# SYNTHESIZING BRITISH ENGLISH RHYTHM – A STRUCTURED APPROACH

**Ian H. Witten*, and Alexandra Smith†**
*Man-Machine Systems Laboratory, Department of Electrical Engineering Science, University of Essex
†Department of English Language, University of Edinburgh

**Abstract**

This paper describes a method of assigning rhythm to synthetic speech, which takes the form of an hierarchical structure of levels. The definition and treatment of the upper levels are based upon Abercrombie's analysis of the rhythmic structure of English. We proceed from an utterance via its feet—the stress-determined, nominally isochronous units of rhythm—to the individual syllables, the initial and final consonant clusters and vocalic nuclei which form these syllables, and the phonetic segments which comprise the clusters.

The many levels in the hierarchy mean that only a small amount of information about the rhythm is needed at each one, and it becomes feasible to store the information in a table, rather than relying on algorithms to approximate it. This appears to provide a mechanism which is flexible enough to incorporate new data supplied from measurements of natural utterances, without affecting the underlying structure of the computation.

## MÉTHODE STRUCTURÉE POUR LA SYNTHÈSE DU RYTHME DE L'ANGLAIS BRITANNIQUE

**Résumé**

La présente communication décrit une méthode permettant de donner un rythme à la parole synthétique, en utilisant une structure hiérarchique de niveaux. La définition et le traitement des niveaux supérieurs sont basés sur l'analyse d'Abercrombie de la structure rythmique de l'anglais. A partir d'une parole dont on détermine les pieds, c.-à.-d. les unités rythmiques nominalement isochrones marquées par l'accent d'intensité, on passe aux syllabes individuelles, aux groupes de consonnes initiales et finales ainsi qu'aux noyaux vocaliques de ces syllabes, de même qu'aux segments phonétiques qui comprennent les groupes.

Le fait que la hiérarchie comporte de nombreux niveaux signifie que chacun de ces derniers ne nécessite qu'un minimum d'information sur le rythme, de sorte qu'il est possible de disposer l'information en tableau plutôt que se fier à des algorithmes pour en obtenir des approximations. Il semble que le mécanisme dont on dispose soit suffisamment souple pour accepter des données supplémentaires obtenues de mesures des prononciations naturelles, sans que la structure sous-jacente de calcul soit modifiée.

# SYNTHESIZING BRITISH ENGLISH RHYTHM - A STRUCTURED APPROACH

Ian H. Witten, Department of Electrical Engineering Science, University of Essex.

Alexandra Smith, Department of English Language, University of Edinburgh.

## 1. Introduction

Rhythm has always been something of an obstacle in automatic speech synthesis. Early programs for synthesis-by-rule used simple context-free table-lookup schemes to determine the durations of phoneme segments, but a facility for over-riding the table values by giving phonemes optional markers was quickly found to be necessary (Holmes et al, 1964). Later systems were implemented which automatically lengthened portions of stressed syllables (Mattingly, 1966; Ainsworth, 1974). Recently, there has been a spate of publications reporting new data on segmental duration in various contexts (for example, Haggard, 1973; Klatt, 1973, 1975; Lehiste, 1973; Oller, 1973; Umeda, 1975), and there is a growing awareness that segmental duration is influenced by a great many factors, ranging from the structure of a discourse, through semantic and syntactic attributes of the utterances, their phonemic and phonetic make-up, right down to physiological constraints (these multifarious influences are ably documented and reviewed by Klatt, 1976).

Our understanding of speech rhythm knows many laws but little order. What seems to be lacking in much of the work cited above is a conceptual framework on to which new information about segmental duration can be nailed. Apart from its theoretical value, such a framework is important practically because of the current surge of interest in speech synthesis systems for computer output. Applications-oriented designers of synthesis software can be excused if they shrink from including rhythm assignment in their programs, for the mass of results which has been reported is difficult to integrate into a sensible, implementable, and - most important - easily extensible package.

This report describes a vehicle for the assignment of duration to synthetic speech, which takes the form of a hierarchical structure of levels. The definition and treatment of the upper levels are based on Abercrombie's work on the rhythmic structure of English. We take as our starting-point the hypothesis of regularly recurring stresses ("isochronous feet"). The syllable structure of each foot is identified and this is used to determine a rhythm for the foot at the syllabic level (Abercrombie, 1965). Once a syllable's duration has been calculated in this way, it is distributed amongst the clusters which comprise the syllable - the initial and final consonant clusters, and the nuclear cluster. This introduction of the cluster level is the chief innovation over earlier work (Witten, 1977). Then the cluster duration is split between the constituent phonemes, and their transition and steady-state times are computed.

The procedures to be outlined have been implemented on a computer as part of a larger speech synthesis program. It was considered especially important that the system degrade gracefully in the face of erroneous or unanticipated inputs, and extensive use has been made of "default" assumptions (which cope in a sensible manner with any situation), overlayed by specific strategies which are triggered by certain inputs. This philosophy allows new results to be incorporated easily by expanding the repertoire of cases which are dealt with by overriding the defaults. So that the system can work with a sensibly small bank of specific strategies, careful attention has been paid to defining an hierarchical structure which allows many cases at one level to be covered by just one exception routine at the next higher level.

A further simple but practically important feature of the rhythm assignment scheme is that most of the decisions it takes can be preempted by appropriate markers in the phonemic input string. Thus, for example, although an attempt is made to divide the utterance into syllables using a procedure described below, this may violate accepted boundaries on occasion because it makes no use of sub-lexical morphemic decomposition. However, syllable boundaries can optionally be placed in the input string, and the procedure will take account of these. This will be useful in case the input is prepared by a morpheme-based text-to-speech program (Allen, 1976; Witten & Pope, 1976), which can easily specify syllable boundaries at the juncture of morphs, leaving the rest of the syllabification to the procedure. Durations of feet, syllables, and even of phoneme segments can all be explicitly specified as and when desired.

Section 2 of the report reviews the theory of isochronous feet, as propounded by Abercrombie. This is incorporated into the foot level of the computer procedure, the operation of which is described next. Then the general structure of syllables is discussed. This leads to algorithms for splitting the phonemic string into its constituent syllables and for identifying the clusters within each syllable. Syllable durations are allocated to phonemes via the cluster level. Section 4 briefly outlines the segmental level rhythm procedures, which check that the segment durations lie between certain maxima and minima, and compute the segment transition times.

## 2. General description of the theory of isochronous feet.

The foot is the basic unit of timing, and is delimited by ictuses or pulses which are felt to recur regularly, other things being equal. A foot begins with an ictus and runs until, but does not include, the next one, which begins the next foot. The ictus is marked in the input to the procedure by a vertical line as follows:

|This is the |house that |Jack |built.

That part of the foot which is not the ictus is called the remiss.

The ictus may be occupied by a syllable, which in that case is salient, or filled by a silent stress (which actually may not be silent, but instead may be filled by a "hesitation noise" or a vowel or other prolongable sound carried over from the preceding syllable). A silent stress is marked by a caret:

|This |^ is the |house that |Jack |built.

The remiss may be occupied by various numbers of syllables, from none to four or five. Syllables filling the remiss are called weak or non-salient.

A foot may be entirely empty of syllables, and consist only of a silent stress.

It may have the ictus filled and the remiss empty, like the foot "Jack" in the example above. In these cases the salient syllable is usually pronounced rather long, filling the whole foot, but the foot is often on the short side compared with feet of more than one syllable.

A foot may have each of its two places filled by one syllable. According to Abercrombie (1965), in the R.P. accent these two-syllable feet may be of three types: with the total duration divided equally between the two syllables, or with a long plus a short, or with a short plus a long. Which type a foot is depends in that accent upon whether there is a word-division within the foot, and if not, upon the structure of the first of the two syllables.

In order to investigate two-syllable feet further, let us call a syllable long-in-quantity if one of two conditions is met:
1) If the vowel is one of the class of "unchecked" vowels, that is, one of (EE, E I, AR, AW, UH U, UU, AR U, AA I, U UH, O I, I UH, E UH, ER, AW UH), (Footnote 1), or
2) If the vowel is followed by two or more consonants.
If neither of these two conditions is met, then the syllable is short-in-quantity.

The three types of two-syllable feet are then:

2A |meadows |trim |∪ — | | 1 : 2 | | $S^S$ . W |
1. There is no word-boundary within the foot.
2. The first syllable is short-in-quantity.

2B |centre |forward |∩ ∩ | | 1 : 1 | | $S^L$ . W |
|sofa |bed
1. There is no word-boundary within the foot.
2. The first syllable is long-in-quantity.

2C |tea for |two |— ∪ | | 2 : 1 | | S # W |
1. There is a word-boundary, marked by "#", within the foot.

Unfortunately, printers' and phoneticians' word-boundaries are not always the same. Some little words behave phonetically as if they were part of another word, that is they are enclitics. This is sometimes reflected in casual spelling such as "pinta" in |pint of |milk (|∩ ∩ | instead of |— ∪ |).

---

Footnote 1.

Unchecked vowels are those which can occur in open monosyllabic words (the so-called "long vowels"): the vowels in bee, bat, bar, paw, mow, boo, bough, buy, boor, boy, beer, bear, burr, bore. Checked vowels are the so-called "short vowels", which must be closed by a consonant in stressed monosyllabic words: the vowels in bit, bet, bat, but, pot, book. Dipthongs, which are classified as unchecked, are treated as sequences of vowels (VV) in our procedure.

A foot may have the rhythmical characteristics of a two-syllable foot while having only one syllable, if the ictus is filled by a silent stress instead of a salient syllable. The rhythmic structure may be:

2C | — ∪ | the more usual type, with a long silent stress followed by a notional word division and then a short weak syllable:
There |once was a |man from Khar|toum, |∧ who |kept two black |sheep ...

2B | ∩ ∩ | with what feels like a sort of syncopation, a silent stress followed by a deliberate, clearly enunciated syllable:
|Want to be a |cave-man? |∧ No |fear!

A foot may have more than one syllable in the remiss. Examples of trisyllabic feet are given by Abercrombie (1965) and investigated in more detail by Sumera (1971).

Trisyllabic feet are of five kinds, depending on the presence or absence of word-boundaries within the foot and upon the quantity of the syllables, as with disyllabic feet.

3A |one for the |road | ∩ ∨ ∨ | | 2 : 1 : 1 | | S # W # W |
|it's incon|ceivable | S # W . W |
1. There is a word-boundary after the first syllable.

3B |little and |small | ∨ ∩ ∪ | | 1 : 3 : 2 | | $S^{S}$. W # W |
1. There is a word-boundary after the second syllable only.
2. The first syllable is short-in-quantity.

3C |after the |war | ∪ ∩ ∨ | | 2 : 3 : 1 | | $S^{L}$. W # W |
1. There is a word-boundary after the second syllable.
2. The first syllable is long-in-quantity.

3D |nobody |knows | ∩ ∨ ∪ | | 3 : 1 : 2 | | $S^{L}$. W . W |
1. The foot contains no word-boundaries.
2. The first syllable is long-in-quantity.

3E |anything |more | ∪ ∪ ∪ | | 1 : 1 : 1 | | $S^{S}$. W . W |
1. There are no word-boundaries within the foot.
2. The first syllable is short-in-quantity.

Feet of four and five syllables are considerably rarer, and also more difficult to analyse and systemize. We know of no orderly attempt to do so. It is hoped that when the differences between two- and three-syllable feet have been adequately described, a method for describing feet of more syllables will be developed from a comparison of the behaviour of weak syllables.

Abercrombie has recommended a further set of distinctions which are useful and should be kept in mind. These concern the terms salient syllable, accent, and stress. A salient syllable is the first syllable of a foot in an actual utterance. Accent is the potentiality of salience, as marked in a dictionary or lexicon. Accent is language- (or even dialect-) specific. In normal speech (but not in shouting), only accented syllables can be salient, although in a given utterance not every accented syllable need be salient. Stress is a general phonetic phenomenon associated with increased muscular activity. In many cases, salient syllables are stressed, but not all speakers do

this in all styles. In shouting, on the other hand, all syllables are usually stressed. "Stress" is often used as an equivalent of "salient" or "accented" or both. "Accent" is sometimes called "word-stress", while "salience" is sometimes called "sentence-stress". If the three terms salient, accented and stressed are carefully distinguished, then one can avoid terms of uncertain definition such as "half-stress" - which often seems to be used for "accented but not salient in this utterance". For instance, the first syllable of the word "very" is accented, that is, potentially salient, but in a sentence as uttered it may not be salient: one can say "|∧ he's |very |good" or "|he's very |good". While it is convenient to be able to make these distinctions, in normal usage (including the rest of this paper), the word "stress" refers to Abercrombie's "salient", i.e. foot-initial.

## 3. The foot level

For the purposes of the suprasegmental analysis to be considered here, the highest level in the hierarchy is the utterance. This comprises one or more intonation units (Halliday's (1967) "tone groups"). For rhythm assignment, the utterance is split into feet, and these are assumed to be already marked in the phonemic input string. A current research topic is the automatic placement of foot boundaries, which in much American literature are called "level-2 stresses" (e.g. Chomsky and Halle, 1968); this is discussed by Witten and Pope (1976). The reason for defining the utterance level, instead of having the intonation unit at the top of the hierarchy as in earlier work (Witten, 1977), is that a rhythmic foot can span an intonation unit boundary.

Each foot has a target duration, currently set at 480 msec, which, if achieved, will produce an exactly isochronous utterance. However, provision is made for the "tonic" foot which occurs at the semantic focus of each intonation unit to be assigned a different target duration from the other feet. This permits the option of heightening the perceptual stress at the tonic by an increase of duration, as well as by a distinctive pitch movement, which has been incorporated into previous rhythm procedures (Mattingly, 1966; Ainsworth, 1974). Note that the target foot duration is only rarely realized, due to the influence of rules at the syllabic and phonetic levels, so that the foot durations will not in fact be exactly equal.

The syllables in the foot are identified (details of the automatic syllable-splitting method are given in the next section), and are classified as long-in-quantity or short-in-quantity according to the criteria laid out above. The default assumption for syllable rhythm is that each syllable is given the same target duration, except that stressed (i.e. foot-initial) syllables which are long-in-quantity are twice as long as each of the other syllables. This default assignment deals sensibly with feet having many syllables. If the foot has one of the two- or three-syllable structures 2A, 2B, 2C, 3A, 3B, 3C, 3D, or 3E discussed above, then the appropriate syllabic rhythm is used instead of the default.

Syllable durations are not completely specified by this procedure, however. There is a preset minimum syllable length, to prevent short syllables in over-subscribed feet from getting lost. What if there are so many syllables in the foot that the ratio algorithm assigns less than the minimum duration to some of them? There are several possible

solutions.

A. Increase the time for the foot until the shortest syllable, when given as its length the appropriate proportion of the foot, becomes as long as the minimum permissable duration. (This will substantially lengthen the foot time).

B. Increase the duration of each syllable whose length falls short of the minimum, by just enough to make it equal the minimum. (This may radically alter the rhythm of the foot.)

C. As a compromise, add to the length of all the syllables that time by which the shortest one falls short of the minimum. (This partially destroys the rhythm, but keeps the foot time quite short.)

Strategy C is the one adopted by the program. We are experimenting with the minimum syllable duration; the current value is 140 msec. Thus a trisyllabic foot of type 3A, whose syllables have duration ratios 2:1:1, will have the actual syllable durations increased from 240, 120, 120 msec to 260, 140, 140 msec, each syllable's time being increased by the same amount. The actual rhythmic ratios will therefore be 26:14:14 = 1.86:1:1, not too different from the target. Note that if strategy B were adopted, the actual rhythmic ratio would be 24:14:14 (1.71:1:1), which is rather more substantially different from the goal of 2:1:1.

It seems clear that, on the whole, natural speech departs from isochrony by making feet with many syllables longer than feet with few syllables. In fact, Halliday (1967) mentions an informal experiment which obtained a ratio of 5:6:7 for the average durations of one-, two- and three-syllable feet. Our procedure will achieve this overall effect for feet with two or more syllables. For example, most types of trisyllabic feet (i.e. 3B, 3C and 3D) will last for 300+220+140 = 660 msec, which is not too far off the ratio of 7:6 over bisyllabic feet. Type 3A feet will last for 540 msec, while type 3E will only last for the standard foot duration of 480 msec, it being unnecessary to extend any of the syllables since none is shorter than the minimum.

Although the procedure takes care to ensure that syllables cannot get too short, it imposes no upper limit on their length. A single syllable in a foot by itself will be given a duration equal to the standard foot time. However, when the syllable duration is apportioned amongst the constituent phonemes, upper limits <u>are</u> imposed, and so the syllable may not in practice achieve its allotted span (see Section 4).

It seems sensible to us to impose lower limits at the <u>syllable</u> level. For if, as Abercrombie (1967) maintains, a syllable is actually related to a chest-pulse, then it will have a physiologically-imposed minimum duration, even though in natural speech the vowel which forms its nucleus may disappear. On the other hand, we know that phonemes can, under certain circumstances, have negligible duration (as when a vowel "disappears" in rapid speech). Conversely, there seems to be no reason to restrain long syllables except insofar as the articulators, having lingered in each posture for "long enough", move on before the target time for the syllable has elapsed.

## 4. The syllable level

### 4.1 Structure of the syllable.

Every syllable has the structure S = Cl N C2, where Cl and C2 are absent, simple, or compound consonant clusters, and the nucleus N is a vowel or diphthong. We call the components in this description the clusters of the syllable, even when a cluster comprises only one element.

Dividing phoneme segments into sonorants (R, W, L, Y, M, N, NG), obstruents (P,T,K,B,D,G,F,V,TH,DH,S,Z,SH,ZH), and vowels, we can write

Cl = O* S*, N = V*, C2 = S* O*

where *, as usual, means repetition zero or more times. Thus

S = O* S* V* S* O* .

Note that the rules implicit in these equations, for instance that Cl cannot contain a sonorant followed by an obstruent, and that C2 cannot contain an obstruent followed by a sonorant, actually do hold true for English.

A syllable cannot be null, although in the above formulation, each of the three primary components Cl, N, C2 can be null. (N is null in the case of a so-called syllabic consonant.) To avoid the necessity for interacting rules to constrain the existence of the primary components, and for other reasons associated with timing of the syllable, we re-define N and C2 as

N = V* S*, C2 = O*,

and insist that N alone be non-null.

### 4.2 Syllabification.

Syllables normally coincide with peaks of sonority, where "sonority" measures the inherent loudness of a sound relative to other sounds of the same duration and pitch. However, difficult cases exist where it seems to be unclear how many syllables there are in a word (Ladefoged, 1975, discusses this problem, with examples such as "real", "realistic", and "reality"). Furthermore, care must be taken to avoid counting two syllables in "sky" because of its two peaks of sonority (the stop "k" has lower sonority than the fricative "s").

Our system takes a rough and ready approach to syllabification, which is justified by the fact that the simple rules work most of the time, especially for carefully-enunciated text ("prononciation familière ralentie", i.e."slow conversational style"), and by the ability to insert difficult syllable boundaries in the input to guide the procedure (Section 1). Firstly, syllable boundaries are placed at word boundaries and either side of silent stress. It is not assumed that foot boundaries mark syllables, in case of errors in the input. (Many untrained people find it quite difficult to place foot boundaries appropriately.) Three levels of notional sonority are defined, based on the structure of the syllable outline above. Obstruents have sonority zero, sonorants one, and vowels two. Syllable boundaries are made to coincide with sonority minima. If only one segment has the minimum sonority, the boundary is placed before it. If there are two segments, each with the minimum sonority, the boundary is placed between them, while for three or more, it is placed after the first two.

These rules produce obviously acceptable divisions in many cases (to"day, ash"tray, tax"free), with perhaps unexpected positioning of the boundary in others (ins"pire, de"par"tment). Actually, people do differ in the placement of syllable boundaries (Abercrombie, 1967).

## 4.3 The cluster sub-level.

We assign durations to the segments of the syllable by apportioning the total syllable duration between C1, N, and C2; and then further subdividing it between the individual segments that make up these clusters.

The percentage of the total syllable duration assigned to C1 is as follows:

| | | |
|---|---|---|
| null cluster | 0% | percentage of syllable duration |
| cluster which contains a voiced obstruent | 25% | |
| any other cluster | 33% | |

If C1 is OOS, then the initial obstruent must be "s". In this case, a fixed duration is used for the initial "s", but this is not counted as part of the total syllable duration. The remainder of the cluster is treated as above. (It may be that the use of "tapping points" (Allen, 1972) will provide a more satisfactory solution to the problem of initial s's in the future.)

Having allocated a duration to C1, the remainder of the syllable duration is divided between N and C2. Although in this first-order theory C1 is considered independently of the nature of the rest of the syllable, there are strong interactions between N and C2 which must be modelled. Specifically, in many accents (including RP) the length of the vocalic nucleus is a strong cue to the degree of voicing of the terminating cluster (Lehiste, 1970). The classification of the nucleus as long or short according as the vocalic element is unchecked or checked (Section 2) is retained, and C2 is classified as voiced or unvoiced. Then the durations of N and C2 are expressed as percentages of the total N C2 time as follows:

| | | | | |
|---|---|---|---|---|
| <short> | <voiced> | 60% | 40% | percentage of N C2 duration |
| <short> | <unvoiced> | 50% | 50% | |
| <long> | <voiced> | 70% | 50% | |
| <long> | <unvoiced> | 60% | 60% | |

Note that the sum of these percentages sometimes exceeds 100% - thus the overall target rhythm of the foot is disturbed by low-level considerations. We feel that effects such as these can be made to account for a great deal of the observed non-isochrony and lack of rigid rhythmic structure of natural speech.

## 4.4 Assigning durations to the segments

To divide the duration of each cluster between its constituent segments, a simple technique of proportions is used again. The default assumption is that the cluster duration is split equally amongst the segments in it. Only in the case of N do we over-ride this default, and then only if N is VV or VS:

| | | | |
|---|---|---|---|
| VV | 67% | 33% | percentages of N duration |
| VS | 33% | 67% | |

It is anticipated that future experimental research will provide grounds for departing from the equal-division assumption in many other cases; even then, however, the default will be useful to deal with rare combinations of phonemes, and erroneous or unnatural phonetic inputs.

## 5. The segmental level

At the segmental level, durations assigned to the phoneme segments by the syllable level are reviewed, and transition times are computed. This corresponds to the imposition of articulatory constraints on the speech rhythm. In addition, any segment durations which are specified explicitly in the input are used to over-ride the value which has been calculated.

A maximum duration is specified for each phoneme segment, in the rule table. (It is anticipated that, in future, the segments will be divided into a small number of classes for this purpose, each class having a uniform maximum duration, but the full flexibility of individual specifications is retained at present for experimentation.) In the case of tonic, utterance final, one-syllable feet, the maximum for each segment is multiplied by a constant (currently set at 200%), to account for the fact that exceptionally long syllables generally occur under these circumstances (Lehiste, 1973).

Phoneme transition times are computed by a method due to David Hill (private communication, 1975). Segments are classified into stops (P,T,K,B,D,G,M,N,NG), approximants (R,W,L,Y), fricatives (F,V,TH,DH,S, Z,SH,ZH), and vowels. A transition into or out of a stop lasts for 20 msec. For diphthongs, the steady-state duration of the second component is set at 20 msec, and the remainder of its allotted time is given to the transition, a minimum of 50 msec being imposed. Transitions into approximants are treated in the same way. In all other cases, a transition lasting 50 msec is used. Throughout, a minimum steady-state time of 20 msec is imposed, even if this increases the total duration calculated for the segment. These rules are intended to model articulatory constraints on phoneme production.

## 6. Conclusion

This paper has described a method of assigning rhythm to synthetic speech. The main emphasis is on an hierarchical structure, proceeding from an utterance to its feet, their constituent syllables, the clusters which form these syllables, and the segments which comprise the clusters. The many levels in the hierarchy mean that only a small amount of information about the rhythm is needed at each one, and it becomes feasible to store the information, as derived from consideration of human utterances, in a table, rather than relying on algorithms to approximate it.

This shift to a table-driven approach represents the chief difference between this and the earlier work reported by Witten (1977). It is made feasible by introducing the cluster level which is absent from most analyses of speech rhythm. We found that much knowledge about the timing of speech is couched in terms of the CVC syllable - for example, one talks about the relative durations of the vowels, and of the terminating consonants, in syllables like "bad"/"bat" - and this knowledge is in fact equally applicable to more general syllable structures. The cluster level forms the bridge between actual syllables and

the CVC archetype.

A further advantage of the procedure over early ones is its robustness. This is achieved by simple algorithmic defaults for assigning durations at each level, which are usually over-ridden by appropriate table-driven rules.

It must be admitted that cut-and-try modifications to the rhythm rules have not yet been attempted. First impressions are that the synthetic speech rhythm generated by the procedure sounds quite good, - occasionally very good, - and is always tolerable. However, some adjustments need to be made. In particular, initial consonants are often too long. One seems to hear a double ("geminated") consonant, which prolongs the perceptual duration of the preceding syllable. It may be that the notion of a "tapping point" after the C1 cluster (Allen, 1972) will have to be reintroduced, so that C1 is treated in a fundamentally different way from the other cluster types.

What we hope will not have to change, however, is the hierarchical structure of levels, with the duration of a construct at one level being used to determine the durations of its constituent sub-constructs at the level below.

Acknowledgements

Grateful thanks are due to our many colleagues in the Department of Linguistics, University of Edinburgh, and the Department of Electrical Engineering Science, University of Essex. In particular, our ideas of rhythm are based on the work of David Abercrombie, and we have been influenced in our design of the hierarchy by Roger Moore. Alexandra Smith was supported by the Joint Speech Research Unit, and the speech research project at Essex is partially supported by the SRC.

References

Abercrombie, D. (1965) Studies in phonetics and linguistics. London, O.U.P.

Abercrombie, D. (1967) Elements of general phonetics. Edinburgh Univ. Press.

Abercrombie, D. (1976) "Stress" and some other terms. In Work in Progress 9, Dept. of Linguistics, Univ. of Edinburgh.

Ainsworth, W.A. (1974) Performance of a speech synthesis system. Int.J.Man-Machine Studies 6(5): 493-511.

Allen, G.D. (1972) The location of rhythmic stress beats in English; an experimental study. Language and Speech 15(1): 72-100 and 15(2): 179-195.

Allen, J. (1976) Synthesis of speech from unrestricted text. Proc.IEEE 64(4): 433-442.

Chomsky, N. & Halle, M. (1968) The sound pattern of English. New York: Harper & Row.

Haggard, M. (1973) Abbreviation of consonants in English pre- and post-vocalic clusters. J. Phonetics 1: 9-24.

Halliday, M.A.K. (1967) Intonation and grammar in British English. Paris: Mouton.

Holmes, J.N., Mattingly, I.G. & Shearme, J.N. (1964) Speech synthesis by rule. Language and Speech 7(3): 127-143.

Klatt, D.H. (1973) Interaction between two factors that influence vowel duration. J.Acoust.Soc.America 54(4): 1102-1104.

Klatt, D.H. (1975) Vowel lengthening is syntactically determined in a connected discourse. J.Phonetics 3: 192-140.

Klatt, D.H. (1976) Linguistic uses of segmental in English: Acoustic and perceptual evidence. J.Acoust.Soc.America 59(5): 1208-1221.

Ladefoged, P. (1975) A course in phonetics. New York: Harcourt Brace Jovanovich.

Lehiste, I. (1970) Suprasegmentals. M.I.T. Press.

Lehiste, I. (1973) Rhythmic units and syntactic units in production and perception. J.Acoust.Soc.America 54(5): 1228-1234.

Mattingly, I.G. (1966) Synthesis by rule of prosodic features. Language and Speech 9: 1-13.

Oller, D.K. (1973) The effect of position in utterance on speech segment duration in English. J.Acoust.Soc.America 54(5): 1235-1247.

Sumera, M. (1971) Aspects of rhythm and verse structure in English. Ph.D. Thesis, Univ. of Edinburgh.

Umeda, N. (1975) Vowel duration in American English. J.Acoust.Soc. America 58(2): 434-445.

Witten, I.H. & Pope, R.J. (1976) Rhythmic stress in synthetic speech-by-rule. Proc.Institute of Acoustic Autumn Conference, Edinburgh, September.

Witten, I.H. (1977) A flexible scheme for assigning timing and pitch to synthetic speech. Scheduled for publication in Language and Speech, September.