Juncture detection[1]

DENNIS NORRIS and ANNE CUTLER

Abstract

The 'units-of-perception' hypothesis has led psycholinguists to concern themselves with units of classification rather than with the more important process of speech segmentation. This paper argues that segmentation rather than classification of the speech signal is the primary prelexical process in speech perception. We claim further that universal juncture detection processes alone can account for processing differences between French and English which might be otherwise taken as evidence that different secondary-level perceptual units are involved in the perception of different languages.

Speech segmentation

The goal of any speech recognition system is to extract meaning from the continuous flow of the speech stream. To achieve this goal it must be able to locate and identify portions of the speech stream which correspond to individual words. However, the problem of segmenting continuous speech into words is far from trivial since word boundaries are seldom explicitly marked.

In principle, recognition could be performed simply by matching the speech signal (suitably transformed and normalized) against the template for each entry in the lexicon. Therefore, the only necessary unit of segmentation and classification is the unit of lexical representation (for convenience we will call this the word, although we recognize that it is a matter of debate whether lexical entries may be smaller or larger than the orthographic word). The obvious drawback of this simple system is that there is no way to be sure of finding out where any match in the signal occurs without carrying out the template matching process on all possible candidates. The signal can only be segmented into individual words by discovering which sections of the signal match lexical templates. In such a

Linguistics 23 (1985), 689–705

0024–3949/85/0023–0689 $2.00

system, segmentation (dividing) and classification (labeling) of the signal are one and the same process.

Of course, in a lexical template matching system a great deal of effort will be wasted on sections of the signal which do not correspond to any of the word templates. This wasted effort could be avoided if the signal could be segmented into lexical units before beginning the matching process. Thus any information in the signal that could assist in locating word boundaries would significantly speed the word recognition process. Even if word boundaries themselves could not be detected reliably, it might be possible to identify points in the signal where word boundaries are more likely to be found. Note that the value of locating these points is logically independent of whether the signal can be classified into units corresponding to intervals between the detected boundary points; all the system actually needs to know is where the boundaries are, not which particular (nonlexical) units are present.

Classification

By going one step further and performing a segmental classification of the signal, the template-matching process could be simplified even further. If the perceptual system could carry out even a partial phonological classification, for instance, then the information this would provide could be used to constrain the lexical search process to a phonologically specified subset of the lexicon. With a complete phonological classification of the input, template matching could be bypassed altogether and access could be based on the phonological specification itself. Similarly, a reliable syllabic classification would allow the processor to base access on a syllabic representation.

This is precisely the rationale of the units-of-perception hypothesis in psycholinguistics. One of the chief projects of psycholinguistics over the past two decades has been to discover whether units such as the phoneme or the syllable, i.e. possible levels of description which can be applied to speech, actually function as levels of representation in speech perception. If speech input can be efficiently segmented and classified into (presumably sublexical) units, it is argued, these units can be used as the basis for direct lexical access. This would then obviate the need for wasteful template matching processes.

However, for some psychologists this emphasis on classification has led to a quest for 'the unit of perception'. It sometimes appears to be assumed that there can only be a single unit of classification; the research aim then becomes to discover what that unit is. The main candidates for such a

perceptual unit have, in practice, been the phoneme and the syllable. Although in both linguistics and psycholinguistics the phoneme has generally received more attention than the syllable, some psychologists have preferred the syllable as a candidate perceptual unit, and of these some have gone so far as to deny the perceptual reality of the phoneme altogether (Savin and Bever 1970). Others have taken a more cautious view and have simply claimed that the syllable is the major unit of perception and that phoneme identification is highly dependent on syllable identification (Mehler et al. 1981).

In the present paper we begin by examining the units-of-perception hypothesis in some detail. In particular we review some of the experimental evidence advanced as support for the syllable as a unit of perception. We will argue that although there is very strong evidence that the syllable plays an important role in the perception of French, structural considerations suggest that the syllable plays a far less important role in a language such as English. Although one could claim that such an argument simply implies that the units of perception are different in different languages, we will suggest that the search for units of classification has drawn attention away from the real problem, i.e. speech segmentation. In the latter part of the paper we will argue that the important feature of units such as the syllable is not their identity, but the location of their boundaries; that is, the junctures between units are more relevant to the segmentation process than the units themselves. Furthermore, we will claim that the process of locating such junctures in the signal is common to all languages, and that the apparent cross-linguistic differences may simply arise because, in different languages, this process interacts with the speech signal in very different ways.

Units of perception

As we see it, there must be at least three prerequisite conditions for a speech segment to function as a 'unit of perception':

1. The segments themselves, at whatever level they are, must be reasonably distinguishable in the speech signal. (Note that it is not necessary that they be MORE distinguishable than words themselves, as long as the set of all possible segments is considerably smaller than the number of words in the lexicon; a slight reduction in distinguishability may trade off against a large reduction in the number of potential candidates.)

2. The whole utterance must be characterizable as a string of the segments in question, with no parts of the utterance unaccounted for.

(Thus although fricative noise satisfies the first requirement, it is not acceptable to propose the interval from one fricative to the next as a 'unit of perception', since utterances may contain no fricatives at all.)

3. The units must correspond in some reliable way to lexical units. The most likely assumption is that the perceptual units are sublexical; then each lexical unit must be made up of one or more whole units at this sublexical level, with the boundaries of the lexical unit being *ipso facto* also sublexical unit boundaries. However, if the perceptual unit in question is not NECESSARILY sublexical, then some simple and predictable translation from the perceptual unit to the lexical unit should be possible.

Primary and secondary units

We would further argue that the potential units of perception are of two kinds: primary and secondary. The only primary unit of representation is the phoneme, by virtue of the fact that it is the smallest linguistic unit into which speech can be sequentially decomposed. The syllable, for example, is a secondary unit, since syllables can be further decomposed into phonemes. Other examples of secondary-level units are those based on prosodic divisions: the stress group, the foot, the mora. Any secondary-level unit can be said to be a less natural candidate for a unit of perception than the phoneme because, as we shall argue below, secondary-level representations will always have to be supplemented by primary-level representations as well. Given the need for primary units, those who favor the syllable or any other secondary-level unit of perception must be able to justify the extra complexity of their theories. The next sections outline the arguments in favor of secondary-level units.

Secondary units of perception: (1) the syllable

Three main lines of argument have been advanced in favor of the syllable as a unit of perception. The first, and perhaps weakest, line has been to argue AGAINST the phoneme by drawing attention to the lack of invariance in the acoustic realization of a given phoneme (e.g. Mehler et al. 1981; Savin and Bever 1970; Wickelgren 1969). However, if the syllables are relatively invariant, then the allophones of the particular phonemes within those syllables must also be invariant. Therefore on this basis there seems little to choose between syllables and phonemes.

A stronger form of this argument against the phoneme is that it does not always satisfy requirement (1) above, in that perception of consonants

may be dependent upon perception of an adjacent vowel (e.g. Liberman, Delattre, et al. 1954), while perception of vowels may be facilitated by the availability of consonantal context (e.g. Strange et al. 1976). This has also been pointed out by proponents of the syllable as a unit of perception (e.g. Mehler et al. 1981).

The second argument has maintained that conscious awareness of the phonemic structure of words seems to be dependent on the acquisition of reading skills. Whereas preliterate children and adult illiterates can readily perform tasks which require explicit awareness of the syllabic structure of words, tasks which require awareness of the phonemic structure cannot be performed without training (e.g. Liberman, Shankweiler, et al. 1974). But, although such evidence does tell us that it is easier to manipulate syllables than phonemes consciously, we should be wary of drawing any inferences from this kind of task to the processes underlying the perception of fluent speech (see also the arguments advanced by Morais, this volume). The natural units of conscious manipulation need not be the natural units of perception. In any task requiring speech to be decomposed into constituent syllables or phonemes, syllables have a built-in advantage. While all syllables can be spoken in isolation, the same is not true of all phonemes: stop consonants, for instance, must have a vowel appended to them before they can be articulated. Therefore, if individual phonemes in a word are to be sounded out, the word must first be broken down into its constituent phonemes, and then some of those phonemes must be modified into a form that can be articulated. The consequence of this process is to obscure the relation between the acoustic realization of the phoneme spoken in isolation and in continuous speech. Perhaps we should not be surprised that this additional process presents a problem for those with no explicit training at phonemic decomposition.

The third and strongest line of evidence for the viability of the syllable as a perceptual unit, though, comes from reaction time studies. It has consistently been shown that syllables can be detected and responded to faster than phonemes (e.g. Savin and Bever 1970). Savin and Bever interpreted the reaction time advantage of syllables over phonemes as evidence that phonemes could only be identified after identification of the corresponding syllable, and that syllables were therefore the primary unit of perceptual analysis. However, Foss and Swinney (1973) demonstrated that words could be responded to faster than syllables. According to Savin and Bever's reasoning the word should be 'the' perceptual unit. Foss and Swinney also drew attention to an experiment by Bever, Savin, and Hurtig (reported in Bever 1970) in which monitoring for words in a list of one-clause sentences was found to be quicker if subjects were told

the entire sentence rather than just the initial word. By analogy with Savin and Bever's argument, the primary unit of perception should then be considered to be the clause.

There is a flaw in this conclusion, however, as Foss and Swinney pointed out. The conclusion concerns the order in which levels of linguistic representation are PERCEIVED; strictly speaking, though, the monitoring result concerns only the order in which the different levels are IDENTIFIED. It is not necessarily the case that the order of identification directly reflects the order of perception; in fact, it is quite plausible to hypothesize that the order in which levels of information become available as the basis of a conscious response may be precisely the reverse of the original order in which the levels are percieved. The ultimate aim of comprehension is awareness of meaning, and any subsidiary tasks such as detection of a target may be forced to wait until this aim is achieved. The more closely the target resembled the overall message, then — by implication, the larger the target — the easier would a matching response be.

Stronger reaction-time evidence for the importance of the syllable as a perceptual unit comes from a series of syllable monitoring studies by Mehler and his colleagues. In one such study (Mehler et al. 1981), subjects were presented with a visual specification of the target and were required to respond as soon as they heard a word beginning with that sound in a list of isolated words. The visual target specifications had either CV or CVC structure, and the target-bearing words also varied according to whether their initial syllable had CV or CVC structure. For example, given the target sequence 'PA' or 'PAL' subjects might hear the words *palace* (PA-LACE) or *palmier* (PAL-MIER). Mehler et al. found that responses were faster when the syllabification of the target-bearing word matched that of the target specification. That is, responses to *palace* were faster with 'PA', whereas responses to *palmier* were faster with 'PAL'. This pattern of results certainly suggests strongly that the syllable is playing an important role in perceptual analysis.

Problems with the syllable as a unit of perception

Suppose, then, that this finding be taken as evidence for reckoning the syllable to be a unit of perception. As pointed out in the preceding section, for a unit of perception to function effectively as a unit of lexical access, it is necessary that it exhibit a reliable correspondence with lexical units. Thus, in the case of the syllable we must assume that boundaries of lexical units and boundaries of syllables should coincide sufficiently often to

make syllabification a useful strategy in lexical matching. It is not a great problem if syllable boundaries occur within as well as between words, though it does increase the number of potential lexical segments which have to be checked. However, where a word boundary occurs within a syllable, a strongly syllable-based system will face severe problems, since it will not be able to achieve the correct lexical segmentation without first carrying out a further analysis at the next level down, i.e. the phonemic level.[2] If it were to be assumed that the syllable is the unique unit of access, a further problem would be created: even if the input has been analyzed into phonemes, it would still have to be recoded into a syllabic form for access. Thus it would appear that merely by virtue of being a secondary level of possible analysis, the syllable faces problems in its candidacy for unit-of-perception status.

However, a strong argument that may be advanced in favor of the syllable as a perceptual unit is that syllabic segmentation is substantially easier than either phonemic or lexical segmentation. The most convincing evidence for the syllable as a perceptual unit comes, as we have seen, from Mehler et al.'s studies conducted in French. This is not a coincidence. French is a language in which syllable boundaries are relatively clear; French speakers show a very high degree of agreement in identifying syllable boundary locations. But this is not true of all languages; in particular, it is not true of languages with stress rhythm. Whereas native French speakers are quite clear that the French word *'palace'* should be syllabified 'pa-lace', English syllable boundaries are far more ambiguous; English speakers are typically unsure where to place the boundary in the corresponding English word. In experimental studies of syllabification, English speakers treat intervocalic consonants, in particular intervocalic liquids and nasals in bisyllabic words like 'palace', as if they belonged to both the first and second syllable (Fallows 1981). Phonologists have resolved this uncertainty by describing intervocalic consonants such as the /l/ in 'palace' as *ambisyllabic*, i.e. belonging to both syllables at once (see for example Anderson and Jones 1974; Kahn 1976). In stress languages, such as English, ambisyllabicity is conditioned by stress — it occurs when the following vowel is unstressed and reduced. For this reason, and because in stress languages syllables are by no means all equal (weak syllables are much less easily perceptible than strong syllables, as will be elaborated below), syllabification tends to be a harder task in a stress language.

Therefore, because English is a stress language, and because consonantal ambisyllabicity is far more widespread in English than it is in French, syllabification is not as easy in English as it is in French. It might therefore be predicted that, at the very least, English listeners should not

show quite the same pattern of syllable monitoring results which Mehler et al. found with French listeners. Accordingly, Cutler et al. (1983; 1986) repeated Mehler et al.'s experiment with English listeners. No trace of the previous syllabification effect was found, either with English materials or with the original French materials. Interestingly, French subjects continued to demonstrate the syllabification effect even when listening to English materials. Cutler et al. (1986) argued that the pattern of results from the English listeners, however, showed them to be using a phonetic segmentation strategy.

These results appear to imply that French listeners are using a particular sentence-perception strategy, namely syllabic segmentation, which English listeners cannot use. This poses a very interesting problem for the psycholinguist. Do speakers of some languages have more segmentation strategies available to them than speakers of other languages? Are the English confined to phonetic analysis without alternative options, for instance, or is there some unit of segmentation above the phonetic level which English listeners can use in the way the French use the syllable?

It might be suggested that a possible answer should lie in precisely that difference between the phonological structure of English and French to which we pointed in predicting that syllabification would not be an efficient strategy in the perception of English. French syllable boundaries tend to be clearer than English syllable boundaries. English syllable boundaries are particularly unclear when the syllable following the boundary is weak; thus clear boundaries in English are to be found only at the onset of stressed (or rather, strong) syllables.

Secondary units of perception: (2) the foot

Suppose we assume that the use of supraphonetic segmentation units is determined solely by the availability of clear boundaries. That is, the syllable is a viable segmentation unit for French listeners not *qua* syllable, but simply because it usually constitutes the interval between one clear boundary and the next. This allows us to postulate a directly comparable segmentation unit in English, in which the interval between one clear boundary and the next will be the interval from the onset of one STRONG syllable to the onset of the next.

This unit is known in current prosodic phonology as the *foot* (see for example Liberman and Prince 1977; Selkirk 1980). It is possible to hypothesize, then, that the foot will act as a segmentation unit for English in precisely the way that the syllable is used in French. Thus we should be

able to construct a foot-monitoring experiment, analogous to the syllable-monitoring experiments described above, in which English listeners respond to a given target faster when the target corresponds to a foot than when it does not. Take, for example, the English words *turbine* and *turban*. The former has two strong syllables, i.e. contains two feet, so that there should be a clear word-medial boundary; the latter has a weak second syllable, so that the word as a whole is one foot rather than two. Are subjects faster to detect and respond to the target TUR- in *turbine* than in *turban*?

In a comprehensive series of experiments we have established that they are not. Neither in words nor in nonwords, in lists of isolated items or in connected speech, have we found any sign of a foot-segmentation strategy used by English listeners. On a strict units-of-perception hypothesis, our findings in these experiments and in the previous experiments on English listeners' identification of syllable targets appear to suggest that English listeners do not have any supraphonetic segmentation unit available to them, despite the indication that French listeners use the syllable in this way.

We will argue, however, that the postulation of such fundamental differences in the way English and French are understood is not the most attractive solution to the present problem. Instead, we would prefer to abandon the 'unit-of-perception' notion altogether. Recall that we have already pointed out that this notion raises considerable problems when the unit in question is not at the lowest sequentially analyzable level. For the syllable, it implies that the signal is FIRST analyzed into syllabic units which can THEN be subdivided into phonemes if necessary; but phonemic analysis will nevertheless be unavoidable whenever syllable boundaries fail to coincide with word boundaries, and if lexical access can ONLY be achieved via syllabic units then phonemes must be recoded into a syllabic representation prior to any attempt at access.

Of course, exactly the same arguments apply to the suggestion that the foot may be a unit of perception in English.[3] Phonetic analysis would be unavoidable whenever the processor encountered a word beginning with a weak syllable; and if access were crucially dependent upon a foot-level representation, recoding would again be necessary. In both cases, that is, the advantages of the perceptual unit appear to be outweighed by the extra processing required when it leads to inappropriate segmentation. In addition, we are left in the rather uncomfortable position of concluding that 'the' unit of perception is quite different in scope in English and French. While such a possibility is difficult to exclude, it would surely be more plausible, and more parsimonious, to suggest that the basic processing requirements of all languages are essentially identical.

In what follows we will suggest an alternative view of the recognition process which abandons the units-of-perception hypothesis but suggests that both the syllable and the foot can play a role in lexical segmentation and phonemic classification.

Segmentation without classification

The units-of-perception approach can be characterized by its emphasis on the process of classification. Speech perception is to be explained in terms of the units into which speech is classified. The problem of how speech is to be segmented into those units is one which, if it receives any consideration at all, is relegated to a secondary role.

An alternative approach is to place the major emphasis on the process of segmentation. We will suggest that segmentation and not classification is the primary function of the speech recognizer. We propose that the speech perception process is designed to utilize any information which can be extracted (either directly or indirectly) from the signal to determine where lexical boundaries might lie. To this end there may be no need to classify any nonlexical segments which are located. The main requirement is simply to detect possible word boundaries. Note that we are not simply suggesting that segmentation should be given equal emphasis with classification. Our claim is that, in the normal course of perception, speech is NEVER classified into secondary-level units[4].

We have argued that any advantage of treating syllables or feet as perceptual units is likely to be overshadowed by the disadvantage of using a unit of classification that will frequently straddle word boundaries. However, the fact remains that, at least in French, the syllable does seem to play a special role in speech recognition. Why should this be so if the syllable is not functioning as a perceptual unit?

The answer is that syllables (in French, at least) may provide valuable junctural information. Their importance in speech recognition arises entirely from the fact that their boundaries can be detected, or at least hypothesized, and can in turn be used to suggest suitable candidates for lexical segmentation. In stress languages, we suggest, feet play a similar role.[5] However, the recognition system need not actually concern itself with the identity of the syllables or feet. In fact there may be no need to know what kind of boundary has been detected. The important information to extract from the signal is simply that there is some kind of juncture, and that the juncture is a likely place for a word boundary to be located. According to this view it is not the identity of the syllables or the feet themselves which is important, merely their boundaries.

Recall the discussion in the introduction to this paper of the goals of speech recognition. Lexical representation of meaning is necessarily in the form of discrete units, whereas speech signals are continuous and do not necessarily contain markers locating the juncture of one lexical unit with another. Thus the location of word boundaries (more properly, lexical unit boundaries) is a task of the utmost priority in speech recognition. Our argument is that the patterning of speech into sequences of syllables or feet can be exploited to provide useful clues as to where such boundaries might be sought. Any information which indicates the presence of a transition from one syllable to another, or from one foot to another, can initiate a hypothesis that the transition in question is also a transition from one word to another.

It is important to emphasize that such information, according to our proposal, does not itself support a mechanism for chunking the signal into units which then form the basis of a lexical access code, but is to be used simply to help locate potential word boundaries. The information in question is, simply, any acoustic, phonetic, or prosodic transition cues that are encoded, either directly or indirectly, in the speech signal. Direct encoding of transitional information, as we have already pointed out, is in fact minimal (although it is possible to cite individual examples, such as insertion of a glottal stop before word-initial vowels in German). Indirectly available information must therefore be the primary basis for juncture detection. The distributional patterns of various sounds with respect to syllable boundaries provide one indirect source of transitional information (see Frauenfelder, this volume, for further discussion of this issue); but we concentrate here on prosodic information, which may be exploited at a very early level. This could possibly be achieved simply by temporal prediction from the preceding rhythm; computation of the average duration of the preceding few syllables or feet could allow the construction of a hypothesis as to the location of the next such unit's onset. Alternatively, gross acoustic features such as the detection of steady-state signals greater than some arbitrary duration could allow the postulation of an occurrence of the unit in question. Although the exact nature of the mechanism is as yet unclear and can only be established empirically, we suggest that in both stress languages (like English) and nonstress languages (like French) prosodic patterns can suggest where junctures should be sought.

Our hypothesis of segmentation without classification does not itself explain why French and English listeners behave differently in the experiments we described above. Nor does it explain why the foot, as the English analogue of the syllable, does not behave like the French syllable. Part of the answer may be that although from the point of view that we

have adopted (i.e. considered as just that which separates two clear junctures) the English foot is the nearest thing to the French syllable, it is nevertheless a very different kind of object. Both the foot and the syllable provide potentially valuable cues to lexical segmentation. However, by definition, the foot begins with a strong syllable; and strong syllables enjoy certain advantages over weak syllables, leading to a difference of phonetic usefulness between them which has no direct parallel in French.

Segmentation in a stress language

The processing of stressed versus unstressed syllables has been extensively investigated. Stressed syllables are typically far more distinct and therefore possess a processing advantage which is clearly exhibited in many speech perception tasks. For instance, monosyllabic words spliced out of context are more recognizable if they were stressed (Lieberman 1963); and clicks which occur on stressed syllables are more accurately located than clicks which occur on unstressed syllables (Bond 1971). In monitoring tasks, response time to phoneme targets is faster if the target is in a stressed syllable (Shields et al. 1974; Cutler and Foss 1977). Evidence from hearing errors (e.g. Garnes and Bond 1975) shows that such slips are LEAST likely to be made on stressed syllables — inaccurate perception occurs most often in unstressed syllables.

This perceptual advantage of the stressed syllable can be put to valuable use by the recognition system. First, the information in the stressed syllable can be assigned more weight than that in unstressed syllables in lexical access — that is, because the perceptual information is clearer and more reliable, a match with a stressed syllable in the access process should be more predictively valuable than a match with an unstressed syllable. Second, the extra clarity of the stressed syllable will increase the probability of performing an accurate phonemic analysis. If the stressed syllable can be readily classified phonemically then this should facilitate recognition by constraining the lexical search space.

Recently, Huttenlocher and Zue (1983) have demonstrated to what extent the segmental information in stressed syllables may be more useful than that in unstressed syllables. They classified words as sequences of broad phonetic categories (vowel, nasal, glide, stop, weak/strong fricative). The average size of the set of potential candidates (in their 20,000-word lexicon) satisfying any sequential description was 2.3; the largest single set contained 210 members. Set size was virtually unchanged (mean 2.6; largest single set 215) when phonetically variable segments (vowel, stop, weak fricative) in unstressed syllables were omitted; and it was not

substantially increased (mean 3.8; largest set 291) when all segmental information from unstressed syllables was discarded. Omitting segmental information from stressed syllables, however, leaving only the unstressed syllable information, increased mean set size more than threefold to 7.7, and allowed single sets of up to 3717 members.

Stressed syllables will therefore have three advantages over unstressed syllables.

1. They will be perceptually clearer.
2. They are phonetically more informative.
3. Because of their clarity and informativeness the information in stressed syllables can be weighted more heavily in any procedure for choosing between alternative phonemic or lexical candidates. Thus stressed syllables may indirectly help to identify the phonemes in unstressed syllables.

Therefore, the stressed syllable in English will provide an island of perceptual clarity which will help both phonemic analysis and lexical access. However, it is still the case that foot boundaries are good candidates for the location of lexical unit boundaries. When the processor postulates a foot juncture in English, therefore, it will not simply be provided with information about a likely lexical boundary; it will also have located a part of the signal where phonemic analysis will reap the greatest benefits.

The benefits of juncture detection

Our proposal is that, instead of the perception of different languages involving different secondary-level units of perception, the perception of any language involves a process of juncture detection. We also propose an equally universal phonetic analysis process. Although the primary aim of the juncture detection process is detection of lexical boundaries, it can also, we argue, exercise the beneficial side-effect of assisting the phonetic analysis process. As those who have argued against the importance of the phoneme in perception have pointed out, there is a great deal of variation in the realization of a given phoneme. The exact form of a phoneme will be strongly influenced by its context. However, if a foot or syllable juncture is postulated or detected at a certain point, it must by definition (because the phoneme is the smallest sequential unit) separate two phonemes. This information will at the very least enable a phonetic analysis process to decide that this particular portion of the signal corresponds to two phonemes rather than one. (Note that this does not contradict the complementary fact that the contextual dependence of

phonemes can itself be perceptually useful. For example, Meltzer et al. [1976] have shown that phoneme monitoring RT can be speeded by moving anticipatory coarticulation forward in time.)

Because of language-specific differences in phonological structure the output of the juncture detector will provide rather different information for the phonetic analyzer in different languages. Thus in French, a juncture detector could provide an output at all syllable boundaries, since locating syllable boundaries is relatively easy. In English, output from a juncture detector will be much less frequent since there are fewer easily locatable boundaries. However, the output should have additional value, in that it will point the phonetic annalyzer in the direction of particularly reliable sections of the signal.

Therefore, although the juncture detection process will be the same for both the French and the English listener, the fact that the juncture detector signals rather different information in the two languages will lead French and English listeners to interpret the output in different ways. That is, although the basic process will be identical in all languages, linguistic differences will cause listeners of different languages to develop different perceptual strategies. Whereas French listeners will come to use the output of the juncture detector primarily to help segmentation, English listeners will also learn to make use of junctural hypotheses to identify informative points in the signal.

On this view the question of secondary-level 'units of perception' is no longer an issue. Where a language has locatable boundaries of any kind, and where those boundaries do not signal confounding information such as differences in intrinsic perceptibility, then the segments thus bounded will appear to function as units of segmentation in speech perception. Nothing in this account, however, requires any process of classification. Segmentation is its own reward.

Conclusion

In this paper we have addressed the question of how speech recognition can be assisted by segmenting and classifying the input at levels other than the word. In the quest to identify 'units of perception' most psycholinguists have concentrated their efforts on locating the units into which speech is classified. In contrast, we have argued that the speech segmentation process may be much more important than the classification process. We have postulated a process of juncture detection which is universal to all languages. The capacity to detect putative junctures will have two important roles in the speech-recognition process. Juncture detection will

help to identify the location of potential word boundaries, and it may also be able to assist the process of phonetic analysis.

By emphasizing the importance of juncture detection, we have relegated the classification process to a secondary role relative to segmentation. In fact, classification of speech into units such as the syllable may have no role outside the context of psychological experiments.

Received 5 October 1984
Revised version received
20 September 1985

MRC Applied Psychology Unit

Notes

1. This paper is based on presentations by the two authors to a European Psycholinguistics Association workshop on 'Cross-linguistic studies of morphophonological processing' held in Paris in June, 1984. We would like to thank the members of the workshop for discussion of the issues involved in this paper. The syllable-monitoring experiments comparing the processing of English with the processing of French, mentioned in the present text, were carried out in collaboration with Jacques Mehler and Juan Segui, and we are particularly grateful to them for useful criticism and stimulating discussions. Financial support for the Cambridge–Paris collaboration was provided by a twinning grant from the European Training Programme in Brain and Behavioural Sciences of the European Science Foundation. Correspondence address: MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, England.
2. Note that this is a problem only for candidate 'units' which are, like the syllable, above the lowest level of sequential representation. At the phonemic level, when a word boundary occurs within a phoneme, no lower level of analysis can be simply consulted to resolve the unclarity because no lower level of analysis is available. Thus it can only be resolved by reference to a rule within the rule set of the current level of analysis. For example, in the English utterance 'Had your dinner yet?', palatalization can apply across the first word boundary, turning the final /d/ of 'had' and the initial /j/ of 'your' into a single affricated consonant. This consonant cannot be split into lower-level units to divide the words from each other, so the phonetic analysis process will have to refer to its knowledge of the palatalization rule, including the fact that it can apply across a word boundary.
3. The present argument, it should be pointed out, refers only to English and to other languages which, like English, have freely varying stress. Many stress languages have fixed stress, i.e. stress which occurs always in the same syllable of a polysyllabic word. In these languages the unit-of-perception requirement (3) is perfectly fulfilled, i.e. the relation between a stress unit boundary and a lexical unit boundary will be entirely predictable. It is possible, therefore, that the processes referred to in our final section produce a different output again in fixed stress languages.
4. It can be argued, of course, that in the limiting case segmentation is always dependent on classification at some level. For example, in order to segment words bounded by silence the input must be classified into silence versus noise. The classification processes we are concerned with, however, concern only classification into LINGUISTIC units.

5. One source of evidence for the special nature of foot boundaries comes from Fowler's (1981) finding that the main coarticulatory effects of stressed vowels are carryover effects rather than anticipatory effects.

References

Anderson, J., and Jones, C. (1974). Three theses concerning phonological representations. *Journal of Linguistics* 10, 1–26.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (ed.), *Cognition and the Development of Language*. New York: Wiley.

Bond, Z. S. (1971). Units of speech perception. *Ohio State University Working Papers in Linguistics* 9, 1–112.

Cutler, A., and Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech* 20, 1–10.

—, Mehler, J., Norris, D., and Segui, J. (1985). A language specific comprehension strategy. *Nature* 304, 159–160.

—, Mehler, J., Norris, D., and Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language* 25.

Fallows, D. (1981). Experimental evidence for English syllabification and syllable structure. *Journal of Linguistics* 17, 309–317.

Foss, D. J., and Swinney, D. A. (1973). On the psychological reality of the phoneme: perception, identification and consciousness. *Journal of Verbal Learning and Verbal Behavior* 12, 246–257.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research* 24, 127–139.

Garnes, S., and Bond, Z. S. (1975). Slips of the ear: errors in perception of casual speech. *Papers from the Eleventh Regional Meeting, Chicago Linguistic Society*, 214–255.

Huttenlocher, D. P., and Zue, V. W. (1983). Phonotactic and lexical constraints in speech recognition. *Working Papers, Speech Communication Group, Research Laboratory of Electronics* (MIT) 3, 157–167.

Kahn, D. (1976). Syllable-based generalizations in English phonology. Unpublished Ph.D. thesis, MIT.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs* 68, 1–13.

Liberman, I. Y., Shankweiler, D. P., Fisher, F. W., and Carter, B. (1974). Reading and the awareness of linguistic segments. *Journal of Experimental Child Psychology* 18, 201–212.

Liberman, M. Y., and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry* 8, 249–336.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech* 6, 172–187.

Mehler, J., Dommergues, J., Frauenfelder, U., and Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior* 20, 298–305.

Meltzer, R., Martin, J., Mills, C., Imhoff, D., and Zohar, D. (1976). Reaction time to temporally displaced phoneme targets in continuous speech. *Journal of Experimental Psychology: Human Perception and Performance* 2, 277–290.

Savin, H. B., and Bever, T. G. (1970). The non-perceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior* 9, 295–302.

Selkirk, E. O. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry* 11, 563–605.

Shields, J. L., McHugh, A., and Martin, J. G. (1974). Reaction times to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology* 102, 250–255.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (1976). Consonantal environment specifies vowel identity. *Journal of the Acoustical Society of America* 60, 213–224.

Wickelgren, W. A. (1969). Context sensitive coding, associative memory, and serial order in (speech) behaviour. *Psychological Review* 76, 1–15.