14
Linguistic rhythm and speech segmentation

A. Cutler

Rhythm

Speech is rhythmic. But this simple statement has many interesting corollaries, among them two which are central to this paper: firstly, rhythmic structures differ across languages, and secondly, rhythm in language is more than just timing.

The first of these statements has long been a linguistic truism, but it took a surprising time for the second to be widely accepted. Consider, for instance, the well-known proposal that some languages exhibit *stress-timing* and others *syllable-timing*. This suggestion was first put forward by Pike (1946 - although he acknowledges similar proposals for English by Classe, 1939); the claim is that in some languages, stresses occur at roughly equal time intervals, while in others, syllables occur at roughly equal time intervals. Pike contrasted English as an example of stress-timing with Spanish as an example of syllable-timing. Abercrombie (1967) added Arabic and Russian to the stress-timed list, and French, Yoruba and Tclugu to the syllable-timed group. A minor literature grew up as other languages were categorised in this way, and *mora-timing* was proposed (for Japanese) as a third category. In particular, the proposal stimulated a very large number of phonetic studies which examined the factual basis of the distinction. Most of these tested the proposal's apparent prediction that very little variation should be found in the duration of the appropriate units - stress intervals in stress-timed languages, syllables in syllable-timed languages, morae in mora-timed languages. Of course such research is extremely difficult, since determining the boundaries of the relevant units is extremely complicated (see Delattre [1966] for an excellent exposition of the problems). Nonetheless, many measurement studies were undertaken, and it is fair to say that the durational hypothesis proved a dismal failure. The absence of perfect or even

approximate isochrony in English stress intervals has been demonstrated over and over again (Bolinger, 1965; O'Connor, 1965; Uldall, 1971; Lehiste, 1973; Faure, Hirst & Chafeouleff, 1980; Nakatani, O'Connor & Aston, 1981). Wenk and Wioland (1982) likewise failed to find isochrony in French syllables. Roach (1982) measured syllable durations in English, French, Russian, Arabic, Telugu and Yoruba, predicting (after Abercrombie, 1967) that (a) syllable length should be more variable in stress-timed than in syllable-timed languages, and (b) intervals between stresses should be more variable in syllable-timed than in stress-timed languages. Neither hypothesis found support, Dauer (1983) measured interstress intervals in English, Spanish, Greek, Italian and Thai; variation patterns were similar in all languages.

The hypothesis of strict isochrony in spoken utterances is, therefore, clearly false. But the notion that languages differ in basic rhythmic pattern is widespread in phonetics and phonology, and supported by listeners' subjective impressions. There is also experimental evidence in its favour. Firstly, there is evidence that rhythmic parameters can be affected by different variables across languages. Delattre (1966) measured syllable duration and amplitude in five minutes of spontaneous speech in each of English, Spanish, French and German, separating closed (CVC) from open (CV) syllables, phrase-final syllables from non-phrase-final, and stressed syllables from unstressed. He found significant inter-language differences: stress, for instance, had a great effect on both syllable duration and amplitude in English, but very little effect on either variable in Spanish; final position in the phrase had a greater effect in French than in the other three languages, with Spanish showing the least effect. Secondly, the same variable (e.g. stress) can exercise different influences on a parameter of rhythm. Thus Hoequist (1983) compared timing in English and Spanish, using reiterant speech to control phonemic content; in English utterances the strongest effect was a shortening of unstressed syllables adjacent to stressed syllables, while in Spanish there was stress-conditioned lengthening, but no such compensatory shortening. Italian differs from English in the same way, as a perceptual study by Bertinetto and Fowler (1989) showed, When vowels were shortened or lengthened in English and Italian words, and the acceptability of the resulting forms tested, both groups disliked lengthening of unstressed syllables; however, English listeners proved to be very tolerant of shortening of unstressed syllables, while Italian listeners were not.

The problem with the isochrony hypothesis was that it focussed too narrowly on durational variation. What we perceive as the characteristic rhythm of a particular language is a complex of features. For instance, much of the strong impressionistic

difference between English and French can be ascribed to a phenomenon very characteristic of stress languages like English, and almost absent in languages like French, namely weak syllable reduction. Part of what the notion of stress-timing attempts to capture is that some languages have two very different types of syllables - strong and weak, or stressed and unstressed - and what happens to one type of syllable in spoken utterances is very different from what happens to the other type. Other languages do not dichotomise syllables in this way.

Dauer (1987) has produced the most comprehensive inventory of variables affecting linguistic rhythm. A wide variety of permissible syllable structures is" characteristic of stress-based languages; many other languages allow only a restricted set of syllable structures (in the extreme case - Polynesian languages such as Hawaiian, for instance - only V or CV are allowed). If a language has phonemic vowel length, it may permit this variation only in stressed syllables. If a language has tonal contrasts, they may be expressed only on stressed syllables. Vowels in unstressed syllables may be centralised; consonants in unstressed syllables may be neutralised. Dauer proposes a check-list for rhythmic categorisation of languages, on eight dimensions. The more positive points a language scores, the more likely it is to have been typically regarded as "stress-timed"; the more negative points it scores, the more likely it is to have been termed "syllable-timed". Endpoint scores are rare since not all dimensions apply to a particular language; neither French nor English, for instance, has phonemic vowel length or tonal contrasts. Nevertheless, English falls towards one end of Dauer's scale, French towards the other.

Thus rhythmic differences between languages form a continuum, whereby some languages make stronger distinctions between syllable types than other languages do.

Let us now turn to the role of rhythm in speech perception. One fact is clear: listeners "lock on" to rhythm. Thus prosodic breaks over-ride syntactic breaks in click location tasks - that is, when prosodic and syntactic boundary location conflict, more clicks are falsely reported to have been heard at the prosodic boundary, indicating that the prosodic structure is more salient at the relevant level of processing (Wingfield and Klein, 1971). If prosodic continuity and semantic continuity conflict, listeners attend to the former (Darwin, 1975). Unsurprisingly, then, the disruption of rhythm impairs performance on many perceptual tasks. Martin (1979), for example, found that either lengthening or shortening a single vowel in a recorded utterance could cause a perceptible momentary alteration in tempo, and increase listeners' phoneme-monitoring

response times. Meltzer, Martin, Mills, Imhoff and Zohar (1976) similarly found that phoneme targets which were slightly displaced from their position in normal speech were detected more slowly. Buxton (1983) found that adding or removing a syllable on a word preceding a phoneme target also increased detection time (although Mens and Povel [1986] have failed to replicate this finding in Dutch). These results seem to suggest that listeners process linguistic rhythm in a rather active way, using it to make predictions about later parts of the speech signal; when manipulations of the signal cause these predictions to be proven wrong, recognition is momentarily disrupted.

It would seem then that listeners And the continuity of speech signals useful in that the rhythm allows them to make useful predictions which presumably lead to an increase in processing efficiency. But there is a severe penalty for continuity in speech, namely the absence of explicit segmentation, i.e. cues which inform the listener how an incoming speech signal may be divided into appropriately recognisable units.

It is a reasonable assumption that whole utterances are only rarely recognisable as single units; most speech recognition must involve separate lexical retrieval of an utterance's component parts. But only rarely do spoken utterances contain reliable cues to the presence of a word (lexical unit) boundary. It is probable, therefore, that human listeners rely on explicit segmentation procedures (or a range of such procedures) which are designed to cope with the necessity of identifying lexical units in the absence of signals which demarcate these units. The next section describes some studies of segmentation procedures across languages,

Segmentation

Segmentation seems, as we listen to continuous speech, to pose no obvious problem. In other words, the segmentation procedures which listeners use are extremely efficient. But not all listeners use the same procedures. In studies of segmentation in English and French - two quite closely related languages within the context of the world's population of languages - my colleagues and I have produced evidence that segmentation procedures for these two languages are very different.

1. Segmentation of French

Mehler (e.g. 1981) and his colleagues (e.g. Segui, 1984) have used a variety of psycholinguistic tasks to demonstrate processing advantages for syllables in speech comprehension. For example, Mehler, Dommergues, Frauenfelder & Segui (1981) had French subjects listen to lists of unrelated words and press a response key as fast as

possible when they heard a specified word-initial sequence of sounds. This target was either a consonant-vowel (CV) sequence such as *ba-* or a consonant-vowel-consonant (CVC) sequence such as *bal-*. The words which began with the specified sound sequence had one of two syllabic structures: the initial syllable was either open (CV), as in *balance,* or closed (CVC), as in *balcon.* Mehler et al. found that response time was significantly faster when the target sequence corresponded exactly to the initial syllable of the target-bearing word than when the target sequence constituted more or less than the initial syllable. Thus responses to *ba-* were faster in *balance* than in *balcon,* whereas responses to *bal-* were faster in *balcon* than in *balance.* Mehler et al. interpreted this result as supporting a syllabically based segmentation strategy.

Other experiments, also conducted in French, further supported this claim. Segui, Frauenfelder & Mehler (1981) found that listeners are faster to detect syllable targets than to detect targets corresponding to the individual phonemes which make up those same syllables. Segui (1984) summarised a number of studies indicating that polysyllabic words, whether they are heard in isolation or in connected speech, are analysed syllable by syllable. Cutler, Mehler, Norris and Segui (1986) found that French listeners even show evidence of syllabic segmentation when listening to a foreign language (English). Thus the evidence from many studies of speech processing by French listeners suggests that their speech segmentation proceeds syllable by syllable.

2, Segmentation of English

A syllabically based segmentation procedure would not seem ideal for English, however. As in all stress languages, syllable boundaries in English are frequently unclear (to native speakers!), and in some words, such as *balance,* a consonant between two vowels seems to be *ambisyllablc,* i.e. to belong to two syllables at once. Of course, where syllable boundaries are hard to detect, division of speech input into syllables would not be a very efficient perceptual strategy; and indeed, Cutler, Mehler, Norris and Segui (1986) found that English listeners do not employ it. Using exactly the same experimental design as Mehler et al. (1981), but English materials (e.g. *balance, balcony)* and English-speaking subjects, they found that response time to CV *(ba-)* and CVC *(bal-)* targets was not significantly different either in *balance-* or balcony-type words. Nor did English listeners show evidence of syllabic segmentation when they listened to French materials (which lend themselves well to such a procedure).

The appropriate segmentation procedure for English appears to be quite different. In a stress language, such as English is, syllables can be either strong or weak; strong

syllables contain full vowels, while weak syllables contain reduced vowels (usually schwa). Cutler and Norris (1988) suggested that this difference could assist segmentation. Their proposal was based on an experimental finding that listeners were slower *to* detect the embedded real word in, say, *mintayf* (in which the second vowel is strong) than in *mintef* (in which the second vowel is schwa). They suggested that listeners were segmenting *mintayf* prior to the second syllable, so that detection of *mint* therefore required combining speech material from parts of the signal which had been segmented from one another. No such difficulty would arise for the detection of *mint* in *mintef,* since the weak second syllable would not be divided from the preceding material.

Cutler and Norris suggested that English listeners take strong syllables to be likely lexical (or content) word onsets, and divide the continuous speech stream at strong syllables so that lexical access attempts can be initiated with the maximum likelihood of immediate success. This procedure appears to be well matched to the structure of the English vocabulary. Cutler and Carter (1987) showed that 73% of all entries in a 33000-word phonetically transcribed dictionary of English had strong initial syllables. But the frequency of occurrence of individual words differs widely; lexical, or content words, are sometimes very common but more often very rare, while some words which in running speech are usually realised as weak syllables - grammatical, or function words, such as *of* or *the* - occur very frequently. Curler and Carter examined a 190,000-word natural speech sample, the *Corpus of English Conversation* (Svartvik & Quirk, 1980), using the frequency *count* of this corpus prepared by Brown (1984); they found that in this corpus 90% of the lexical words have strong initial syllables. However, the grammatical words in the coipus were actually in the majority, and they were virtually all weak monosyllables. Cutler and Carter computed that about three-quarters of all strong syllables in the sample were the sole or initial syllables of lexical words; while more than two-thirds of all weak syllables were the sole or initial syllables of grammatical words. Thus a listener encountering a strong syllable in spontaneous English conversation would seem to have about a three to one chance of finding that strong syllable to be the onset of a new lexical word. A weak syllable, on the other hand, would be most likely to be a grammatical word. English speech therefore provides a good basis for the implementation of a segmentation procedure in which strong syllables are assumed to be the onsets of *lexical* words.

Evidence that listeners may indeed use such a procedure in the segmentation of continuous English speech is found in segmentation *errors,* i.e. the way in which word boundaries tend to be misperceived. Butterfield and Cutler (1988) examined both

spontaneous and experimentally elicited misperceptions, and found that erroneous insertions of a word boundary before a strong syllable (e.g. "disguise" ':>eing heard as "the skies") and deletions of a word boundary before a weak syllable (e.g. "ten to two" being heard as "twenty to") were far more common than erroneous insertions of a boundary before a weak syllable (e.g. "variability" being heard as "very ability") or deletions of a boundary before a strong syllable (e.g. "in closing" being heard as "enclosing"). This is exactly what would be expected if listeners are dealing with the segmentation problem by applying a strategy of assuming that strong syllables are likely to be word-initial, but weak syllables are not. Segmentation in English, therefore, appears to be based on the opposition of strong and weak syllables.

Rhythm **and** Segmentation

The experimentally demonstrated segmentation procedures for French and English mirror each language's characteristic rhythmic structure, and hence the classic rhythmic contrast between these two languages. The use of the opposition between strong and weak syllables in segmenting English reflects the English language's characteristic stress-based rhythmic pattern, and the use of the syllable in segmenting French reflects the characteristic syllable-based rhythm of French.

My colleagues and I certainly believe that linguistic rhythm may be the key to speech segmentation (see Cutler, Mehler, Norris & Segui, 1986). One aspect of acquiring a native language would then be learning how the language's characteristic rhythmic pattern interacts with the structure of the vocabulary, and developing segmentation heuristics based on that knowledge. Thus in English, for example, one would learn that there are strong and weak syllables, that these tend roughly to alternate in continuous speech, and that the initial syllables of lexical words are much more likely to be strong than weak. Out of this would grow relatively efficient procedures for dealing with the continuity of spoken utterances.

Just as linguistic rhythm is not a simple matter, however, neither is its exploitation via processing procedures. The listener cannot simply expect rhythmic units to occur with temporal regularity, since, as we saw above, rhythm is not just regular timing. Thus in English the principal component of rhythm seems to be the distinction between strong and weak syllables, and we have seen that listeners exploit this distinction in segmenting speech. The importance of this distinction in segmentation seems to imply that listeners will treat it as a categorical decision: a given syllable is eitlier strong or

weak. But it is not easy to express the strong-weak distinction in terms of acoustic-phonetic parameters; any attempt seems to give a continuous rather than a categorical distribution. However, a recent study by Fear (1990) suggests that quasi-continuity in the acoustic-phonetic distribution is no bar to categoricality in perception. Fear examined the production of the initial vowels in sets of words such as *audience, auditorium, audition* and *addition* - that is, vowels bearing primary stress, vowels bearing secondary stress, unstressed non-schwa vowels and reduced (schwa) vowels. Measurements of vowel duration, pitch and intensity showed that the four vowels were not distributed continuously along any of these dimensions - for instance, the durations of vowels vvith primary and secondary stress differed by much less than any other pair. However, the distributions of the four vowel types differed on each prosodic dimension - thus with standard deviation of pitch across the vowel (a measure of pitch movement), the two most similar vowel types were unstressed and schwa,

The listener, therefore, is faced with a distribution of English vowels which differs according to the dimension under consideration. On what basis under such circumstances can the strong-weak distinction be drawn? Fear tested this by cross-splicing all the vowels in each set of four, and assessing the perceptual acceptability of the result. The listener judgements were clear - schwa belonged to a different category from any of the full vowels, even the unstressed one. All the cross-spliced words involving either substitution of schwa for another vowel or substitution of another vowel for schwa were rated as less acceptable than all others; moreover, the acceptability ratings for the cross-spliced words *not* involving schwa did not differ significantly either from each other or from the ratings for unaltered words.

Cutler and Norris (1988) suggested that detection of strong vowels could be implemented in a model of speech processing in several different ways: for instance, detection could occur upon the occurrence in the input of one of the set of full vowels (if the model involved a phonemic level of representation), or of a high-energy steady-state portion of a specified minimum relative duration (if the model involved no phonemic representation). Other implementations are also conceivable, such as one in vhich the English system more closely approximates to the French system via a syllabic level of representation (which for English only would be categorised by the perceiver into strong versus weak syllables). Thus although the relationship between rhythm and segmentation seems to be quite complex, and to involve concepts which belong more to the realm of phonology than acoustic phonetics, these factors present no obstacle to the conclusion that rhythm plays a clear functional role in human speech processing.

References

Abercrombie, D. (1967). *Elements of General Phonetics,* Edinburgh University Press, Edinburgh.

Bertinetto, P.M. & Fowler, C.A. (1989). On sensitivity to durational modifications in Italian and English. *Rivista di Linguistica,* 1, 69-94.

Bolinger, D.L. (1965). Pitch accent and sentence rhythm. In *Forms of English: Accent, Morpheme, Order,* Harvard University Press, Cambridge, MA.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the *London-Lund Corpus of English Conversation, Beh. Res. Meth., Instr. & Comp.,* 16, 502-532.

Butterfield, S. and Cutler, A. (1988). Segmentation errors by human listeners: Evidence for a prosodic segmentation strategy. *Proc. SPEECH 88* (Seventh symposium of the Federation of Acoustic Societies of Europe), Edinburgh; 827-833.

Buxton, H. (1983). Temporal predictability in the perception of English speech. In Cutler, A. and Ladd, D.R. (eds.), *Prosody: Models and Measurements,* Springer, Heidelberg.

Classe, A. (1939), *The Rhythm of English Prose,* Blackwell, Oxford.

Cutler, A. & Carter, D.M, (1987). The predominance of strong initial syllables in the English vocabulary. *Comp, Sp. Lang.,* 2, 133-142.

Cuder, A., Mehler, J., Norris, D. & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. /. *Mem. Lang.,* 25, 385-400.

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

Darwin, C.J. (1975). On the dynamic use of prosody in speech perception. In Cohen, A. and Nootcboom, *S.G,* (eds.), *Structure and Process in Speech Perception,* Springer, Berlin.

Dauer, R.M. (1983). Stress-timing and syllable-timing reanalyzed. *J. Phon.,* 11, 51-62.

Daucr, R.M. (1987). Phonetic and phonological components of language rhythm. *Proc, llth Int. Cong. Phon. Sci.,* Tallinn; Vol. 5, 447-450.

Delattre, P. (1966). A comparison of syllable length conditioning among languages. *Int. Rev. Appl. Ling.,* 4, 183-198.

Faure, G., Hirst, D.J. & Chafcouloff, M. (1980). Rhythm in English: Isochronism, pitch and perceived stress. In Waugh, L.R. and van Schooneveld, C.H. (eds.), *The Melody of Language,* University Park Press, Baltimore.

Fear, B. (1990). *Perceptual and Phonetic Distinctions of Syllabic Categories.* MPhil Dissertation, Cambridge University.

Hoequist, C.E. (1983). Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica,* 40, 203-237.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *J. Acoust. Soc. Amer.,* 54, 1228-1234.

Martin, J.G. (1979). Rhythmic and segmental perception are not independent. *J. Acoust, Soc. Amer.,* 65, 1286-1297.

Mehler, J. (1981). The role of syllables in speech processing. *Phil Trans. Roy. Soc,* B295, 333-352.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U. & Segui, J. (1981). The syllable's role in speech segmentation. *J. Verb. Learn. Verb. Beh.,* 20, 298-305.

Meltzer, R.H., Martin, J.G., Mills, C.B., Imhoff, D.L. and Zohar, D. (1976). Reaction time to temporally displaced phoneme targets in continuous speech. *J. Exp. Psy,: Hum. Perc. Perf.,* 2, 277-290.

Mens, L. and Povel, D. (1986). Evidence against a predictive role for rhythm in speech perception. *Quart J. Exp. Psy.,* 38A, 177-192.

Nakatani, L.H., O'Connor, K.D. & Aston, C.H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica,* 38, 84-106.

O'Connor, J.D. (1965). The perception of time intervals. *University College Phonetics Laboratory Progress Report,* 2, 11-15.

Pike, K.L. (1945). *The Intonation of American English,* University of Michigan Press, Ann Arbor.

Roach, P. (1983). On the distinction between "stress-timed" and "syllable-timed" languages. In Crystal, D. (ed.), *Linguistic Controversies,* Arnold, London.

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing. In Bouma, H. and Bouwhuis, D.G. (eds.), *Attention and Performance X,* Erlbaum, Hillsdale, N.J.

Scgui, J., Frauenfelder, U. & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *Brit. J. Psy.,* 72, *411-411.*

Svartvik, J. & Quirk, R. (1980). *A Corpus of English Conversation,* Gleerup, Lund.

Uldall, E.T. (1971). Isochronous stresses in RP. In Hammerich, L.L., Jakobson, R. and Zwirner, E. (eds.), *Form and Substance:* Akademisk Forlag, Copenhagen.

Wenk, B.J. & Wioland, F. (1982). Is French really syllable-timed? *J. Phon.,* 10, 193-216.

Wingfield, A. and Klein, J.F, (1971). Syntactic structure and acoustic pattern in speech perception. *Perc. Psychophys.,* 9, 23-25.