Proceedings of the Institute of Acoustics

ON THE PHONETIC INTERPRETATION OF RHYTHM IN NON-SEGMENTAL SPEECH SYNTHESIS

John Local

Department of Language and Linguistic Science, University of York, UK

1. INTRODUCTION

One of the enduring problems in achieving natural sounding synthetic speech is that of getting the rhythm of the speech output right. Usually this problem is construed as the search for appropriate algorithms for altering durations of segments under various prosodic conditions (eg in stressed versus unstressed syllables). Van Santen [1] identifies a number of different approaches employed to control timing in synthesis applications and provides and overview of their relative strengths and weaknesses. All the approaches he deals with rest on the assumption that there is a basic unit to be timed, that it is some kind of (phoneme-like) segment and that rhythmic affects are assumed to 'fall-out' as a results of segmental-level timing modifications. As is obvious on listening to extant text-to-speech systems implementing these duration models this is not the case. Recently, Campbell and Isard [2] have suggested that a more effective model is one in which the syllable is taken as the distinguished timing unit and segmental durations accomodated secondarily to syllable durations. However, rhythm is relational we are not simply dealing with strings of units (segments, syllables, demisyllables or whatever) but with relations between units in particular pieces of linguistic structure. Any successful account of the rhythmic organisation of language requires representations which will permit the expression of hierarchical structure of varying domains.

In large part the pervasive problem of rhythm in synthesis trouble can be seen to arise from the adherence by researchers to representations which are based on concatenated strings of consonant and vowel segments and which allocate those segments uniquely to a given syllable. In the approach sketched here, we reject string-based data structures in favour of hierarchically structured, non-segmental representations which admit structure sharing. These provide for the felicitous expression and representation of relationships necessary to generate synthetic versions of polysyllabic utterances which faithfully mimic the rhythmic organisation of natural speech.In order to focus what I have to say discussion will be confined to a consideration of polysyllabic words. I will consider how syllables are joined together in phonological representation to construct representations for such words and how those representations are given phonetic interpretation

2. STRUCTURE AND TIMING IN NON-SEGMENTAL SYNTHESIS

The YorkTalk speech generation system (Coleman [3], Local [4], Ogden [5]) is a Prolog-based computer program which creates synthesis parameter files from non-segmental phonological representations based on Firthian Prosodic phonology (Firth, [6]). There are two main components to the system: phonotactic and metrical parsers and a phonetic interpreter.

The parsers are employed to construct non-segmental phonological representations which are structured, directed acyclical graphs, rather than the more usual strings of segment symbols. Figure 1, below, provides a simplified example of a such a graph which reflects our non-segmental analysis of generalised English monosyllables: (*heavy / light* represents the structural distinction between syllables with branching rimes and/or branching codas and those with non-branching rimes and codas. The phonological units *y / w* operating at the syllable node represent the phonological distinction between forms such as *pit* and *put*; or *geese* and *goose*; the features *h* / ¬ *h* represent the distinction operating at onset between forms such as *pit* and *bit*, and at rime between forms such as *bit* and *bid*. – / ¬ – operating at the rimal structural constituent represent the distinction between forms such as *bet* and *bent* or*bed* and *bend*; *q* / ¬ *q* represent the distinction between checked and unchecked rimes; *i / ɛ / a* represents the terms in 'height' contrastivity system.

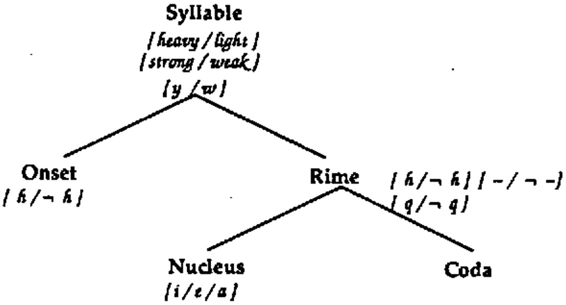


Figure 1: Partial phonological representation for generalised English monosyllables

These graphs is that the constituents are unordered and there is no distinguished type of phonological constituent and phonological information is distributed over the entire structure and not concentrated at the terminal nodes. A graph of this kind makes it possible to represent phonological contrastivity wherever it is needed in the structure - at phrase domain, word domain, at syllable domain, at constituent of syllable (onset, rime etc) for instance. The graphs must be phonetically interpreted in order to generate parameter files. This interpretation is of two kinds: temporal interpretation and parametric phonetic interpretation. Temporal interpretation establishes timing relationships which hold across the constituents of the graphs. Parametric interpretation (exponency) instantiates time-value parameter strips for any given piece of structure (any feature or bundle of features at any particular node in the graph). These parameters are ones which provide the information necessary to drive the Klatt cascade-parallel formant synthesiser. These non-segmental representations allow in a straightforward fashion for a rather different approach to matters of timing and parameter instantiation than do the distinguished unit or segment-sequence approaches.

Consider the following standard segmental model timing interpretation of a syllable-graph: onset_start = syllable_start; onset_end = rime_start; nucleus_start = rime_start; nucleus_end = coda_start; coda_end = rime_end. Such an interpretation proposes that the nodes can be instantiated with phonetic parameters which are well-ordered, concatenated sequences of 'phonetic objects' of some kind and treats all parts of the the structure as having the same status. By contrast, we recognise explicitly that the phonetic interpretation of some parts of structure is dependent on the phonetic interpretation of some other parts of structure (ie onsets on rimes). This leads us to propose a rather different view of the timing relationships thus: onset_start = syllable_start; rime_start = syllable_start; onset_end = onset_start + onset duration. From these constraints alone (and there are others, of course) it is possible to deduce that, in our model, the interpretation of the onset starts at the same time are the interpretation of the rime. They are not concatenated, but 'co-produced'. The co-production model of speech production (Fowler, 1980) accounts, amongst other things, in a straightforward way, for the so-called coarticulatory effects observed between onsets and rimes.

3. BUILDING POLYSYLLABIC STRUCTURES

In YorkTalk, single syllable structures of the kind pictured above in Figure 1, are defined straightforwardly by means of phonotactic phrase structure grammar of English. But what happens when we want to build larger structures such as disyllabic words? How are the syllable timing relations extended to handle multi-syllabic structures? Our approach to phonology is declarative (non-derivational) and constrained by the principle of compositionality which states that the 'meaning' of a complex expression (eg a syllable) is a function of the meanings of its parts and the rules whereby the parts are combined. We phonetic interpretation is compositional and consistent. Any feature, or bundle of features, at a particular place in a phonological representation is always interpreted in the same way. Within this approach, it should not be neccessary to invent novel, independent categories to deal with polysyllabic words. These larger structures should be composable from existing smaller ones employing a hierarchical constituency of a similar kind. Thus, if intervocalic consonantal portions can be treated as simply the concatenation of possible onsets and codas so that every such portion can be analysed as end of one monosyllable and beginning of another, it should be possible to build polysyllabic words simply by concatenating well-formed monosyllabic structures. However, apparently not all combinations of legal codas and legal onsets are permitted. For instance, in non-compound words we do not find intervocalic consonantal portions such as -lptfr- although monosyllables such as *sculpt,* with -lpt at their ends exist, as do monosyllables such as *frown* with fr- at their beginnings but these clusters do not seem to occur intervocalically in English. This 'exceptional' intervocalic portion is misleading, however, because there are no well formed codas in English such as -lpt. Such structures are to be treated as binary branching codas with an 'appendix' ie a additional piece of (morphological) structure which is immediately dominated by the word node. This provides a felicitous account of why they do not occur in intervocalic position within a word.

Another important constraint on intervocalic consonant pieces is that they must not be 'overlong'. Consider the following well-formed syllables which can be defined by a grammar of English phonotactics: ɒpt and tɪk, ɒsp and preɪ. Straightforward joining of such syllables, however, yields the ill-formed poylsyllabic strings ɒpttɪk and ɒsppreɪ. This provides additional evidence that the task of defining the set of well-formed polysyllables is more complex than simply concatenating syllables and fitting them to lexical or metrical structures. I will show below that the 'overlong' consonant-piece constraint is one of the problems which can be elegantly dealt with by the construct of ambisyllabicity.

4. SYLLABIFICATION AND AMBISYLLABICITY

With the machinery of phonotactic parser of the kind referred to above, it is possible to determine structured representations for well-formed monosyllables. However, even with this resource, there are clearly a number of ways in which syllabification of words greater than a single syllable can be achieved and all have found some support in the literature. Syllabifications may, for instance, differ depending on whether we are dealing with a phonotactic structure with (a) a single intervocalic C, (b) more than one C intervocalically and/or (c) whether or not the first syllable has a phonologically long or short head.To illustrate this consider possible syllabifications of the word *hammer* (the subscript numbers at the edges of the brackets indicate syllable affiliation: (a) $[_1$ham $]_1$ $[_2$ ə $]_2$ (maximal coda); (b) $[_1$ha $]_1$ $[_2$ mə$]_2$ (maximal onset); (c) $[_1$ha $[_2$m$]_1$ ə $]_2$ (maximal coda and maximal onset: ambisyllabicity)

Of these (a) is likely to be deemed the least problematic, though there are problems with the interpretation of the phonetic parametric join of the intervocalic C and the final vocalic portion. Syllabification (b) is phonologically problematic in that it runs counter to the observation that stressed monosyllables with 'short' rimal heads must be of closed syllable types. The syllabification in (c) might also be deemed problematic in that it countenances what has been referred to as 'ambisyllabicity' wherein the intervocalic C is taken to be at one and the same time the coda of the first syllable and the onset of the second. However, this final syllabification provides the key representational mechanism within YorkTalk for ensuring that the phonetic interpretation of polysyllables is appropriate.

Ambisyllabicity is structural way of treating syllables in contact. An important part of its motivation derives from observations about the nature and variability to be found in portions of utterance. Ambisyllabic portions may have characteristics which differ from 'the same' phonological unit in initial or final position or which mix exponency characteristics of both initial and final position (eg the t r cluster in *petrol* where the intervocalic closure portion may have coincident glottal closure (as in final position) while the post alveolar release portion may have the voicing, temporal and other characteristics associated with its co-occurrence (syllable-initially) with voiceless apicality and plosivity). However, the most compelling motivation for the recognition of ambisyllabicity, then, is that it does away with the need to posit novel objects in our analysis; it removes need to formulate a phonotactic sub-grammar specifically for word-internal clusters. It also obviates the problem of 'overlong' intervocalic consonantal portion. It is possible to enforce the constraint that the left and right parts of an intervocalic cluster are a coda and onset respectively, and that repetition is prohibited, by enforcing maximal ambisyllabicity.

Ambisyllabicity within the YorkTalk non-segmental model explicitly involves *structure sharing* rather than simply the sharing of some terminal (segmental) element. This allows us (a) to avoid syllabifications which would violate the short nucleus-open syllable constraint (b) to preserve a thorough-going compositional account of the building of polysyllables and (c) to provide just the right phonetic interpretation to the medial consonantal portions of polysyllables (eg timing, resonance affiliations, aspiration, glottalisation as well as the appropriate exponents of syllable rhythm and quality).

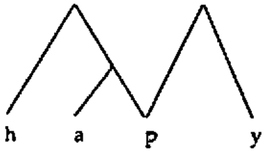
5. RHYTHM IN ENGLISH DISYLLABLIC FEET: 'SQUISH' IN SYNTHESIS

Abercrombie [7] provides a seminal description of rhythmic-quantity configurations in disyllabic feet in English. In particular he draws attention to two different kinds of rhythmic patterns which can be observed in initially accented disyllables. The first, which he labels 'short long' is found in disyllabic structures where the first syllable can be analysed as 'light' (ie a Rime with a phonologically short nucleus and non-branching coda). The second kind of rhythmic patterning, 'equal-equal' is found in disyllabic structures having a 'heavy' first syllable (ie a Rime with a phonologically short nucleus and branching coda or a phonologically long nucleus irrespective of coda structure). This analysis accounts for the observed differences in word pairs such as *whinny* versus *windy* and *filling* versus *filing*. Notice that the first pair of words allow us to see that the phonetic exponents of a syllables strength and weight (which in large part determine its rhythmic relations) include many more features than traditional 'suprasegmentals' and are not simply local to that syllable. The final vocalic portions in these words have different qualities depending on whether the first syllable is light or heavy.

In YorkTalk, such rhythmical effects are primarily modelled by the temporal interpretation function 'squish' - a unit of temporal compression. Squish is a means by which the system calculates the duration of any given kind of syllable in a given context. It depends on the structural piece in which the syllable occurs eg its position in the metrical foot; in strong syllables, it depends on the weight of the syllable; typically in foot-medial syllables, on the distribution of the [voice] feature in the Onset and Rime and in foot-final syllables, on the distribution of the [voice] in the Rime, the weight of the syllable, and contents of the Rime. This last constraint ensures a Squish which will yield a percept of syllabic nasals and liquids. The appropriate temporal interpretation of final weak syllables depends on having available information about the weight of the preceding syllable; once the appropriate Squish is employed, the percepts of tenseness/laxness and vowel quality differences fall out automatically.

6. THE PHONETIC INTERPRETATION OF AMBISYLLABICITY

Given a graph representation of the following kind for a word such as *happy*. How is this ambisyllabic piece of structure to be phonetically interpreted? (The ambisyllabic portion is indicated in outline font.)



One important aspect of ambisyllabic portions is that if the utterance is to sound natural the portion must 'coarticulate' properly with its flanking vocalic portions. In terms of the non-segmental approach to phonetic interpretation employed in the YorkTalk model, coarticulation is not a mechanical effect which arises from some kind of 'temporal smearing' of 'neighbouring segments', as is so often asserted. Rather, it is part of the phonetic interpretation of the phonological domains Onset and Coda. For intervocalic consonantal portions to coarticulate with their flanking vocalic portions, then, we need to ensure that they are interpreted just like codas with respect to the preceding portion, and just like onsets with respect to the following portion. As well as making them sound right, such a step allows for the use of the existing definition of the exponency function to generate intervocalic consonantal portions using appropriate parts of the parametric data for codas and appropriate parts for onsets. However, we cannot simply represent intervocalic consonants as a concatenative sequence of phonologically similar coda and onset, as the phonetic interpretation of codas includes characteristics appropriate for a syllable final release. What we require is that the phonetic interpretation of the intervocalic consonant should start off like a coda and then 'evolve' into an onset. In a segmental model implementation of such an interpretation is unlikely to be straightforward. In the non-segmental YorkTalk model, where the phonetic interpretation is parametric exponency of partial phonological structures the process is tractable. It can be accomplished as follows. Observe that in the case of coda plosivity we have an internal sequential structure: a closing phase: *Closing_Coda*, a closure phase: *Closure_Coda*, and a release phase: *Release_Coda*. Similarly onset plosivity has an internal sequential structure: a closure phase: *Closure_Onset*, and a release phase: *Release_Onset*. In voiceless stops the closure phase is acoustically silence. We hypothesise that up to closure, intervocalic stop consonants are like codas, and after closure they are like onsets. The internal temporal structure of an intervocalic stop consonant, then, is *Coda_Closing, Coda_Closure, Onset_Closure,* and *Onset_Release*. A straightforward way of instantiating the parametric exponents of these ambisyllabic pieces of structure is to construct the relevant first syllable parameters up to the coda closure, construct those for the second syllable from onset closure and then to overlay the parameters for the second syllable on those of the first at an appropriate point. By doing this it is unnecessary to invent new parametric exponents and we can preserve the thorough-going compositional phonetic account of ambisyllabicity. Figure 2 below gives a schematic representation of syllable overlaying.

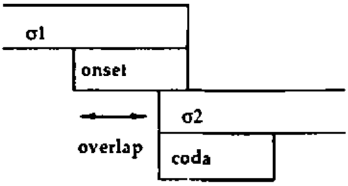


Figure 2: Temporal relations between two overlaid syllables

The extent of temporal overlaying is an empirical issue and actual proportions vary with different bits of phonological structure. Different amounts of overlap can be used to model in, part, the difference between ambisyllabic and geminate structures. Compare the intervocalic portions in *holy* and *holly* with that in *wholly*. In the first pair of words the laterality expones ambisyllabic structure and the interpretation of the appropriate parameters is dependent on both syllable1 and syllable2 and their Coda and Onset. Typically in these words the period of laterality is shorter (we have greater overlap) and has quite different resonance characteristics from that in the word *wholly*. In *wholly* where laterality expones gemination, interpretation of the parameters is not bi-dependent. This has striking consequences for the duration of the period of laterality, (it is typically longer) and its resonance characteristics (it is typically darker). The two spectrograms in Figure 3 show the effects of different amounts of syllable overlap in two synthetic versions of the word *silly*. The second of the pair has less syllable overlap. Notice the difference this makes to the temporal and spectral characteristics of the ambisyllabic portion (particularly observable in the second formant).

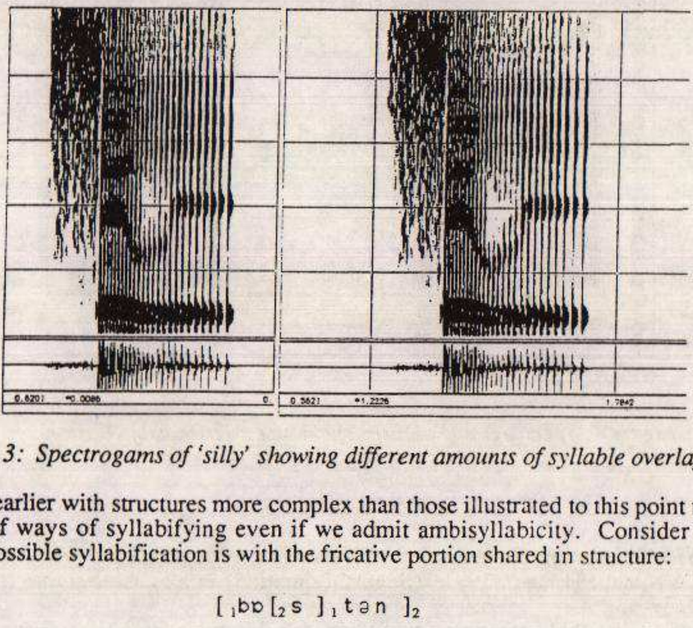


Figure 3: Spectrogams of 'silly' showing different amounts of syllable overlap

As I indicated earlier with structures more complex than those illustrated to this point there may be a number of ways of syllabifying even if we admit ambisyllabicity. Consider the word *Boston*. One possible syllabification is with the fricative portion shared in structure:

$$[_1\text{bɒ}\,[_2\text{s}\ ]_1\,\text{tən}\ ]_2$$

Under this analysis the strong first syllable is 'light' and thus the predicted rhythm is short-long. However, this is not what is observed. The rhythm of *Boston* is Abercrombie's equal-equal (ie heavy first syllable). Thus we require a syllabification which analyses the first syllable as 'heavy' such as:

$$[_1\text{bɒs}\,[_2\text{t}\ ]_1\,\text{ən}\ ]_2 \quad \text{or} \quad [_1\text{bɒ}\,[_2\text{st}\ ]_1\,\text{ən}\ ]_2$$

These last two representation while providing for an appropriate equal-equal rhythm make different predictions for the phonetic interpretation of the word-internal consonantal portion. The first with shared, ambisyllabic t predicts an aspirated (ie syllable-initial) release of the apicality and plosion. The second, with ambisyllabic **st** predicts an unaspirated release of the plosivity (ie a syllable initial clustered C). This structure, which gives just the right rhythmic and articulatory/phonatory phonetics, is the one provided by the principle of maximal ambisyllabicity with syllables 'squishes' appropriate to the structure. In Figure 4 below spectrograms of the word *Boston,* synthesised with YorkTalk's compositional parametric phonetic interpretation, show the different acoustic consequences of the three different parses.

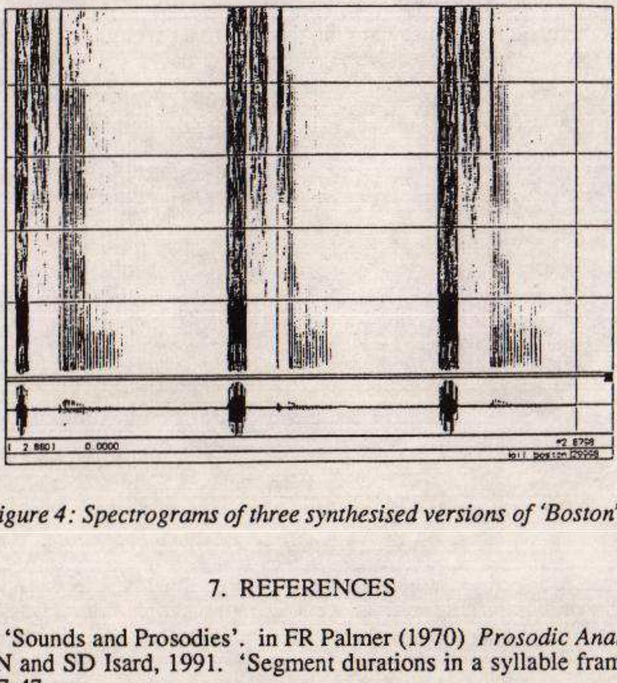


Figure 4: Spectrograms of three synthesised versions of 'Boston'

7. REFERENCES

Firth, JR, 1937. 'Sounds and Prosodies'. in FR Palmer (1970) *Prosodic Analysis.*

[1] Cambell, WN and SD Isard, 1991. 'Segment durations in a syllable frame'. *Journal of Phonetics,* 19, 37-47.

[2] Van Santen, JPH, 1992. 'Deriving text-to-speech durations from natural speech'. In G Bailly, C Benoit (eds), *Talking Machines.* Amsterdam: North-Holland, Elsevier. 275-287.

[3] Coleman, J, 1992. '"Synthesis-by-rule"' without segments or rewrite-rules'. In G Bailly, C Benoit (eds), *Talking Machines.* Amsterdam: North-Holland, Elsevier. 43-60.

[4] Local, JK, 1992. Modelling assimilation in a non-segmental, rule-free phonology. In GJ Docherty and DR Ladd (eds), *Papers in Laboratory phonology II.* Cambridge: CUP. 190-223.

[5] Ogden, R, 1992. 'Parametric Interpretation in YorkTalk'. *York Papers in Linguistics,* 16, 81-99.

[6] Abercrombie, D, 1964. 'Syllable Quantity and Enclitics in English' in *In Honour of Daniel Jones.* London: Longmans.