

# TEMPORAL INTERPRETATION IN PROSYNTH, A PROSODIC SPEECH SYNTHESIS SYSTEM.

Richard Ogden, John Local & Paul Carter  
*Department of Language & Linguistic Science*  
*University of York*  
*York, UK, YO10 5DD*

## ABSTRACT

ProSynth is an approach to speech synthesis which takes a rich linguistic structure as central to the generation of natural-sounding speech [3]. This paper outlines the model of temporal interpretation employed in ProSynth in generating polysyllabic utterances, and the phonological structures used to drive the synthesis. We start from the assumption that the speech signal is informationally rich, and that this acoustic richness reflects linguistic structural richness. The primary timing unit is the syllable, situated within a prosodic hierarchy. Two mechanisms are used for timing: (1) Syllables are joined by overlaying one over another (2) Syllables are temporally compressed to produce the correct rhythmical effects.

## 1. INTRODUCTION

ProSynth is a linguistic model for speech synthesis. In this paper, we describe the timing model used in ProSynth, which is built on YorkTalk [4, 5, 6, 7, 8], and the linguistic structures used in temporal interpretation.

The phonological structure used in ProSynth has units at the following levels: syllable constituents (Onset, Rhyme, Nucleus, Coda); Syllable; Foot; Accent Group; Intonational Phrase. Each lower-level unit is dominated by a unit at the next highest level (the Strict Layer Hypothesis). Constituents at each level have a set of possible attributes, and relationships between units at the same level are determined by the principle of headedness.

## 2. SYLLABLES.

ProSynth uses a non-segmental phonological model. Linguistic information is distributed across Directed Acyclic Graphs (a form of tree structure), with phonological contrasts available at non-terminal nodes.

In our view, segments are acoustic events which result from clusters of parametric change in the signal. We do not discuss parametric interpretation in this paper (see [4, 5, 7]). Where we have given segment durations, these are durations measured by hand from the acoustic output of a formant synthesiser using the ProSynth model to generate the timing using the same criteria as we would apply to natural speech. A graph for the syllable [sit] is presented in Fig. 1.

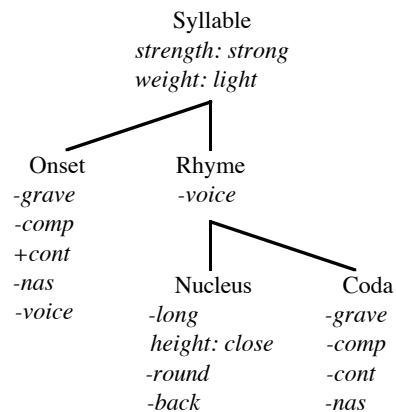


Fig. 1. Partial representation of the syllable /sit/.

### 2.1. Syllable weight.

The attribute <weight> has two values, *heavy* and *light*. Because of our view of syllable structure in polysyllabic structures, heavy syllables are syllables with either a branching Nucleus (i.e. a diphthong or a long vowel) or a branching Coda (i.e. a coda cluster), or both. All other syllables, including syllables of the shape CVC are treated as light. We do not use extrametricality in ProSynth, so that heavy syllables can also be weak, as in the final syllable of *statement*.

### 2.2. Syllable strength.

The attribute <strength> has two values, *strong* and *weak*. Syllables are immediately dominated by Feet, and Feet are left-headed. The head of a Foot is always a strong syllable.

### 2.3. Rhythm

One of our goals is to model English rhythms accurately. Abercrombie [1] describes three kinds of rhythm for disyllabic feet, of which two are important for disyllabic words: short-long and equal-equal. These two types correspond to the type of Foot-initial Syllable. If the first syllable is light, the rhythm is short-long: *happy, funny, city*. If the first syllable is heavy, the rhythm is equal-equal: *hamper, funding, seedy*. (These rhythms are true for Standard Southern British English, but the details for other dialects are different.) This rhythmical relationship can be modelled easily in a model which incorporates Foot structure; but one implication of Abercrombie's insight is that in interpreting the second Syllable of any Foot, reference has to be made also to the first Syllable of that Foot.

## 2.4. Ambisyllabicity.

In polysyllabic utterances, the join between syllables is crucial. We are testing the view that syllables within a word share as much information as possible: we pursue the hypothesis that intervocalic consonants are ambisyllabic as far as possible: any consonant which forms both an Onset and a Coda is parsed, wherever possible, as the Coda of one Syllable, and also the Onset of the next. Such consonants carry the feature [+ambisyllabic]. Ambisyllabicity means that structural information between syllables is shared: there is token-identity between the Coda of one Syllable, and the Onset of the next Syllable (Fig. 2).

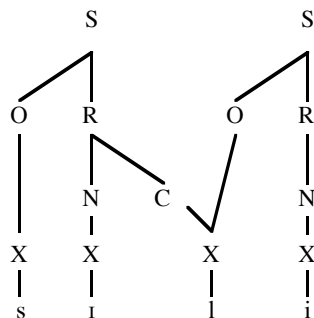


Fig. 2. Ambisyllabicity in the word *silly*.

The motivation for our view of ambisyllabicity is:

- No words in English have short, open Rhymes: syllabification of eg. *silly* as /sɪ•li/ gives an ungrammatical syllable /sɪ/. Ambisyllabicity means that a consistent analysis of syllables is possible.
- Ambisyllabicity means that on-glides can be consistently treated as properties of Codas, and off-glides into vowels as properties of Onsets.
- The secondary resonance of intervocalic consonants has properties of both the preceding and subsequent vowel [9]. Ambisyllabicity models this.

Constituents are [+ambisyllabic] wherever this does not result in a breach of constraints. *Loving* comprises two Syllables, /lʌv/ and /vɪŋ/, since /v/ is both a legitimate Coda for the first Syllable, and a legitimate Onset for the second. *Loveless* has no ambisyllabicity, since /vl/ is neither a legitimate Onset nor a legitimate Coda. Clusters may be entirely ambisyllabic, as in *risky* (/rɪsk/+/ski/), where /sk/ is a good Coda and Onset cluster; partially ambisyllabic (i.e. one consonant is [+ambisyllabic], and one is [-ambisyllabic]), as in *selfish* /sɛlf/+/fɪʃ/, or non-ambisyllabic as in *risk them* (/rɪsk/+/ðəm/).

Ambisyllabicity makes it easier to model certain segmental effects, such as coarticulation and plosive aspiration in intervocalic clusters. For instance, in the word *crispy*, the parse /krɪ/+/spi/ would give an ill-formed syllable; the parse /krɪs/+/pi/ gives the rhythm of a light initial syllable, and wrongly predicts aspiration for the plosive; the parse /krɪs/+/spi/ also gives the rhythm of a light initial syllable, which is wrong; the parse with maximal

ambisyllabicity, /krɪsp/+/spi/, gives the rhythm of an initial heavy syllable and non-aspiration of the plosive.

We interpret certain kinds of morphological and word joins as non-ambisyllabic, in particular joins between two Germanic (level 2) morphemes, where long consonants may occur (as in *soul•less*) and between two words (as in *free • lice*, *feel • ice*, *feel • lice*).

## 3. TEMPORAL INTERPRETATION.

ProSynth's model of Temporal Interpretation is based on [2, 4, 5, 6, 7, 8]. Temporal interpretation is sensitive to the linguistic structural information set out above. There are two ways in which rhythm can be controlled in ProSynth: syllable overlay, which is the phonetic interpretation of ambisyllabicity, and temporal compression. The primary timing unit in ProSynth is the syllable. We begin by presenting its temporal interpretation.

### 3.1. Temporal interpretation of Syllables

The temporal interpretation of Syllables is carried out by satisfying a set of temporal constraints (Fig. 3), such that the temporal extent of the Syllable is the same as that for the Head of the Syllable, the Rhyme, and the Head of the Rhyme, the Nucleus. Onsets and Codas are 'overlaid' onto syllables. Interpreting consonant constituents so that they are overlaid onto Nuclei produces the effects of coarticulation.

|  |
|--|
| Syllable Start = Rhyme Start = Nucleus Start |
| Syllable End = Rhyme End = Nucleus End       |
| Onset Start = Syllable Start                 |
| Coda End = Syllable End                      |

Fig. 3. Constraints on the temporal interpretation of Syllables.

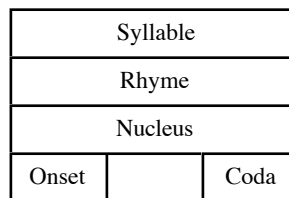


Fig. 4. Temporal structure of Syllables.

### 3.2. Overlay.

The phonetic interpretation of syllable joins is *overlay*. Overlay is a statement of the temporal relationship between the end of one syllable and the start of the next. The temporal relation between adjacent Syllables can be expressed as:

$$\text{Syllable}_n \text{ Start} = \text{Syllable}_{n-1} \text{ End} - \text{Overlay}$$

A low value for Overlay results in a comparatively short proportion of the resulting acoustic segment being attributable to the Coda of Syllable<sub>n-1</sub>. A high value for Overlay means that less of the Coda is 'masked', and produces an intervocalic consonant with longer duration: see Fig. 5.

Fig. 5 shows the temporal interpretation of *tieless* and *tileless*. In *tieless*, /l/ is [+ambisyllabic], whereas in *tileless* (which has a morphological boundary), /l/ is [-ambisyllabic]. Overlay is sensitive to the constituency of both Onsets and

Codas, and to ambisyllabicity, and this structural difference is reflected in the amount of Syllable Overlap: it is smaller for the [+ambisyllabic] case, and larger for the [-ambisyllabic] case. In *tieless*, less of the exponents of the Coda are audible than in *tileless*. The result is, in the case of *tieless*, a shorter, clearer lateral, and in the case of *tileless* a longer, darker lateral.

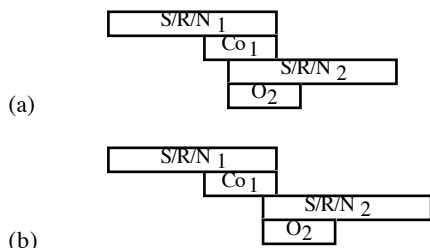


Fig. 5 (a) Overlay of the syllables /taɪl/+/lɔs/ [+ambisyllabic], *tieless* and (b) of the syllables /taɪl/+/lɔs/ [-ambisyllabic], *tileless*.

Making reference to ambisyllabicity, it is possible to lengthen or shorten intervocalic consonants in relation to linguistic structure. There are morphologically bound differences which can be modelled in this way, provided that the phonological structure is sensitive to them. For instance, the Latin prefix *in-* is [+ambisyllabic] and is fully overlaid with the stem to which it attaches, giving a short nasal in *innocuous*, while the Germanic prefix *un-* is [-ambisyllabic] and is not overlaid to the same degree, giving a long nasal, as in *unknowing*.

### 3.3. Temporal compression.

Our model of temporal compression allows the statement of relationships between syllables at different places in metrical structure, using a knowledge database. For instance, the syllable /man/ in the words *man*, *manage*, *manager* and in the utterance *She's a bank manager* all have different degrees of temporal compression which are related to the metrical structure as a whole.

Monosyllabic utterances are not compressed, and therefore have a compression factor of 1. Syllables in other contexts are temporally compressed as a proportion of their interpretation as monosyllabic isolates. *Distance* is an expression of the temporal separation between the end of an Onset and the start of a tautosyllabic Coda, and is used to calculate the temporal compression factor for any given syllable; it relates to the notion of "minimum duration" for the phonetic interpretation of vowels in segmental TTS systems.

Temporal compression applies across the whole syllable: Onset, Rhyme (and the daughters of the Rhyme) are all compressed. Plosives and affricates are compressed less than other constituents. Part of our ongoing work is to model compression so that it is sensitive to individual features and their attributes, such as [±continuant] and [±voice]. We envisage in future work compressing constituents within a given proportion of their range of compressibility.

Temporal compression is sensitive to structural information at a number of levels in the Prosodic Hierarchy. Relevant factors are:

- Nucleus: short or long; diphthong or monophthong; /aɪ ɔɪ aʊ/ vs. other diphthongs.
- Coda: simple or branching.
- Rhyme: heavy or light; voiced or voiceless
- Syllable strength: strong or weak
- Position in Foot: initial, medial, or final
- The weight and strength of adjacent Syllable(s)

## 4. CURRENT WORK

The details of the temporal model are being constructed from work on a database of recorded speech, produced by a single speaker of Southern British English. The database has been designed to exemplify a subset of possible prosodic structures, including Feet which differ according to the weight of the Head Syllable. The data have been recorded and labelled, and for each speech file, an XML [10] file generated containing an XML representation of the utterance including a phonological parse, and a set of acoustic-phonetic labels. A search algorithm has been developed which makes it possible to conduct structure-driven searches through the database of natural speech, and to make enquiries about segment durations in relation to structure. For instance, it is possible to search through the database and find all instances of the vowel /ɔ/ in the second syllable of a disyllabic foot where the first Syllable is light and the second syllable contains a voiceless plosive in the Onset and no Coda, as in the word *supper*. One of our long-term goals is to use the XML files for the database to produce by rule synthetic speech whose temporal properties approximate those of the original natural-speech data.

## 5. PERCEPTUAL TESTS.

### 5.1. Hypothesis

The hypothesis we are testing in ProSynth is that having hierarchically organised, prosodically structured linguistic information should make it possible to produce more natural-sounding synthetic speech which is also more robust under difficult listening conditions. As an initial test of our hypotheses about temporal structure and its relation to prosodic structure, an experiment has been conducted to test whether the categories set out in Section 2 make a significant difference to listeners' ability to interpret synthetic speech. If the timings predicted by ProSynth for structural positions are perceptually important, listeners should be more successful at interpreting synthetic speech when the timing appropriate for structure is used than in the case where the timing is inappropriate for the linguistic structures set up.

The data consists of phrases from the database of natural English generated by MBROLA [11] synthesis using timings of two different kinds: (1) the segment durations predicted by the ProSynth model taking into account all the linguistic structure outlined in Section 2 (2) the segment durations predicted by ProSynth for a different linguistic structure. If the linguistic structure makes no significant linguistic difference, then (1) and (2) should be perceived equally well (or badly). If temporal interpretation is sensitive to linguistic structure in the way that we have suggested, then the results for (1) should be better than the results for (2).

## 5.2. Data

12 groups of structures to be compared on structural linguistic grounds were established (eg "light ambisyllabic short initial syllable" versus "light nonambisyllabic short initial syllable"). Each group has two members (eg *robber/rob them* and *loving/loveless*). For each phrase, two synthetic stimuli were generated: one with the predicted ProSynth timings for that structure, and another one with the timings for the other member of the pair. Files were produced with timing information from the natural-speech utterances, and an approximation to f0 of the speech in the database. The timing information for the final foot was then replaced with timing from the ProSynth model. This produced the 'correct' timings. In order to produce the 'broken' timings, timing information for the rhyme of the strong syllable in this final foot was swapped within the group so, for example the durations for *ob* in *robber* were replaced with the durations for *ob* in *rob them* and vice versa.

The stimuli have segment labels ultimately from the label files from the database, f0 information from the recordings in the database, and timing information partly from natural speech and partly from the ProSynth model.

As an example, consider the pair (*he's a*) *robber* and (*to*) *rob them*. The durations (in ms.) for *robber* and *rob them* are:

|   |     |   |     |
|---|-----|---|-----|
| ɒ | 120 | ɒ | 110 |
| b | 65  | b | 85  |
| ə | 150 | ð | 60  |
|   |     | ə | 120 |
|   |     | m | 135 |

Stimuli with these durations are compared with stimuli with the durations swapped round:

|   |     |   |     |
|---|-----|---|-----|
| ɒ | 110 | ɒ | 120 |
| b | 85  | b | 65  |
| ə | 150 | ð | 60  |
|   |     | ə | 120 |
|   |     | m | 135 |

## 5.3. Experimental design.

22 subjects heard every phrase once at comfortable listening levels over headphones, presented by a Tucker-Davies DD1 digital analogue interface. The signal-to-noise ratio was -5dB. The noise was cafeteria noise, i.e. different background noises like voices and laughter. Subjects were instructed to transcribe what they heard using normal English spelling, and were given as much time as they needed. When they were ready, they pressed a key and the next stimulus was played.

Each subject heard half of the phrases as generated with the ProSynth model, and half with the timings switched. The subjects heard six practice items before hearing the test items, but were not informed of this.

## 5.4. Results

The phoneme recognition rate for the correct timings from the ProSynth model is 77.5%, and for the switched timings it is 74.2%. Although this is only a small improvement, it is

nevertheless significant using a one-tailed correlated t-test ( $t(21) = 2.21, p < 0.02$ ).

Examples of the stimuli and further details of the results of the experiments (including updates) are available on the world wide web [12].

## 5.5. Discussion

The results show a significant effect of linguistic structure on improved intelligibility. The results are for the whole phrase, including parts which were not switched round: excluding these may result in improved results. The MBROLA diphone synthesis models durational effects, but not the segmental effects predicted by our model and described in more detail in Section 3: for example, the synthesis produces aspirated plosives in words like *roast*[<sup>h</sup>]*ing* where our model predicts non-aspiration. It uses only a small diphone database. The rather low phoneme recognition rates may be due to the quality of the synthesis was problematic, or the cognitive load imposed by high levels of background noise. Further statistical analysis will group the data according to foot-type, and future experiments will use a formant synthesiser.

## 5.6. Future work

Future work will concentrate on refining the temporal model so that it generates durations which approximate those of our natural speech model as well as possible. The work will be checked by more perceptual experiments, including presenting the synthetic stimuli under listening conditions that impose a high cognitive load, such as having the subjects perform some other task while listening to synthesis.

## ACKNOWLEDGEMENTS

We are grateful to acknowledge financial support from the Engineering and Physics Research Council of the United Kingdom, Grant Number GR/L51829. We acknowledge the work of Sarah Hawkins and Sebastian Heid at the University of Cambridge in helping to set up and conducting the perceptual tests.

## REFERENCES

- [1] Abercrombie, D (1964): Syllable quantity and enclitics in English. In D Abercrombie et al. (eds.) *In Honour of Daniel Jones*. London: Longman, 216-222.
- [2] Coleman, John (1992): The phonetic interpretation of headed phonological structures containing overlapping constituents. *Phonology Yearbook* 9 (1), 1-44.
- [3] Hawkins, S, J House, M Huckvale, J Local, R Ogden (1998): ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia. 1707-1710. Available on CD-ROM: ICSLP-98 Paper #538.
- [4] Local, J. K. (1992): Modelling assimilation in a non-segmental rule-free phonology. In G J Docherty and D R Ladd (eds): *Papers in Laboratory Phonology II*. Cambridge: CUP, 190-223.
- [5] Local, J. K. (1993): The 'Segmental' intelligibility of the YorkTalk non-segmental speech synthesis system. *York Research Papers in Linguistics*. YRPL 93-01.
- [6] Local, J. K. & R. Ogden (1997): A model of timing for nonsegmental phonological structure. In Jan P. H. van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (eds.) *Progress in Speech Synthesis*. Springer, N. Y. 109-122.
- [7] Ogden R (1992): Parametric Interpretation in YorkTalk. *York Papers in Linguistics* 16, 81-99.
- [8] Ogden, Richard & John Local (1996): YorkTalk Annual Report 1991. *York Research Papers in Linguistics* YRPL 96-01.
- [9] Ohman, S E G (1966): Coarticulation in CVC utterances: spectrographic measurements. *JASA* 39, 151-168.
- [10] <http://www.w3.org/TR/1998/REC-xml-19980210>
- [11] <http://tcts.fpms.ac.be/synthesis/mbrola.html>
- [12] <http://www-users.york.ac.uk/~lang19>