Phones and Phonemes are Conceptual Blends, Not Cognitive Letters

Robert Port (port@indiana.edu)

Departments of Linguistics and Cognitive Science, Indiana University
5975 Handy Rd, Bloomington, Indiana 47401 USA

May 1, 2011

Abstract

Despite our strong intuitions that language is represented in memory using some kind of alphabet, phones and phonemes appear to play almost no psychological role in human speech perception, production or memory. Instead, evidence shows that people store linguistic material with a rich, detailed auditory and sensory-motor code that tends, in its details, to be unique for each speaker. The obvious phonological discreteness of languages reflects conventional categories of pronunciation but not discrete symbols. In learning to read, we all master the Speech-Letter Blend, so that letters can be effortlessly interpreted as speech when reading. This mapping between letters and speech, requiring many years of training, is apparently achieved in the Visual Word Form Area of cortex. The notion of a phoneme is actually a conceptual blend of letters and speech.

Keywords: speech processing, speech perception, phonology, phones, phonemes, linguistic memory, rich memory, complex adaptive system, categorization, symbols.

Linguistic Intuitions

Linguistics, for at least the last century, has almost invariably subscribed to the following assumption:

> *Speech in any language consists of a sequence of words which are composed from a sequence of phonemic (or phonetic) symbols*.

This is the standard view of language as the archetypal symbol system and is shared by neighboring disciplines, such as psychology of language and language development. Of course, this assumption is shared by lay people as well since we all have strong intuitions that words and phonemes are just the right units for description of any language. This paper will argue that these intuitions primarily reflect conventions about alphabetical orthographies that have been assumed to be true of spoken language as well, but without sufficient examination (Port, 2007; 2010). In fact, many kinds of familiar data have been incompatible with this assumption for at least 60 years (Pisoni, 1997). But linguists and others have refused to consider seriously the idea that speech may demand much higher dimensionality and thus much richer memory for utterances than has been presumed by the standard view. Interestingly, engineers, despite similar intuitions, began over 40 years ago to turn toward speech recognition systems that seek models of whole words and phrases (rather than phonemes) and to specify them in terms of spectral detail rather than discrete symbols (Jelinek, 1969; Huckvale, 1999). There have been a few attempts to apply these insights to models of human speech perception, such as Klatt's LAFS (`Lexical Access From Spectra') program (1979), but such models have not found favor over the years.

This paper will point out some of the evidence against a compositional, low-dimensional, discrete-time description of language. Then I will suggest a new, high-dimensional view of linguistic memory that is supported by many straightforward properties of linguistic behavior. Because we were sure that words had to have a ``spelling'' of some kind, we have been trying to find a fixed and uniform description that does not exist. Phonemes live as categories in the speech patterns of a community, but are not represented identically in the brain of each speaker and are not timeless, discrete and differentiable from each other by a small number of features.

Rich Memory for Language

Although many linguists treat phonetics as the discipline that provides an inventory of discrete segments or feature vectors ``available'' for the use of all languages in the world (Chomsky & Halle, 1968), phoneticians have typically been quite coy about whether the phonetic options available to languages are unlimited (Maddieson & Ladefoged, 1969, pp. 369; IPA 1999, pp. 32-38). Indeed some phoneticians have explored very rich (high-dimensional or high bit-rate) continuous-time descriptions of speech for various languages (e.g., Browman & Goldstein, 1992; Hawkins & Nguyen, 2004; K. Johnson, 1997). In fact, there is almost no clear evidence supporting any role for an ``efficient,'' fixed-size inventory of serially ordered speech symbols for any language (Port & Leary, 2005; Port, 2007, 2010), although there is evidence for abstract, adaptable categories of speech sounds.

Here are 4 kinds of evidence that are incompatible with the idea of low-bitrate segmental linguistic memory,

i.e., memory for words that employs a small set of letter-sized units, whether phones or phonemes.

Continuously variable pronunciations.

First, if there were a small universal phonetic alphabet, then there should be noticeable discrete jumps in pronunciation within and between speakers. Yet, every utterance in a language can be pronounced with tiny phonetic variations – small changes in vowel quality, place of articulation, degree of voicing, pitch, etc. – that other speakers can imitate depending on the circumstances. These small variations can lead over time to significant changes in pronunciation over the generations (Labov, 1963). No small or discrete alphabet can account for how these gradual changes along continuous dimensions could be learned by speakers or spread across a population without the employment of a very large number of continuous phonetic variables (Bybee, 2001). Nor is there any evidence suggesting that speech sound types are actually discrete but suffer some amount of superposed noise.

Speech timing.

Second, speech production in all languages exhibits various conventional timing patterns that cannot be modeled with the letter-sized units that must be relied upon by the segmental view. Thus, a long (or geminate) consonant in Japanese is not simply either 2 or 3 singleton consonants in duration (Hirata, 2004) and English voiced and voiceless consonants in pairs like *rabid-rapid* exhibit a compensatory timing change involving the duration of both the stressed vowel and the stop closure (Lisker, 1984). Another interesting case is `mora timing' in Japanese, a tendency for vowel onsets (the most salient time points during speech) to begin at integer-spaced intervals, either one time unit later (e.g., in *kono),* or almost exactly two time units after the first vowel onset (e.g., in *chotto).* Thus the distance between the two vowels in *chotto* is twice the interval between the two /o/ onsets of *kono* (Port, Dalby & Odell 1987; Han, 1994). (The /t/ closure, on the other hand, is almost 3 times as long in *chotto* as it would be in *choto*.) None of these timing effects can be captured for perception or production using only consonant and vowel segments. Instead, speakers must control timing details that extend over many segmental time intervals. Such variable continuous-time patterns again must rely upon very rich memories for speech material. These memories support our ability to imitate the temporal details of the speech of others.

No physical invariance.

Third, most letter-sized ``sound units,'' such as the stops, fricatives and nasals, do not have acoustic correlates that are invariant across contexts the way letters are, but rather have highly context-sensitive acoustic shapes that differ widely depending on the neighboring sound (Liberman, et al, 1957). Thus, unlike letters, they do not tend to have an invariant shape. This has led to various attempts to use the articulatory invariance (of, e.g., the tongue tip closure common to /di/ and /du/) to replace acoustic invariance (Liberman, et al, 1957). But only acoustic invariance can account for listeners' ability to recognize CV syllables just from exposure to the acoustic signal. This is further evidence that listeners must employ rich, detailed representations of speech exhibiting different place cues from context to context. The main reason we want to capture the /d/ when comparing *Dee* and *dew* rather than just represent them as 2 different syllabic gestures beginning at the same place is that our writing system isolates the /d/.

Concrete memory for words.

Fourth, if words were stored in memory only in abstract, letter-like form (i.e., in a phonetic or phonemic alphabet) with no concrete auditory detail, then recognizing the repetition of a word in a verbally presented list should be equally difficult whether the repeated word was in the same voice or in a different voice. We make this prediction since nothing about the voice should be stored with the words. In fact, however, listeners always do better if the same voice is used (Palmeri, et al.,1993). This surprising result is further evidence that speakers retain richly detailed auditory representations of speech in memory, not simply conventional segmental representations.

Altogether, these results are strong evidence against the traditional view espoused by most linguists and many speech scientists. As noted above, most of the speech recognition community gave up on phones and phonemes for specifying words long ago. That should have been a hint for psychologists and linguists. The results reviewed here imply that the regularities we refer to as phones or phonemes cannot be simple symbol tokens. They must be far richer in information. In fact, they must be rather like categories.

Categories are Not Symbols Tokens.

Although linguists usually treat the terms `linguistic symbol' and `linguistic category' as basically synonymous, it is essential to differentiate them. One definition of a **category** is a group of things treated by community members as being sufficiently alike in some respect that they can be considered ``the same.'' This definition is essentially statistical and implies richness in the number of variables that may be relevant. And, of course, there is always some uncertainty about whether any speech instance belongs to one category or

another because category descriptions always exhibit uncertainty (Barsalou, 2005, p. 419). A **symbol token**, in contrast, is a member of a fixed-size list of discrete tokens, differentiable with very few degrees of freedom. The *alphabet* and the *numbers* are the ancestral, indeed archetypical, symbols of western culture. They are cultural inventions and a very important kind of technology. But, one might ask, aren't letters also highly variable and, especially in cursive form, non-discrete? Certainly, but letters are still drawn from a list of 26, so when one sees a strangely drawn letter, you can still be confident that you are looking at one or another of the 26 discrete letters. This is not true of speech sounds where there is, I claim, no fixed inventory. So speech sounds may belong to a variety of categories, but they present a very different perceptual problem from letters which belong to the technology of symbol tokens.

There are some partly discrete and symmetrical aspects of phonologies, such as the tendency for very similar speech sounds to recur in different contexts. Thus, there are often sets of words with very similar vowels (like *mad, pack, pass, land),* alliterative sets (*pat, pit, pot, putter)* and matching vowel sets *(beat, bit, bet, bat; mean, Min, men, man; peal, pill, Pell, pal),* reflecting conventions about how to pronounce words. These phonological conventions about syllable onsets and codas, consonant clusters and rhymes, etc., comprise categories of similar sounds. Speakers apparently can adapt their expectations about what these categories should sound like depending on details of the context. It has been shown speakers can generalize rapidly from deviant pronunciations by one speaker to the same sound categories by that speaker in novel contexts (Norris, et al. 2003). Such results clearly rule out a primitive exemplar model of memory for speech that can only categorize based on sounds present in memory (e.g., Hintzman, 1986). However these results still do not require a segmental model that is based on discrete-time speech symbols like phonemes. They show that speakers have the ability to generalize their categorization criteria across syllabic contexts.

How could a community as a whole create such structures of speech sounds? Communities act like *complex adaptive systems* (Beckner et al, 2009) that are capable of creating a language with relatively discrete sound categories and lexical descriptions generally without awareness (nor explicit representation) by the agents in that system. A *category* is a class of things that the culture treats as the same – such as a particular word with many variable pronunciations. No physical invariant should be expected for all instances of sounds categorized as a /t/ or an /æ/, since ``phonemes'' are simply the same by convention. Similarly, the members of the categories ``tree'' and ``game'' are whatever English speakers conventionally call instances of a ``tree'' or a ``game''. There are no defining traits or necessary and sufficient conditions.
Linguistic categories (such as words, phonemes, etc.) can only be stored in memory as statistical regularities in speech along with their conventional categorization. It is typically impossible to assign one stretch of speech to one category and the adjacent stretch to another (as one can with letters). But can language work with no real symbols? Yes. Speakers do not need them since their memories tend to be largely auditory and somatosensory, and are typically categorized (Barsalou, 2010). A language is a system of regularities or conventional speech patterns shared by a community. These patterns require much phonetic and auditory detail for their specification although they can be approximately described as nested categorical units (lexical ones, phonological ones, etc.) but individual speakers are certain to differ in the categories they employ as well as in their category definitions.
Thus, a language is a kind of social institution, an inventory of conventional speech signals – phonological, lexical and supralexical (such as collocations and idioms). Such a system evolves over many generations. Each speaker has an idiosyncratic version of it and constantly makes tiny changes in their own speech patterns. The produced patterns that result have enough nested and overlapping categories, that they may *resemble* a system composed from brief discrete components – at least the patterns resemble it well enough that a simple letter-based technology will work well enough for writing and reading. Presumably spoken languages evolve toward nested structures with many symmetries to gain the benefits of discreteness such as, at least, greater robustness in noise and to provide the resources to easily invent new vocabulary (see Abler, 1989). Of course, it cannot actually *be* a discrete system in the technical sense, since it is only maintained by social convention within an unlimited space of continuous phonetic variables, by children and young adults who imitate the speech of those who know the language better or whose speech they admire.

This hypothesis builds on the notion that the brain of each speaker is not the only complex adaptive system (Holland, 1995) that is relevant to language. The `community of speakers' itself is a complex adaptive system, one that has evolved (in most cases) through thousands of generations. Each culture creates various technologies – for nutrition (e.g., hunting, fishing, farming), for competition with other communities (e.g., martial technologies) and for coordinating the behavior of community members through speech. The language of a community is simply part of its culture and, like the rest of culture, evolves slowly on its own depending on how the group situation changes. It seems that many cognitive scientists have tended to presume that all the problems of cognition must be explained by psychology or neuroscience – by studying individual brains and behaviors. In fact, our cognition also

reflects our culture. This includes literacy training as one major component. The existence in us of literacy skills can directly influence some basic properties of our cognition.

The Speech-Letter Blend

The ideas above may sound implausible and, at the very least, non-intuitive to many readers. But when it comes to phones and phonemes, we must consider how our intuitions about language might be shaped by the skills we acquired for reading, that is, for converting letters into speech (where speech is understood to include time-locked temporally-extended patterns of both speech gestures and speech sound). Letter-like units seem overwhelmingly natural and appropriate for the description of language. My proposal is that the phone and the phoneme reflect a particular kind of *conceptual blend* (Fauconnier & Turner, 1998) that is inculcated in us beginning in childhood as we develop reading skill.

Conceptual blends come in many forms. Hutchins (2005) points out that a ``queue'' as a cultural practice is a conceptual blend of a line (a series of people in a row) with a trajector along the line producing an ordering with one end being first in the queue and the other end being the last. Some blends are humorous: in a cartoon, a worker turns away disappointed from the office coffee machine and says to an approaching coworker ``It's out of toner.'' Behaviors related to a coffee machine are blended with behaviors related to a laser printer. Part of why it's funny is that machine coffee often tastes like it was made from something as black and unpleasant as the toner of a printer. But some conceptual blends comprise important intellectual achievements. For example, the notion of negative numbers was viewed until the 18th century as logically impossible and intellectually suspect (Lakoff & Nuñez, 2000). But eventually mathematicians began to blend the notion of `counting numbers' with a line. Then zero becomes a point on the middle of the line and adding negative numbers is motion in one direction and adding a positive number is the opposite motion. The conceptual blend of the `natural' numbers with points on a line allows addition and subtraction to be understood simply as opposite movements. This blend made negative numbers intuitive and easily comprehensible.

The proposal made here is that phones and phonemes represent a conceptual blend as well. But this blend serves one specific purpose, to facilitate skilled reading. When we think about phones and phonemes, that is, think about the so-called ``sounds'' from which words are constructed, we blend some properties of speech (i.e., actual speech gestures or physical speech sound) with some properties of letters, (i.e., being discrete graphic figures selected from a short list and arrayed serially). We overlook (in fact, ignore) the little problem that many phonemes are unpronounceable in isolation.

Linguists (including me) often speak of ``speech sounds,'' in the plural, as if speech were manifestly divisible into separate letter-sized units. But we linguists and phoneticians have long known that neither the sound nor the speech gestures are divisible this way – at least since the first spectrograms were made over 60 years ago (Fant, 1962). One cannot take a sound spectrogram (or a tape recording) and divide it in time into stretches that correspond, plausibly, to letters. Yet we continue talking about ``speech sounds'' using this familiar blend anyway.

It seems to me that such a blend of speech and letters would be an expected consequence of becoming skilled at reading and writing with an alphabet. The distinctive physical shape of letters provides a concrete, material basis for thinking and reasoning about the complexities of continuous speech (Hutchins, 2005). Of course, letters are a massive oversimplification, at least, if we want to understand speech perception, production and processing. But this blend probably plays a practical role in our ability to read words we have never seen. What has happened to linguists and modern speech scientists is that we have taken the Speech-Letter Blend to be, not an idealized relationship derived from our alphabet technology – an achievement of our nervous system resulting from many years of practice using this technology – but rather taken it to describe actual cognitive tokens (based on the letter side of the blend) that words are psychologically ``spelled'' from. This blend is why we have such powerful intuitions about a letter-like cognitive representation of language. But the evidence consistently shows that memory representations of language show no evidence of being letter-like.

Neural Rewiring

How could such a drastic mistake be made without our noticing? How could we be misled so casually into a powerful conviction of the existence of something that does not exist? The basic reason is that learning to read one's language using an arbitrary set of graphic shapes is intrinsically very difficult (Ryner et al, 2001; Dehaene, 2009). It demands a systematic training program for at least several years to achieve practical skill at reading. Most of us academics probably never stop refining our literacy skills. But recent neuroscience research has revealed that one consequence of a decade or more of reading practice is the creation of a specialized region in the literate brain, normally in the left ventral occipito-temporal cortex, that expedites the linkage of visual patterns to speech pronunciation. It has come to be known as the Visual

Word Form Area (VWFA, McCandliss et al., 2003). Damage to this area of the brain can destroy reading ability (Damasio and Damasio, 1983). It seems likely that this region, created only over hundreds of hours of reading practice, may play a role in our powerful intuitions about speech having a letter-like basis. In fact, it is likely that *we cannot help thinking about language in terms derived from our orthography experience*. That is to say we tend to ``hear'' spoken language as being discrete and segmented like our written language. Although Chomsky advised linguists to trust their intuitions and interpret language in terms of the intuitions, we cannot, in fact, trust them, at least not our intuitions about speech as having discrete, letter-like form. Of course, one difficulty here is that social convention can support auditory-articulatory patterns that are *approximately* discrete units – discrete *enough* that we are able to use a discrete alphabetical writing system as a practical means to represent language on paper.

Conclusions

So, my conclusions are that:

1. There is no evidence that speakers make use of an abstract, speaker-independent, context-independent, serially-ordered, i.e., letter-like representation of language, such as that implied by all phonetic and phonological transcription schemes. There are only our powerful intuitions that they do.
2. All the supposedly ``discrete'' linguistic structures of spoken language (e.g., *distinctive features, phones, phonemes, words, sentences*, etc.) are only approximately discrete, since they are merely conventions, i.e., socially created categories, not actual psychological tokens. These structures are created by communities of speakers, but each individual speaker has only dim awareness of these categorical patterns (unless they are literate).
3. It is not only individual brains that are complex adaptive systems dealing with language. The community of speakers itself is an independent actor.
4. The units called phones and phonemes which offer a segmental, alphabet-like description of speech in any language, are not true cognitive units, but rather are conceptual blends, the Speech-Letter Blend that results from refinement of reading and writing skills.

Acknowledgements

Thanks to Carol Fowler, Ed Hutchins, Tom Schoenemann and the anonymous reviewers for helpful comments.

References

Abler, W. ``The particulate principle in self-diversifying systems. *Journal of Social and Biological Structure* 12, 1-13, 1989.

Barsalou, L. ``Abstraction as dynamic interpretation in perceptual symbol systems'' In L. Gershkoff-Stowe & D. Rakison (eds.) *Building Object Categories* (389-431) Carnegie Symposium Series, Majwah, NJ: Erlbaum, 2005.

Barsalou, L. ``Grounded cognition: past, present and future'' *Topics in Cognitive Science 2* 716-724, 2010

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T. (2009) Language Is a Complex Adaptive System: Position Paper. *Language Learning, 59*, 1-26.

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

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

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

Damasio, A. R. & H. Damasio ``The anatomic basis of pure alexia,'' Neurology 33, 1573-1583, 1983

Dehaene, S. *Reading in the Brain: The New Science of How We Read* New York, Peguin 2009.

Fant, G. ``Descriptive analysis of the acoustic aspects of speech.'' *Logos 5*, 3-17, 1962.

Fauconnier, G and M. Turner ``Conceptual integration networks'' *Cognitive Science 22*, 133-187, 2000.

Han, M. "Acoustic manifestations of mora timing in Japanese.," *Journal of the Acoustical Society of America*, vol. 96, pp. 73-82, 1994.

Hawkins, S. 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.

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

Hirata, Y. "Effects of speaking rate on the vowel length distinction in Japanese," *Journal of Phonetics*, vol. 32, pp. 565-589, 2004.

Holland, J. *Hidden Order: How Adaptation Builds Complexity*, Cambridge, Mass: Perseus Books, 1995.

Huckvale, M. "Ten things engineers have discovered about speech recognition." *NATO ASI Speech Pattern Processing Conf*, Jersey, pp. 1-5, 1997.

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

Hutchins, Edwin The material basis of conceptual blends. Journal of Pragmatics 37, 1555-1577, 2005.

Jelinek, F. "A fast sequential decoding algorithm using a stack," *IBM Journal of Research and Development*, vol. 13, 1969.

Johnson, K. "Speech perception without speaker normalization: An exemplar model," *Talker Variability in Speech Processing*, K. Johnson and J. Mullenix, eds., pp. 145-166, London: Academic Press, 1997.
Klatt, D. "Speech perception: A model of acoustic phonetic analysis and lexical access," *Journal of Phonetics*, vol. 7, pp. 279-342, 1979.
Labov, W. "The social motivation of a sound change," *Word*, vol. 19, pp. 273-309, 1963.
Lakoff, G. and R. Nuñez *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York, Basic Books, 2000.
Liberman, A. M., K. S. Harris, H. Hoffman *et al.*, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, pp. 358-368, 1957.
Lisker, L. "``Voicing" in English: A catalogue of acoustic features signalling /b/ vs. /p/ in trochees," *Language and Speech*, vol. 29, pp. 3-11, 1984.
McCandliss, B. L. Curran & S. Dehaene ``The visual word form area: Expertise for reading in the fusiform gyrus.'' *Trends in the Cognitive Sciences 7*, 293-299, 2003.
Maddieson, I. and P. Ladefoged, *The Sounds of the World's Languages*, Oxford: Blackwell, 1969.
Norris, D., J. McQueen and A. Cutler. ``Perceptual learning in speech. *Cognitive Psychology* 47, 204-238.
Palmeri, T. J., 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.
Pisoni, D. B. "Some thoughts on `normalization' in speech perception," *Talker Variability in Speech Processing*, K. Johnson and J. Mullennix, eds., pp. 9-32, San Diego: Academic Press, 1997.
Port, R. "How are words stored in memory? Beyond phones and phonemes," *New Ideas in Psychology*, vol. 25, pp. 143-170, 2007.
Port, R. "Language as a social institution: Why phonemes and words do not have explicit psychological form " *Ecological Psychology* 22, 304-326, 2010.
Port, R., J. Dalby, and M. O'Dell, "Evidence for mora timing in Japanese," *Journal of Acoustical Society*, vol. 81, pp. 1574-1585, 1987.
Port, R. F. and A. Leary, "Against formal phonology," *Language*, vol. 81, pp. 927-964, 2005.
Rayner, K., B. Foorman, C. Perfetti *et al.*, "How psychological science informs the teaching of reading," *Psychological Science in the Public Interest*, vol. 2, pp. 31-74, 2001.