. Our alphabet experience accounts for the persuasiveness of our intuitions but segments (phonemic or phonetic) are probably not important units in the psychological representation of language.

One of the many reasons for studying second-language acquisition is that it contributes to our understanding of `phonetic' and `phonological' spaces. What are the ultimate cognitive dimensions of speech sounds? Jim Flege's work, in looking closely at the details of patterns in foreign accented speech, offers not just information about how languages are learned, but also provides important insights into the inherent dimensionality of phonetics and phonology. What kind of phonological objects can there be? Results from language acquisition provide powerful evidence of the inadequacy of describing speech in terms of any fixed set of phonetic or phonological units such as segments or segmental features (Chomsky, 1964; IPA, 1999). Flege's careful measurements of articulatory and temporal detail in the speech of people who speak more than one language (Flege & Hillenbrand, 1986; Flege, Munro, & Skelton, 1992; Flege, Munro, & Mackay, 1995) reveal that foreign accented speech exhibits rich and complex speech patterns (in production and perception) that are not, for example, simply the sum of the two phonologies, but involve a complex interaction between two completely different and typically incompatible phonological and phonetic systems. It seems that a major implication of phonetic studies of second-language acquisition is that real speakers employ a far richer phonetic and phonological system than is implied by the minimal number of segmental phonetic features that linguists hope to be able to get away with (Bloomfield, 1933; Chomsky, 1964; Chomsky & Halle, 1968; Jakobson, Fant, & Halle, 1952).

Further evidence of linguistic memory as far richer, more detailed and more redundant than was imagined on the traditional semiotic or speech-as-code view, comes from speech perception research. Low-dimensional semiotic descriptions of the speech signal (Bloomfield, 1933, 1939; Chomsky & Halle, 1968; Jakobson et al., 1952) have proven completely unsuccessful. They simply cannot be implemented computationally. Instead, the weight of research evidence suggests to me a speech perception process that employs a great many redundant patterns coming in a wide range of overlapping sizes from fragments of a segment to patterns that are several syllables in length. The units may range from single features to entire syllables or even feet and will overlap each other profusely. The dimensionality of the relevant space is not 50 to 80 phonetic features but probably thousands of degrees of freedom especially in audition. The representations that support speech perception will be whatever auditory patterns have enough statistical predominance in speech that they can be learned by the appropriate neural networks

(Bod, Hay, & Jannedy, 2003; Diehl, Lotto, & Holt, 2004; Grossberg, 2003; Hawkins, 2003). We should assume that different speakers of the same dialect will have analytical units that differ widely in detail. Of course, speakers in the same community will, at a gross level, talk and listen to speech very similarly (that is, they will have the same ``accent''), but they each have different histories of exposure to various changing auditory and linguistic environments. The greater the detail at which you look, the greater should be differences between individual nervous systems in the `coding' of speech. Even in a single speaker, there will not necessarily be a single representation for each word, but a distribution of particular exemplars in a very high-dimensional space. Possibly speakers do store a ``summary representation'' of some kind for each word (or word-like unit), but they apparently also store much additional information – enough that they can be sensitive to the likelihood of particular variants in a particular context. My conclusion is that we must abandon the goal of finding any minimal or ``efficient'' (in a symbol counting way) word representations in our memory for linguistic utterances. Speech perception depends on a rich perceptual learning process beginning before birth and continuing throughout life.

I think the evidence overwhelmingly supports a description of speech perception along these lines. But we linguists and phoneticians seem to be strongly biased against accepting that this might be all there is to it. We cannot avoid trusting our intuitions about how we perceive speech. And our intuitions overwhelmingly support an important role for segmental speech units. We are sure that phonetic segments or phonemes or ``speech sounds.'' represent a critical stage or primitive level in the process of listening to words. The idea that speech can and should be represented with segmental units – that is, with units that are easily represented with letters or a short vector of phonetic feature – is very appealing. It seems intuitively clear that the words *cat, bin, set* and *mop* have three segments each, while *scat, cats, lamp* and *spin* have four.[2] The goal of this essay is to suggest a reason why our intuitions about segmental descriptions of speech are sometimes very persuasive and clear. Too many linguists believe that once they have recorded enough phonetic detail to produce an alphabetic description, they need look no further into the phonetics of a language. To many, a segmental description seems like a sensible place to stop exploring the potential morass of phonetic detail. Who needs zealous phonetic experts? asked Bloomfield (1933, p. 128). However, I claim **there is no clear evidence for a universal phonetic inventory of letter-like sound types** (Port & Leary, 2005). There is no level of detail where an apriori justification can be found for ceasing to look closer at the speech signal. This implies there is no reliable basis for the cross-language comparisons that modern phonology relies on. How do we know this? In part, because we find second-language learners inventing their own ways of combining two sets of phonological skills (Flege et al., 1995). But we also know it from 50 years of experimental phonetics research documenting the critical role of temporal patterns, formant trajectories, articulatory movements and various other structures that are

[2] On the other hand, my own intuitions about the number of segments are not clear at all for my pronunciations of words like: *purr* [pəˑ, pər], *fire* [faˈr, fajəˑ, faˈjəˑ], *chose* [čoz, tʃoʊz], *Iowa* [ajwə, aɪəwə], *Hilary* [hɪlri, hɪləˑij], *police* [plis, pəlis] and *tree* [tʳi, tr̥i, tʃri], and many others.

inadequately described in segmental terms. A fully developed argument for these claims is presented in Port and Leary, 2005. The primary goal of this paper is limited to enquiring why linguists, phoneticians and other scientists find descriptions of speech in terms of consonant and vowel segments so natural and satisfactory.

In many cases, it is unclear how many segments are appropriate. Thus, orthographic *mite* could have 3 segments, 4 segments, or 5 segments, depending on the conventions of interpretation we assume for the letters. We could spell the word as [mit] (where the phoneme /i/ is interpreted as a diphthong, roughly as in Chomsky-Halle, 1968) or as [mayt] (cf. Bloomfield, where the diphthong is written with 2 letters) or even as [may[closure][t-burst]] (an acoustically inspired description that treats the closure and burst as separate segments). However I suspect most people believe there is a fact of the matter about whether the stop in *spew* is a [p] or a [b] or some third type of labial stop and whether *spew* contains a [j] segment.[3] On the conventional view, every word in a particular dialect has **some** specific phonological spelling. This essay suggests that these compelling intuitions about how to describe speech are, to a considerable degree, a **consequence of our lifelong practice using alphabets and not a necessary psychological fact about speech**. This is not really a new suggestion, as we shall see, but we linguists need to understand our segmental intuitions far better than we do. The basic story is that we who were trained in an alphabet-based literacy tradition find it natural to employ letters to describe speech sounds. Correspondingly, those not trained this way will not find segments as natural or compelling. Alphabetic notation for speech is a cultural tradition acquired (with considerable difficulty in many cases) at an early age. The phonological patterns of speech have structure at many different levels only one of which sometimes resembles the segments used in phonetic or orthographic writing.

What are words really made of? They seem clearly to be made of parts – pieces that are reused in many other words – making phonological patterns. Is there some reason to conclude in advance that the pieces have a certain size or cannot overlap in arbitrary ways? Languages obviously differ greatly in what pieces they use. Speech sounds come in many overlapping sizes and can differ from each other in very minute increments.

Attending to Speech Sounds.

In preliterate times it must have been very difficult for those who might wonder about language to attend consciously to speech sounds, either in perceptual or articulatory terms. We learn to talk very early, seemingly with no effort, and tend to spit out syllables

[3] In this paper, I take for granted that English orthography is based on the `phoneme concept' even though standard orthographic spellings exhibit less perfect representations of an ideal phonemic spelling system than, say, the orthography of Spanish or Swahili. Words like *sit, set, sat,* and *sin, send, sand* suggest the phoneme idea quite transparently while *through* and *exception* admittedly obscure that idea. But I presume that nevertheless literate speakers of English will in most cases be able to grasp the underlying notion of a phoneme from learning to read and write English orthographically.

with heedless abandon. People typically have no awareness of what they are doing. One major exception is the early Indian grammarians who did develop a quite accurate phonetic description of Sanskrit and transmitted this knowledge orally for several centuries via Panini's summary before it was committed to written form. But aside from this case, motivated by a religious mission, knowledge of the speech organs and speech sounds has remained quite unknown to most human populations until the present day (aside from certain professionals like academic linguists, musical voice pedagogs, and some reading teachers).

To appreciate the difficulty of analyzing speech in an unfamiliar language, I recommend listening several times to a phrase spoken in a language you don't know at all. It seems impossible to imagine reproducing those sounds exactly. (You may understand the sounds much better when spoken by someone of your native language speaking the language as a second language.) It's also difficult to imagine doing a phonetic transcription of it. But, in fact, much the same problem must arise listening to our own voice as we let our syllables fly in casual speech, at least if you do not know alphabetic writing.

There are two main reasons why accurate introspective analysis of our own speech is very difficult without external help:

1. Speech sounds are articulated very quickly. Most so-called `speech sounds' are short enough that 10 to 15 of them can be uttered within one second. In contrast, typical consciously controlled human movements – those like walking, wrestling, moving checker pieces, slicing bread, riding a bicycle – are much slower with roughly 1-2 complete movements per second. Only musical performance – think of piano, violin or clarinet performers – and touch-typing seem to involve movement rates that resemble those of speech where. All these skills other than speech require extensive, concentrated practice at slower rates to develop high performance speeds. But all children learn a language presented to them at normal rates with little evidence of slower speech for early learners. Speech also produces very complex patterns in time (Port, 2003; Port & Leary, 2005). But the recognition or manipulation of rapid patterns in time is intrinsically difficult for humans. To appreciate the problem, imagine trying to recognize a Touchtone telephone signal auditorily when it is redialed automatically by your phone. You hear 11 short tones within about a second and a half. It would take a great deal of practice – probably several hours – to reliably recognize even one such tone pattern much less recognize an arbitrary phone number from the sound [4] (Cooper, Liberman, & Borst, 1951; IPA, 1999; A. M. Liberman, Delattre, Gerstman, & Cooper, 1956; Watson, Kelly, & Wroton, 1976). In contrast to arbitrary auditory patterns, we can recognize speech sounds at roughly this rate, however, because our perceptual system models them with an articulation-like dynamic code – the phonology of the language. The trouble is that this gesture-based code only runs very fast! We cannot generally articulate (or perceive) speech at, say, a quarter of normal rate. But the normal rate is too fast to

[4] Actually, Touchtone signals are each a pair of sinusoidal tones. Numbers in the same row (e.g., 1, 2, 3) share the same lower tone and those in the same column (e.g., 3, 6, 9) share the same upper tone.

grab by its component parts – just as with the automatically generated Touchtone phone signal. In short, there is an apparent mismatch apparently between human attention and the natural rate of speech sound production. The way humans deal with this problem is through rapid perceptual learning at an early age. Child speaker-hearers learn to recognize many of the auditory patterns relevant to their language within the first year of life before actually producing any words themselves and, in the process, lose sensitivity to acoustic properties that are not relevant to their language (Kuhl & Iverson, 1995; Werker & Tees, 1984). These biases facilitate rapid speech perception but seriously impede perception of speech in a language other than the native language.

2. A second problem for introspective analysis of speech is that most body parts whose movements are important for speech, especially the tongue body, the velum and the vocal folds, lie hidden within the head and neck where they are almost impossible to observe visually. It is primarily the lips and jaw, and sometimes the tongue-tip, that can be observed directly. Most people have no awareness of the actual shape of their tongue or the very existence of the soft palate and have no idea what their larynx is physically like or what it does during speech. Thus, articulatory introspection is severely limited.

How can these problems be solved? What is required for humans to have reliable phonetic introspections? The answer, of course, is a practical notation system.

Scaffolding and Writing.

Human intelligence exhibits some remarkable skills aside from speech perception – the ability to recognize a person's face after many years, the ability to separate a single voice from an audio signal of many voices, the ability to understand language spoken at a high rate of speed or to throw a ball accurately, the ability to run and leap across complex terrain, etc. However, as Andy Clark notes, there are some cognitive skills that we humans are *not* very good at (Clark, 1997). These include reproducing the exact wording of a speaker, producing a complex plan for our time management, keeping a record of intermediate results while working on a multistep arithmetic problem, remembering a list of 10 or more items, etc. For these tasks, where bare human intelligence finds itself at a disadvantage, we humans have developed many ways to use external instruments to supplement our native intelligence. Such external techniques have come to be known in cognitive science as ``cognitive scaffolding'' (Clark, 1997, 2004; Hutchins, 1995; Kirsh, 1995).

Good examples of cognitive scaffolding can be found in arithmetic when counting using fingers or an abacus, as well as in the process called `long division.' The algorithm we learned in elementary school for dividing a number with multiple decimal places into another is useable only if we have a pencil and scratch paper so the problem can be written down in the right format and the results of intermediate steps can be recorded for further manipulation (and for proofing). The pencil, the scratch paper and the graphic numerals are scaffolding in this case. Other forms of scaffolding include the use of blueprints for architectural projects, date-book calendars and phone directories. A further example is the habit of some cooks to take all the ingredients for, say, a cake recipe out of the cupboard and fridge at the beginning of food preparation, so their physical presence

in the workspace serves as a reminder to include them (Kirsh, 1995). In each of these cases, people use external objects, including marks on paper, to help with a task where our memory might fail us.

Arguably the most important form of external scaffolding for modern humans is writing, specifically, in many cases, alphabetic writing. As argued by Ong (Ong, 1982) and others (Goody, 1977; Goody & Watt, 1968), it is literacy that made possible the careful study of linguistic utterances. Ong argues plausibly that philosophy, science and law, as we know them, are all possible only because written texts make critical study and focused interpretation of specific, concrete linguistic behavior possible. This is why civilization could only arise after the advent of literacy. Civilization implies a culture where each generation can build on the accomplishments of its predecessor. It is only writing that makes rapid buildup over generations of technical, philosophical and scientific knowledge. Civilization and language phenomena are essentially social structures, not necessarily structures within individual brains.

How does alphabetic writing work? Let us consider the key properties of letters so we can compare phones and phonemes to them. Aside from their reference to speech gestures and auditory impressions, letters are:

1. A small set of **discrete graphic images** that are reliably differentiable from each other. One of the visual symbols is the **blank letter space** used to separate words.
2. The letters are arrayed in **serial order without overlap** for spelling (or encoding) words.
3. **Letters are static**. Because of their graphic nature, they do not involve any change.

This kind of alphabet is not important just for the development of orthographic systems. Another consequence is the development of specialized notational systems for logic and other branches of formal mathematics. Formal languages were developed from idealized letters and printed words. Propositional calculus was historically based on idealized written sentences of natural language (by Aristotle) and was assimilated during the 19th and 20th centuries into the more general notion of a formal language. Chomsky's famous hierarchy of formal grammars was an elaboration of these concepts and became the basis of computer languages and the basic ideas of computer science. The spectacular power of formal mathematics and computer programming is achieved by using representations that depend on discrete tokens in serial order – very much like letters and words. In computers the letters are recoded into binary digits which are simpler versions of discrete, ordered tokens and easier to implement in hardware.

The phones and phonemes of 20th century linguistics are generally taken to be formal objects, much like letters – except that they are not graphic or visual. Actually, phones and phonemes are a *conceptual blend* (Fauconnier & Turner, 2002; Lakoff & Nunez, 2000) of a graphical concept (the letter) and articulatory concepts (the speech gestures). Although phoneticians and linguists always insist that the phonetic symbols and their features have articulatory definitions, the fact is phonetic symbols never lose the staticness, discreteness and the strict serial ordering of conventional letters. The difference between the **phone** (or a `speech sound', as it is often called) and the **phoneme** is just a matter of how one resolves the competing goals of greater detail (leading to a more precise but non-minimal representation, suitable for a non-speaker of the language)

versus a maximally efficient and minimal representation of words (suitable especially for those who already speak the language). In the architypal cases, the phonetic letter represents a single articulatory target, such as a vowel position or combination of consonantal properties. But many times segments describe articulatory motions, such as in glides like [w] and [j] or diphthongs, or complex sequences of articulatory states, such as in aspirated stops (with a closure, burst and voicing lag) or affricates (with a stop followed by an approximately homorganic fricative), etc. So in cases like the word *coach* [$k^{h}o^{\upsilon}\check{c}$] or *twine* [$t^{h}wa^{ɪ}n$], all the phonetic symbols are supposed to **represent simultaneously both phonetic states** (since they are nonoverlapping, serially ordered segments) **and also dynamic articulatory or acoustic events**, since each segment represents a complex gesture many of which must overlap each other. Amazingly, the paradox built into this blended notion of segment as both a letter and a gesture has not been seen to be a theoretical problem. Why not? Because we all learned to `just get over' these paradoxes when we learned how to read in childhood.

The proposal here is that during historical times, and increasingly for the past 3 millenia, some human communities have exploited static, spatial and graphic models for speech. This spectacularly successful set of notational conventions transformed quasi-continuous, overlapping, time-distributed and highly variable speech sounds into conventionalized, discrete, ordered graphic tokens where each word has a single spelling. This blended representation has the effect of freezing many of the degrees of freedom in speech acoustics and speech gestures. As long as one looks at a single language at a time, letters conveniently provide a specification that is detailed enough for practical indication of how a speaker should pronounce a word. The rapid and variable movements and complex time patterns are converted to a series of graphic marks that stand still indefinitely. But if we ask what the speaker's patterns for controlling motor gestures are like or what the listener's perceptual units are, we will not be able to find identity (or even much similarity) with phones or phonemes. Motor commands and the speech perception systems have quite different constraints than apply to writing language down on paper. It is clear that non-overlap, context independence and strict serial-order may be useful for writing because these properties provide orderly presentation and reduce the number of visual distinctions required. But they are not going to do what needs to be done either for speech motor control or for speech perception.

The ability to describe and understand human speech in terms of such graphic tokens was first achieved by the Phoenicians and Greeks. It has continued to provide a convenient description of speech and an influence on all literate people in the western cultural tradition up to the present day. The question is to what degree our ``native-speaker intuitions'' about the nature of human speech are influenced by our practical skill using alphabetic writing.

Biassing Intuitions.

How exactly could alphabetic writing influence our perceptual intuitions about speech sounds? It could have worked like this. First, languages do have phonological systems in the sense that words tend to differ from each other within the language in a very restricted set of ways. This creates a situation where a small inventory of graphic tokens might adequately represent many languages well enough for a speaker to identify most words. The restricted variations can be illustrated by starting with the English word

block or [blɑk]. We can minimally change it to possible (but nonexistent) words like *brock* or *glock* or *slock*, but not to *dlock* or *mlock*. And commuting the vowel can yield real words like *bleak, Blake, bloak, black*, etc. But there is a maximum number of possible vowel contrasts in the context *bl_k* similar to the maximum number in other contexts (like *h_d* as in *heed, hid, head, had,* etc.). Presumably our human brains are equipped with the necessary competence to produce and recognize rapidly produced words defined in this way. Of course, these capabilities do not imply that we have any conscious awareness whatever of the components used – any more than we are aware of how we walk, stand on one foot or ride a bicycle. Nevertheless, the restricted options in phonological differentiation mean that a very limited inventory of symbol tokens *could* succeed in keeping words distinct within any language. That is to say, looking just within a language, only a small set of distinctions in speech sounds need to be maintained.

The development of the technology of alphabetic writing about 3 thousand years ago tamed the complex gesture space by focusing on static points, or extremes of gesture, that is, on `targets'. Of course, static targets do not always work because many so-called ``speech sounds'' are movements or complex sequences of gestures. But this problem can be dealt with by going ahead to assign affricates, glides, semivowels, etc, to single letters anyway. The inventors of the alphabet exploited the fact that the sound spelled with letter ‹ l › in *block* is quite similar to (if not exactly the same as) sounds in *ball, pillow, slap, lick,* etc. A context independent similarity was noted and represented graphically with a serially ordered symbol. This context-independent use of the graphical‹ l › symbol which suppressed many differences in articulation greatly simplifies the problem of noticing and remembering the serial order of tokens. Comparing, say, the purely auditory impression of our pronunciation of the word *pearl* with *pillar*, *Prell* and *plural* (which could be written phonemically as [pəˠl̩, pɪləˠ, prɛl, pləˠl̩]) must have been very difficult without the use of alphabetic writing to keep track of the sounds during attentive listening. But it is not so hard once you can write the words down and look at your spellings. Without an alphabet to draw our attention to commonalities and serial order differences, the sound of spoken multisyllabic phrases, such as *What the hell are you doing?* (which might sometimes sound like [wət̪ðɛləˠjɪˈdʊ̃n] and sometimes more like [t̪ˠðɛjɪˈdʊ̃n), for an illiterate (or, more precisely, any non-alphabet-literate person) will be largely an auditory blur as far as conscious awareness goes. Too many sounds happen too quickly and every word or phrase tends to be pronounced in many so-called ``linguistically equivalent'' variants[5]. But once a phonological writing system based on recording the sequence of critical articulatory states is conventionalized, each word acquires a standard form and it becomes possible to write down consistent spellings ignoring all the incidental variations of detail in how they are pronounced.

[5] What does `linguistically equivalent' mean here? Apparently it means only that if we were writing the text down orthographically, we ought to use the same `canonical' or `lexical' forms for all the variants. That is, it means the variant pronunciations are orthographically the same. Any other interpretation is strictly speculative.

Thus writing a language with an alphabet imposes a conventional structure on the sound system. We shoehorn the language into the letters and typically ignore whatever doesn't fit. It greatly helps us as scientists of language to be able to lean on a visual and spatial representation of words – even if it means we must ignore all temporal properties and phonetic details (see Port and Leary, 2005). The point is that, on becoming literate, we gain a tool, not just to supplement memory for specific utterances, but also a tool for two other important roles: for regularizing and standardizing (thus simplifying) the language and also for paying attention to the sound of words. With cultural experience and personal cognitive development, the letters on paper come to be the psychologically natural units for description of words. During our years of early schooling, the method for representing words graphically is thus gradually internalized. Alphabetic notation of speech, once learned, becomes so natural and convenient that it comes to be the normal way for a literate person to think about and talk about speech sounds.

Given the pervasiveness of alphabetic representations of speech, it seems fair enough to say that literate adults really do actually ``hear letters'' – not because speech sounds really are formally equivalent to letters (as we have mistakenly thought for over a century) but because we are skilled at thinking of speech sounds using the blended cognitive scaffolding provided by letters. As Öhman points out (Ohman, 2000), the phoneme was invented, not discovered. The phone *is* a kind of letter and needs to be understood as an invention. The use of letters for speech depends on negotiating a compromise between properties of speech sounds (sufficient to suggest the right pronunciation) and the properties of graphic images (which must be serially ordered and visually distinctive). This invention, refined by Phoenicians and Greeks from over a thousand years of previous experiments with graphical representations of language, is a wonderfully simple system for storing speech. Once the alphabet idea was established, there was a straightforward way to engineer a writing system for almost any language and also to make available a cognitive scaffold for paying attention to speech sounds. The disciplines of linguistics and phonetics were probably inevitable eventually as soon as the alphabet was firmly established since the necessary scaffolding was available.

So whatever unit-like sounds might be composed to make the words of a language, they need to have *some* of the properties of letters (since occasionally permutations are possible – viz. *tan, Nat, ant* or *pat, tap, apt*) but the cognitive units should lack the complete permutability of letters. The conviction that, nevertheless, human speech consists of strings of letter-like segments which so many linguists and phoneticians (including this author) have found persuasive is not based on the actual psychological structure of auditory or articulatory patterns. The powerful impression we have of serially ordered static tokens reveals more about our bias to lean on visual and spatial metaphors for speech whenever we find a way to do so, than it reveals about the speech sounds themselves. Without noticing our cognitive tricks, we have continued to mistake our supplementary cognitive scaffolding for the inherent structure of language.

Now if the phoneme is a consequence of, rather than the explanation for, the written alphabet, one might ask, what actually happened when scientists thought they ``discovered'' the phoneme around a hundred years ago? What happened, it seems to me, is that many European language scientists began to look for theoretical underpinnings for linguistic capacity. They sought to incorporate Peirce's notion of the *sign* into a psychological account of language (Saussure, 1916). Languages, they

supposed, have meaningless sound units (rather like letters) because the sounds are *signifiers* for some *signified information* (the meaning of the word). The pioneers of modern linguistics had begun the long (but still incomplete) progression toward a psychological theory of language by interpreting letters as models for hypothesized cognitive tokens. Conceiving letters as theoretical objects, they tended to dismiss the physical properties of graphic letters as incidental rather than essential. The role of a penlike tool applied to a paperlike medium in the use of the alphabet was ignored or overlooked. There must be objects in the mind, they thought, that are somehow just like letters only not graphical. Just as writing represents the spoken word for a reader, the phonemic spellings were hypothesized to represent words to the mind. The neologism *phoneme* was adopted to describe these hypothetical objects that would have all the invariances of letters but are written in the mind, not on paper. Both phones (or, as they are often called, speech sounds) and phonemes are derivative concepts but linguists (and others with education based in European languages) are so thoroughly practiced in their use, they have become second nature and so highly salient to us that they overwhelm any other possible perceptual description.

Phoneticians and linguists found phonemes very compelling. Indeed, once proposed, there was fairly rapid acceptance of the phoneme as the intrinsic sound unit of language. As noted by Twaddell in his famous review in 1935, phoneticians (e.g., Jones and Sievers) as well as linguists (e.g., Troubetzkoy, Jakobson and Bloomfield) endorsed the phoneme but there were some major differences in what they thought it was (Twaddell, 1935). Most researchers in the era between the world wars interpreted the phoneme as a psychological or mentalistic concept (Troubetzkoy, Sapir, Jakobson,), that is, as an ``intention'' of the speaker or ``auditory impression'' of the hearer. A few, like Bloomfield insisted there were actual physical commonalities to the various variant sounds that belonged to a phoneme even though it was not yet known just what they were (Bloomfield, 1933). Twaddell evaluated both psychological theories and the theories claiming common physical properties, dissecting and rejecting both in favor of the view that a phoneme is merely a convenient fiction. I must confess this conclusion annoyed me when I first read this paper as a graduate student. I was confident that a psychological account would eventually succeed, but today his conclusion seems honest and insightful. It seems now that the reason for the inability to find a satisfactory definition of the phoneme by our linguistic forebears is that it is actually a chimera. It is a compelling blend of graphical and auditory-articulatory properties.[6]

---

[6] Despite their intuitive persuasiveness, phonemes are surprisingly difficult to be explicit about. For example, how many phonemes does any particular dialect of English have? Amazingly, few if any phonologists will make a claim about this for any dialect. And they rarely argue about such an issue. It doesn't seem important. We all just *know* that English (and any dialect thereof) has an integer number of phonemes – whether or not we know what that integer is. There just *has* to be a simple answer to the question of the phoneme list, just as there is a simple answer for how many distinct letters are employed in this essay. For letters, you just make a list and count them. Printed dictionaries, of course, must take a

This hypothesis is radical but it is not new. The notion that phonemes are somehow profoundly dependent on letters has been raised several times over the years. I am aware of J. R. Firth (Firth, 1948) who decried the ``apotheosis of the sound-letter in the phoneme'' and complained that ``the roman alphabet has determined a good deal of our phonetic thinking in Western Europe." More recently, Alice Faber (Faber, 1992) suggested that phonemic segmentation was a epiphenomenon resulting from our familiarity with alphabetic writing. Even more recently Sven Öhman made arguments similar to those of this paper claiming that ``the so-called `segmental principle' must be regarded as a principle governing the structure of alphabetic writing … rather than speech itself'' (Ohman, 2000). Peter Ladefoged has also expressed suspicions along this line, suggesting that ``accounts of human behavior in terms of phonemes are nearly always examples of what has been called the psychologist's fallacy – the notion that because an act can be described in a given way that it is necessarily structured in that way. As far as I can see, phoneme size units play only a minor role in … normal speaking and listening''(Ladefoged, 1984). There may well be others who have expressed concern about the confusion of alphabets and orthography with linguistic structure.

Some Evidence Against Segments as Basic.

If this radical story is on the right track, there should be plenty of evidence and there is. Port and Leary (2005) review a variety of technical arguments against any formal analysis of linguistic sound systems, but in this essay, only a few common-sense arguments will be reviewed. In fact, I am aware of no evidence providing clear support for segments as the primary or exclusive units of speech.

The first kind of evidence is something that should have been obvious all along: the massive over-generation of strings that is implicit in phonemes. Letter-like tokens are inherently perfectly commutable. Graphic symbol tokens can obviously be permuted without limit, just as beads can be put in any order on a string. This affordance is intrinsic in their simply being beads or in their ``letter-nature.'' Indeed, if you can write the string *left*, you can also write *felt* with the same tokens reordered, and *flet* (not a word but it is similar enough to *fleet, flit*, etc. that it probably *could* be a word). Unfortunately, further permutation just as easily yields *tfel, ftel, lfte*, *etfl*, *letf*, etc. and none of these could possibly be words in English.

---

stand on this issue and many, for example, employ different numbers of phonemes for English. Usually between 44-47. What is the correct phonemic spelling of an arbitrary English word? There are an amazing number of ambiguities. For example, take the word *spear* in my dialect. Is it /spɪəˑ/, /spir/ /sbɪr/, /sbir/ or something else? Again, phonologists have taken various points of view, but actually settling the issue of cognitively correct spellings is not taken very seriously within phonology. Why isn't there an obvious and unambiguous answer to any phonemic spelling question, such as there is for orthographic spellings? I think the reason is that we still do not have a definition of phoneme that tells us clearly how words should be spelled – the very problem that concerned Twaddell in 1935.

How serious is the problem of over-permutation of phonemes? Using the Webster Pocket Dictionary of English which employs 46 phoneme symbols, we can ask how many strings of, say, 5 or fewer phonemes are possible (allowing reuse of letters for cases like *mama*). The answer is about 228 million.[7] How many of these 1-to-5-phoneme sequences are actual English words? The Webster's Pocket Dictionary lists only about 8,000. That is, roughly 2 out of every 100,000 possible permutations is an actual word. If we include longer words, the problem gets far worse very quickly. For example, there are almost 10 billion 6-phoneme permutations but the dictionary lists only about 3,000 words with 6 phonemes – roughly one word for every 100 million possibilities. Postulating phonemes as context-independent tokens is obviously far too strong a hypothesis to entertain seriously for human cognition. Yet we do more than simply entertain this hypothesis, we can hardly think about speech sounds in any way other than using precisely this model!

This awkward property – the fact that only an infinitesimal fraction of the permutations of the letters used in standard spellings are actual (or even possible) words – is inherited by the phoneme from its graphical prototype, the letter. Both letters and phonemes are completely permutable in principle even though speech segments in any language can be permuted only to a minute extent. Of course, the standard way to address this obvious problem is to immediately divide the alphabet into subtypes, such as consonants and vowels (or consonants, vowels, nasals and semivowels, etc.), and to talk about various ``phonotactic constraints'' on their sequencing. According to Chomsky, the entire mission of grammar construction is to constrain overpermutation. But we will need more and more subcategories of segment types and complex statements of constraint. In the end we still will not be able to distinguish what is a so-called ``possible but nonexistent'' English word (like *flet*) from a so-called ``impossible'' word (like *ftel*). This serious problem is entirely a consequence of the assumption that letters (and their cognitive analogs phonetic segments) are the basic units of speech. The right way to solve the problem is to throw out these segments and seek units that will be far larger in number but are entire gestures and include only the patterns that actually occur in the language (Port & Leary, 2005). Of course, if the units are only fragments of patterns observed, another problem is raised: what about the fact that people can still recognize and even invent novel words? These abilities are presumably based on statistical similarity to the set of existing words in memory but described in terms of non-minimal gestural and auditory components (Pierrehumbert, 2003).

The second kind of evidence against segments arose from the early data on speech perception. From studies of speech spectrograms and experiments on speech synthesis beginning in the 1950s, it was found that the cues for speech segments were encoded in a way that prevented context-free specifications for individual segments (A. M. Liberman, Delattre, Gerstman, & Cooper, 1968). This was described as the `coarticulation problem' and led to notions like the `motor theory of speech perception' Other early research on

---

[7] Using an alphabet of n tokens combined in ordered sets of k tokens that include reuse, the rule is that the total number of strings is n^k. So with an alphabet of 46 and word size of 2, the total number of possible words (assuming the complete permutability implied by context-free letters) is $46^2 = 2{,}304$.

speech cues showed the importance of timing patterns distributed over entire syllables for the specification of segmental features like consonant voicing (Lisker, 1957; Lisker & Abramson, 1971; Port & Leary, 2005). These problems are consequences of trying to match up actual speech gestures and acoustic trajectories with segment-sized, context-invariant units.

A third kind of important evidence against phone or phoneme-sized units is found in the difficulty in learning to read using a phoneme-based spelling system that is experienced by a significant fraction of normal, intelligent children educated in alphabetical cultures. The term ``phonemic awareness'' has been employed since the 1970s to describe the ability to perform segment-dependent tasks like adding an initial consonant to a word (e.g., to change *no* into *snow*) or a final consonant (changing *bye* into *bide*) or deleting a consonant (changing *pant* into *pat*) (Hempenstall, 1997; I. Y. Liberman, Shankweiler, Fischer, & Carter, 1974). It is now well accepted that performance on phonological-awareness tasks is a very good predictor of reading ability in children. Good readers find these tasks easy and poor readers find them very difficult indeed (Lyon, 1998). In addition, adult illiterates have been found to do very poorly at such tasks (Morais, Cary, Alegria, & Bertelson, 1979) and adult literates who know only nonalphabetic writing (such as Chinese who have not been taught a romanization of Chinese) also perform very poorly (Cheung & Chen, 2004; Read, Zhang, Nie, & Ding, 1986). The evidence is strong that there is nothing natural or inevitable about alphabetic writing. Learning to read apparently does not depend on ``becoming aware'' of the true phonemic nature of our cognitive linguistic representations. It requires learning to impose letters uniformly on whatever sound and gesture structures languages actually use. Linguists and phoneticians have been imposing phonemes and phones on languages for a couple centuries now without noticing the poorness of the fit.

Finally, phonological games designed to obscure speech, along the lines of pig-Latin, are found in many languages. But they depend most often on insertion or deletion of syllable-sized units. Games are sometimes based on half-syllable units (like pig-Latin) but segment-based games are found only in alphabet-literate cultures (Botne & Davis, 2000). This is further evidence from folklore that segments or phones are a culturally derived imposition.

Altogether, one finds that segmental descriptions of speech, whether phonemic or phonetic, are very useful and are not so arbitrary as to be unlearnable. However they are certainly not the only way to describe speech.

Is this the end of linguistics as we know it?

For some readers, this paper may seem deeply skeptical. But its goal is only to clear up some of our biases about the interpretation of speech and language. Phones and phonemes are excellent units for describing speech if you have been trained in alphabetic reading and writing – which happens to be true of all of us language professionals. These units are useful for talking and thinking about speech (so the IPA alphabet will naturally continue its central role in academic communication), but the vividness of our intuitions about segments must not be taken as evidence that languages really are based on alphabetic units. On the other hand, of course, this story about the role of alphabetic literacy is not evidence, in itself, **against** the possibility of letter or segment-sized units in any particular language either. What units are employed is an empirical question that can

be answered by the study of phonological systems as manifested in the behavior of speakers of each language. Some questions can be answered by conventional auditory transcription (as long as the potential bias involved in using any segments is kept in mind) but many descriptive issues will require experimental studies of various kinds (see Port and Leary, 2005, for further discussion). Exactly what the phonology of the future will look like will have to grow out of attempts to provide motivated descriptions of specific languages and specific analytic problems.

Conclusions

The primary goal of this paper has been to explore the reasons why segmental descriptions of language are so compelling and satisfying to us. The argument is that speech consists of inherently difficult patterns for humans to attend to. The relevant bodily movements are mostly impossible to observe directly and both the movements themselves and the resulting auditory patterns are, in any case, very rapid relative to the capabilities of our conscious attention. Despite these difficulties, our cultural ancestors were able, over many centuries, to develop a graphical notation system for representing the sound patterns of words that was easy to learn. Alphabetic letters seem to be well characterized by the 19th century semiotic notion of a Sign, as described by Peirce and Saussure: a graphical token conventionally (and arbitrarily) associated with a property of speech. In the past century linguists and phoneticians have been able to describe the properties in primarily auditory or articulatory terms. A writer deploys the graphic tokens and a reader interprets the distinct ordered tokens in terms of speech sounds. Alphabetic writing systems eventually conquered most of the world and were adopted by the majority of the world's language communities. With the development of printing, more sophisticated reading skills became possible where larger patterns of letter groups and whole words and phrases might be recognized as perceptual units.

Phonetic segments as a blend. A development of quite a different sort came in the late 19th century when language scientists began to suspect that something analogous to letters might underlie human language as it is processed ``in the mind.'' Thus the notion of the ``phoneme,'' and roughly simultaneously, the more detailed and language-independent unit, the ``speech sound'' or ``phone,'' seem to have been nearly universally adopted in the scientific community within a generation. Now, a century later, nearly all linguistic theories are still predicated on the notion of segments of speech sound, known as either ``phonemes'' or ``phonological segments.'' And most phoneticians depend on ``phones'' serving as fundamental descriptive units for speech. What was overlooked when graphic letters were recast and transformed into psychological phones and phonemes was that in psychologizing the letter, they psychologized the spatial and visual properties of letters as well. Phones, it was thought, could be assumed to be discrete and serially ordered – an assumption that has had tremendous importance for theories of language in the 20th century. Thus one could, for example, always count how many segments were in a word (as when saying that *mop* has 3 segments), or assign each segment to one or another syllable (as in, e.g., describing *breakfast* as [brɛ́k-fəs]). The discreteness of phonemes justified the assumption that words too must be discrete since they are spelled from phonemes (and just as words are discrete and countable in a printed text since orthographic words are defined by sequences of letters). Also, of course, their discreteness justified assuming they could be parsed into discrete components, such as

the distinctive phonetic features of Chomsky and Halle (Jakobson, Fant & Halle, 1952; Chomsky and Halle, 1968; Port & Leary, 2005).

Phones and phonemes were a new conceptual blend and a hypothesis about human linguistic cognition. On one hand they are defined by associated articulatory or auditory properties, like *voiced, labial, glide, sibilant,* etc., but on the other hand, they retain key properties of letters, such as being *static, serially ordered* and *discretely different* from each other. There is nothing wrong with blending different spaces, but, if we wish to construct cognitive theories on their basis, it is important to understand when we are doing it. In order to understand the actual psychological properties of human speech we need to let go of the segment as the intrinsic organizing unit of language.

Consequences of Literacy. In addition to providing half of the conceptual blend of the phoneme as letter and as hypothetical cognitive `symbol', there were, of course, other consequences of literacy for history and cultural development. First and most importantly, of course, the development of alphabetic writing was a technology that made it possible to store specific utterances for indefinite periods of time. Written language, in turn, supported critical thinking leading to philosophy and eventually the sciences.

But there is a less obvious consequence of alphabetic writing that would be important for the eventual development of a science of language. Conventional spellings hide from us readers the wide variations in language between various contexts and between speakers. A standardized visual form of each language was established for various speech communities. The importance of this for the development of a linguistic science is that the existence of a standard written form for each word makes the social invariant visible (just as the approximate sound units of speech are made visible by writing). A literacy convention thus offers implicit support for the speculation that there might be an invariant abstract form for each word in a language shared by all speakers in the community. That is, an orthographic standard seems to endorse or authorize the possibility of a cognitive standard. All of us who speak, read and write in, say, English agree on how various words are spelled and printed. Maybe that implies we also agree on a common cognitive structure, that is, a common grammar. So the proposal that there is a single unitary linguistic structure, a grammar, with independent existence in the mind of each speaker of the language may seem like an assumption that is almost a sure bet. Such a unitary notion is central to modern linguistic thinking, and is a train of reasoning that seems to plays a large role in the Chomskyan school of linguistics. Thus, it seems likely that the vividness of our intuitions about phones ultimately provides the rationale for the notion of a shared grammar of discrete components.

Finally, there is a third consequence of alphabetic writing directly relevant to the goal of this paper. Writing technology itself provides a scaffolding that enables the scientifically inclined with an interest in the process of speech production and perception (such as linguists, phoneticians, psychologists, language teachers, etc.) to focus their attention on speech in a practical, organized way. Such a notational system, especially if modified to be more consistent and rational, makes it possible to think carefully and objectively about the rapid and mysterious sounds of human speech. The notation developed by the International Phonetic Association around a century ago provides a visual model for speech where time is converted into a spatial axis. By interpreting the letters in terms of articulatory or auditory labels (e.g., ``a /b/ is a bilabial, voiced obstruent''), some aspects of the structure of phonological systems can be exposed to

scientific scrutiny. Research on language during the past hundred years has employed these conceptual tools and put them to good use. At the same time, of course, there is much other structure in speech sounds that has been overlooked or ignored within linguistics although somewhat less ignored by phoneticians (Port & Leary, 2005). But now many researchers in language science are beginning to see the limitations of these conceptual tools and trying to develop new methods, based on new theoretical frameworks, that can take us further toward understanding the remarkable linguistic abilities of our species. Jim Flege's work revealing the unusual skills of those who speak multiple languages has contributed to a more complete understanding of the true richness of human speaking skills.

References

Bloomfield, L. (1933). *Language*. New York, New York: Holt Reinhart Winston.

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic Linguistics*. Cambridge, Massachusetts: MIT Press.

Botne, R., & Davis, S. (2000). Language games, segment imposition and the syllable. *Studies in Language, 24*, 319-344.

Cheung, H., & Chen, H.-C. (2004). Early orthographic experience modifies both phonological awareness and on-line speech processing. *Language and Cognitive Processes, 19*, 1-28.

Chomsky, N. (1964). Current Issues in linguistic theory. In J. Fodor & J. Katz (Eds.), *The Structure of Language: Readings in the Philosophy of Language* (pp. 50-118). Englewood Cliffs, New Jersey: Prentice-Hall.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again.* Cambridge, Mass.: Bradford Books/MIT Press.

Clark, A. (2004). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford, England: Oxford University Press.

Cooper, F., Liberman, A., & Borst. (1951). The interconversion of audible and visible patterns as a basis for research on the perception of speech. *Proceedings of the National Academy of Science, 37*, 318-325.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149-179.

Faber, A. (1992). Phonemic segmentation as epiphenomenon: Evidence from the history of alphabetic writing. In P. Downing, S. Lima & M. Noonan (Eds.), *The Linguistics of Literacy* (pp. 111-134). Amsterdam: John Benjamins.

Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities.* New York, New York: Basic Books.

Firth, J. R. (1948). Sounds and prosodies. *Transactions of the Philological Society*, 127-152.

Flege, J., & Hillenbrand, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *Journal of the Acoustical Society of America, 79*, 508-517.

Flege, J., Munro, M. J., & Skelton, L. (1992). Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin and Spanish. *Journal of the Acoustical Society of America, 92*, 128-143.
Flege, J., Munro, N., & Mackay, I. (1995). Effects of second-language learning on the production of English consonants. *Speech Communication, 16*(1), 1-26.
Goody, J. (1977). *Domestication of the Savage Mind*. New York: Cambridge University Press.
Goody, J., & Watt, I. (1968). The consequences of literacy. In J. Goody (Ed.), *Literacy in Traditional Societies* (pp. 27-68). New York, New York: Cambridge University Press.
Grossberg, S. (2003). The resonant dynamics of speech perception. *Journal of Phonetics, 31*, 423-445.
Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*, 373-405.
Hempenstall, K. (1997). The role of phonemic awareness in beginning reading: A review. *Behavior Change, 14*, 201-214.
Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, Massachussetts: MIT Press.
IPA. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, England: Cambridge University Press.
Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to Speech Analysis: The Distinctive Features*. Cambridge, Massachusetts: MIT.
Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence, 73*, 31-68.
Kuhl, P., & Iverson, P. (1995). Linguistic experience and the ``perceptual magnet effect''. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.* (pp. 121-154). Timonium, Maryland: York Press.
Ladefoged, P. (1984). `Out of Chaos comes order': Physiological, biological and structural patterns in phonetics. In M. P. R. V. d. Broeke & A. Cohen (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences* (pp. 83-95). Dordrecht: Foris.
Lakoff, G., & Nunez, R. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York, New York: Basic Books.
Liberman, A. M., Delattre, P., Gerstman, L., & Cooper, F. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J. Experimental Psychology, 52*, 127-137.
Liberman, A. M., Delattre, P., Gerstman, L., & Cooper, F. (1968). Perception of the speech code. *Psychological Review, 74*, 431-461.
Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation of the young child. *Journal of Experimental Child Psychology, 18*, 201-212.
Lisker, L. (1957). Linguistic segments, acoustic segments and synthetic speech. *Language, 33*, 370-374.
Lisker, L., & Abramson, A. (1971). Distinctive Features and Laryngeal Control. *Language, 44*, 767-785.
Lyon, R. (1998). *Overview of Reading and Literacy Initiatives*: Report to Committee on Labor and Human Resources, United States Congress.
Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7*, 323-331.

Ohman, S. (2000). Expression and content in linguistic theory. In M. Gustafsson & L. Hertzberg (Eds.), *The Practice of Language*. Dordrecht: Kluever.
Ong, F. J. (1982). *Orality and Literacy:The Technologizing of the Word* (1st ed.). London: Routledge.
Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probability Theory in Linguistics*. Cambridge, Mass.: MIT Press.
Port, R. (2003). Meter and speech. *Journal of Phonetics, 31*, 599-611.
Port, R., & Leary, A. (2005). Against symbolic phonology. *Language, to appear*.
Read, C., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition, 24*, 31-44.
Saussure, F. d. (1916). *Course in General Linguistics* (W. Baskin, Trans.). New York: Philosophical Library.
Twaddell, W. F. (1935). On defining the phoneme. *Language, Language Monograph 16*.
Watson, C., Kelly, W., & Wroton, H. (1976). Factors in the Discrimination of Tonal Patterns II: Selective Attention and Learning Under Various Levels of Stimulus Uncertainty. *Journal of Acoustical Society, 60*, 1176-1186.
Werker, J., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49-63.