OSU Working Papers in Linguistics 50 (101-113)

The auditory/perceptual basis for speech segmentation *

Keith Johnson
kjohnson@ling.ohio-state.edu

Abstract: Language is temporal in two ways. Words and sentences occur in time, each utterance having a beginning and end. But, also the learner's experience of language occurs over time, the items that are crucial for defining linguistic structure are experienced over the course of years. These two observations are addressed in an exemplar model of phonological learning and word recognition. Major features of the model are described and its operation is illustrated in two simulations.

Introduction

Language unfolds slowly over time. Of course, sentences and words are temporal events to be segmented and analyzed. But in addition to this local temporal structure, the experience of language as a whole occurs over time. Only rarely is an explicit phonological contrast demonstrated to the child. Most elements of language structure if they are to be learned must be extracted from memory. That is to say, the contrasts and similarities, the crucial comparisons from which linguistic structure emerges, are based on remembered instances of linguistic objects.

Consider the role that the linguist's 3x5 cards, notebook, or relational database plays in producing a linguistic analysis. The cards are used to write transcriptions of words, which are drawn from work with consultants who teach the linguist how to say things in a given language. These records are the starting point for linguistic analysis. They are culled and compared, stacked according to similarities. Words with very similar pronunciations but very different meanings reveal phonemes, the minimally contrastive sounds in the language, and words with slightly divergent pronunciations but similar meanings form paradigms, revealing patterns of inflection or word derivation.

The point is that, for both child and linguist, linguistic structure - the analysis of language into its combinable elements - crucially relies on a pre-analytic store of linguistic items in memory.

This is one of the considerations which has led me to explore a class of models called instance-based or exemplar models of linguistic memory and speech recognition. Before going on to describe some simulations of the process by which linguistic structure emerges from specific instances in memory I will briefly outline some further considerations which point to an instance-based model of speech recognition.

* This paper was presented as a part of the OSU Seminar on Human Communication sponsored by the Cognitive Science Program, Department of Linguistics, Department of Psychology, and Department of Speech and Hearing Science, on March 31, 1997. I am grateful to the organizers of the seminar, Osamu Fujimura, Rob Fox, and Neal Johnson for inviting me to speak. For helpful discussions on various aspects of this work I would like to thank Anne Cutler, Mary Beckman, Mark Pitt, Rob Fox, and the members of the OSU Linguistics Laboratory. Thanks also to Kim Ainsworth-Darnell for helpful comments on an earlier version of the paper.

Exemplars in speech processing

A traditional concern in the theory of speech perception is that phonemes vary quite considerably across talkers and contexts. That is, the acoustic cues for phonemes lack invariance, and consequently pose a difficult problem for theories of speech recognition (and for automatic speech recognition systems).

It turns out that variation across talkers can be reduced by normalization schemes (see figure 1). For example, Potter & Steinburg (1952) noted that the ratios of vowel formant frequencies show much less variation across talkers than do their absolute values. Observations such as this have led researchers (Bladon, Henton & Pickering, 1984; Miller, 1989; Syrdal & Gopal, 1986; Sussman, 1995; Traunmüller, 1981) to assume that linguistic categories are recognized by reference to 'higher order invariants' like formant ratios. (Gibson, 1966, was especially influenced by this argument.)

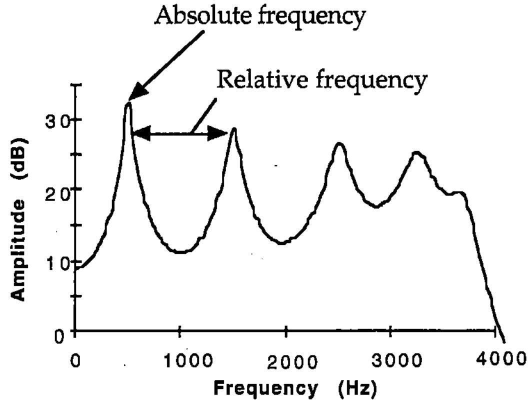

Figure 1. Formant ratios as higher order invariants. Distances between formants show less between-talker variability than to absolute formant frequencies.

So, the basic scheme for recognition in this view has two stages, (1) normalization then (2) comparison with category prototypes. Simply put, this approach doesn't work. When you implement it you get recognition performance that doesn't come close to human performance, and then you have to build separate mechanisms to recognize speakers, dialects, styles, and so on. Miller (1989, see figure 2)) showed that the shapes of the category regions in a derived 'higher order' perceptual space are quite irregular, and adequate performance is best achieved by demarking category regions by reference to exemplars. The vowel regions that Miller (1989) presented show a multimodal structure even in the 'higher order' space.

This "normalize and compare" scheme doesn't work because Potter & Steinburg's observation is only approximately true. Talker differences are only partly eliminated in higher-order invariants and the remaining differences are enough to disrupt recognition by reference to prototypes. This is true even for very constrained laboratory speech such as in the Peterson & Barney (1952) database. If we consider even small variations in speaking styles (isolated words versus words in carrier phrases) we see further overlap and multimodal distributions.

Another consideration is the fact that recent research has found that prior exposure to an utterance facilitates later recognition. For example, Goldinger (1997) found evidence for the retention of word exemplars in tests of implicit memory. If the identity of the talker was the same across repetitions of a word in successive blocks of word recognition trials

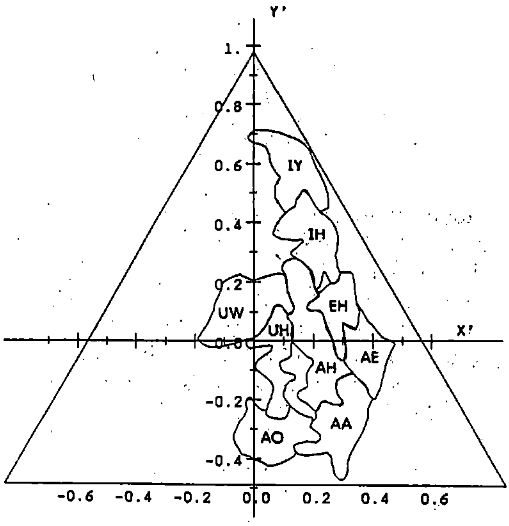

Figure 2. Miller's (1989) auditory/perceptual space. Category regions in a 'higher order invariant' representation are irregularly shaped; category boundaries are determined by exemplars near the boundary.

(even if the blocks occurred one week apart) listeners were able to recognize the word more accurately than if the repeated word was spoken by a different talker. A gain in word recognition accuracy across repetitions in the two blocks of trials occurred in all nine conditions in Goldinger's study, varying over delay intervals and number of talkers. These results (and others like them) show that low-level acoustic details of word presentations are retained and have an effect on later processing.

Finally, neurophysiological studies of memory show that single events alter synaptic strengths, and even the number of synapses, in the hypocampus. I don't want to make too much of this other than to note that I take these studies to indicate that long-term changes in neurological organization though perhaps small may result from single events. These findings lend a bit of plausibility to an exemplar-based model of speech recognition, however indirectly relevant they may be in other ways.

Whole-word exemplars

The Goldinger (1997) study and and other work along the same lines suggest that remembered instances of speech are acoustically detailed as would also be expected on psychophysical grounds. Additional evidence suggests that not only are speech exemplars acoustically detailed, but at least during language acquisition they are also unanalyzed whole words.

Bregman (1990) outlines several principles of primitive auditory scene analysis, which break an auditory array into objects - the fan of the projector, the cough at the back of the room, the utterances of the talker, and so on. In this view, these auditory objects have beginnings and ends, but no internal structure. Assuming that speech recognition begins with auditory scene analysis, we then would have to say that speech recognition starts with unanalyzed wholes.

The 'holistic' stage in language acquisition shows that children at first learn words without internal structure. One reason to believe that this is the case is that during acquisition there is a period of rapid vocabulary growth (often called an 'explosion') which suggests that the child has learned to analyze auditory objects into recombinable articulatory primitives. This sudden change in behavior is evidence that at an earlier stage language was not so organized.

A related observation is that phonological inventories and alternations are position specific. For example, the inventory of contrastive vowels in English depends on context; in my dialect there are 10 contrasting vowel qualities in [hVd] context but only 6 in [hVr]. Also, in many languages consonant voicing contrasts are 'neutralized' in coda position, and in all languages the acoustic cues for consonants differ between onset and coda. These facts are relevant to the view that speech exemplars are unanalyzed wholes because they show that similarity and contrast depend upon temporal location within utterances.

Assumptions

In summary the work described in this paper starts from three assumptions. First, speech is recognized by reference to stored instances (exemplars). Second, these exemplars have no internal structure, rather they are unanalyzed auditory representations. And third, they are word-sized chunks, as a result of primitive auditory scene analysis where isolated word productions form the basis for word recognition in running speech.

The reader may prefer to think of these assumptions as relevant for the development of linguistic representations during language acquisition, but the evidence suggests that adults also use exemplars in word recognition. So keep in mind as we discuss some simulations that these properties may be active in the mature recognition system.

How is speech 'analyzed' into segments?

Given these assumptions, in particular that remembered instances of speech are stored as unanalyzed wholes, how can speech be analyzed into segments?

One answer is that it isn't, that the segmentation of speech is a figment of the imagination fired by orthography. Some reasons to believe that this stance is incorrect have already been mentioned. The 'lexical explosion' argument, for example, is evidence for both preanalytic representation and of segmentation. Three additional observations suggest that segmentation is not merely an invention.

Listeners and talkers experience the speech stream as a sequence of separate words, any one of which can be repeated or replaced. Though a model that assumes word-sized exemplars may readily handle such segmentation (see Johnson, 1997), it should be noted that primitive auditory scene analysis does not. To achieve word-level segmentation in running speech we must posit a system in which word-sized exemplars support a cognitive scene analysis in which the recognition system segments the speech stream into words.

We can also note briefly that writing systems generally reflect analyses of speech into recombinable units such as segments or syllables, and their very existence suggests the psychological reality of sublexical units at some point in history, though not necessarily the use of these units in on-line speech processing. Also, segmental speech errors suggest that segmental organization is used in speech production.

The remainder of this paper describes an exemplar-based model of auditory word recognition - focusing particularly on behavior of the model which is related to segmentation. These simulations explore the degree to which linguistic structure may be an emergent property of recognition based on remembered auditory representations of speech.

The model

The basic operation of an exemplar model (Nosofsky, 1986) is to categorize perceptual objects by evaluating the similarity between the item to be categorized and a set of stored category exemplars. Within-category variation is explicitly represented in the set of exemplars (which substantially out-number the categories). Similarity between exemplars and the unknown is an exponential function of auditory distance (1) where d_{ij} is the Euclidian distance between exemplar *j* and the unknown object *i*. and κ is a sensitivity parameter. Word activation is the weighted sum of similarity (2) where W_{jc} is the connection weight between exemplar *j* and word *c*..

$$sim_{ij} = exp(-\kappa d_{ij}) \qquad (1)$$

$$act_c = \Sigma sim_{ij} * W_{jc} \qquad (2)$$

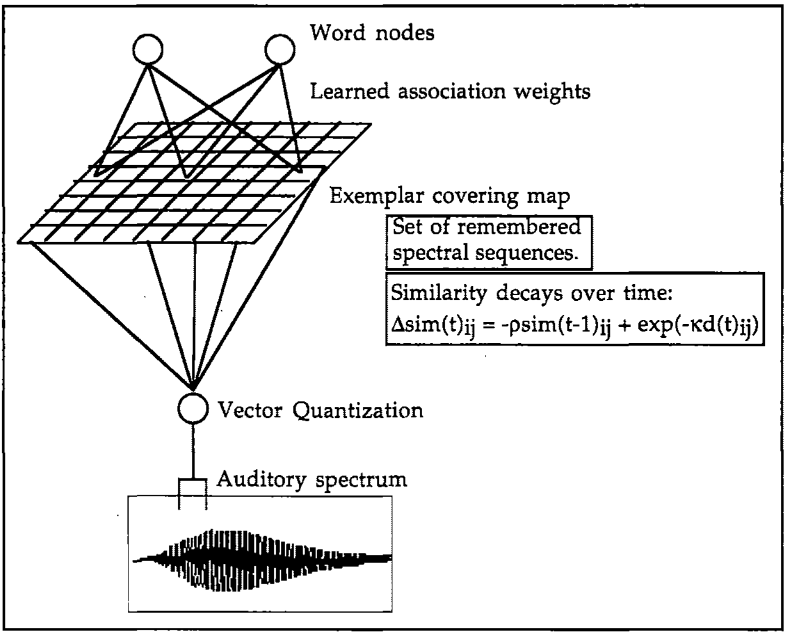
Word nodes
Learned association weights
Exemplar covering map
Set of remembered spectral sequences.
Similarity decays over time:
Δsim(t)ij = -ρsim(t-1)ij + exp(-κd(t)ij)
Vector Quantization
Auditory spectrum

Figure 3. An exemplar model of auditory word recogntion. Processing proceeds from the bottom to the top of the graph. Each 23 ms frame of speech is processed by an auditory model, vector quantized, and compared with the set of remembered sequences in the exemplar covering map. Word node activation is the product of similarity to the covering map location and the learned associations between that location and the word node.

The model has three stages of processing (see figure 3). The first stage converts the speech wave form into a sequence of auditory spectra. I use a very simple psychoacoustic critical-band filtering routine (Johnson, 1990), with a frame rate of 43 Hz. The auditory spectra are then coded as most similar to one of a set of stored spectra. This vector quantization stage is not strictly necessary for the simulations that I discuss here and introduces some noise, but is necessary for systems that store a large number of exemplars because with vector quantizing each spectrum in an exemplar can be stored as a single integer rather than as a vector of real numbers. The vector-quantizing stage uses adaptive resonance theory (Carpenter & Grossberg, 1989). The in-coming spectrum is compared with the spectra stored in the vector quantizing codebook and if it is not similar to any of them it is added to the codebook, given a code number, and that number is returned. If it is similar to one of the stored spectra, that spectrum's code number is returned and the stored spectrum is shifted slightly to be more similar to the in-coming spectrum. The degree of noise introduced by vector quantizing is thus determined by a 'vigilance' parameter which determines when spectral templates will be added to the VQ codebook. This approach is also used in the third stage during training but not during test to build an exemplar covering map.

In the third stage the sequence of auditory spectra is compared with the sequences stored in an exemplar covering map (Kruschke, 1992). As with vector quantizing, if the in-coming sequence is unlike any existing exemplar it is added to the map. Weights connecting locations in the exemplar covering map and word nodes are learned during training by counting the number of times that the covering map location is an instance of the word. The weight connecting the closest exemplar in the covering map and the 'correct' word is incremented by one for each training token. It should be noted that though the architecture shown in figure 3 is similar to Kruschke's (1992) ALCOVE model, the use of exemplars is more similar to Nosofsky's (1986) GCM.

The model assumes that similarity is evaluated one time-frame at a time starting at the onset of the auditory objects being compared, and that activation decays over time as a function of a decay parameter ρ:

$$sim(t)_{ij} = -\rho sim(t-1)_{ij} + exp(-\kappa d(t)_{ij}) \qquad (3)$$

Simulation of the recognition of 'cap'.

This model was trained to recognize eight words (Table 1) spoken in list-reading style by a single male speaker. The utterances were recorded directly to computer disk at a 22 kHz sampling rate, using 16 bit samples. The words were chosen to illustrate segmental contrasts in CVC words, and a case of a 'phantom' word. 'Catalog' and 'battle-log' may be confusing for a word recognition system because right context '-alog', '-le-log' distinguishes the short words 'cat' and 'bat' from the longer words 'catalog' and 'battle-log'. The system being tested in this simulation is not time invariant, but rather assumes that the beginnings and endings of the words are known. Nonetheless, we will see some interesting behavior in the segmentation of the longer words.

Table 1. Words used in the first simulations.

bap	cap
bat	cat
battle-log	catalog
beet	keep

The model was trained on the first 10 of 13 repetitions of each word. During training the codebook and exemplar covering map were constructed and weights between exemplars and words were established in one pass through the first 10 repetitions of the words. This very simple training algorithm led to 96% correct recognition of the remaining three repetitions of the words.

Figure 4 shows word activations as a function of time (which is given in frame number) during the presentation of 'cap' to the model. A spectrogram of the instance of 'cap' being recognized is shown in approximate alignment with the frames. At frame 1 all of the words starting with [k] are more activated than are the [b] words. At frame 3 activation of 'keep' drops out, giving us a set of activated words that start [kæ]. At frame 5 all of the words with [æ] as the first vowel show increasing activation. At frame 10 activation for 'catalog' drops off (as did 'battle-log' at frame 7), perhaps because of vowel duration mismatches. At frames 13-16 the words that end in [p] show increasing activation, though this increase is only slight for the correct answer 'cap'.

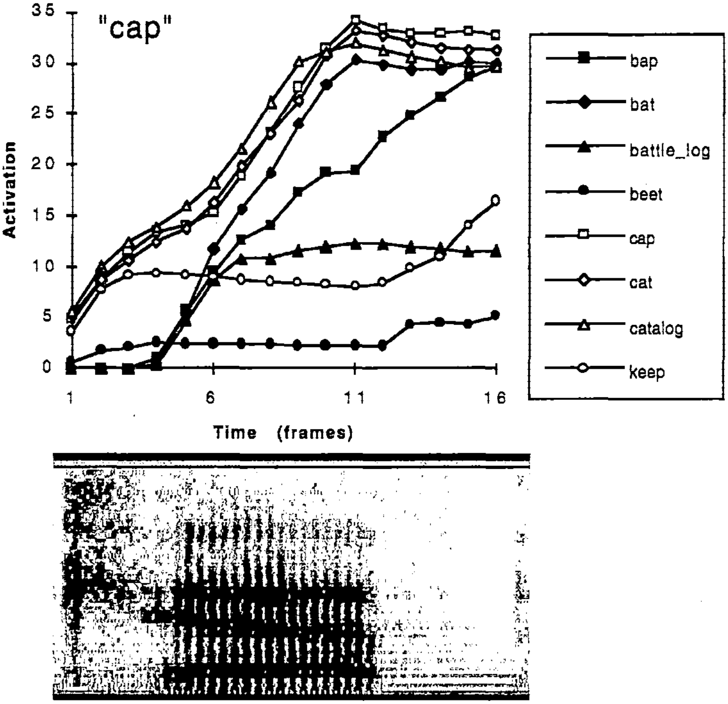

Figure 4. Spectrogram of the word 'cap' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

One interesting aspect of this simulation (and of the one to follow) is that right context cues seem to exist in the onset syllable 'catalog'. Because pronunciations of a sequence like [kæt] differ phonetically in monosyllabic and disyllabic words, a model such as the one discussed here which is sensitive to acoustic detail begins detecting the difference between them even during the initial CVC sequence. One argument supporting lexical competition in models like TRACE (Elman & McClelland, 1986; McClelland & Elman, 1986) and Shortlist (Norris, 1994) is that the initial sequence of phonemes in 'catalog' is no different than the sequence 'cat'. This simulation suggests though that word pairs such as this which seem to require right context for disambiguation are not completely ambiguous without right context. Hence, some of the work done by lexical competition is accomplished by attention to phonetic detail in the exemplar model.

These patterns of activation also show three emergent segments. /k/ emerges in the first frame as the subset of lexical items which begin with [k], /æ/ emerges at frame 5 as the set of words having first vowel [æ], and /p/ emerges near the end of the word as the words with [p] codas show increasing activation. This analysis suggests that phonemes are defined in terms of subsets in the set of exemplars which have time-aligned similarities in their auditory/perceptual representations.

This account of the emergence of segments from unanalyzed exemplars has some interesting properties. The emergent segments are based on auditory similarity and are position-specific. They are not, however, Wicklephones (Wicklegren, 1969). That is, we see here evidence that all words with [æ] show increasing activation during the vowel, not just those that have the same consonantal context. (We will see below a case in which a context sensitive allophone emerges.)

Though I am interpreting the pattern of activation in 'cap' as having emergent segments, there are no segmental representations in memory being activated in response to the signal. That is, no recombinable units exist in the memory representations of the words. This is because I have stored exemplars as containing auditory descriptions only, as formalized in (4) where E_O is a set of exemplars defined by auditory properties A. If we assume that the child's own productions are exemplars that contain both auditory and motor descriptions, as formalized in (5) where E_S is a set of exemplars defined by both auditory properties A and motor commands M, we can speculate that the sequence of activated lexical subsets that we have just seen gives rise to the activation of a sequence of articulatory gestures.

$E_O = <A>$ exemplars produced by others (4)

$E_S = <A,M>$ exemplars produced by self (5)

Simulations of the recognition of 'catalog' and 'battle-log'

We turn now to simulations of the recognition of 'catalog' and 'battle-log' using the same vocabulary and trained model that were just described.

Figure 5 shows word node activation levels over time in response to an instance of the word 'catalog'. Many of the segmental phenomena that we saw in the 'cap' example are evident here as well. For example, as before in the first frame all of the [k] initial words show increased activation in response to the word 'catalog'. Also in frame 6 all of the [æ] words show increasing activation.

But in addition to these segmental phenomena we see at the end of the word (frame 16 and after) that both 'catalog' and 'battle-log' show increasing activation at about the same rate over time. In this case the unit of linguistic structure which is being defined by a lexical subset is a syllable. A hierarchical structure of syllables and segments emerges from the activity of the model.

This is apparent also in the word activations in response to the word 'battle-log' which are shown in figure 6. Some segmental phenomena are seen during the first syllable while over the course of the second syllable both 'catalog' and 'battle-log' show increasing activation.

Figure 6 also shows the context sensitive allophonic response that was mentioned earlier. In the first frame only the three [bæ] words are activated. The word 'beet' remains virtually unactivated during the entire course of the word 'battle-log', despite the fact that they both start with 'b'. This is a topic for future investigation, but this simulation does suggest that there may be circumstances in which the subsets of activated lexical items generated by the model define allophones rather than phonemes or syllables.

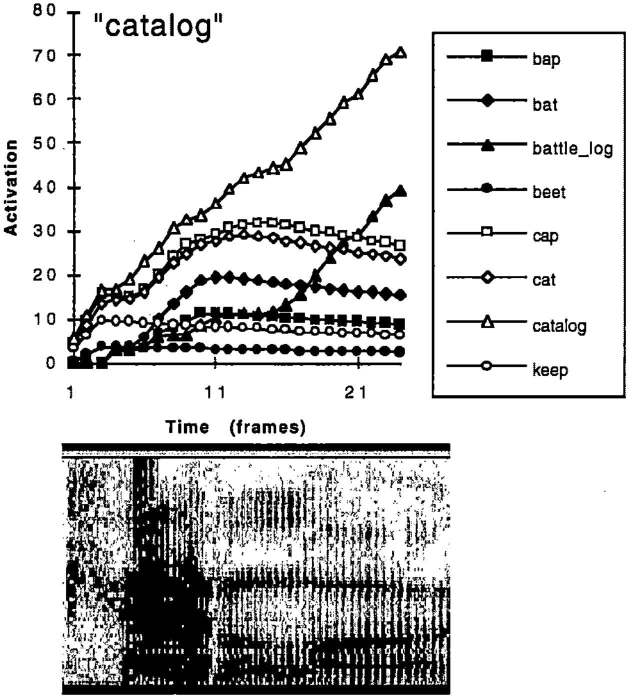

Figure 5. Spectrogram of the word 'catalog' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

Simulating the metrical segmentation strategy

The final simulations use a different database of training tokens and then replicates Cutler and Norris' (1988) finding that it is easier to spot the word 'mint' in a nonword with a strong-weak metrical structure like 'mintuf' ['mɪntəf] than it is in a nonword with a strong-strong metrical structure like 'minteif' ['mɪn,teɪf]. The model was trained on the words listed in Table 2 as before and then tested on the nonwords. Each word in table 2 was repeated eight times in isolation by a single male talker and recorded directly to computer disk with 22 kHz, 16 bit sampling.

Table 2. Words used in the second simulations.

training words:		test nonwords:	
mint	men		
minty	rented	mintuf	minteif
mints	minted		
retain	maintain		

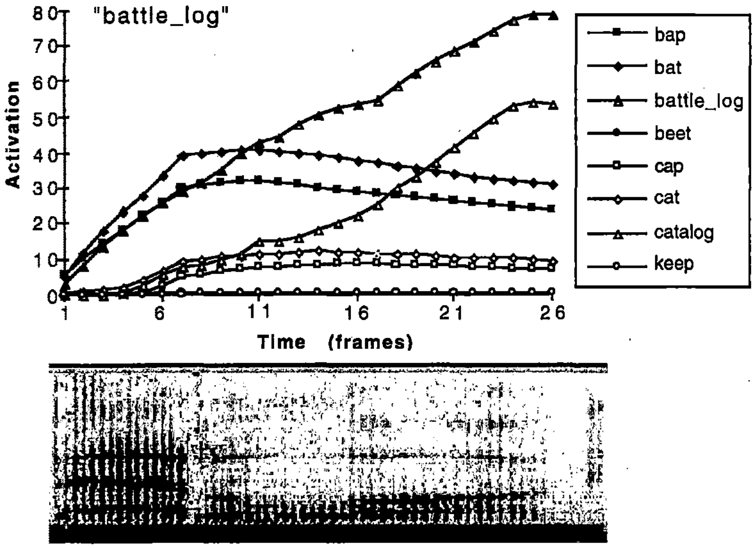

Figure 6. Spectrogram of the word 'battle-log' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

Figure 7 shows the activations of the eight words in the lexicon in response to the non-word 'mintuf', which is shown in the spectrogram. In frame 1, activations of the words that start with [m] are greater than those that start with [r]. The most highly activated word in the eight word lexicon is 'minted', and during the first syllable 'mint' and 'mints' show relatively high activation (near a value of 25). By the end of the word there is a cluster of relatively activated words having activation somewhat less than 'minted'; these were 'mint', 'mints','minty', and 'maintain'.

Now compare this with the pattern of activations prompted by 'minteif' (Figure 8). Some segmental phenomena are apparent in this simulation. As before, in frame 1 words starting with [m] are more activated that words starting with [r]. Also, as in figure 7 the final fricative in 'mintuf' seems to have partially overlapped with the final fricative in 'mints' indicating that the model is sensitive to mid-class phonetic similarity (Dalby, et al., 1986).

The most highly activated word in the lexicon was 'maintain' which like this production of 'minteif' has two metrically strong syllables. The other words which show fairly high activation in response to 'minteif' are the two syllable words in the lexicon which have a strong first syllable. Interestingly, 'retain' which was pronounced by this speaker with a weak first syllable only showed increasing activation during the second syllable of 'minteif'.

Finally, note that the activation of 'mint' peaks at about 15. Given that the activation of 'mint' reached 25 in response to 'mintuf' we would predict that it would be easier for the model to spot 'mint' in 'mintuf' just as it was for Cutler and Norris' subjects.

This simulation, in addition to modeling Cutler and Norris' result without explicitly segmenting the speech stream into metrical feet, shows that like segments and syllables, metrical units may emerge as sets of activated lexical items in an exemplar-based recognition model.

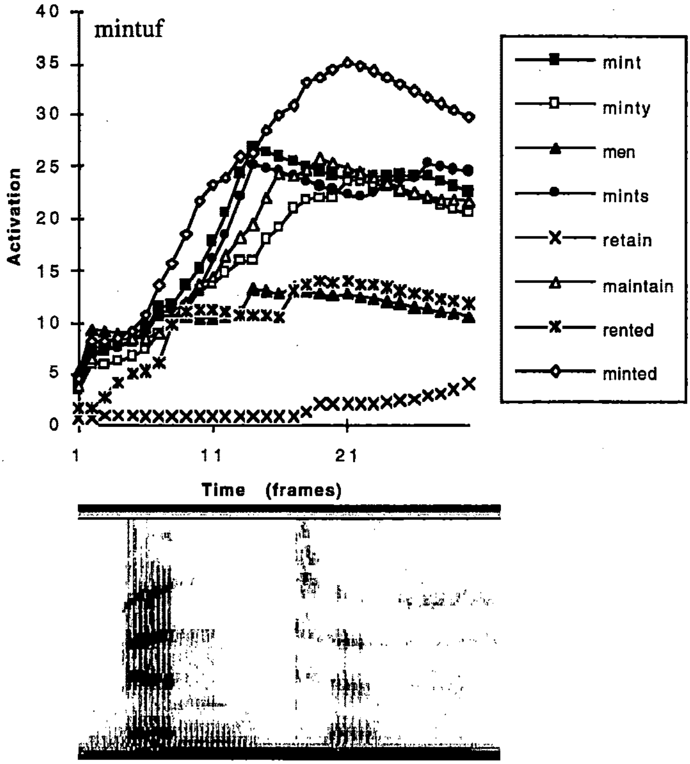

Figure 7. Spectrogram of the nonword 'mintuf' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top)..

Conclusion

One way to describe these results is say that they describe a developmental process by which a child learning language might build representations of abstract linguistic units like segments, syllables, or metrical feet. Indeed, I was inspired in this line of research by a talk by Jan Edwards on phonological disorders in language acquisition.

However, if you accept any of the arguments supporting the view that adult speech recognition is an exemplar-based process, then we can raise the interesting possibility that abstract phonological structure is a fleeting phenomenon - emerging and disappearing as words are recognized.

This may explain what we mean when we say that the speaker/hearer has implicit or unconscious knowledge of phonological structure. Abstract phonological structure in this view is never explicitly stored or detected, though the subsets of lexical items which define these abstract entities, for both the language user and the linguist, are implicitly linked through their auditory/perceptual similarities.

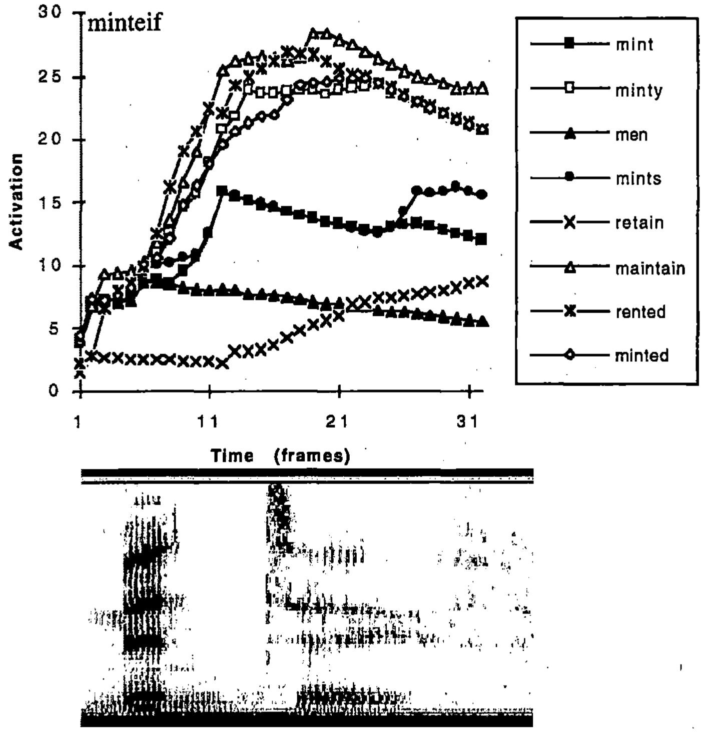

Figure 8. Spectrogram of the nonword 'minteif' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

References

Bladon, A., Henton, C. & Pickering, J.B. (1984) Towards an auditory theory of speaker normalization. *Language Commun.*, 4, 59-69.

Bregman, A. (1990) *Auditory Scene Analysis*. Cambridge: MIT Press.

Carpenter, G.A. & Grossberg, S. (1989) Search mechanisms for adaptive resonance theory (ART) architectures. *International Joint Conference on Neural Networks*, Washington, DC, June, 18-22, 1989 (Vol I, pp. 201-205). Piscataway, NJ: IEEE.

Cutler, A. & Norris, D. (1988) The role of strong syllables in segmentation for lexical access. *J. Exp. Psych.: Hum. Perc. & Perf.*, 14, 113-121.

Dalby, J., Laver, J. & Hiller, S.M. (1986) Mid-class phonetic analysis for a continuous speech recognition system. *Proceedings of the Institute of Acoustics*, 8, 347-354.

Elman, J. & McClelland, J. (1986) Exploring lawful variability in the speech waveform. In S. Perkell & D.H. Klatt (Eds.) *Invariance and Variability in Speech Processing* (pp.

360-385). Hillsdale, NJ: Erlbaum.

Gibson, J.J. (1966) *The Senses Considered as Perceptual Systems*. Boston: Houghton-Mifflin.

Goldinger, S.D. (1997) Words and voices: Perception and production in an episodic lexicon. In Johnson, K. & Mullennix, J.W. (Eds.) *Talker Variability in Speech Processing* (pp. 33-66). NY: Academic Press.

Johnson, K. (1990) Contrast and normalization in vowel perception. *J. Phon.* **18**, 229-254.

Johnson, K. (1997) Speech perception without speaker normalization. In Johnson, K. & Mullennix, J.W. (Eds.) *Talker Variability in Speech Processing* (pp. 145-166). NY: Academic Press.

Kruschke, J. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psych. Rev.*, **99**, 22-44.

McClelland, J. & Elman, J. (1986) The TRACE model of speech perception. *Cog. Psych.*, **18**, 1-86.

Miller, J. (1989) Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am.*, **85**, 2114-2134.

Norris, D. (1994) Shortlist: a connectionist model of continuous speech recognition. *Cognition*, **52**, 189-234.

Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *J. Exp. Psych.: Gen.*, **115**, 39-57.

Peterson, G.E. & Barney, H.L. (1952) Control methods used in a study of the identification of vowels. *J. Acoust. Soc. Am.*, 24, 175-184.

Potter, R. & Steinburg, J. (1950) Toward the specification of speech. *J. Acoust. Soc. Am.*, **22**, 807-820.

Sussman, H.; Fruchter, D. & Cable, A. (1995) Locus equations derived from compensatory articulation. *J. Acoust. Soc. Am.*, **97**, 3112-3124.

Syrdal, A. & Gopal, H. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.*, **79**, 1086-1100.

Traunmüller, H. (1981) Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.*, **69**, 1465-1475.

Wickelgren, W.A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psych. Rev.*, **76**, 1-15.