Forget about *phonemes*:

Language processing with rich memory

Robert Port

Departments of Linguistics and Cognitive Science, Indiana University

port@indiana.edu, May 3, 2010

Abstract

This paper argues that phones and phonemes play almost no psychological role in human speech perception, production or memory. Instead, people store language in memory with a rich, detailed auditory and coupled sensory-motor code that is idiosyncratic to the speaker. The evidence is overwhelming that linguistic memory consists of rich, highly redundant (and idiosyncratic) memories of heard language. Now if only statistical definitions are possible for the minimal linguistic units of language, then the rest of language also cannot have a fixed inventory of units, whether words or syntactic units. The engineering implications are that language processing systems that hope to emulate human performance should seek ways to store large amounts of high-dimensional data about speech and find ways to use these rich memories.

Index Terms: speech processing, speech production, speech perception, exemplar memory, rich memory, language engineering, complex adaptive system.

1. Introduction

Speech always consists of a discrete sequence of words which are composed from a discrete sequence of phonemes (or phones). This is the standard view of language in linguistics, shared by adjacent disciplines, such as speech engineering, psychology of language, language development, etc. But these properties are actually conventions for alphabetical orthographies and have simply been assumed to be true of spoken language as well (Port, 2007). In fact, many kinds of data have been incompatible with this assumption for at least 50 years (Pisoni, 1997). But linguists, in particular, have refused to take seriously idea that speech demands much higher dimensionality and vastly more memory for utterances than is presumed by the standard view. Engineers, on the other hand began 40 years ago to turn toward speech recognition systems using whole-word models (rather than phonemes) specified in terms of spectral detail (Jelinek, 1969). There were other attempts to apply these insights to models of human speech perception, such as Klatt's LAFS (Lexical Access From Spectra) program (1979). Much more recently, in the same spirit, there is a far broader model for all of natural language processing that relies on a massive memory for heard linguistic material (Daelemans & van den Bosch, 2005). In my view these approaches, that work directly from raw data patterns, are more likely to be successful than the long-standing attempts to identify some common phonological and grammatical units that are physically represented in all speakers of a language.

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 misunderstood where a language ``lives'', we have been trying to find a description (for what the speaker ``knows'') that does not exist. Phonemes and words, etc. are in the speech patterns (the corpus) of a community, and not represented identically in each speaker's brain.

2. Why rich memory for language?

Although linguists frequently treat phonetics as a matter of a discrete inventory of segments or feature vectors ``available'' for the use of all languages in the world (Chomsky & Halle, 1968), phoneticians have typically assumed that the phonetic options available to languages are basically unlimited (IPA 1999). 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 evidence supporting any role for an ``efficient,'' low-bitrate, fixed-size inventory of discrete speech sounds for any language (Port, 2007; Port, in press).

3. Why reject low-bitrate phonological memory

Here are some of the reasons to reject the idea of low-bitrate segmental linguistic memory, i.e., memory for words that employs a small set of letter-sized units (i.e., phones or phonemes).

1. *Continuously variable pronunciations.* Every utterance in a language can be pronounced with small variations – small changes in vowel quality, place of articulation, degree of voicing, pitch, etc. – that other speakers can imitate, if so inclined. 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 variables could be learned by speakers or spread across a population (Bybee, 2001).
2. *Speech timing.* Speech production in all languages exhibits various timing constraints that cannot be modeled with letter-sized units. 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). One 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 mora time unit later,

e.g., in *kono*, or almost exactly two time units away from the first vowel onset, e.g., in *chotto* (so the distance between the two vowels in *chotto* is twice the interval between the two /o/s of *kono* (Port & Dalby, 1987; Han, 1994). None of these timing effects can be captured for perception or production using simple letter-like segments. Instead, speakers must employ representations extending over some time interval. Such continuous-time representations again point to very rich and redundant memories for speech. These memories support people's abilities to imitate the temporal details of the speech of others.

3. *No physical invariance.* Most letter-sized sound units, such as the stops [b,d,g, p,t,k], do not have invariant acoustic correlates, but rather context-sensitive ones that vary widely depending on the neighboring vowel or other sound type (Liberman, et al, 1957). Thus, unlike letters, they do not have an invariant sensory shape. This has led to various attempts to use the articulatory invariance to replace acoustic invariance (Liberman, et al, 1957). But it is acoustic invariance that is necessary to account for listeners' ability to recognize CV syllables just from exposure to the acoustic signal. Here is further evidence that listeners employ rich, spectrally detailed representations of speech.
4. *No abstract word memory.* If words were stored in memory in abstract form (i.e., in phonetic or phonemic form) with no auditory details, then recognizing the repetition of a word in a long list should be equally difficult whether the repeated word was in the same voice or in a different voice (since nothing about the voice is 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, and do not simply *reuse* the identical words or phones each time they come up.

Altogether, these results are strong evidence against the traditional view espoused by most linguists and speech scientists. If human speakers do not use these abstract representations, then speech engineers need not be concerned with them either. In fact, most of the speech recognition community gave up on phones and phonemes long ago. That should have been a hint for psychologists and linguists, but it was generally ignored (see Huckvale, 1997). The results reviewed here imply that the regularities we call phones or phonemes do not comprise an alphabet that includes everything required to specify words. Phones and phonemes are only statistical invariants, i.e., conventions about speaking, and not the psychological equivalent of bit-strings. Where do the conventions come from? They are apparently created by whole communities over many generations.

Presumably these concrete memory representations can be employed for speech perception by comparing an incoming speech stimulus to many many exemplars and identifying the syllable-, word-, or phrase-sized chunks with the category identities of the closest match (cf. Hintzman, 1986; Grossberg, 2000). A *category* is a class of things that the culture treats as the same – e.g., as multiple pronunciations of a particular word. Thus, no physical invariant should be expected for a /t/ or an /a/, since phonemes, like words, are simply the same by convention. Similarly, the members of the categories ``tree'' or ``game'' are whatever English speakers agree to call instances of a ``tree'' or a ``game''. There are no defining traits or necessary and sufficient conditions. It is strictly a matter of convention, but the relevant categories are linked as part of the exemplar –like memories.

4. Where does a language `live'?

Linguistic units, i.e., linguistic categories, therefore can only be stored in memory as statistical regularities in speech together with their conventional categorization. They are not physically definable tokens. But can language work with no physical symbols? Yes. Speakers do not need them since their memories are largely concrete but categorized. A language is a set of regularities or conventions shared by a community, and not a set of physically definable symbols.

Thus, a language is a kind of social institution, an inventory of ``agreed upon'' speech signals, phonological, lexical and supralexical. Such a system evolves over generations. Many speakers make tiny changes in the speech patterns of their language such that the resulting patterns *resemble* a componential system (see de Boers, 2000). Presumably it evolves this way to achieve the benefits of discreteness for its community of speakers. But, of course, it cannot actually *be* a discrete system (in the mathematical sense) since it is only maintained by convention – by children and young adults imitating the speech of those who know the language better than they do.

These ideas rely on the notion that speakers' brains are not the only *complex adaptive system* that is relevant to language (Holland, 1995). The community of speakers is also a complex adaptive system, one that has evolved (in most cases) through thousands of generations, creating various community technologies for finding food (hunting, fishing, farming), for defense/offense and for coordinating the behavior of community members using speech. Thus the language of a community is just part of its culture and, like the rest of culture, evolves slowly on its own depending on the situation. For too long cognitive science has presumed that all the problems of cognition must be solved by a common human psychology. In fact, much of cognition is what we learn from our cultural training, a training that includes literacy as a major component as well as an important training tool.

5. What role does written language play?

Understanding the ideas in the last section may be difficult because our logic and our intuitions about language have been very strongly shaped by our experience using the written language. Alphabetical writing represents graphically a particular perspective on the spoken-language conventions of a people. Thus hand gestures, many facial expressions, intonation and speech timing are ignored by most orthographies – even though they are not clearly distinguishable from the conventions that orthographies do represent. Because we have all learned to internalize our orthography vividly, we tend to think linguistic units are internal and part of human knowledge.

About 3k years ago, the first full alphabet was engineered by modifying the Phoenician writing system (a variant of the consonant-only Semitic writing in use for a thousand years by that time). This early Greek writing system represented the consonant and vowel phonemes of spoken Greek fairly well. Variants of this alphabet were soon adapted for use in many human communities.

But learning to read one's language via an arbitrary set of graphic shapes is intrinsically difficult (Ryner et al, 2001). It takes systematic training for several years, sometimes beginning with `alphabet blocks' at age 2, to become skilled with reading and writing. In fact, most of us never stop sharpening our literacy skills. But the consequence of decades of practice is that our intuitions about what a language is and what components it has are strongly shaped by our training with our discrete orthography. *We cannot help thinking about language in terms that lean heavily on our orthography.* This means we assume spoken language is as neat and discrete as our written language. But we need to take into account the effects of literacy on our intuitions and look again at the evidence about spoken language units without strong assumptions about what we will find. Doing this, it becomes obvious that spoken language is not discrete nor can it be described using tokens that have just serial order but no continuous time. The difficulty is that social convention can support auditory-articulatory patterns that are *approximately* discrete – discrete *enough* that we are able to use a discrete graphical writing system to represent them. But mere social convention cannot produce genuinely discrete units of sound. One needs paper and pencil – additional technology – for that.

So, my conclusions are that:

1. There is no evidence that speakers make use of an abstract, speaker-independent, context-independent serially-ordered representation of their language, such as that implied by all phonetic and phonological transcription schemes. There are only our powerful intuitions that they do.
2. All the ``discrete'' linguistic structures of spoken language (e.g., *distinctive features, phones, phonemes, words, sentences*, etc.) are only approximately discrete, since they are only conventions, i.e., socially created structures, not generally psychological ones. These structures are created by communities of speakers, but each individual speaker has only dim awareness of these categorical patterns (unless literate).
3. It is not only individual brains that are complex adaptive systems dealing with language. *The community of speakers is itself an independent actor*. Indeed, it creates the linguistic conventions, such as what can be called a ``game'' or an instance of /t/, as well as whether /t/ shares a [- voice] feature with [s].
4. If segments and syllable types are social conventions, then so must the rest of a language be: its lexicon, its grammar, etc. This implies lack of discreteness and temporal extension. So syntax (i.e., serial structure) cannot be neatly separated from realtime patterns.
5. It is not just important for cognitive science to understand language as embodied, but it is also important to see that the social group as a whole is a complex adaptive system that creates and maintains structures of many kinds, including systems of phonology, tense systems, case systems and, indeed, lexicons and supralexicons, consisting of patterns of speech conventions.

All these conclusions have great consequences for speech technology. First, engineers who seek to emulate the behavior of individual language users should abandon attempts to find a place for phonemes and explicit syntactic rules in memory. Instead, they should focus on how to store heard utterances in a form that is useful for comparison to incoming utterances. Second, they should be concerned with how sentence-sized fragments can be stored so they can support comparison with incoming sentences.

6. Engineering implications of rich memory

Our conclusion has been that speech is stored in full auditory detail and that abstract so-called ``speech sounds'' like phones and phonemes play no role. On this view, each ``word'' (or whatever category) has a large number of representations in the episodic memories of each speaker as well as across speakers (depending on what corpus fragment of speech each has heard). So how could the rest of language – the utterance understanding aspects – be any different? If rich memory episodes are always being stored, then processing the rest of language, i.e., words, phrases, utterances, etc., must surely exploit the information available in each speaker's personal linguistic corpus.

This author is not aware of an engineering approach to speech perception or production that implements these ideas about speech perception, although work along these lines may already exist. LAFS was, after all, an early model along these lines (Klatt, 1979). If speech recognition could be dealt with by employing a large corpus of detailed, categorized utterances, then one could address higher problems such as understanding utterances using variants of the same method. That is, given that words could be reliably recognized, then a novel utterance could be interpreted by seeking similar phrases in similar contexts and using those to guide interpretation of a novel utterance.

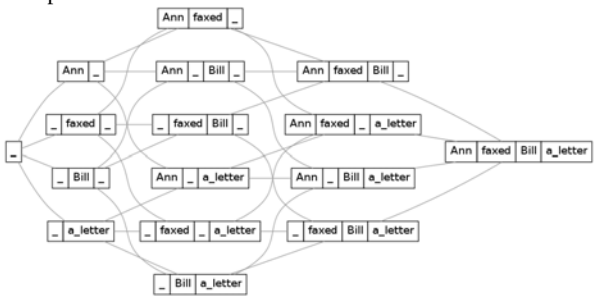

Figure 1. *A pattern lattice treatment of a sentence with the Pattern Lattice software (Kuroda (2009)*

Similarly, doing language processing of written language, then discreteness can be assumed (and the speech perception problem does not arise) but similar questions can be asked about how to combine information from a large number of utterances. Addressing this problem, the study by Kuroda (Kuroda, 2009) seems like a good beginning for development of a model for text understanding that is compatible with the ideas presented above about phonology. If (a) creative language use and (b) understanding novel utterances require comparing and combining utterance fragments with various shapes and sizes, then people must have a way to do rapid comparisons of the fragments of one utterance with fragments of a great many others. Kuroda developed his Pattern Lattice model to break each written utterance into fragments of various sizes (usually words in specific locations), to serve as indices into memory. These fragments will be kept in storage until a stimulus wakes them up. Figure 1 illustrates the breakdown of ``Ann faxed Bill a letter.'' This display uses standard English orthography, although a model emulating human performance should use patterns with a continuous time axis.

Because each utterance is stored as a collection of partial templates, they partially match many other utterances that are only partially similar, such as, say, *Ann googled Bill a picture*, even if nothing like it exists in the analyzed corpus. Notice that Kuroda's model will effectively store speech chunks that are the size of word-like lexical items or items that are larger than words – the `supralexical' units. This seems an essential move. The size of chunks used by speakers needs to be highly variable since there can be no assumed distinction between a lexical item and any larger (or smaller) unit.

In sum, the beginning of Kuroda's work on this problem is promising and exploits many implications of my phonetic and phonological results regarding speech.

7. Conclusions

Symbolic models of human cognition and language in particular have frequently pointed to the efficiency obtained by using abstract symbols as coding units for memory and cognition (Chomsky, 1965). But these supposed efficiencies presume that there are severe constraints on human ability to remember detailed episodic information. It turns out that there are seemingly no absolute restrictions on episodic memory. This has been repeatedly demonstrated over the past 40 years for all modalities, including vision and audition. Many linguists, however, continue to evaluate their analyses by ``efficiency'' – how complex the notating description is.

In engineering, as well, for entirely different reasons, memory capacity has become much less of an issue. So engineers should now be taking advantage of cheap memory to exploit what is now clear about human linguistic behavior – that speakers can rely on richly detailed memory for speech to interpret linguistic constructions that they hear or see.

8. Acknowledgements

Thanks to Kow Kuroda for inspiration to write this paper and to Pierre Divenyi and Nobuaki Minematsu for helpful discussion of these issues.

9. References

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

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

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

[4] J. Coleman, "Phonetic representations in the mental lexicon," *Phonetics, Phonology and Cognition*, J. Durand and B. Laks, eds., pp. 96-130, Oxford: Oxford University Press, 2002.

[5] W. Daelemans, and A. v. d. Bosch, *Memory-Based Language Processing*, Cambridge, UK: Cambridge University Press, 2005.

[6] B. de Boers, "Self-organization in vowel systems.," *Journal of Phonetics*, vol. 28, pp. 441-465, 2000.

[7] A. Goldberg, *Constructions at Work: The Nature of Generalization in Language*, Oxford: Oxford University Press, 2006.

[8] S. Grossberg, and C. W. Myers, "The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects," *Psychological Review*, vol. 107, pp. 735-767, 2000.

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

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

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

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

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

[14] M. Huckvale, "Ten things engineers have discovered about speech recognition." pp. 1-5.

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

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

[17] K. Johnson, "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.

[18] D. Klatt, "Speech perception: A model of acoustic phonetic analysis and lexical access," *Journal of Phonetics*, vol. 7, pp. 279-342, 1979.

[19] K. Kuroda, "Pattern lattice as a model for linguistic knowledge and performance." pp. 278-287.

[20] W. Labov, "The social motivation of a sound change," *Word*, vol. 19, pp. 273-309, 1963.

[21] A. M. Liberman, 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.

[22] L. Lisker, "``Voicing" in English: A catalogue of acoustic features signalling /b/ vs. /p/ in trochees," *Language and Speech*, vol. 29, pp. 3-11, 1984.

[23] 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.

[24] D. B. Pisoni, "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.

[25] R. Port, "The graphical basis of phones and phonemes," *Second Language Speech Learning: The Role of Language Experience in Speech Production and Perception.*, M. Munro and O.-S. Bohn, eds., pp. 349-365, Amsterdam, Holland: John Benjamins, 2006.

[26] R. Port, "How are words stored in memory? Beyond phones and phonemes," *New Ideas in Psychology*, vol. 25, pp. 143-170, 2007.

[27] R. Port, "Language as a social institution: Why phonemes and words do not have explicit psychological form.," *Ecological Psychology*, in press, 2010.

[28] R. Port, J. Dalby, and M. O'Dell, "Evidence for mora timing in Japanese," *Journal of Acoustical Society*, vol. 81, pp. 1574-1585, 1987.

[29] R. F. Port, and A. Leary, "Against formal phonology," *Language*, vol. 81, pp. 927-964, 2005.

[30] K. Rayner, 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.