Language and its two complex systems*

Robert Port
Indiana University

April 22, 2008

1. Definition and Scope

Language has been a topic of great interest at least since the Greek philosophers. Clearly speech is one of the most sophisticated and complex of human capabilities and has remained a major puzzle through the centuries. Complex systems promise to provide analytical tools that will help break through some of the mystery. From this perspective, it appears that there are three complex systems relevant to speech. The first is natural selection, the slow influences on the gene pool of the species that give us whatever innate skills we need to learn at least one language. However, we will have nothing further to say about this issue here since the time scale involved is so long relative to human history. The remaining two complex systems are highly relevant. They are, first, the language itself, viewed here as a particular type of social institution, a set of cultural practices that present the language learner with a broad range of linguistic materials with which to communicate with others. This system clearly evolves as history and the cultural practices of the community change. It sometimes requires only a couple hundred years (a dozen or so generations) for a language to split into mutually incomprehensible dialects. The final complex system is the one that the individual speaker develops as he/she gradually becomes competent in the use of the language. This system begins before birth and continues to learn and adapt right up until death as the language environment of an individual changes. It involves controlling the many muscle systems of the vocal tract and respiratory system as well as specializations of the auditory system that customize it for hearing and understanding speech in the ambient language or languages.

* To appear in *The Encyclopedia of Complexity and System Science*, (2008) Springer-Verlag, Heidelberg, Germany.

There is one historical development of central importance that greatly complicates our ability to apply our intuitions to language. This is the 3 thousand-year-old tradition of training our children to read and write using an alphabet. This training, involving hundreds of hours over many years, typically results in our hearing speech as though it consists of a sequence of discrete letter-sized sounds, the consonants and vowels. The lifelong training and practice interpreting letter strings as continuous speech and interpreting speech as a string of letters has, in this author's opinion, encouraged unwarranted confidence in the reliability of our intuitions about the structure of language. Thus, one goal of this essay is to show that our intuitions about the segmental structure of speech are untrustworthy.

2. Introduction – a self-demo

What kind of structures are exhibited by a natural spoken language? Lay people think first of *words*. Words are the archetypal units of language. It seems every language displays sequences of words and must have a dictionary, its foundational list of words. Linguists agree with this intuition and add another unit that may be somewhat less obvious to laymen: the speech sound, the *phone* or phoneme, that is, the fragment of speech that is represented by a single *letter* from some alphabetical orthography or from the International Phonetic Alphabet (IPA, 1999). To begin the study of language by using letters and a list of words seems very natural but turns out to be problematic if our goal is to understand linguistic cognition because letters (and phonemes) provide inadequate models of speech. The main problem is that letters (and phonemes) provide too few bits of information to represent what speakers need to know about speech. One unfortunate consequence of this supposed efficiency is that only the ordering of letters can represent the temporal layout of speech events. Let us begin then by making observations, not about our intuitive description of language structure in terms of an alphabet, but rather by making observations of some specific realtime linguistic behavior.

Demonstration.
In order to provide an intuitive example of realtime linguistic behavior, the reader is urged to participate in a tiny experiment on oneself. The reader is encouraged to repeat the following phrase out loud: *Take a pack of cards*. Please repeat this phrase over and over – aloud (a whisper will work fine), between 5 and 10 times. Now. After cycling through the phrase, notice the rhythm pattern that you adopted unconsciously. When a person repeats a short phrase like this over and over, it seems the easiest way to do it is to slip into some rhythmic timing pattern that is the same from repetition to repetition. A remarkable fact is that there is only a small number of ways that most people will arrange the timing of a cycled phrase like this one. To discover which rhythm you used, repeat the phrase again just as you did before but tap a finger in time with the stressed syllables of the phrase. You should find yourself tapping with only some of the syllables of the phrase, probably at least the first and last words, *Take* and *cards*. Now there appear to be only 3 rhythmic patterns for reading phrases like this that are readily available to English speakers, and probably speakers of some other languages as well (Cummins and Port, 1998). The three preferred patterns are shown in Example 1 where underline and bold face mark stressed syllables that should align with finger taps. The reader is urged

to produce a few cycles of the three discrete patterns in Example 1. The notations |: and :| are borrowed from music notation to indicate cycling of the pattern between them.

Example 1.

(a) |: **Take** a pack of **cards**, **Take** a pack of **cards**, :|
(b) |: **Take** a pack of **cards** ***[rest]***, **Take** a pack of **cards** ***[rest]***, :|
(c) |: **Take** a **pack** of **cards**, **Take** a **pack** of **cards**, :|

Pattern (1a) has 2 beats per cycle, a beat on *Take* and another on *cards* with the next beat on *Take* again. There is a finger tap for each beat (but none on *pack).* Pattern (1b) has 3 beats (and taps) per cycle: the first on *Take*, the second on *cards* and the third is a beat that has no syllable, just a longer pause than (1a) has – equivalent to a musical `rest'. Patterns (a) and (b) must have a long period so that *Take a pack of* can be spoken in a single beat-cycle. Example (1c) also has 3 beats, on *Take, pack* and *cards*, and, as in the waltz rhythm, the repetition cycle begins immediately with another *Take*. Most Americans first discover pattern (a) of Example 1, but most people can produce all three without much difficulty. To those familiar with music notation, we could say Example (1a) has a 2-beat measure with *Take* and *cards* beginning at each beat and dividing the measure into equal halves. Both of the others have a 3-beat measure but (1b) has a rest on beat 3. The unstressed syllables seem to be squeezed or stretched to assure that the salient target syllable onsets begin at the proper time, that is, as close as possible to phase zero of one or both of the coupled oscillators generating the metrical time structure.

Other timing patterns can be achieved with a little training. Try pronouncing the sample phrase with equal stress on all syllables. Be sure to repeat the phrase several times as in Example 2.

Example 2 |: **Take** **a** **pack** **of** **cards**, ***[rest]*** :|
(1 2) (3 4) (5 6)

The easiest way (for English speakers) to repeat the phrase so that you put equal stress on all 5 syllables is to use a 6-beat meter (probably with emphasis on beats 1, 3, and 5) and with a rest on beat 6. For another attractor, one can leave out the rest on beat 6 but to immediately return to Take after cards. This creates a 5-beat pattern. (If leaving out the rest on beat 6 proves difficult for the reader, then count, **1**, 2, **1**, 2, 3, **1**, 2, **1**, 2, 3 a few times, then switch to the test phrase.) English speakers can learn to produce these more exotic 5- and 6-beat meters, but it takes some effort and practice to achieve them. There is some anecdotal evidence, however, that both Japanese and Finnish speakers, for example, are able to effortlessly produce the 5-beat pattern. A few experimental tasks like these have been looked at with speakers of several languages – at least Arabic, Japanese, German and Finnish (Tajima & Port, 2003; Zawaideh, et. al., 2002). Everyone tested so far seems to find at least one of the temporal patterns of Ex. 1 to be very natural and almost unavoidable for many short phrases that are cycled in any language.

Since it is likely that differences between languages in preferred rhythm patterns will continue to hold up, it is likely that rhythmic patterns are an intrinsic aspect of any

language, and thus that research on the phonology of a language must begin to rely on continuous time data. These language-specific constraints on speech timing (as well as many subtle phonetic differences between dialects and languages) show that it is reckless to begin the study of language by first encoding all speech data into a small set of uniform letter-like tokens that have only serial order instead of real time. The problem is that, even though our intuitions about language tell us letters are obvious units of language, letters are actually a culturally-transmitted technology that humans have learned to apply to (and, if necessary, impose upon) spoken language. Letters and their psychological counterparts, phones and phonemes, have very weak experimental support as psychological units used for cognitive ``spelling'' of words (Pisoni, 1997; Port, 2007; Port & Leary, 2005; Pisoni & Levy, 2006).

3. Speech and Meters

Given the intuitive descriptions of the timing of repeated short phrases, an experimental study is needed to verify that these three patterns are the primary ones that English speakers can reliably produce. But first, for rhythmically produced speech, an important question arises: just what in the acoustics are subjects aligning with the finger taps? That is, what is the physical definition of a ``beat'' or auditory pulse? Most scientists working in this area of speech agree that it is approximately the onset of a vowel (Allen, 1972; Sophie Scott, 1998, Cummins and Port, 1998; Kochanski & Orphanidou, 2007). This is usually located by low-pass filtering of speech (at about 800 Hz) and integrating the remaining audio signal over about a 250 ms time window and looking for positive peaks in the second derivative that indicate a rapid increase in the energy in the low frequencies. That is where people tap their finger in time to a syllable.

In the experiment, speakers were asked to cycle short phrases like *Dig for a dime* and *King for a day* and to put the final noun (eg, *dime*) at specific phase lags relative to the *Dime-Dime* interval (Cummins & Port, 1998). The target phase angle was indicated by a metronomic pattern of 2 tones repeated 10-15 times per trial. Can speakers begin the word *dime* anywhere we ask within the cycle? It was expected the participants would find it easier to locate the onset at the simple harmonic fractions discovered above, ie, one half, one third and two-thirds of the cycle. These phase angles might be accurately reproduced and other phase lags would be less accurately reproduced. Subjects were played a metronomic pattern of 2 tones, one High and one Low which they listened to for several cycles. Then they were to repeat the indicated phrase 10 or 12 times in a row aligning the Low tone with *Dig* and the High tone with *dime*. Now in the stimulus patterns, the phase angle φ by which the High tone (i.e., for *dime)* divided the Low-Low interval (ie, *Dig-Dig)* was varied randomly and uniformly over the range from $\varphi = 0.20$ to 0.70. The overall period of *Dig-Dig* was also varied inversely with the ratio. The earlier the target phase angle, the slower the tempo of the repetition cycle, so as to leave a constant amount of time for pronouncing the phrase.

If the American participants were able to imitate those phase angles without bias, as they were asked to do, then they should produce a flat, uniform distribution of the onset of the vowel of *dime* relative to *Dig* onsets over that range – thereby replicating the pattern of

the stimulus metronome. But Figure 1 shows the observed frequency histogram of the median phase angle of the onset of *dime* for each trial across 8 speakers. The histogram represents about 1400 trials where each trial consisted of 12-15 cycles of a test phrase said to a single metronome pattern.

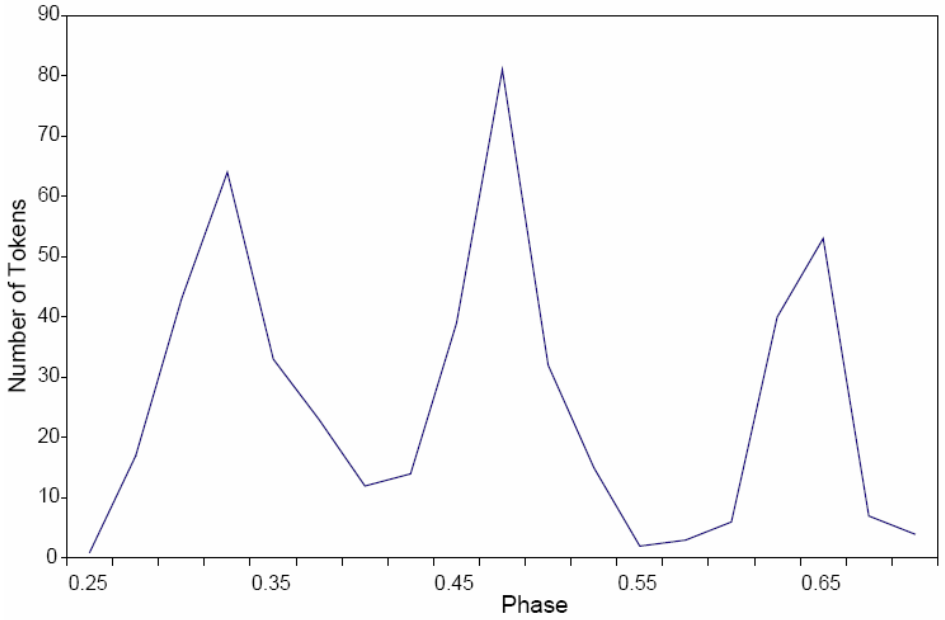

Figure 1. The distribution of the vowel onsets for the last stressed syllable by 8 speakers saying phrases resembling *Dig for a dime* with target phase angles specified by a metronomic sequence of 2 tones. Figure from Port, 2003.

Clearly the participants exhibited a strong tendency to locate their syllable onsets near just the 3 points: 1/3 of the cycle, ½ of the cycle or 2/3 of the cycle corresponding to the readings of Example (1b, 1a) and (1c) respectively. But they do not show evidence of any preference for fifths of the repetition cycle – no hint of peaks at 0.4 or 0.6. The subjects were trying to do as they were told, but when the high tone fell at, e.g., $\varphi = 0.4$, the speakers tended to produce φ closer either to 0.5 or 0.33.

Dynamical Model of Meter. These timing phenomena can be accounted for with dynamical models for neural oscillations correlated with auditory patterns (Kelso etal, Large-Jones, Port, 2005). A natural interpretation for the three peaks is that they represent attractors of a cyclical metric pattern specified by 2 coupled oscillators locked into one of 2 patterns, either a 2-beats-per-cycle pattern or 3-beats-per-cycle pattern. An important point about meter is that it is audition, the auditory sense, that has the strongest link to meter. Of course, we cannot tell from these data whether this bias toward the 3 attractor phase angles comes primarily from the participants' perception systems (so, e.g., $\varphi = 0.4$ simply *sounds* like $\varphi = 0.5$) or if the distortion reflects the influence of their motor system which might find it easier to consistently produce the auditorily salient pulse at one of the attractors.

Notice that these patterns involve cycles at two time scales nested within each other as shown in Figure 2. The slower cycle is the repetition cycle of the whole phrase, represented in the upper panel in the figure. Its phase zero coincides with the onset of *Take*. For all the phrases in this experiment, there is also a second cycle (thereby implying a state space that is a torus although a torus is inconvenient for display). The second one is faster and tightly coupled with the repetition-cycle oscillator. It cycles either 2 or 3 times before its phase zero aligns with phase zero of the slower oscillator (Port, 2003). If we assume that the temporal location of any phase zero is an attractor for a prominent auditory onset, then such a system of 2 coupled oscillators could define several musical meters and account for the 3 powerful patterns of Example 1.

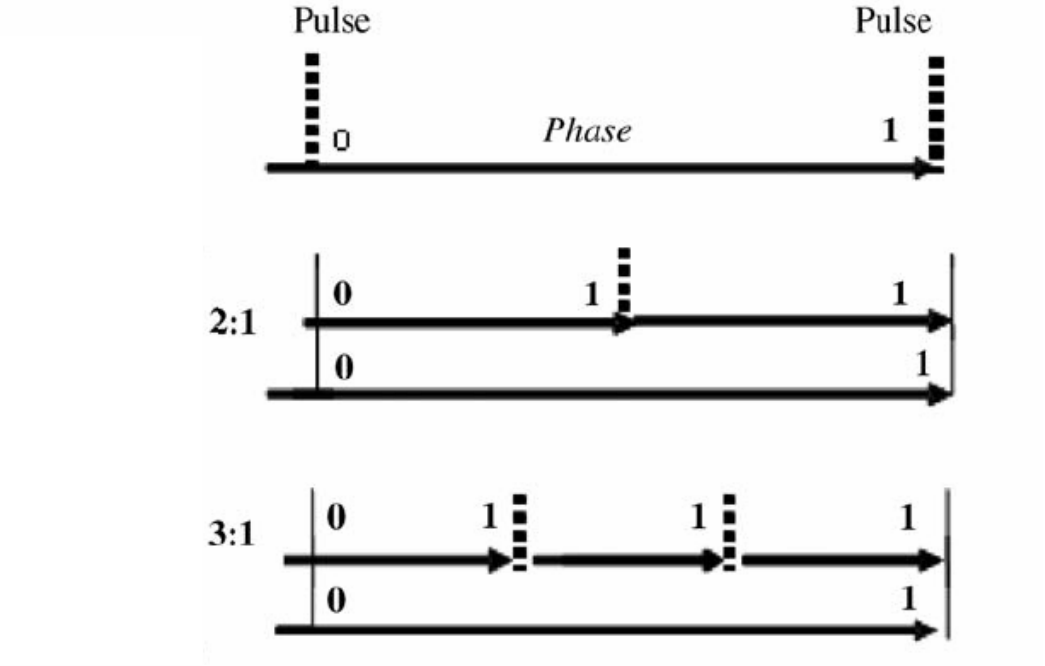

Figure 2: Simple meters represented as circles opened out into a line. Phase 0 and Phase 1 are, of course, the same point on the circle. The upper panel shows a single oscillator with periodic phase zero serving as an attractor for perceived or produced energy onsets. The second and third panels show two ways that a second oscillator can be coupled with the first. Each produces either 2 cycles or 3 cycles before coinciding with slower cycle. It is assumed that all phase zeros attract energy onsets both in speech and nonspeech sound. The 3 patterns of Example 1 can be accounted for by the 2:1 and 3:1 coupled oscillator or meter systems. Figure from Port, 2003.

Of course, subject behavior in ``speech cycling'' experiments is not the only evidence for a preference by humans for periodic speech. Seemingly all human communities have developed traditions of religious, martial or commercial chants and songs as well as other genres of vocal music that build on the metrical principles exhibited by people performing basic speech repetition. So the phenomena reviewed here point toward several important things about how to analyze human speech behavior.

The first lesson is that human speech performance, like so many other kinds of dynamical systems, finds periodic cycling easy and natural to do even though there is no mass-

bearing object (such as a leg or finger) oscillating at the observed frequencies. The speech periodicities reflect oscillations that must be occurring in various regions of the brain stem, cortex and cerebellum (see Mayville, et al, 2002; Patel, 2003; Zanto et al, 2006). Humans, in just about any language, can warp their speech into singsong or chant very easily. We seem to enjoy entraining our speech to each other (see Cummins, 2003 and to external auditory patterns (Patel, et al., 2005). Just above, the reader was asked simply to repeat a phrase and within a couple repetitions, the performance slipped into one of a small number of periodic attractors that this particular fragment of English can settle into. Once there, the speech production system tends to stay there until you force yourself out of the attractor (e.g., by trying to obey impertinent instructions from the author about assigning stress to bold-face syllables and tapping a finger).

The most important lesson from the phenomena reviewed so far is that we must be very careful about taking alphabetical forms of language as the basic data for a study of human speech. We may do so (as most linguistic research has done for a century) for convenience but must keep in mind that alphabetical representation makes constrained timing relationships like those revealed in the demonstration in Section 1 completely invisible, since letters and words in their written form exhibit only serial order, not continuous time. It seems like for this speech cycling task the speech is riding on the dynamics of low frequency oscillations. Examples 1a, 1b and 1c are, of course, identical in terms of alphabetical description, but they are far from identical in timing. Speakers of different languages exhibit different rhythmic effects in speech cycling and in other aspects of timing. Clearly it makes little sense to define language as somehow completely independent of these powerful speech behaviors.

4. Further Problems with an Alphabetical Model of Language

The invisibility of real time is one reason to reject representations of language using an alphabet. There are many more reasons. For example, the results of the demonstration and the laboratory experiment show that many properties of human speech resemble those of familiar dynamical systems in which oscillatory patterns are frequently observed (Abraham & Shaw, 1983). But within linguistics and much of psychology, language has been defined for the past century or so as a `symbol system' or a `code' shared by the speaker and hearer (Saussure, 1916; Bloomfield, 1933; Chomsky, 1965; Harris, 1981; Love, 2004). These timing patterns cannot be symbolically coded since they are defined by measurements in continuous time. The tendency of people to warp speech timing to fit patterns of integer-ratio time intervals is inexplicable by any model consisting only of serially ordered symbols – like words or the consonant and vowel segments. There are many kinds of temporal patterns in various languages that are inexpressible with any alphabet. So before approaching the study of speech and language in dynamical terms, we need to clarify that we will address a domain that is much broader than `language' as most linguists and psychologists have considered it.

The domain of phenomena for linguistic research needs to be ***speech events in continuous time*** of people in contextualized linguistic performance. This kind of raw data can easily be recorded, thanks to recent technologies) on audio and video or partially

displayed as a sound spectrogram (where time is mapped onto one spatial axis). But transcription or reduction of raw speech sound into any alphabet cannot be presumed, whether a hypothetical psychological alphabet such as of (language specific) phonemes, or an orthographic alphabet (such as the form you see printed on this page), or a technical phonetic alphabet (e.g., the International Phonetic Alphabet used by linguists for all languages for over a century now). Letters throw away far too much of what is critical for the dynamics of language as used by speakers.

In addition to speech rhythm, a second major problem with the traditional view is that it predicts invariant acoustic correlates for each segment. But it has been known for almost 50 years, that speech perception relies on formant trajectories and details of spectral shape and spectral change that cannot be captured by any invariant, context-free acoustic cues (e.g., Liberman, et al 1957, Liberman et al, 1968). Speech cues for phonetic segments simply cannot be defined so as to do the job of specifying phones or phonemes independently of the identity of neighboring vowels and consonants, speaker identity, speaking rate, etc. despite many attempts to do so (see Stevens & Blumstein, 1979; Kewley-Port, 1983). Thus, the shape of the second formant that specifies [d] before an [u] (as in English *do*) shows a falling trajectory whereas the second formant rises for [d] before an [e^{I}] (as in *day*). Such examples are found throughout the domain of speech cues. The acoustic cues for each phonetic segment tend to be different for each context, different enough so that a general definition can not be formulated that works across a range of different contexts. One consequence of these failures is that engineers working on speech recognition long ago abandoned attempts to recognize words by first recognizing segments or segmental features as a way of specifying words. Instead, the engineers try to specify whole words and phrases directly in terms of acoustic spectra rather than try to detect the kind of segmental units that linguists continue to insist are essential for word specification. By representing words using a codebook of spectrum slice descriptions, successful artificial speech perception systems effectively store context-specific versions of the phones and phonemes (Jelinek, 1988; Huckvale, 1997).

A third difficulty with the view that spoken words have a phonemic spelling for memory is that it makes the wrong prediction about the results of recognition memory experiments, where a subject indicates whether each word in a list is repeated (that is, has occurred earlier in the list) or not. The data show that words repeated by the same voice have a significant advantage in recognition accuracy over words that are repeated in a different voice. This shows that subjects do retain speaker-specific information (often called ``indexical information'') in memory. This is difficult to reconcile with the claim that words are stored using only representations like phones or phonemes that are abstracted away from speaker-specific properties indicating the speaker's voice (Palmeri et al, 1993; Goldinger, 1996; Pisoni, 1997; Port, 2007).

Finally, if words are perceived, produced and stored using a discrete list of distinct, abstract phones or phonemes, then linguistic variation of all kinds – dialect variation, historical changes in pronunciation, speaker idiosyncrasies, etc. – should all show discrete jumps in pronunciation between these discrete sound types. Speech should not show continuous variation in parameters like vowel height or backness, degrees of lip

rounding, place of articulation, voice-onset time, segment durations, etc. All these parameters should exhibit audible jumps when the transcription changes, e.g., from [t] to [d] or from [i] to [I] (except for possible noisy variation in production). However, generations of phonetics research reveals no evidence of discrete changes in any of these parameters as far as I can tell (Lisker & Abramson, 1964; Ladefoged & Maddiesen, 1996; Cho & Ladefoged, 1999; Port, 2007, 2008). The apparent target values of vowel quality show no signs of discrete jumps in target in a plot of the first formant against the second formant. And for voice-onset time, the time lag between the release of a stop and the onset of voicing exhibits many apparently continuous values of VOT target depending on the degree of stress, the following vowel or consonant identity, etc. (Port, 2007). Nowhere is there evidence that speech sounds are discretely different from each other in the same way that letters are. Nor is there evidence that the gestures are nonoverlapping in time (quite the opposite, they overlap a great deal, Browman, 1992). So, the hypothesis that words are cognitively spelled, that is, represented in memory, in terms of discrete, alphabet-like units simply finds no experimental support. It is extremely unlikely that the standard view of lexical representations used in linguistics will turn out to be correct.

Could there be some reason we have been overlooking for why the nervous system might not rely on these abstract units like phones and phonemes? One problem is probably that a vast amount of essential information about the speech would be lost in the transcription process. In addition to all temporal properties of speech (such as segment durations, rhythmic patterns, etc), information about coordination details of the speaker motor activity, details of voice quality and speaker identifying properties, etc. are all discarded when speech is reduced to alphabetical form. Many linguists have suggested that an alphabetical transcription is a ``more efficient'' representation than an audio or video recording (Jakobson, et al., 1952; Bloomfield 1933; Clark 2006). But more efficient for what? Is it more efficient as a form of word memory that is useful to a human speaker? This is, of course, an empirical question. The answer depends on how memory for words actually works. For a memory system optimized for dealing with a non-discrete world where non-discrete sound patterns are the norm, perhaps storing as much information as possible works better and might actually be ``easier''.

Like most modern linguists, Chomsky (1965, Chomsky & Halle, 1968) followed his predecessors in the ``structural linguistics'' movement (e.g., Saussure, 1916; Hockett, 1968, Bloomfield, 1933) in presuming that words have a representation in a psychological alphabet that is small enough that each segment type can be assumed to be distinct from all the others. Every language is expected to exhibit this kind of static `representation'. Words are presumed to be distinguished from each other by letter-like segments (unless they happen to be homonyms, like *two, too, to,* and *steal, steel, stiel*). Yet the evidence reviewed above shows that speakers **do** encode much phonetic detail in their speech perception and in speech productions. Consider all that a good actor encodes into speech: subtle, constantly changing moods, and the region of the speaker's origin, their social class, etc. as well as the text of the screenplay. Listeners, too, are able to interpret the speaker's mood, feelings, region, etc. The traditional picture of language as represented in memory in alphabet-like form finds no more support from these familiar phenomena about speech than it found in experimental studies.

Origin of Alphabetical Intuitions

Where might these intuitions come from if, in fact, words are not stored in the abstract way that we supposed? Why are we so comfortable and satisfied with alphabet-like descriptions of speech? The answer should have been obvious. For roughly the past 3k years, the educated minority in the Western cultural tradition have had a powerful technological scaffold for thinking about language (Clark, 1997, 2006). They have had an alphabetical orthography for their language. Writing captures a few important and relatively distinctive properties of speech that can be encoded (by someone with appropriate training) into letters and individual words, as in either an orthographic or a phonetic transcription. Please consider that every reader of this page has spent many hours a week since age 4 or 5, practicing the conversion of letters into words and words into letters. These hundreds of hours of training with an alphabetical orthography surely play a major role in shaping the intuitions that all of us share about language, and surely biased the very definition of language, and thus of linguistics. In the linguistic tradition since Saussure, including the Chomskyan tradition, knowledge of a language is assumed to consist in large part, of linguistic representations of words in memory. Since we are talking about `knowledge', each word is stored just once – just as in a published dictionary of a language. In addition to the set of lexical tokens, there are supposed to be rules in the ``grammar'' that manipulate the linguistic representations as simple tokens (moving segments or segment strings around, deleting or inserting them). This symbol-processing mechanism constitutes a powerful assumption about highly abstract representations of words. But why is it so easy for people to make such elaborate assumptions about language? The traditional linguistic representations for phonemes and words, etc., turn out to get many of their properties by generalization from the technology of conventional orthographic letters and words. Because of our experience using the technology of writing, it has come to seem natural to assume that something like this would be used by speakers mentally.

An orthographic system is a collection of customs about how to write a language. Development of a useable orthography requires deciding exactly what the alphabet will be and its application to sounds spoken in the language, which fragments of speech constitute a `word' and how each word is to be spelled. Other orthographic conventions address what patterns count as a sentence and what words a sentence may or may not begin with. For a learner of English orthography (including linguists and foreign language teachers), one problem is that the relation between letters and sounds is far from consistent (Rayner, 2001). Thus, we no longer pronounce either the **k** or the **gh** in *knight*. The letter **i,** which in most languages sounds like the vowel in *pea*, often sounds like *my* in English, and the vowel [ɛ] can be spelled either **ea** (in *head*), **e** (in *bed*) or **ai** (in *said*). Because of these well-known inconsistencies in many languages of Europe, scientists of speech together with second-language pedagogues developed the first version of the IPA phonetic alphabet about 1890. This system employs primarily Roman letters plus many others to describe speech sounds as pronounced with a certain level of detail. But it takes some training to be able to accurately hear some sounds that speakers use. This method pays no attention to `intended' speech sounds or to orthographic spellings. It can be seen that now the 3k year-old technology of an alphabet was employed for two very different purposes: It plays its traditional role of providing the basis for a useful orthography for

various languages but also very recently now provides a consistent notation system for scientific recording of speech gestures and speech sounds in many languages.

It seems likely that the development of the IPA alphabet influenced theories of language. It was immediately recognized that one could transcribe to varying levels of detail, so the distinction between *broad* vs. *narrow* transcription was discussed. Within 20 years of the release of the first IPA alphabet, some language scientists (in particular, a Polish linguist named Beaudoin de Courtenay) began speaking of `*phonemes'* (see Twaddell, 1935 for the early history of the term). These were to be a minimal set of abstract speech sounds, the smallest set that seems adequate to write a language. So the phoneme appeared as a kind of *idealized letter* that is hypothesized to describe something in the head of speakers. To see how the phoneme is an idealized letter, consider first the properties of letters. Letters in an alphabetical orthography:

1. *are discretely* different from each other.

2. There is a *fixed small set* of them available in any orthography.

3. They are *serially ordered* and graphically *nonoverlapping*.

4. Each letter has an *invariant shape* across different writers. (Although handwriting and font may differ, the letters are still discretely identifiable.)

5. Each word has a *single canonical spelling* in the `dictionary' (that is, in the inventory of the linguistic representations of the language).

Notice that because in orthographies words are spelled only from letters, words will always be discretely distinct as well. In fact, even syntax is assured of inheriting the discreteness of the orthographic alphabet. The usefulness and efficiency of letters (plus their intimate familiarity to literate scientists) makes it very plausible to hypothesize psychological counterparts to letters. It seems quite natural to imagine a *phone* or *phoneme* as the psychological analogue of a letter. Thus linguists routinely assume that there are:

1. *discrete* differences between *phones* or *phonemes* which are

2. *drawn from a small set* of phoneme types for each language,

3. and are *serially ordered* and *nonoverlapping in time*.

4. Each phone or phoneme has an *invariant physical form* across contexts, speakers, rates of speech, etc.

5. The words of a language have a *single ``canonical'' representation* in memory, that is, some prototype version to which incoming tokens are compared.

From this traditional view, linguistic memory is primarily a dictionary with every word having its distinctive spelling in a small, abstract alphabet so it can be differentiated discretely from all non-homophones in the language. A major problem has been Assumption 4 about an invariant physical form for each phoneme. A century of phonetics research on phones and phonemes has shown unambiguously that ***there is no way to give these abstract linguistic segments concrete specification***. Successful acoustic definitions for all linguistic segments and features have not been found – in fact, there are no definitions for *any* of the segments or features that are effective across neighboring contexts, speaking rate, etc. Thus there is an incommensurability between letter representations and the actual representations in memory. Very simply, there are no ``acoustic letters'' for speech. We phoneticians spent a generation or two trying to provide physical acoustic definitions of linguistic units like particular features and segments, but we have not been able to. It almost certainly cannot be done.

So before starting to analyze the dynamics of human speech, we must accomplish some mental self-cleansing. It is important to really understand that letters in the phonetic alphabet are just one kind of model for speech. This model is very useful – especially to those of us with literacy education. It is convenient for a writing system (because it requires learning only a small number of letters). However, such a representation is not, it turns out, closely related to the way speakers, literate or illiterate, store words in memory, recognize them or produce them.

A century ago, the phonetic alphabet provided the best technology available for representing speech for scientific study. But for up to half a century now continuous-time technical displays like audio waveforms, sound spectrograms and smoothed electromyographic plots have been available to show us vividly that letter-like units of speech sound are ***not*** what people use for representing speech. We have been making excuses for this counter-evidence for over half a century. It is time to give up the illusion that words have a cognitive spelling into units that resemble letters. There are only two places where they really have such a spelling: on paper when people use orthographic letters to write words down, and in the conscious conceptualization of language by literate people.

The conclusion is that we must reject letter-like phones or phonemes as actual units that define spoken language. But then how can we begin to analyze any particular language or human language in general? We must begin afresh with a clean slate, looking at language without the biases that come from our cultural history. This literate culture demands that we develop great skill at interpreting speech sound as letters and reading letters as speech sound.

5. Two Levels of Complex System

A very different basic cut of the phenomena surrounding human speech is required, one that distinguishes the two primary complex systems that support human speech communication. It is proposed, following Smith, Brighton and Kirby 2003, that we separate:

(A) the properties of the ***language patterns of a community of speakers***, that is, what we might call the social institution of language, on one hand, from

(B) the skills of ***real-time speech processing,*** on the other.

The ``grammar of a language'' should be understood as a description of patterns across a community of speakers and contexts summarized over some historically brief time window (of, say, a generation). Linguistics is (or should be) studying the patterns and the cultural categories that are found to define the unit-like chunks in any language. Linguists should describe the most common patterns and the categories of sounds, and make whatever generalizations they are able to, stated in terms of whatever descriptive vocabulary for timing and articulatory state seem to work. But, if we look into the details, it will be clear that all speakers in a community discover their own pattern definitions.

This implies that approximation across speakers is the best that can be done. There is an unavoidable uncertainty about the precise definitions of the auditory micro-features that ``spell'' words (or word-like units) for any individual. In addition, each speaker has been exposed to a different subset of the corpus of speech in that language (differing in regional and social dialects, foreign accents, etc.). Thus any given language may have a gross description, the kind linguists and phoneticians are trained to produce using the IPA alphabet or one that is provided by an orthographic representation. But the closer we look, the more variation will be found. Spoken languages do not have a consistent or uniform finely detailed description. Very simply, there exists no universal vocabulary with which to describe it. The reason for this is, again, that the fine description in each speaker is created independently by developmental processes based on the language-exposure history of each speaker. This is why phonological observations about the sound patterns of any language exist only at the group level and are not necessarily ``in'' the individual speaker (nor even in *any* speaker). The idea that phonology only exists in the corpus of a community and not in the individual speaker's representations simply acknowledges that speakers can and will differ from each other to varying degrees.

The acquisition of the skills of real-time speech processing can now be understood in outline. The child, well before birth, begins learning the auditory patterns (human generated and otherwise) (Jusczyk, 1997). During the first year, the child learns to recognize the typical speech sound-gestures of his ambient language and classifies them into categories (Werker & Tees, 1984; Kuhl & Iverson, 1994). On this theory, the child increasingly develops his auditory analysis system to perceive speech in his ambient language and then to produce it. Notice that this system has to work without any apriori description of what the speech patterns of the community actually are. The learner has no idea in advance what ``the grammar'' really is, so the perceptual system learns as best it can to predict the patterns based on very concrete descriptions. It appears the learner makes progress by storing lots of detail about specific utterances that have been heard. Counter-intuitively (at least to literates like us), the real-time speech system makes no use whatever of the abstract segments or letter-like units that are so prominent in our conscious experience of language. Similarly, speech perception takes place with no necessity of recovering any low-dimensional alphabet-like description along the way.

The story told here implies several observations: (1) discrete categories of linguistic patterns are not the same as discrete symbol structures, (2) the abstract and partly discrete structure of phonology exists in a speaker only as a structure of categories. But phonological categories are sets of speech chunks specified in a rich code that are taken to be `the same' as members of some category. Their discreteness is never guaranteed because they are fairly high dimensional. Plus some events will found where category assignment is difficult to impossible (or ambiguous). Finally note that (3) speakers can exhibit patterns in their behavior that they have no explicit representation of (and thus the patterns are not appropriately described as `knowledge' of any kind). The child learns to perceive and produce speech that is compatible with the way others talk in his or her community but each does so using idiosyncratic components (because each is controlling only his own vocal tract and exploiting his own auditory system whose detailed structure reflects the speaker's personal history).

So then, the two complex systems of human speech are, first, the cultural institution, the set of linguistic habits, patterns and category types exhibited by the community of speakers (what we usually call the phonology, grammar and lexicon but also cultural style, etc.), and, second, the individual's working system for speech production (controlling a vast number of articulatory degrees of freedom) and speech perception (employing the speaker's partly idiosyncratic speech sound analysis system). This realtime system seeks sufficient compatibility with the community's usage patterns that the speaker can speak and be understood. The two systems each have their own development processes on very different time scales. The social institution is shaped slowly over time by changes in the statistical patterns of a community of speakers, but the speaker continues to develop new linguistic skills throughout life. We now look closer at each of these systems.

6. Language as a Social Institution

The first level of complex structure is in the **language as a system of shared speech patterns produced by speakers in a community**. When coming into the world, the child begins to have exposure to a corpus of the language of his community in various social contexts, first from the parents and later from siblings, peers, teachers and pop singers. The system of regularities in the corpus of language a child is exposed to has been self-organized by the speech behavior of the community of speakers of the language over many generations. The result is the existence of patterns that seem to underlie most word shapes, including certain ``speech sound'' types (typical consonants and vowels of the language), syllable-structure patterns, lexical items and ``grammatical patterns.'' The child learns to perceive and to produce adequate versions of this system of composite patterns as he becomes more skilled at listening to and producing speech. (Werker &Tees. 1984; Kuhl & Iverson, 1995)

There is plenty of evidence of discreteness in these phonological categories. For example, the vowels in the series *bead, bid, bade, bed* seem to be the same as the vowels in *mean, min, Maine, men* and in *peal, pill, pail, Pell* (at least in many English dialects). These similar vowel contrasts suggest a set of categories of words in English (e.g., the category that includes *bead, mean* and *peal* versus a different category that includes *bed,*

men and *Pell*). But the claim that these are best understood psychologically as categories and not as symbol types implies greatly weakening the predictions that can be made about them. There are many ways to be assigned to a mere category, some rule-based and easily brought to awareness and others quite arbitrary and explicable only historically. Linguistic categories (such as /i/ or /t/, and what pronunciations we recognize as productions of the word *and*, etc.) are just aspects of the culture of a community. We cannot expect that *bead, mean* and *peel* will be identical in respect to any particular acoustic or auditory property. Speakers do not necessarily represent these three words as sharing any specific feature. They share only the category that we represent with the IPA symbol /i/. We English speakers (or probably primarily educated English speakers) think of them as ``sharing the same vowel.'' It may be alright to speak this way, but it is important to keep in mind that they are not represented by or spelled in memory with any specific symbol token. The same analysis applies to the variants of English /t/ as found in *teach* [t^h], *stop* [t], *butter* [ɾ] and *cat* [$t^?$]. They are considered by all of us to be /t/s at the same time that we know they are not the same but differ greatly. Ultimately, the main reason to call them all varieties of /t/ is that orthographically we conventionally spell them all with a **t** (and, of course, they were once pronounced with more similar **t**-like sounds a few centuries ago).

Small changes in statistical distributions by many individuals gradually have the effect of facilitating a continual self-organization process in the community. For reasons that are not yet clear, this process tends to lead eventually to such typical phonological patterns as (a) the use of a similar vowel inventory in different consonantal contexts, (b) the symmetrical patterns of voiced and voiceless stop and nasal consonants (eg, [b d g, p t k, m n ŋ]), that appear in many languages, (c) the constraints or preferences on the form of syllables in each language, and so on.

Although languages have been known for almost two centuries now to change gradually over time, research has yet to provide a convincing explanation of how or why these changes occur. Since it is very difficult to study such slow changes in real human communities, various attempts have been made to simulate socially defined language-like systems in computational systems employing various kinds of linguistic agents (Kirby, 2001; Smith et al., 2003; Cangelosi & Parisi, 2002, Steels, 2007; Smith et al., 2003). This work, much of it based on the techniques of artificial life, seems very likely to contribute to our understanding of the process of language evolution.

7. Realtime Language Processing.

The second level of complex system that is central to human speech is the **psychological system for speech production and perception** – the neural control system that manages an individual's physiological speech system, the lips, tongue, velum, larynx and respiratory system plus the cognitive mechanisms that accomplish speech perception and understanding in realtime. The individual speaker acquires these skills in developmental time, that is, within the speaker's lifetime. However, the units of the social institution of language are not very useful for representation in memory or linguistic processing of speakers. Speech perceivers apparently rely on detailed episodic representations – much

more like an auditory description of a spectrotemporal pattern than like something spelled using a small, abstract alphabet.

It is now clear that dynamical systems play a central role in speech production and perception skills (van Gelder & Port, 1995). The individual speaker-hearer must discover ways to coordinate all their muscular systems for the production of speech and also somehow to simulate the auditory-phonetic patterns heard in the speech of others in a realtime perceptual system. These skills require several years to achieve normal competence at basic speaking and listening. The more successful scientific efforts to model the real-time processing of speech production and speech perception have focused in recent years on dynamical system models, most often implemented as ``neural networks'' (Grossberg, 1995, 2003; Guenther, 1995). These models simulate the realtime processing of language but do not rely much on the letter and word based description.

Speech Production Modeling.
There has been a great deal of research and modeling effort in the study of speech production. The challenge here is that speech production is a formidable motor problem that seems to tax the limits of human motor capabilities. It demands impressive feats of coordination and skilled control of a great many articulatory variables. There are many domains for modeling aspects of human speech performance. But, in order to give some feeling for the kind of dynamical models that appear to be successful, the DIVA model of Frank Guenther will be sketched here.

This model addresses both the problem of skilled speech production as it occurs in realtime, as well as the problem of how such skills could be acquired by the child language learner. In addition, this is one of the first models of speech perception for which specific neural sites can be proposed for each functional component of the model (although the neurophysiology will not be discussed here). The model gains inspiration from much previous work in neuroscience demonstrating, for example, that the brain contains many `maps' of cells arranged in two or three-dimensional arrays that are topographically arranged, meaning that specific regions of the map are devoted to specific content (or actions) and that neighboring regions are devoted to similar content (or action). Thus, in vision, neighboring regions of visual cortex typically correspond to neighboring regions of the visual field while in audition, neighboring regions of auditory cortex correspond to similar frequencies of sound. Typically some other features (such as frequency rise vs. fall will be represented by the axis perpendicular to the frequency axis).

In the DIVA model, we can begin the overview with the Speech Sound Map (SSM, near the upper left region of Figure 3) of an adult speaker, where each speech sound type (which might be a specific gesture, syllable or even phrase) is postulated to have a small area whose activation (and its inhibition of competing regions/gestures) sets off the pattern of neural events that give rise to the skilled production of the sound type. There are three output mappings from the Speech Sound Map. Going downward in the figure is the Feedforward Control Subsystem. Neural paths from the SSM to the motor cortex

produce the gesture sequence that the speaker learned for producing this sound type. In addition, there are 2 sets of sensory expectations generated: what the sound of this speech gesture should be like in auditory terms and what sensory patterns in the vocal tract should be like. Thus, there is a mapping from the SSM to a somatosensory target region where a pattern of expectations is generated for a specific temporal sequence of sensations from the vocal tract (lips, tongue, palate, glottis, etc.) as the gesture is produced. (These expectations were learned during an earlier babbling stage and from previous speech production experience.) A similar set of predictions is produced about the expected sequence of auditory patterns. Then as the speech gestures are carried out (due to the Feedforward Control Subsystem), they produced actual auditory and somatosensory patterns which are compared to the expectations in the Auditory and Somatosensory Error Maps. If the inhibition from the realtime sensory feedback matches the pattern of excitations from the SSM, then the error maps register zero. If there are deviations (because something expected did not occur or if something unexpected occurs in either stream, then the error is detected and some form of motor correction is applied.

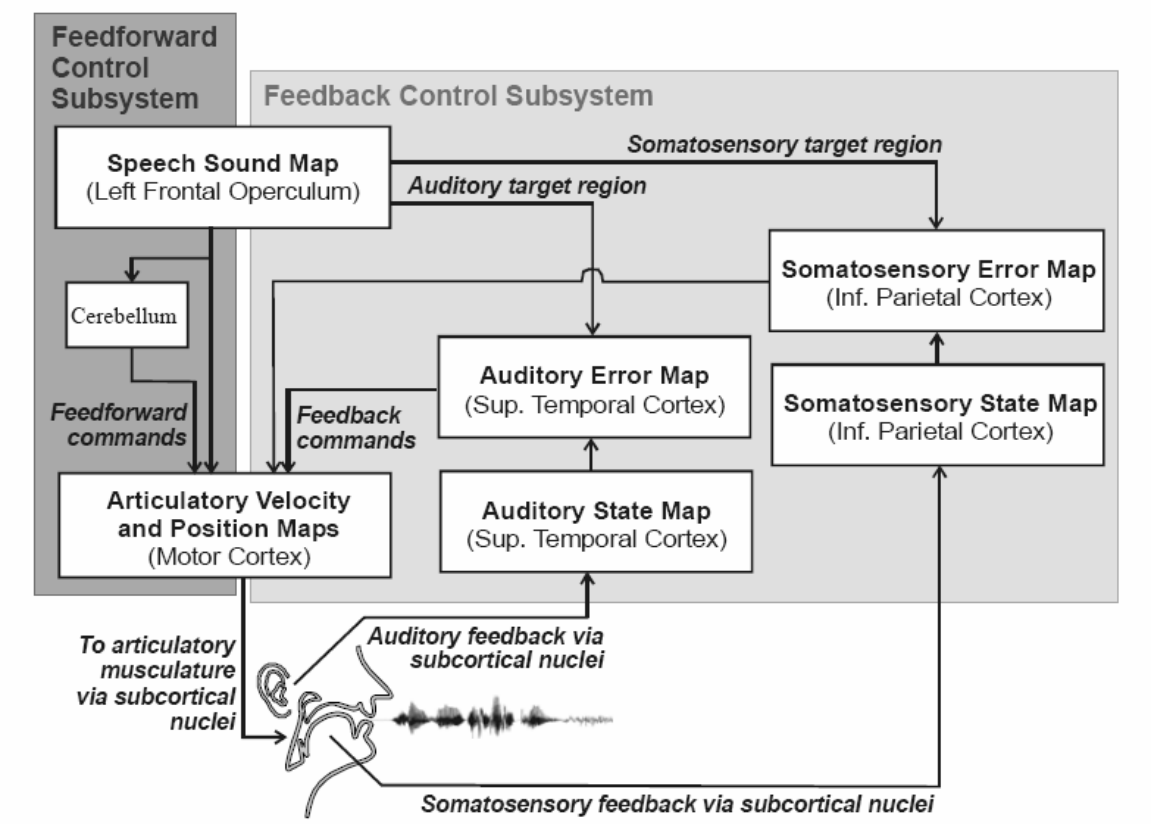

Figure 3. Schematic diagram of the DIVA model of speech production (from Guenther, 2004).

During the acquisition of speech skills, beginning with random babbling, there are many errors and the speaker learns gradually to produce better versions of feedforward gestural control, but during normal, skilled speech, there will be very few errors or deviations.

It can be seen that this system gathers whatever rich sensory information it can, both for audition and for proprioception and predicts the sensory patterning in continuous time. As long as predictions are fulfilled, all proceeds as an approximately `open loop' system. But clearly speech production is not simply a matter of issuing commands. There is much realtime control keeping the gesture on track at every instant.

Speech Perception Models.

For audition of speech, the primary problem, rather than the control of many motor variables, it is recognition of the vast number of variants that can arise in ordinary speech. Audition skills begin early since hearing is possible in the womb. Infants at birth are already familiar with much about the auditory environment their mother lives in and even with her voice (Jusczyk, 1997). Infants gradually learn the auditory patterns of their language and the many variants that speech exhibits. Increasing perception skills lead to recognition of ``chunks'' of speech that are statistically predominant (Grossberg, 2003) so speech perception results in a kind of parsing of the speech signal into recognizable chunks of variable size.

Literacy has a dramatic influence on us. Somewhere between about 12 months and 5 years, children in many communities are taught the names of the letters and proceed over the next few years (often for many years) to become more and more literate and familiar with the conventions of the written language styles of the culture as well. But the written orthography has much that is unknown in the spoken language. The written language presumes discrete words, separated by spaces, enforces fixed conventional spellings, imposes sentence structure on texts, marks proper names and paragraphs, and can even support tabular layouts of text (Harris, 2000). Literate people become very skilled at interpreting letter strings as speech and interpreting speech as letter sequences. These skills are the reason for our powerful intuitions that speech comes automatically and uniformly in letter-like form. But the description of language offered by alphabetical orthographies, despite its enormous influence on how linguists conceptualize their domain, plays almost no role whatever beyond its role in our conscious experience of language.

8. Future Directions

The discipline of linguistics has been severely hampered by the inability of linguists to escape from their intuitive understanding of language, an understanding that has been shaped by the literacy training of linguists. Because of our lifelong experience using alphabetical representations of language, it has been nearly impossible to look beyond our intuitions to see what the experimental data show. The conceptualization of the problem of language presented here says that a language is a social product and not a cognitive one. Speakers cobble together methods that are suitable for talking and hearing within the conventions and customs of their linguistic culture, but this does not imply (as we thought earlier) that the apparent units of language (the sound types, lexical entries, etc.) are themselves cognitive units used and manipulated by speakers.

Simulation work on the social aspects of language, i.e., how a language comes into being in a community and how it evolves over time, is well under way. And the study of ``sociolinguistics'', studying language variation in its social context has recently become important. Similarly, there has been much work done on speech production and perception. But the view that intuitive linguistic units, like phones, phonemes, syllables, morphemes, words, etc., must comprise stages in the realtime processing of language has proven to be a red herring that has led us astray for at least a century. It is time to let this go and to focus on the two complex systems, one communal and one personal, that underlie human speech.

Bibliography:

[1] W. Abler, "On the particulate principle of self-diversifying systems," *Journal of Social and Biological Structures,* vol. 12, pp. 1-13, 1989.

[2] R. Abraham and C. Shaw, *Dynamics: The Geometry of Behavior, Part 1*. Santa Cruz, California: Aerial Press, 1983.

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

[4] L. Bloomfield, *Language*. New York, New York: Holt Reinhart Winston, 1933.

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

[6] A. Cangelosi and D. Parisi, "Simulating the Evolution of Language," New York: Springer-Verlag, 2002.

[7] T. Cho and P. Ladefoged, "Variations and universals in VOT: Evidence from 18 languages.," *Journal of Phonetics,* vol. 27, pp. 207-229, 1999.

[8] N. Chomsky, *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press, 1965.

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

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

[11] A. Clark, "Language, embodiment and the cognitive niche," *Trends in Cognitive Science,* vol. 10, pp. 370-374, 2006.

[12] F. Cummins, "Rhythmic grouping in word lists: Competing roles of syllables, words and stress feet," in *Proceedings of the 15th International Conference on Spoken Language Processing*, Barcelona, Spain, 2003, pp. 325-328.

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

[14] 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.

[15] S. Grossberg, "The neurodynamics of motion perception, recognition learning and spatial atttention.," in *Mind as Motion: Explorations in the Dynamics of CognitioN*, R. Port and T. v. Gelder, Eds. Cambridge, MA: MIT Press, 1995.

[16] S. Grossberg, "The resonant dynamics of speech perception," *Journal of*

Phonetics, vol. 31, pp. 423-445, 2003.
[17] F. Guenther, "Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production," *Psychological Review,* vol. 102, pp. 594-621, 1995.
[18] F. Guenther, S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language,* vol. 96, pp. 280-301, 2006.
[19] F. Guenther and J. Perkell, "A neural model of speech production and supporting experiments.," in *From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*, Cambridge, Massachusetts, 2004.
[20] F. H. Guenther and M. Gjaja, " The perceptual magnet effect as an emergent property of neural map formation," *Journal of the Acoustical Society of America,* vol. 100, pp. 1111-1121, 1996.
[21] H. Haken, J. A. S. Kelso, and H. BUnz, "A theoretical model of phase transitions in human hand movements.," *Biological Cybernetics,* vol. 51, pp. 347-356, 1985.
[22] R. Harris, *The Language Myth*. London: Duckworth, 1981.
[23] R. Harris, *Rethinking Writing*. London: Continuum, 2000.
[24] C. Hockett, *The State of the Art*. The Hague: Mouton, 1968.
[25] M. Huckvale, "10 things engineers have discovered about speech recognition," in *NATO ASI Speech Patterning Conference* Jersey, England, 1997.
[26] IPA, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, England: Cambridge University Press, 1999.
[27] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features*. Cambridge, Massachusetts: MIT, 1952.
[28] F. Jelinek, "Applying information theoretic methods: Evaluation of grammar quality," in *Workshop on Evaluation of Natural Language Processing Systems*, Wayne, PA, 1988.
[29] P. Jusczyk, *The Discovery of Spoken Language*. Cambridge, Mass.: MIT Press, 1997.
[30] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants.," *Journal of Acoustical Society of America,* vol. 73, pp. 322-335, 1983.
[31] S. Kirby, "Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity.," *IEEE Journal of Evolutionary Computation,* vol. 5, pp. 102-110, 2001.
[32] S. Kirby, "Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity.," *IEEE Journal of Evolutionary Computation,* vol. 5, pp. 102-110, 2001.
[33] G. Kochanski and C. Orphanidou, "What marks the beat of speech?," *Journal of Acoustical Society of America,* 2008, in press.
[34] P. Kuhl and P. Iverson, " Linguistic experience and the ``perceptual magnet effect"," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.*, W. Strange, Ed. Timonium, Maryland: York Press, 1995, pp. 121-154.
[35] P. Ladefoged and I. Maddieson, *Sounds of the World's Languages*. Oxford, U.K.:

Blackwell, 1996.
[36] A. M. Liberman, P. Delattre, L. Gerstman, and F. Cooper, "Perception of the speech code," *Psychological Review,* vol. 74, pp. 431-461., 1968.
[37] A. M. Liberman, K. S. Harris, H. Hoffman, and B. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology,* vol. 54, pp. 358-368, 1957.
[38] L. Lisker and A. Abramson, "A cross-language study of voicing in initial stops: acoustical measurements," *Word,* vol. 20, pp. 384-422, 1964.
[39] N. Love, "Cognition and the language myth," *Language Sciences,* vol. 26, pp. 525-544, 2004.
[40] J. Mayville, K. Jantzen, A. Fuchs, F. Steinberg, and J. S. Kelso, "Cortical and subcortical networks underlying syncopated an synchronized coordination reveals using fMRI," *Human Brain Mapping,* vol. 17, pp. 214-229, 2002.
[41] 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.
[42] A. Patel, "Language, music, syntax and the brain," *Nature Neuroscience,* vol. 6, pp. 674-681, 2003.
[43] A. Patel, J. Iverson, Y. Chen, and B. Repp, "The influence of metricality and modality on synchronization with a beat," *Experimental Brain Research,* vol. 163, pp. 226-238, 2005.
[44] A. Patel, A. Lofquist, and W. Naito, "The acoustics and kinematics of regularly timed speech: A database and method for the study of the P-center problem.," *14th International Congress of Phonetic Sciences,* 1999.
[45] D. Pisoni and S. Levi, "Some observations on representation and representational specificity in spoken word processing," in *Oxford Encyclopedia of Psycholinguistics*, G. Gaskell, Ed. Oxford, UK: Oxford University Press, 2006.
[46] 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.
[47] R. Port, "Meter and speech," *Journal of Phonetics,* vol. 31, pp. 599-611, 2003.
[48] R. Port, "What are words made of?: Beyond phones and phonemes," *New Ideas in Psychology,* 2007.
[49] R. Port, "All is prosody: Phones and phonemes are the ghosts of letters," in *Prosody2008*, Campinas, Brazil, 2008.
[50] R. F. Port and A. Leary, "Against formal phonology," *Language,* vol. 81, pp. 927-964, 2005.
[51] 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.
[52] F. d. Saussure, *Course in General Linguistics*. New York: Philosophical Library, 1916.
[53] S. Scott, "The point of P-centres," *Psychological Research,* vol. 61, pp. 4-11, 1998.
[54] E. Smith and D. Medin, *Categories and Concepts*. Cambridge, Mass: Harvard University Press, 1981.

[55] K. Smith, H. Brighton, and S. Kirby, "Complex systems in language evolution: The cultural emergence of compositional structure.," *Advances in Complex Systems,* vol. 6, pp. 537-558, 2003.
[56] L. Steels, "Experiments on the emergence of human communication," *Trends in Cognitive Sciences,* vol. 10, pp. 347-434, 2006.
[57] K. Stevens and S. Blumstein, "Invariant cues for place of articulation in stop consonants," *Journal of Acoustical Society of America,* vol. 64, pp. 1358-1368, 1978.
[58] K. Tajima and R. Port, "Speech rhythm in English and Japanese," in *Phonetic Interpretation: Papers in Laboratory Phonology* vol. 6, J. Local, R. Ogden, and R. Temple, Eds. Cambridge: Cambridge University Press, 2003, pp. 317-334.
[59] W. F. Twaddell, "On defining the phoneme," *Language,* vol. Language Monograph 16, 1935.
[60] T. van Gelder and R. Port, "Its about time," in *Mind as Motion: Explorations in the Dynamics of Cognition*, R. Port and T. v. Gelder, Eds. Cambridge, Mass.: MIT Press, 1995, pp. 1-44.
[61] J. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life.," *Infant Behavior and Development,* vol. 7, pp. 49-63, 1984.
[62] T. Zanto, J. Snyder, and E. Large, "Neural correlates of rhythmic expectancy," *Advances in Cognitive Psychology,* vol. 2, pp. 221-231, 2006.
[63] B. Zawaydeh, K. Tajima, and M. Kitahara, "Discovering Arabic rhythm through a speech cycling task," in *Perspectives on Arabic Linguistics*, D. Parkinson and E. Benmamoun, Eds.: John Benjamins, 2002, pp. 113 ff.

Glossary:

phone: a `minimal' speech sound whether consonant or vowel, the unit that is represented by a single letters of the International Phonetic Alphabet. A phone is invariant across syllable positions, neighboring context, speaking rate, speaker, etc.

phoneme: an abstract speech sound unit in a particular language, typically consisting of several phone types that are treated as the same (despite any differences) by speakers of the language. Thus, English /t/ (slashes indicate use of a symbol as a phoneme, not a phone) includes both an **allophone** (a particular phoneme variant) that is aspirated (as in the word *take*), another allophone that is a glottal stop (in, the usual American pronunciation of *cotton*) and another that is a flap or tap (as in *butter*).

phonology: the branch of linguistics that studies the nature of the speech sound patterns of particular languages, such as the inventory of phonemes, the patterns of syllable construction, stress patterns, etc. Thus for the English word *stops*, the *st-* is the onset of the syllable, the vowel is the nucleus and the coda is the *–ps*. Phonology should be concerned with intonation and speech timing as well, although these are not traditional interests.

meter: an underlying temporal pattern of a fixed number of `beats' within a larger cycle. Thus, there are meters consisting of 2 beats per measure, 3 beats or 4, 5 or 6. Speech often aligns itself with such patterns, e.g., in chant or song, by locating vowel onsets close in time to the pulse of each cycle.

orthography: the set of conventions about how to write a language. This includes conventions about which letters are used and what sounds they represent in the ideal case. The conventions also include standard spellings for every word, plus conventions of capitalization, punctuation, the definition of a sentence and so forth.
