THE ROLE OF STRUCTURAL CONSTRAINTS IN AUDITORY WORD RECOGNITION

H. C. Nusbaum and D. B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405, USA.

In the past, much of the research on human speech perception has focused on the recognition of acoustic-phonetic properties of isolated CV and CVC syllables. The tacit assumption of this research has been that our understanding of auditory word recognition is contingent upon solving the problems inherent in phoneme perception. By this assumption, auditory word recognition is equivalent to visual word recognition carried out one letter at a time. Indeed, most current theories of auditory word recognition directly reflect this sequential pattern matching approach to word recognition. However, a different perspective is that word perception may be approached as a problem of "weak" constraint satisfaction, in which the structural properties of words in the lexicon interact to specify the identity of an utterance. We will present the results of several analyses of the phonotactic constraints of word patterns that suggest the type of constraints that may be used by human listeners to mediate spoken word recognition.

RECOGNITION IN THE CONTEXT OF THE LEXICON

Context exerts an undeniably strong influence on perceptual processes. However, it is interesting to note that "context" is defined in almost all speech research by whatever stimulus information is presented immediately prior to or subsequent to a target stimulus. Thus, a phoneme is perceived in the context of a syllable, a syllable is perceived in the context of a word, and a word is perceived in the context of a sentence. In all cases, there are objectively definable physical dimensions to the context that is typically investigated. But there is another context that affects word perception as well: the implicit context of the mental lexicon. Beyond the listener's explicit knowledge about words, the structure and organization of the sound patterns of lexical entries may serve as an implicit context within which recognition occurs.

Marslen-Wilson and Welsh (1978) called attention to the potential importance of the structural properties of words with the cohort theory of word recognition. According to this theory, the initial sounds in a stimulus word activate all the words in the lexicon beginning with those sounds. Inappropriate candidates in the cohort are then deactivated when a mismatch occurs in comparing the left-to-right order of subsequent segments in the stimulus with the structures of activated candidates. The word that is ultimately recognized is the candidate that remains after all the other incompatible candidates have been deactivated.

According to cohort theory, the activated cohort of word candidates in the lexicon forms the mental context for spoken word recognition. However, unlike the sentential context that may precede a spoken word, this context has no physical dimensions that can be directly measured or analyzed. In the past, this has posed a problem for investigating the role of the lexicon in word recognition. However, several computer-readable databases of orthographic and phonetic representations of words have recently become available for analyzing the structural properties of words in the lexicon. The database used for all the analyses we will describe contains orthographic, phonetic, and syntactic information for 243,000 words (see Crystal, Hoffman, & House, 1977). Proper names and possessives were excluded from the analyses, leaving about 126,000 words that were examined in the database.

PHONOTACTIC PATTERNS IN THE LEXICON

Although the listener may be presented with spoken words as a temporally distributed sequence of segments, a recognition process need not compare these segments to lexical representations in a strict left-to-right order as claimed by some theories. Indeed, it is unclear how serial pattern matching strategies can recognize a word if the initial segment of the input is obscured, degraded or ambiguous. Since this initial segment is treated as the index into the lexicon, recognition could not proceed without a well-defined access point. An alternative approach is to view auditory word recognition as a constraint satisfaction process, in which the propagation of a number of weak constraints is used to specify the recognized word. When viewed as a constraint satisfaction process, a number of constraints may simultaneously be applied to the lexicon to refine the set of word candidates. Even if one constraint is inappropriate or uninformative, the intersection of the other constraints may still specify the correct word. Given this view, it is important to determine precisely which constraints are actually used during word perception.

The approach that we have taken to investigate structural constraints on human auditory word recognition was motivated by several recent studies that investigated the relative heuristic power of various classification schemes for large vocabulary word recognition by computers. Zue and his colleagues (Huttenlocher & Zue, 1984; Shipman & Zue, 1982) have shown that a partial phonetic specification of every phoneme in a word results in an average candidate set size of about 2 words for a vocabulary of 20,000 words. The partial phonetic specification consisted of six broad phonetic manner classes. Thus, with this approach, a recognition system need not accurately identify the phonemes in spoken words. Instead, only the most robust manner information must be coded. Using a slightly different approach, Crystal et al. (1977) demonstrated that increasing the phonetic refinement of every phoneme in a word from four broad phonetic categories to ten more refined categories produces large improvements in the number of unique words identified in a large corpus of text.

It is important to note that these computational studies examined the consequences of partially classifying every segment in a word. Thus, they actually employed two constraints: the partial classification of each segment and the broad phonotactic shape of each word resulting from the combination of word length with patterned phonetic information.

The analyses that we have carried out used a large lexical database of 126,000 words to study different constraints that might be appropriate for describing human auditory word recognition. This work extends the previous research of Zue and his colleagues to a much larger set of words. In addition, since human listeners are capable of recognizing much more phonetic information than just six manner categories, we have carried out analyses based on the assumption that human listeners will be able to identify some segments completely, while other segments will be unanalyzed.

The results of these analyses are quite revealing about the recognition constraints provided by the structural properties of spoken words. For the coarsest level of segmental analysis, that is knowing only the length of a word in number of phonemes, the search space is reduced from 126,000 words to 6,342 words. Clearly, word length is a very powerful constraint for reducing the candidate set in the lexicon by about two orders of magnitude, even without any detailed segmental phonetic information. Furthermore, the length constraint is strongest for relatively long words. If the length of a word is 21 segments, there are only two candidates out of 126,000 words. Thus, as word length becomes extreme, less detailed segmental information is needed to identify a word.

By simply classifying each segment as either a consonant or vowel (i.e., two categories), without providing any more detailed phonetic description, the reduction in the search space beyond the the length constraint phonotactic constraint is enormous. The number of candidates is reduced by an order of magnitude to 109 words averaged across different word lengths. Furthermore, it is interesting to note that much of this reduction in the candidate set is due to the specific phonotactic constraints provided by the ordering of consonants and vowels. If the segments in a word are classified with just two categories, as consonants or vowels, but the order information is removed, there are 1196 words in the average candidate set. This means that the phonotactic order information in the pattern structure of a spoken word accounts for an order of magnitude reduction in the candidate set size compared to just knowing the number of consonants and vowels, but not their arrangement.

Increasing the amount of phonetic detail for each segment to the six manner classes used by Zue and his colleagues reduces the search space by another two orders of magnitude from the CV classification scheme that maintains phonotactic order information. Using six categories for classifying every segment in each word reduces the average candidate set size to about 5.5 words from 126,000 words in the lexicon. This result agrees very well with the results reported by Shipman and Zue (1982) for a 20,000 word lexicon, indicating that this broad classification scheme is very powerful in reducing the number of word candidates in the search space. Increasing the lexicon by an order of magnitude from 20,000 words to 126,000 words only results in a tripling of the number of candidates from 2 to about 6 words. By any metric, partial information about every segment is an extremely effective constraint on the candidate set.

However, human listeners are capable of resolving much more phonetic detail than just six broad categories. One issue that can be raised then, concerns the constraint provided by complete phonetic information about some of the segments in a word compared to partial information about every segment in a word. Classifying every segment in a word provides two types of information: (1) partial phonetic information about every segment, and (2) the phonotactic "shape" of the entire word. By comparison, complete classification of some of the segments provides: (1) detailed phonetic information about a few segments, and (2) partial information about the phonotactic shape of a word. Based on the previous demonstration of the power of phonotactic shape with just two categories (i.e., consonant or vowel), it seems reasonable to predict that partial classification of every segment in a word should be more effective than complete classification of some of the segments in a word.

To test this prediction the following analyses were carried out: (1) the phonetic information in first half of every word was classified completely leaving the remaining segments unclassified, (2) the phonetic information in the last half of each word was classified completely leaving the first half unclassified, (3) only the consonants were phonetically classified leaving the vowels unlabeled, and (4) the vowels were phonetically classified leaving the consonants unlabeled. The results demonstrate that complete information about some of the segments in a word provides a more powerful constraint on the candidate set than partial classification of every segment. Classifying the beginning of words completely reduces the search space from 126,000 words to 1.7 words and classifying the last half of words reduces the candidate set to 1.9 words. By comparison, classifying only the consonants exactly and leaving the vowels unclassified yields a set size of 1.4 words, while classifying the vowels only yields a set size of 3.2 words. In each analyses, complete phonetic information about some of the segments in a word constrains the search space much more than partial classification of every segment. These results demonstrate that detailed phonetic information about some of the segments in a word provides enough constraint, in general, that other segments can be completely obscured or ambiguous without significantly impairing recognition. Moreover, to the extent that some phonetic information is available about other segments, the candidate set will be reduced further, probably to the extent of uniquely specifying the correct word.

CONCLUSIONS

The view of word recognition that emerges from these analyses differs substantially from serial pattern matching approaches. As more of a stimulus word is heard, the listener progressively narrows the candidate set based on the development of a phonotactic specification for the input. Over time, acoustic information in the stimulus is successively refined into more detailed phonetic representations. In some cases, only a broad phonetic description of segments may be computable and the phonotactic structure is used to further narrow the candidate set. This approach, called Phonetic Refinement Theory, is currently being implemented as a model of the recognition process. Although further research is needed, it is clear that computational analyses of the sound patterns of words can provide new information about the processes that mediate speech perception.

REFERENCES

Crystal, T. H. , Hoffman, M. K., & House, A. S. (1977) Statistics of phonetic category representation of speech for application to word recognition. Princeton, NJ: Institute for Defense Analysis.

Huttenlocher, D. P. , & Zue, V. W. (1984) A model of lexical access based on partial phonetic information. Proceedings of ICASSP-84, New York: IEEE Press, Volume 2.

Marslen-Wilson, W. D., & Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.

Shipman, D. W., & Zue, V. W. (1982) Properties of large lexicons: Implications for advanced isolated word recognition systems. Proceedings of ICASSP-82, New York: IEEE Press.