# All is prosody:
# Phones and phonemes are the ghosts of letters

*Robert F. Port*

Departments of Linguistics and Cognitive Science
Indiana University, Bloomington, Indiana, USA
www.cs.indiana.edu/~port

## Abstract

The standard way to represent language within linguistics is using letter-like units, i.e., with consonant and vowel phones and phonemes. But this representation, despite its powerful intuitive appeal is not supported by any experimental data. There are 4 straightforward predictions implied by this model. All are shown to fail. One of the main problems is the inability of segments to permit representing actual values of time. One realtime property of speech is the periodically produced speech found, e.g., in song and chant. Several audio clips of spontaneously produced rhythmic speech will be analyzed and speaker performance on the related laboratory task of speech cycling will be reviewed. In this phrase-repetition task speakers quickly adopt nested periodic timing patterns that are surprisingly rigid in time. Letter-based descriptions of speech make such rhythm invisible and irrelevant to anything `linguistic'. This is further evidence that phones and phonemes are only categories of patterns in the speech of a community, and not the psychological symbol tokens employed in the realtime production and perception of speech.

## 1. Phones and Phonemes are Letters in Disguise

The conventional idea about the representation of language in the mind by professional linguists and psychologists as well as by laymen is based on letters – either orthographic letters or letters in a technical alphabet such as the IPA alphabet. For over a century scientists have assumed that speakers employ phones and phonemes, the psychological counterparts of letters, as the cognitive units of linguistic perception, production and memory. These units are intuitively compelling and seem guaranteed to be the units from which English and all other languages are constructed. Still, despite their intuitive naturalness and their undeniable efficiency, and despite their consistent endorsement by the fathers of our field (Saussure, 1916; Jones, 1918, p. 1; Jakobson, Fant and Halle, 1952, Ladefoged, 1972; Chomsky & Halle, 1968; IPA, 1999), there are many strong reasons to be suspicious of letter-like units as the form that words embody in memory. Only a few have dared to doubt them (Twaddell, 1935; Faber, 1992; Browman and Goldstein, 1992; Port & Leary, 2005). Phones and phonemes are always assumed to be invariant across adjacent contexts (so the same consonant occurs in *do* as in *day* and same vowel in *red* as in *wreck*), abstracted away from any speaker's voice and across speaking tempos and the segments are always discretely distinct from each other (thus making reliable identification possible). And most problematically, the segments exist only in discrete time (like integers) not in continuous time (like real numbers).

The goal of this essay is to show that the realtime processes of language, that is, speech production, perception and memory, do not, in fact, rely on such phone or phoneme-like representations. Instead information about words and the rest of language is stored and processed in a much richer and more concrete auditory form. And the patterns that are learned are all defined in continuous time. On the other hand, the structures of phonologies (e.g., the phones, phonemes, etc.) are category structures that are created at the level of a social institution, as part of culture. Phonology and grammar exist at the level of a culture and evolve over historical time as a product of social behavior. (Whereas realtime processes evolve on the developmental timescale. You learn it and pass it to the next generation.) A linguist looks across a corpus of utterances by a group of people and describes what the language is like. This is what linguistics should be studying, the phonologies and grammars of linguistic communities, using whatever descriptions do the job. But the current theoretical basis for research is misguided since linguistic structures are *not*, in fact, cognitive ones. A phonology is generally not a psychological structure, something ``stored'' or ``represented'' in the brain. The social patterns of phonology play almost no role in real time language use, except when someone has been given literacy training. The generalizations of phonology are about statistical properties of the linguistic corpus of some community. So a `theory of phonology' is a theory about a social institution, and not a theory about human cognition. If words are stored using abstract, discrete tokens like vowels and consonants, as claimed by most language scientists, then many straightforward predictions should follow. For example, it should be predicted that:

(1) both synchronic variation and diachronic sound change should always exhibit discrete phonetic jumps when one segment type is replaced with a different one or when some discrete feature is changed. Of course, adding noise could make discrete differences harder to observe, but evidence should be found somewhere.

(2) Each segment or distinctive feature should have a single invariant physical specification across all its contexts

(according to ``the invariance condition'' of Chomsky and Miller, 1963). And the sequence of physical cues should agree with the sequence of segments (Chomsky and Miller's ``linearity condition''). This link to the physical world is essential for linguistics to be an empirical science. Chomsky understood that one cannot obtain physical measurements from a phonetic transcription, but the most raw data of any science needs to be physical.

(3) Our memory for utterances should show evidence that the stored segmental representations are invariant across speakers. This implies that the memory representations of a word spoken by, e.g., Bill would be the same as the representation of the same word spoken by Mary.

(4) Since segments can only exploit serial position to represent time, there should be no temporal effects observed in languages that cannot be described in terms of ordered segments.

These are all simple consequences of assuming that the acoustic signal of speech contains units directly analogous to letters and that these letter-like units are the basic representational tokens for language in memory. But unfortunately, as will be shown, *none* of these expectations are upheld. In many cases the counter-evidence has been around for many decades so that a variety of attempts have appeared over the years to explain away some of these problems. There is not enough space here to consider all these attempts although many have been addressed elsewhere (e.g., Port & Leary, 2005). My hope is that readers will attend to all arguments together and look objectively at the big picture. The traditional theory clearly claims that an abstract, canonical representation of each word (consisting of a segmental transcription using consonants and vowels) is used by speakers for recognizing and remembering what someone said. After all there is no alternative representation conceived of within linguistics.

## 1.1. Prediction 1: Discrete phonetic jumps are observed synchronically and diachronically

As for this prediction, generations of literature from experimental phonetics have shown continuous variations in parameters like voice-onset time, vowel quality, consonant place of articulation, the duration of vowels and consonants, etc. (e.g., Peterson & Lehiste, 1960; Lisker and Abramson, 1964, 1967; Local, 2001). Phonetics research over the past century offers no suggestion that the phonetic space is discretely differentiated as the widely used Chomsky and Halle (1968) theory of phonetics presumes. It is sometimes claimed (citing Lisker & Abramson, 1964) that voice-onset time (VOT), the time lag between release of a stop consonant and the beginning of voicing, shows only 3 (or 4) values (Chomsky & Halle, 1968; Stevens, 19nn), but this is a misinterpretation of their data. They display a histogram of VOTs of word-initial stops across 10 languages that does indeed show 3 clear modes which they label `prevoiced,' `short-lag' and `long-lag' (or aspirated). But the data for specific languages show considerable variation in their apparent VOT target values even within the phonetic category of aspirated. Thus the mean VOT found by Lisker and Abramson for the English long-lag [$k^h$] was 80 ms (ranging from 55-135 ms) while Korean [$k^h$] was 126 ms (ranging from 85-200 ms). Most linguistic targets differed considerably across languages. Furthermore, their experiment looked only at word-initial cases, but VOT varies depending on the position of a syllable in a word, on speaking rate, on the following vowel or consonant, etc. (e.g., Lisker & Abramson, 1967; Port & Rotunno, 1979). So there is no support here for any claim that voicing categories are universally discrete.

Of course, Stevens (1989) has pointed to a few places where there are flat spots in the relation between articulation and acoustics. It may be that some languages exploit these regions since inaccuracy in articulation would result in less acoustic variation here than in neighboring regions. But these nonlinearities do not begin to support a claim that the phonetic space is fully discrete as claimed by Chomsky and Halle (1968). Nor is there any reason to suggest that the phonetic space is closed or limited in size. There is no reason not to expect languages to occasionally invent new sounds that have never occurred in any language before. Would anyone claim that there is a fixed inventory of sound patterns that a violin can produce? I doubt it. Why would anyone suppose that the human vocal tract could be limited to any fixed set?

Furthermore, students of dialect variation (e.g., Labov, 2001; Bybee, 2001 and Local, 2007) have also found no hint of discreteness and no limit to the variety of regional accents, foreign-accented pronunciations and speaker idiosyncrasies. Many people can imitate these accents in their own speech. It is widely appreciated that all speakers vary their own pronunciations along many phonetic continua depending on various subtleties of social and pragmatic context (Foulkes & Docherty, 2006). How could these minute phonetic differences be employed in production and perception if we stored language using only a very low-dimensional phonological or phonetic representation based on a small number of distinctive phonetic features, or if we have only limited linguistic control of our vocal tracts? Since speakers recognize many subtle phonetic and temporal differences and can control many of them in their own speech, they obviously must have a way of perceiving and representing these variants. It seems obvious that speakers have richly detailed phonetic memories for speech, not the supposedly ``economical'' ones using a very small set of letter-like allophonic units.

## 1.2. Prediction 2: Each phonetic unit has its invariant acoustic cues.

As for the prediction that there will always be invariant acoustic specifications for phonetic or phonological units (whether segments or features), almost all of the experimental observations related to speech perception over the past century fail to support this prediction. It is known that almost all speech cues are highly context-sensitive rather than context-free. And many cues overlap in time and may be distributed widely across the speech stream (e.g., Lisker, 1984; Hawkins, 2004). It has long been known that cues for place of articulation and other features are very different depending on the specific consonant and vowel neighbors of any segment (Liberman et al, 1968). Thus the ``problem of coarticulation'' seemed to be formidable: how can people hear speech in

terms of nonoverlapping, context-free segments when the auditory stimulus exhibits a great deal of temporal overlap and context-sensitive variation that seems to be inaudible (in the sense that the variations are inaccessible to consciousness, Liberman, et al., 1968; Kent and Minifie, 1977).

But it appears that the coarticulation problem, as an aspect of speech perception, exists primarily for the conscious experience of language by literate people – people who spent hundreds of hours a year for decades perfecting their skills at translating speech into letters and letters into speech (Port, 2006). Research on the ``phonological awareness'' of illiterates (both preliterate children and adults lacking alphabet-literacy education) has shown that the segmental intuitions that allow us to easily add or remove a single segment from a word are found only in those who have had alphabet training (Morais, et al, 1979; Rayner, et al, 2001; Ziegler & Goswami, 2005). The entire coarticulation issue disappears as a problem once we consider the possibility that speakers employ a rich and detailed auditory memory for language. Of course, the problem that remains to be answered below is: ``Why do different syllables (e.g., *do* and *day*) seem to begin with ``the same sound''?

The experimental data are generally compatible with a view of word storage that is concrete and detailed such as that proposed by ``exemplar models of memory'' (Hintzman, 1986; Pierrehumbert, 2001; K. Johnson, 2006; viz. the special issue *The Linguistic Review*, 2007). On this view, speakers store large amounts of phonetic detail for specific utterances. But these rich representations are categorized into various lexical and phonological classes, into the categories that we formerly thought were abstract linguistic symbols. It appears that speech perception does *not* need to extract an abstract, context-free invariant for each nonoverlapping consonant and vowel. At least, it does not until one has had training in alphabet literacy. Alphabet-literate people do consciously describe speech to themselves using a small phonological (or orthographic) inventory of sound categories, but for all real-time tasks (e.g., production, perception, recall and recognition), experimental data suggest that they rely on a fairly rich audition-based linguistic memory and subtle motor control skills that far exceed the very gross gestures described by linguistic distinctive features. Apparently, ordinary human speakers have the ability, when they hear some linguistic stimulus, to quickly access a memory bank that is rich enough that it acts like a database of categorized utterances and (1) locate a closest match and (2) recover its category memberships. So the answer to the question above is that, although syllables like [de$^{I}$] (*day*) and [du] (*do*) do not share any unit in linguistic memory, *contra* Liberman, et al., 1968 and most of modern linguistic thought, they are nevertheless categorized the same by English speakers. That is, the two words are classified as sharing the same initial phone or phoneme even though they may be acoustically very different.

The context sensitivity of phones and phonemes and subtle variation of human speech imply that real-time psychological processing of language cannot be dependent on the abstract, low-dimensional, static and non-overlapping descriptions proposed by traditional linguistics and as suggested to us by our literacy-based intuitions. The data imply a rich speech memory that retains large numbers of speech episodes and massively redundant speech material – rather than a dictionary-like representation with a single form for each word. These episodes must be encoded by a set of rich auditory and articulatory trajectory fragments through the speech auditory space, rather than a representation that is maximally efficient for a small set of graphic tokens like those we draw on paper. It seems speakers remember whatever details they can about specific utterances and about the linguistic categories that each exhibits. Given a rich memory for examples, then generalizations and abstractions about linguistic categories (such as features, phonemes, syllable onsets and codas, etc.) can be extracted as needed from memory. For example, Hintzman (1986) models each remembered event as a long vector of all co-occurring features of an episode and models long-term memory as a matrix of such vectors. Retrieval from memory depends on the similarity of a (short) memory probe vector to all features in the stored vectors. Generalizations about a category can be found by activating many exemplars of the category and noticing which features receive the most activation. Thus apples are found to be most often `red.' This model is not a plausible direct model of memory, but it exhibits some of the right behaviours.

If this proposal for rich linguistic memory strikes the reader as implausible, consider that exemplar theories of memory have found support from research in many modalities in experimental psychology (Nosofsky, 1986; Shiffrin and Steyvers, 1997) and from the human ability to remember randomly collocated events on a single exposure. One example is the reader's own memory for specific events that happened earlier today or last week (Tulving, 2002). There is still much that is not understood about human memory and learning and there is experimental support both for memory of concrete exemplars as well as for the gradual acquisition of abstract generalizations about stimulus patterns. But this much can be said with confidence: Linguistics throughout the 20$^{th}$ century and up to the present was based on the firm assumption that memory representations for language must be as ``efficient'' as possible, requiring as few bits as possible for representations, and achieving as much abstraction as possible. But the evidence cited here shows that this is quite misguided. Linguistic memory is apparently very rich and detailed and able to store complex episodes as a whole.

## 1.3. Prediction 3: Memory for heard speech relies on abstract phones and phonemes.

The third prediction is that our memory for what people have said should show evidence of an abstract coding that lacks all details that are not encoded into phonological or phonetic features. One particularly clear example of evidence against this prediction is the recognition memory studies of Goldinger, Pisoni and colleagues (Goldinger, 1996; Pisoni, 1997). In one experiment, they played a continuous list of auditorily presented words to subjects who indicated by a button press when a word was repeated in the list (Palmeri, et al, 1993). The list was prepared from list readings by from 2 to 20 speakers, so in conditions with many speakers, every word seems to be spoken by a different voice, although the participants were told to ignore any change in voice and focus on word identities. They designed the lists so it was possible to compare performance on words that were repeated in the same voice vs. repeated in a different voice. The prediction

that most linguists would make (including me at the time) is that it should make no difference whether the voice was the same or different, since we won't remember the voice details. Speakers should just remember the words using an abstract phonological or phonetic code that should, therefore, be identical for different speakers of the same dialect.

The surprising result (replicated several times) was that the ability to recognise that a word repetition was consistently 10% better when the word was repeated by the original voice. Of course, it was the exact same recording. If linguistic memory employs a Hintzman-like exemplar system, same-voice repetitions will share more features of phonetic detail than different-voice repetitions, so recognition performance should be better when the voice is the same. This only confirms what we should know from personal experience: we can indeed remember something about individual speaker's voices and pronunciations and thus must somehow store that indexical information in memory (Pisoni, 1997).

Of course, this result by itself does not prove there could not also be an abstract ``linguistic'' representation as well, but it does show that the voice-specific details remain in memory for some time to facilitate recognition. The problem is that I have found no experimental results that demand abstract representations. Where is the evidence for a segmental code like the one that we have such strong intuitions about?

### 1.4. Prediction 4: Speech timing patterns are restricted to serial order patterns

Finally, the fourth prediction is that only serially-specified temporal patterns can be linguistic. That is, any timing patterns observed in language must be accounted for using serially ordered tokens. Thus, long vs. short vowels might be represented as 2 segments vs. 1 segment or they might be 2 different segments whose ``phonetic implementation'' differs in duration. But this is implausible since there are a great many complex temporal aspects of speech in any language (e.g., Klatt, 1976; Port & Leary, 2005) and these can differ in small and subtle details of timing (e.g., in patterns of voice-onset time, characteristic vowel durations, etc.), as noted above. But languages also differ in the typical rhythmic patterns that each language exhibits. Before looking closer at some rhythmic forms of speech, it is important to note right here that, since a serially ordered, letter-like representation of speech can only represent timing differences in the order (or choice) of symbol tokens, it cannot provide a description of any periodic patterns that are defined by a fixed period in seconds or milliseconds. They support only patterns that are specified in terms of serial order. This means that the string *[ABC] = [A BC] = [AB C]* but *≠ [ACB]*. But in the next section, it will be shown that actually speech frequently exhibits periodicities defined by fixed intervals in time. Altogether, the limits on the abilities of phonetic segments (i.e., phones or phonemes) to model real speech behavior are so severe, it is difficult to see why phoneticians and linguists (including this author) have clung to them for so long. We turn now to periodic patterns that violate the constraint against patterns that depend on particular intervals of time.

## 2. Periodic Speech Patterns

Every linguistic community, as far as we can tell, has some speaking styles based on periodic timing. What we call singing or chanting are just some of the many genres of speech that exhibit highly regular periodicity. How do speakers achieve the regularity? And what is it exactly that is periodically spaced in time? One of the most important discoveries relative to the latter issue has been known for many years although its importance has been generally underestimated. It seems to be George Allen (1972) who first discovered that if English speakers are asked to tap a finger ``in time with'' a spoken list of short words, they locate the fingertaps close to the onset of the stressed vowel. Thus, when they tap in time with, say, *stop cop, mop, shop*, the finger taps occur close to the vowel onset (the location of the strongest energy onset), not close to the onset of the /s/ or /m/ or to the burst of the /k/ (which may be phonetically salient but involve much less energy). In later research, this time point was called the ``perceptual center'' (or ``P-center'') of a syllable (Morton, Marcus & Frankish, 1976; Pompino-Marschall, 1989) but here it will be referred to as an auditory pulse. Signal processing methods for locating this pulse have been refined over the years (Scott, 1993; Kochanski & Orphanidou, 2006 unpublished). Methods for finding the auditory pulse all rely on smoothing the energy contour of speech over several hundred milliseconds and locating maxima in the first derivative.

Human cultures around the world have invented forms of chanting and unison singing, from a Buddhist mantra like ``*Om mani padme hum*'' that is repeated rhythmically for extended periods, Gregorian chant and modern chants for athletic teams like the 4-beat phrase ``*Go I U, Go I U, ...*''. Poetry is, of course, text that was composed so as to exhibit periodicity in its performance (although 20th century styles of poetry reading in English try to avoid any strong suggestion of periodicity). Japanese has long been said to exhibit regular periodic, timing of moras (Port, et al 1987; Warner-Arai, 2001; Hirata & Witton, 2005). For many examples of Japanese speech, the automatically extracted auditory pulses are very regular. But a related timing pattern can be seen in various short audio clips consisting of, such as, TV journalist signoffs and punchlines in a lecture or sermon (e.g., ***God will never leave*** *us* ***or*** *for****sake*** *us*) or other kinds of speech performance. What is periodic in all of these cases appears to be the temporal spacing of the auditory pulses, correlating fairly closely with vowel onsets, as I will demonstrate with several examples in my oral presentation (although many of my audio examples can be found on my webpage).

Experimental methods for observing metrical speech performance have also been developed, such as speech cycling where participants repeat a short phrase over and over. In one experiment using this method (Cummins & Port, 1998), English speakers were asked to repeat a phrase like *Take a pack of cards* (or *Buy a dog a bone)* over and over. Speakers tend naturally to repeat these with timing that is strongly periodic with *pack* dividing the interval from *Take* to *Take* into integer-ratio fractions (e.g., so *pack* begins at either one third, one half or two-thirds of the whole repetition cycle from *Take* to *Take*). In an experiment to explore this phenomenon, a 2-tone metronome was created where the interval from low-tone to low-tone was defined as 1.0, with the higher tone occurring at equiprobable random phase angles between 0.20 to 0.70 of

the full cycle. If the participants were able to do just what was asked, they would initiate the phrase (e.g., the auditory pulse of *Take*) on the first low metronome tone and adjust the timing in such a way that they say the target word (e.g., the pulse of *cards*) at whatever phase angle was indicated by the higher tone. Thus the phase of the target word, when binned into small ranges of phase angle (40ths of the cycle in this case), should exhibit the same flat distribution as the metronome pulses. But, as shown in Figure 1, the phase of the target-word onset was very strongly biased toward one of just three phase angles. 1/3, ½, and 2/3. These data (from Cummins & Port, 1998) show the median phase angle of the produced target word for 8 speakers over about 1400 trails (each containing 24 repetitions) of the test phrases that have the target phase angles distributed uniformly over the range 0.20 to 0.70. But the speakers showed a powerful bias to produce the target syllable onsets close to 0.33, 0.50 or 0.67.

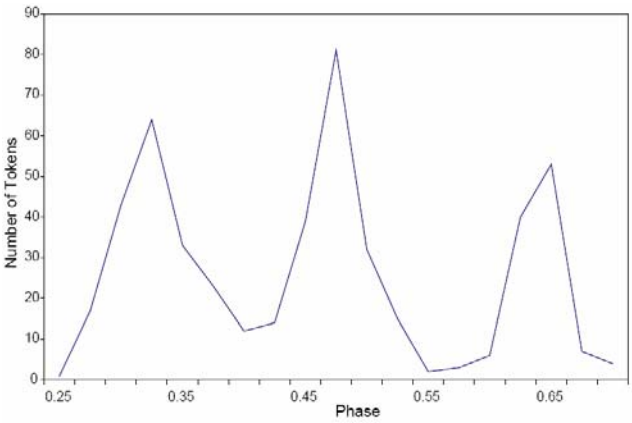


Figure 1. *Frequency histogram of the auditory pulse of the final monosyllabic word in various short test phrases for 8 speakers listening to a complex metronome pattern (Figure from Port (2005) replotted from results of Experiments 1-2, Cummins & Port, 1998). The pattern provided a target tone for the final stressed syllable of the phrase at randomly chosen phase lags uniformly distributed over the range 0.20 - 0.70. As the target phase was shortened, the repetition cycle was lengthened to keep the speaking rate required for the phrases constant.*

What could account for the overwhelming bias toward just these phase angles? It has been proposed (Port, 2005) that the participants generated a metrical timing pattern by coupling two cognitive oscillators, a slower oscillator (cycling once for each repetition of the phrase) and a faster one cycling either 2 or 3 times for each phrase repetition (Large & Jones, 1999; Port, 2003). The two oscillators were tightly coupled so that both passed through phase 0 together (where the onset of *Take* was located) and the faster one passed through its phase zero again at either 0.33 and 0.67 (for the 3:1 frequency ratio) or 0.50 (for the 2:1 ratio). Presumably all of these phase zeros serve as attractors for the auditory pulses of the test phrase, thus making it fairly easy for participants to do the task. It is an important property of this meter system that it is tied very strongly to audition. That is, auditory pulses (as in speech and music) are much easier to entrain to a meter than, say, visual events or gesture onsets.

Such a system of oscillators appears to underlie many kinds of musical meters. The more basic single-oscillator version might underlie the production of the preacher who said `***God will never leave*** *us* ***or*** *for****sake*** *us'* and also underlie the perception of the periodicity of this utterance by listeners (see Jones, 1976; Large & Jones, 1999). It probably also accounts for similar short stretches of periodic speech that occur frequently in the speech of many English speakers and in many other languages as well. When Japanese speakers exhibit mora timing, they may also be activating an oscillator that attracts vowel onsets. But to relate this discussion to the primary argument of this paper, the existence of temporal attractors for speech events displaying precise measurements of time reveals another aspect of the utter inadequacy of segment-based psychological representations of speech.

Many linguists will respond that these rhythmic phenomena may be interesting but lie outside the domain of linguistics. But what has determined the domain of linguistics? I claim that because of its reliance on alphabetical representations, linguistics has effectively defined itself as ``studying whatever can be represented with letters.'' This is why both prosody and experimental phonetics are often viewed by linguists as marginally linguistic at best or completely outside the field. But it is not a good practice to allow a notation system to determine the domain of a discipline.

## 3. Concluding Discussion

The reader will have noticed that phones and phonemes have not been differentiated in this essay. This is because both uses of the alphabet suffer from the same defects. Both are segmental and ignore continuous time. Both are abstracted across neighboring contexts and across speakers and speaking rates. And both were developed by highly literate scientists who modified traditional alphabetical orthography to make it more consistent and suitable for scientific discourse. They differ in both their size and their representational purpose. For phonemes or ``phonological segments,'' the goal is to make the set of tokens as small as possible (perhaps to make an optimal writing system or to prepare teaching materials for adult learners), while phonetic transcription strives for articulatory and auditory accuracy (and uses a much larger alphabet). There is no reason to reject the use of either notational system. We who were trained in the use of an alphabet early in life find it an indispensable tool for the description of speech. I continue to teach phonetic transcription in my classes and to rely on it for professional communication. But it is important for us to keep in mind that the linguistic structures we model by using an alphabet (when either using orthography or when doing linguistic phonological analysis) are social structures or statistical patterns across a large corpus. They are not psychological units or structures directly represented in ``the mind''. We must appreciate that our linguistic intuitions about speech and about phones and phonemes have almost surely been shaped by our extensive experience using alphabets. Thus we should not rely on these intuitions over more objective descriptions of the sound of speech.

Letter-based symbolic descriptions of language do not come from nowhere. They exist in our culture because they were engineered over several thousand years of

experiments with graphical representations of language and then taught to each new generation through intensive drill. The physical symbols – the graphic alphabets for words and numbers – that provide the key external scaffolding for our symbol manipulation skills (Clark, 1997) are a cultural invention. We scientists all have considerable competence with the symbolic technologies of literacy, mathematics and, these days, computing. And because of its very early acquisition, we forget that alphabetical writing is a culturally transmitted technology – and one that we have been internalizing since we were handed alphabet blocks at age two.

The criticism raised in this essay is that phones and phonemes receive their intuitive grip from our experience at reading and writing. But, of course, segments and phonological distinctive features are not being denied. They exist, of course, and are shaped (and changed) over time by generations of speakers. They are classes of events grouped into categories by members of some speech community. The [d] in *day* and in *do* describes a set of different (but usually similar) word fragments that (a) the community of English speakers treat as a category and that (b) turn out to be conveniently represented with the letter token ``**d**''. What about the [d] in *stop*? It sounds like a [d] to me and, although there is no contrast with a [t] in this context (following a word-initial [s]), it is not considered to be a member of the ``**d**'' category. Probably this is primarily because *stop*, like *stoop* and *steer*, has been written with ``**t**'' in English orthography for over a thousand years. The English phonological category /d/ is just as difficult to define precisely and non-arbitrarily as most other culturally determined categories, such as ``table'' or ``mountain.'' Some things are prototypical ``tables'' and ``mountains,'' and some things are difficult to assign or may be assigned arbitrarily. Other things in the world may be called ``table,'' like, say, a flat-topped mountain, even though its similarity to a prototype table is remote.

Similarly, the category types of phonology and linguistics in general are established by generations of speakers. They are units in the social institution called Language L. Since we usually learn the orthographic alphabet very young, the orthography helps to fix for us which words exemplify each sound type. These sound categories may gradually take on for us the properties of symbol tokens – properties just like those of the graphic letters themselves. But the speech sounds are not symbols. It has been clear for over half a century that they are not invariant, nonoverlapping and abstract symbolic units like those that might be manipulable by a hypothetical cognitive symbol processor. They are just messy categories whose exact boundaries are established by the conventions of our culture and the community of speakers of our language. Plus many of them have a conventional correspondence with letters which really *are* symbols.

What all this means for a conference on prosody is that linguists used to believe that language is composed from `phonological segments', the true linguistic units – discrete, very limited in number and psychologically real. These were what linguists were concerned with. Then there were other properties of speech called `prosodies' or `suprasegmentals' – all the properties of speech that are much less discrete, seem to be overlaid on the truly linguistic segments, and not easily represented graphically nor supported by our intuitions of the serial order of phonemes. Thus, in the conventional view, these properties do not appear to be linguistic properties at all. But it turns out that phonetic segments are not as different from prosodies as we thought. The segments too overlap in time and are not discrete and thus not easily represented accurately with serially ordered letter tokens. Hence, the title of this essay: It turns out that everything about speech is like speech prosody.

## 5. References


[1] G. Allen, "The location of rhythmic stress beats in English: An experimental study I," *Language and Speech,* vol. 15, pp. 72-100, 1972.

[2] C. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica,* vol. 49, pp. 155-180, 1992.

[3] J. Bybee, *Phonology and Language Use*. Cambridge, UK: Cambridge University Press, 2001.

[4] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper and Row, 1968.

[5] N. Chomsky and G. Miller, "Introduction to the formal analysis of natural languages," in *Handbook of Mathematical Psychology* vol. 2, R. D. Luce, R. R. Bush, and E. Galanter, Eds. New York: Wiley and Sons, 1963, pp. 323-418.

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

[7] F. Cummins and R. Port, "Rhythmic constraints on stress timing in English," *Journal of Phonetics,* vol. 26, pp. 145-171, 1998.

[8] A. Faber, "Phonemic segmentation as epiphenomenon: Evidence from the history of alphabetic writing," in *The Linguistics of Literacy*, P. Downing, S. Lima, and M. Noonan, Eds. Amsterdam: John Benjamins, 1992, pp. 111-134.

[9] P. Foulkes and G. Docherty, "The social life of phonetics and phonology," *Journal of Phonetics,* vol. 34, pp. 409-438, 2006.

[10] S. D. Goldinger, "Words and voices: Episodic traces in spoken word identification and recognition memory," *Journal of Experimental Psychology: Learning, Memory and Cognition,* vol. 22, pp. 1166-1183, 1996.

[11] S. Hawkins and N. Nguyen, "Influence of syllable-final voicing on the acoustic onset of syllable-onset /l/ in English.," *Journal of Phonetics,* vol. 32, pp. 199-231, 2004.

[12] D. L. Hintzman, "`Schema abstraction' in a multiple-trace memory model," *Psychological Review,* vol. 93, pp. 411-428, 1986.

[13] Y. Hirata and J. Whiton, "Effects of speaking rate on the single/geminate stop distinction in Japanese," *Journal of Acoustical Society of America,* vol. 118, pp. 1647-1660, 2005.

[14] IPA, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, England: Cambridge University Press, 1999.

[15] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features*. Cambridge, Massachusetts: MIT, 1952.

[16] K. Johnson, "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," *Journal of Phonetics* vol. 34, pp. 485-499, 2006.

[17] D. Jones, *An Outline of English Phonetics*. Leipzig, Germany: Teubner, 1918.
[18] M. R. Jones, "Time: Our lost dimension," *Psychological Review,* vol. 83, pp. 323-355, 1976.
[19] R. Kent and F. Minifie, "Coarticulation in recent speech production models," *Journal of Phonetics,* vol. 5, pp. 115-135, 1977.
[20] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence.," *Journal of the Acoustical Society of America,* vol. 59, pp. 1208-21, 1976.
[21] G. Kochanski and C. Orphanidou, "What marks the beat of speech?," 2006, submitted.
[22] W. Labov, *The Principles of Linguistic Change, Social Factors, Volume II*: Wiley-Blackwell, 2001.
[23] P. Ladefoged, *A Course in Phonetics*, 1 ed. Orlando, Florida: Harcourt Brace Jovanovich, 1972.
[24] E. W. Large and M. R. Jones, "The dynamics of attending: How we track time-varying events," *Psychological Review,* vol. 106, pp. 119-159, 1999.
[25] A. M. Liberman, P. Delattre, L. Gerstman, and F. Cooper, "Perception of the speech code," *Psychological Review,* vol. 74, pp. 431-461., 1968.
[26] L. Lisker, "``Voicing" in English: A catalogue of acoustic features signalling /b/ vs. /p/ in trochees," *Language and Speech,* vol. 29, pp. 3-11, 1984.
[27] L. Lisker and A. Abramson, "A cross-language study of voicing in initial stops: acoustical measurements," *Word,* vol. 20, pp. 384-422, 1964.
[28] L. Lisker and A. Abramson, "Some effects of context on voice-onset time in English stops.," *Language and Speech,* vol. 10, pp. 1-28, 1967.
[29] J. Local, "Phonetic detail and the organization of talk-in-interaction," in *XVIth International Congress of Phonetic Sciences* Saarbruecken, Germany: 16th ICPhS Organizing Committee, 2007.
[30] J. Morais, L. Cary, J. Alegria, and P. Bertelson, "Does awareness of speech as a sequence of phones arise spontaneously?," *Cognition,* vol. 7, pp. 323-331, 1979.
[31] J. Morton, S. Marcus, and C. Frankish, "Perceptual centers (P-centers)," *Psychological Review,* vol. 83, pp. 405-408, 1976.
[32] R. Nosofsky, "Attention, similarity and the identification-categorization relationship," *Journal of Experimental Psychology: General,* vol. 115, pp. 39-57, 1986.
[33] T. J. Palmeri, S. D. Goldinger, and D. B. Pisoni, "Episodic encoding of voice attributes and recognition memory for spoken words," *Journal of Experimental Psychology, Learning, Memory and Cognition,* vol. 19, pp. 309-328, 1993.
[34] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," *Journal of the Acoustical Society of America,* vol. 32, pp. 693-703., 1960.
[35] J. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Frequency Effects and the Emergency of Linguistic Structure*, J. Bybee and P. Hopper, Eds. Amsterdam: John Benjamins, 2001, pp. 137-157.
[36] D. B. Pisoni, "Some thoughts on `normalization' in speech perception," in *Talker variability in speech processing*, K. Johnson and J. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 9-32.
[37] Pompino-Marischal, 1989.
[38] R. Port, "Meter and speech," *Journal of Phonetics,* vol. 31, pp. 599-611, 2003.
[39] R. Port, "The graphical basis of phones and phonemes," in *Second Language Speech Learning: The Role of Language Experience in Speech Production and Perception.*, M. Munro and O. Schwen-Bohn, Eds. Amsterdam, Holland: John Benjamins, 2006, pp. 349-365.
[40] R. Port, J. Dalby, and M. O'Dell, "Evidence for mora timing in Japanese," *Journal of Acoustical Society,* vol. 81, pp. 1574-1585, 1987.
[41] R. Port and R. Rotunno, "Relation between voice-onset time and vowel duration," *Journal of the Acoustical Society of America,* vol. 66, pp. 654-662., 1979.
[42] R. F. Port and A. Leary, "Against formal phonology," *Language,* vol. 81, pp. 927-964, 2005.
[43] K. Rayner, B. Foorman, C. Perfetti, D. Pesetsky, and M. Seidenberg, "How psychological science informs the teaching of reading," *Psychological Science in the Public Interest,* vol. 2, pp. 31-74, 2001.
[44] F. d. Saussure, *Course in General Linguistics*. New York: Philosophical Library, 1916.
[45] S. Scott, "The point of P-centres," *Psychological Research,* vol. 61, pp. 4-11, 1998.
[46] R. Shiffrin and M. Steyvers, "A model for recognition memory: REM: Retreiving effectively from memory," *Psychonomic Bulletin and Review,* vol. 4, pp. 145-166, 1997.
[47] K. N. Stevens, "Acoustic correlates of some phonetic categories," *Journal of Acoustical Society of America,* pp. 836-842, 1980.
[48] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics,* vol. 17, pp. 3-46, 1989.
[49] E. Tulving, "Episodic memory: From mind to brain," *Annual Review of Psychology,* vol. 53, pp. 1-25, 2002.
[50] W. F. Twaddell, "On defining the phoneme," *Language,* vol. Language Monograph 16, 1935.
[51] N. Warner and T. Arai, "Japanese mora timing: A review," *Phonetica,* vol. 58, pp. 53-87, 2001.
[52] J. Ziegler and U. Goswami, "Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory.," *Psychological Bulletin,* vol. 131, pp. 3-29, 2005.