How Listeners Find the Right Words

Anne Cutler

Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

Abstract: Languages contain tens of thousands of words, but these are constructed from a tiny handful of phonetic elements. Consequently, words resemble one another, or can be embedded within one another, a coup stick snot with standing. The process of spoken-word recognition by human listeners involves activation of multiple word candidates consistent with the input, and direct competition between activated candidate words. Further, human listeners are sensitive, at an early, prelexical, stage of speech processing, to constraints on what could potentially be a word of the language.

THE VOCABULARY OF A LANGUAGE

Languages differ in how they construct words: some use a rich system of inflections, some use none; some exploit suprasegmental as well as segmental contrasts; some allow lexical forms to vary according to the context in which they are produced while others avoid such variation. But it is safe to say that every natural human language has a vocabulary running into the tens of thousands.

No language has a phonemic inventory running into even the low hundreds. The largest phonemic inventory size listed by Maddieson [1] for the UCLA Phonological Segment Inventory Database is 141, and the mean and median in that database both lie around 30. English, with an inventory in the forties, is thus in the top quartile of the languages surveyed. Nonetheless, the size of the phonemic repertoire is obviously trivial in comparison to the size of the vocabulary. And not only are there few phonemes; strict constraints rule the order in which they may occur to constitute a word. Thus the string *string* contains five phonemes, but of the 120 orderings that are conceivably possible for a string of five elements, only one is allowed by the phonotactic constraints of English (a couple more are pronounceable, but are ruled out by voicing assimilation constraints). Inevitably, in such a situation, the words of a language resemble one another strongly, and are often found to be embedded within one another. McQueen and Cutler [2] report the (very high) statistics on embedding for the vocabulary of British English. Cutler, McQueen, Baayen and Drexler [3] extended these analyses to a corpus of naturally spoken English; a representative result of their investigation is shown in Figure 1.

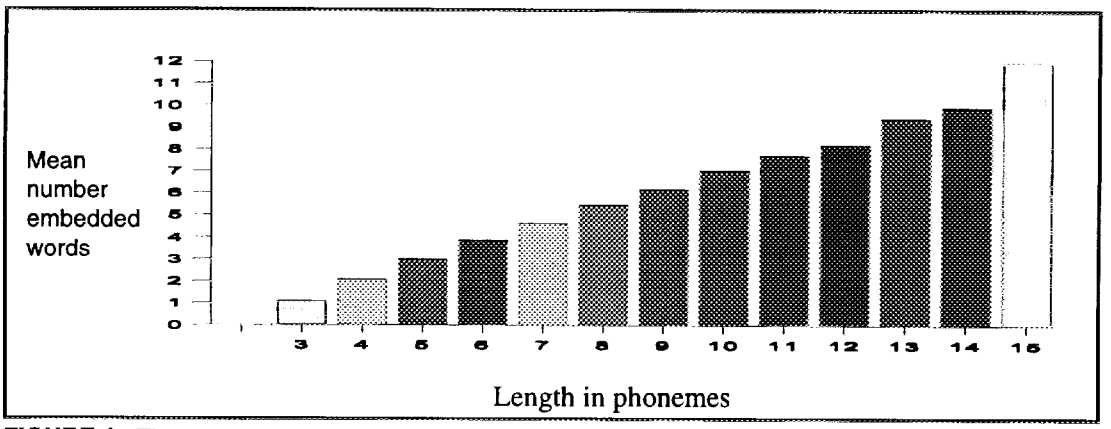

FIGURE 1. The mean number of embedded words within words occurring in the MARSEC corpus of naturally spoken British English, as a function of word length in phonemes.

Thus most speech signals effectively contain phantom words, i.e. words which are embedded within the real words uttered by the speaker. However human listeners clearly manage to deal with this situation and understand spoken language rapidly and without noticeable effort. The well-known problem of the continuity of the speech stream, and the absence of reliable signals corresponding to the boundaries of lexical units, cannot prevent human listeners from efficiently segmenting speech signals into their component lexical units. Psycholinguistic investigations of how this is achieved have produced, in recent years, some fascinating results.

THE WORD-SPOTTING TASK

This is a laboratory task in psycholinguistics which was developed for the specific purpose of studying the segmentation of words from a speech context. It resembles word-spotting as the term is used by engineers working on automatic speech recognition, in that the listener's task in a word-spotting experiment is to find any known lexical element which may occur anywhere in an unpredictable context. However, in most word-spotting experiments the context is minimised so that confounding factors can be ruled out when some crucial aspect of the context is manipulated. Thus each stimulus item might consist of just one or two syllables, and may or may not contain a real word: *crinthish*, *obzel*, *lunchef*.... As soon as subjects spot any real word, they press a response key, and then say aloud the word they have spotted - in the above example, they should respond to *lunch* in *lunchef* (Fig. 2). Their keypress responses yield a measure of response latency; their spoken responses are recorded to ensure that the intended word was spotted. Since some contextual manipulations are predicted to make word-spotting difficult, miss rate (failure to spot an embedded word) can in some experiments be as informative as response latency.

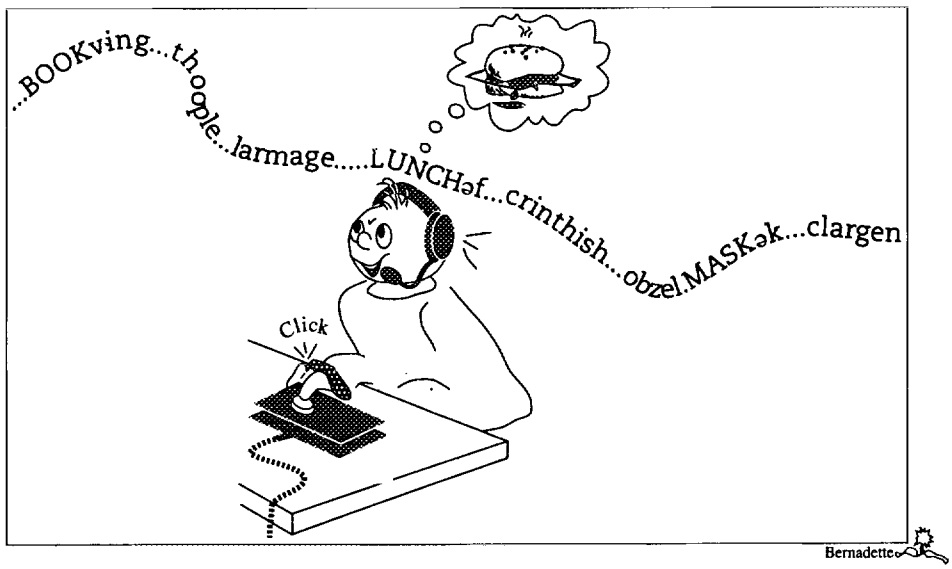

FIGURE 2. Listener in the word-spotting task receives auditory input comprising a sequence of nonwords; whenever a real word is spotted within one of the heard nonwords, the listener presses the response key and then speaks the word aloud.

The task was first used by Cutler and Norris [4], who added a VC context to CVCC words such as *mint* or *jump*, with the added VC containing either a full vowel (*mintayf, jumpoove*) or the reduced vowel schwa (*mintef, jumpev*). Cutler and Norris found that the CVCC words were spotted much faster with the following reduced-vowel context than in the full-vowel context, which they explained as the effect of a segmentation strategy employed by English-speakers whereby syllables with full vowels were assumed to be word-initial. Thus the second syllable of *min-tayf* would be segmented from the first and detection of the embedded word would be slowed by the necessity of recombining material across a point at which the (putatively automatic) segmentation procedure had triggered. Corpus analyses demonstrated that such a strategy would be highly efficient for natural English spoken texts [5], and corroborating experimental evidence arose in studies of juncture perception [6].

CONCURRENT ACTIVATION AND INTER-WORD COMPETITION

Current models of human spoken-word recognition (such as TRACE [7]; Shortlist [8]; or the latest version of the Cohort model [9]) assume that words which are compatible with the input are automatically activated and compete with one another for recognition. In the Shortlist model [8], competition is instantiated as interactive activation including lateral inhibition between units at the same level. At the word level, any candidate word competes with other, "phantom",

words which incorporate portions of the same input - thus in the example *acoustics notwithstanding*, the first word will compete with *coup*, *stick*, *tick*, etc., the second word with *knot*, *stand*, etc., and at least one candidate, *snot*, will be activated which contains part of each word in the input (Fig. 3). Words with partial support (*cool*, *tickle*, *wither* etc.) will be temporarily activated but will drop out of the competition process as incoming information does not match their full form. The more a given word is activated by the incoming speech, the more it is able to compete with - inhibit - other activated words; as words are inhibited, they lose activation and therefore become less able to inhibit other words. Thus in Figure 3, *notwithstanding* enters the shortlist after only its first four phonemes have been heard; its activation is temporarily reduced by competition from other candidate words such as *snot* and *stand*, but recovers once subsequent information (plus the ensuing competition) rules out those competitors. This process will eventually lead to only one successful competitor for each part of the input; at the end of the string, the two most highly activated candidate words are *acoustics* and *notwithstanding*, corresponding to the actual content of the input string.

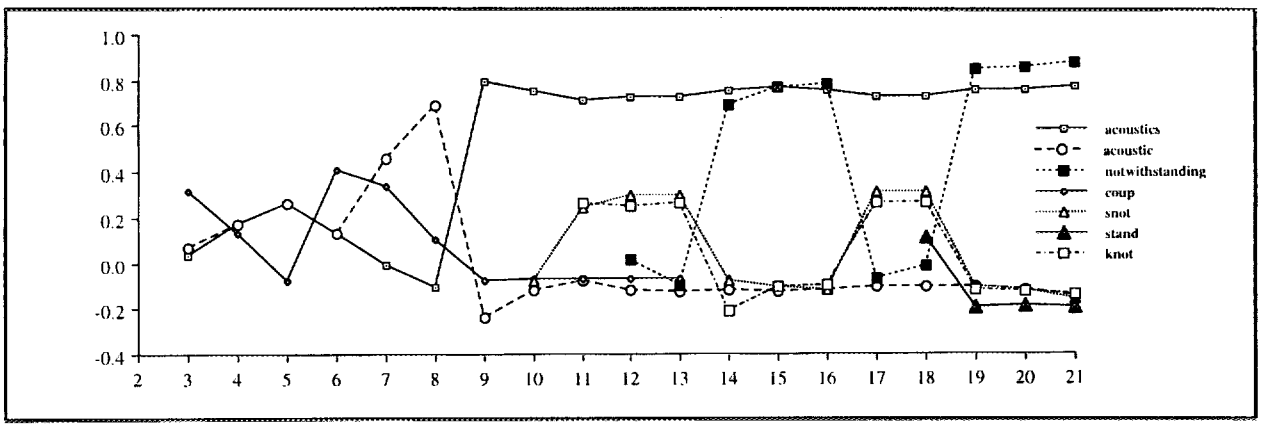

FIGURE 3. Shortlist simulation (using a 26000-word lexicon) of activation patterns given as input the 21-phoneme string *acoustics notwithstanding*. A subset of the most highly activated candidate words in the shortlist is shown.

The Shortlist model is able to operate with a realistically sized lexicon of tens of thousands of words, and the full phoneme set of a language, so that it can support sensitive simulations of the results of experiments on word recognition by human listeners, e.g. with the word spotting task. It is not specific for English, of course, but runs on any language for which a phonetic lexicon is available. Word-spotting experiments have provided clear evidence of active competition between simultaneously activated words: words are harder to spot if the remainder of the string partially activates a competing word. Thus listeners presented with the auditory input [nɒmɛs] will correctly and rapidly spot the presence of a real English word (*mess*) in that string; if, however, the string is [dɒmɛs], word-spotting will be slower and less accurate, because the latter string, although it also contains *mess*, is the beginning of another word - *domestic* - and presumably activates that word as well as *mess*, leading to competition [10]. In Shortlist simulations also, *mess* is more highly activated in the input [nɒmɛs] than in [dɒmɛs]. Shortlist also captures with detailed accuracy other experimental findings which show that the more competitor words are available in the vocabulary for activation by a given input, the greater the inhibitory effect on an embedded-word target [11, 12].

THE POSSIBLE-WORD CONSTRAINT

In a word-spotting study by Norris, McQueen, Cutler and Butterfield [13], English listeners were presented with words like *egg*, embedded in nonsense strings like *fegg* and *maffegg*. In *fegg*, the added context [f] is not a possible word of English - there are no English lexical items consisting of a single consonant. In contrast, the added context *maff* in *maffegg*, although it is actually not a word of English, might conceivably have been one - *mat*, *muff* and *gaff* are all English words. Listeners were faster and more accurate in detecting real words embedded in possible-word than in impossible-word contexts, whether the context preceded (*fegg*) or followed the target (*sugarth*); in other words, they found it hard to detect a word if the result of recognising it was to leave a residue of the input which was unparseable into words.

In Shortlist, these data were simulated by reducing the activation of any candidate word which leaves no vocalic segment between the edge of the word and the nearest known boundary in the input. (In this experiment, of course, the

nearest known boundary is the silence at each end of the stimulus string. But the constraint can be implemented in a more general way to capture other boundary effects known to be exploited by listeners, such as Cutler and Norris' [4] finding of segmentation at the onset of syllables with a full vowel, or McQueen's [14] finding that phonotactic sequencing constraints can trigger segmentation.) Again, Shortlist simulations accurately captured the pattern revealed in the listening data.

Further investigations of the possible-word constraint addressed its language-specificity versus universality. Languages differ in the precise constraints which apply to what may or may not be a word. Yet all human listeners, whatever their language, deal effectively with the phantom words and partially activated words which occur in any speech input as the inevitable result of a large vocabulary constructed from a small phonemic repertoire. They do so by drawing on highly effective mechanisms which allow words to compete with one another for the input, and further constraints which include an early filter to rule out any segmentation which would result in a residue of the input which could not be a possible word.

ACKNOWLEDGEMENTS

Thanks to Dennis Norris and James McQueen who collaborated in all the work described above, to Harald Baayen and Sally Butterfield who collaborated in further particular projects, and to Peter Roach for making available the MARSEC corpus of spoken English.

REFERENCES

1. Maddieson, I. *Patterns of Sounds*. Cambridge: Cambridge University Press, 1984.
2. McQueen, J.M. and Cutler, A. . "Words within words: Lexical statistics and lexical access," *Proceedings of the Second International Conference on Spoken Language Processing*, Banff, Canada,Vol. 1, 221-224, 1992.
3. Cutler, A., McQueen, J., Baayen, H. and Drexler, H. "Words within words in a real-speech corpus," *Proceedings of the 5th Australian International Conference on Speech Science and Technology*, Perth, Vol. 1, 362-367, 1994.
4. Cutler, A. and Norris, D.G. "The role of strong syllables in segmentation for lexical access," *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113-121 (1988).
5. Cutler, A. and Carter, D.M. "The predominance of strong initial syllables in the English vocabulary," *Computer Speech and Language*, **2**, 133-142, (1987).
6. Cutler, A. and Butterfield, S. "Rhythmic cues to speech segmentation: Evidence from juncture misperception," *Journal of Memory and Language*, **31**, 218-236, (1992).
7. McClelland, J.L. and Elman, J.L. "The TRACE model of speech perception," *Cognitive Psychology*, **18**,1-86, (1986).
8. Norris, D.G. "Shortlist: A connectionist model of continuous speech recognition," *Cognition*, **52**, 189-234, (1994).
9. Gaskell, M.G. and Marslen-Wilson, W.D. "Integrating form and meaning: A distributed model of speech perception," *Language and Cognitive Processes*, **12**, 613-656 (1997).
10. McQueen, J.M., Norris, D.G. and Cutler, A. "Competition in spoken word recognition: Spotting words in other words," *Journal of Experimental Psychology: Learning, Memory and Cognition*, **20**, 621-638, (1994).
11. Norris, D.G., McQueen, J.M. and Cutler, A. "Competition and segmentation in spoken word recognition," *Journal of Experimental Psychology: Learning, Memory and Cognition*, **21**, 1209-1228, (1995).
12. Vroomen, J. and de Gelder, B. "Metrical segmentation and lexical inhibition in spoken word recognition," *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 98-108, (1995).
13. Norris, D.G., McQueen, J.M., Cutler, A. and Butterfield, S. " The possible-word constraint in the segmentation of continuous speech," *Cognitive Psychology*, **34**, 191-243, (1997).
14. McQueen, J.M. "Segmentation of continuous speech using phonotactics," *Journal of Memory and Language*, (in press).