

All-Prosodic Speech Synthesis

Arthur Dirksen
John S. Coleman

ABSTRACT We present a speech synthesis architecture, IPOX, which allows the integration of various aspects of prosodic structure at different structural levels. This is achieved by using a hierarchical, metrical representation of the input string in analysis as well as phonetic interpretation. The output of the latter step consists of parameters for the Klatt synthesizer. The architecture is based primarily on YorkTalk [Col92, Col94, Loc92], but differs in that it uses a rule compiler [Dir93], which allows a clean separation of linguistic statements and computational execution as well as a more concise statement of various kinds of generalizations.

8.1 Introduction

A major problem in speech synthesis is the integration of various aspects of prosodic structure at different structural levels. We present an architecture in which this problem is addressed in a linguistically sophisticated manner. Our system, IPOX, is based on the idea that it is possible to generate connected, rhythmically appropriate speech from a hierarchically structured representation, a prosodic tree. This metrical representation is assigned by parsing an input string using declarative, constraint-based grammars with a standard parsing algorithm. Each node in the metrical representation is then assigned a temporal domain within which its phonetic exponents are evaluated. This evaluation is done in a top-down fashion, allowing lower-level prosodic constituents to modify the exponents of higher-level nodes. The phonetic exponents of adjacent nodes in the metrical tree are allowed to overlap with one another. Also, the order in which constituents are evaluated depends on the prosodic configuration in which they appear. Within the syllable, heads are evaluated before nonheads, allowing metrically weak constituents such as onset and coda to adapt to their strong sister constituents (rime and nucleus, respectively) with which they overlap. Across syllables, the order of interpretation is left-to-right, so that each syllable is "glued" to the previous one. After all phonetic exponents have been evaluated, a parameter file for the Klatt formant synthesizer is generated.

The architecture of IPOX is rather similar to that of YorkTalk [Col92, Col94, Loc92], on which it is based, but is different in a number of respects:

- YorkTalk representations are implemented as arbitrary Prolog terms. In IPOX, metrical structure is made explicit in the representation [DQ93]. This has made it possible to define general algorithms to process these structures in various ways, whereas in YorkTalk such algorithms are spelled out on a case-by-case basis.
- The YorkTalk morphological and phonotactic parser is a Prolog DCG (definite clause grammar). IPOX, on the other hand, uses a rule compiler, which forces the developer to keep linguistic rules separate from the control logic, which is fixed.
- IPOX includes a facility to state feature co-occurrence restrictions separately from the phrase structure rules [Dir93].

More generally, IPOX aims to further formalize and extend the YorkTalk architecture, making it more flexible and easier to adapt to different languages, which is one of our long-term goals. Also, IPOX integrates all functions, including synthesis and sound output, in a single executable file (which runs under Windows on a PC with a standard 16-bit sound card), using graphics to display analysis trees, temporal structure, phonetic interpretation, and audio output waveforms. However, the system is still under development. Currently, there is no interface between morphosyntactic structure and phrase-level prosodic structure, although grammars for each of these modules have been developed separately. As a consequence, speech output temporarily suffers from certain linguistic limitations. We call this architecture “all-prosodic” because the phonological approach is based on metrical theory, making extensive use of distinctive features at nonterminal nodes, and the approach to phonetic interpretation is based on phonetic parameters computed in parallel rather than a sequence of concatenative units.

In this chapter, we discuss the various components of IPOX, illustrated with examples from British English that have been generated using rule sets adapted from (an earlier version of) YorkTalk.¹ First, in section 8.2, we present the basic architecture of IPOX, illustrated with the analysis and generation of isolated monosyllables. Section 8.3 discusses the use of overlap and compression in the generation of polysyllabic utterances and demonstrates how various kinds of vowel reduction can be obtained by varying the speech rhythm. Section 8.4 discusses the generation of connected speech, illustrated with a detailed consideration of the sentence “*This is an adaptable system.*”

¹All examples discussed in this chapter are provided on the CD-ROM (see Appendix) in the form of audio files generated by IPOX.

8.2 Architecture

8.2.1 Analysis

The analysis component of IPOX uses declarative, constraint-based phrase structure grammars to analyze input text and assign a metrical-prosodic representation. A declarative grammar is one in which the rules define the well-formedness of grammatical representations without specifying procedures for constructing such representations. In this section, we discuss how the internal structure of English syllables is defined in terms of such a grammar.

Basic syllable structure is assigned by the following two—presumably universal—rules:

```
syl --> (onset / rime).
rime --> (nucleus \ coda).
```

In these rules, the slash is used to encode which of two nodes is the prosodic head: the forward slash (/) indicates the pattern *weak-strong*, the backward slash (\) indicates *strong-weak*. Further (often language-specific) rules are used to define the internal structure of onset, nucleus, and coda (including the possibility that onset or coda may remain empty).

One of the properties that must be specified by the syllable grammar is syllable weight, which is an important factor in the distribution of stressed versus unstressed syllables in quantity-sensitive stress systems such as that of English. For example, the following rule forces heavy syllables (and light syllables that precede a light syllable) to be the head of a foot:

```
foot --> (syl \ syl:[-heavy]).
```

As is well known, syllable weight is usually determined only by the internal structure of the rime, disregarding the onset. For English, we assume that a syllable is heavy if the nucleus or the coda branches.² Thus, we might annotate the above rules for syllable and rime as follows (capital letters are used to indicate shared variables within a single rule):

```
syl:[heavy=A] --> (onset / rime:[heavy=A]).
rime:[heavy=A] --> (nucleus:[branching=A] \ coda:[branching=A]).
```

This works if we also write rules that assign the feature specification [+branching] to branching nuclei and codas,³ while leaving nonbranching nuclei and codas unspecified for this feature. For example:

```
coda:[+branching] --> (cons \ cons).
coda --> cons.
```

²This particular formulation hinges on our assumption of maximal ambisyllabicity (subsection 8.3.1).

³Note that the name of a feature does not by itself imply any interpretation whatsoever and serves only a mnemonic purpose.

If we set up the grammar this way, light syllables remain unspecified for the feature *heavy*, and are accepted anywhere. Heavy syllables, on the other hand are specified as [+heavy]. They cannot appear as the weak node of a foot, as this would involve conflicting feature specifications, a situation not allowed in a declarative system.

A better approach, however, is to encode feature co-occurrence restrictions separately by means of *templates*, which are general descriptions of phrase structure rules. Advantages include the following: linguistic universals and language-specific rules are separated better, generalizations across rules can be stated just once, and readability of grammars is improved.

In the present example, we need the following templates:

```
[heavy=A] --> ([ ] / [heavy=A]).
[heavy=A] --> [branching=A], [branching=A].
[+branching] --> [ ], [ ].
```

The rule compiler applies every template to every rule with which it unifies. The first template is applied to every phrase structure rule that introduces the *weak-strong* pattern. The second and third templates are applied to every phrase structure rule that introduces a binary-branching structure, whether *weak-strong* or *strong-weak*, because there is no slash in the template. However, the feature unifications specified in templates are instantiated only to the extent that categories are defined for the relevant features. So, if only the categories *syl* and *rime* are defined for the feature *heavy*, and the feature *branching* is limited to nucleus and coda, the above templates derive the same effect as the phrase structure rule annotations they replace.

The syllable grammar also defines “spreading” of vocalic place features, which is how we model allophonic variation due to coarticulation. In our grammar, both vowels and consonants are defined for vocalic place features, which are complex feature structures of the following form:

```
voc:[+/-grv, +/-rnd, height=close/mid/open]
```

The idea is that *voc* features encode the primary articulation of vowels and glides, and the secondary articulation of consonants. The term “V-place” has recently been proposed by some other phonologists for roughly the same purposes. Vowels and glides are inherently specified for *voc* features, whereas nasals and obstruents are not. Liquids have inherent specifications as well; however, /l/ is “clear” (i.e., [-grv]) in onset position; “dark” (i.e., [+grv]) in coda position; and /r/ is unspecified for rounding.

To the extent that *voc* features are unspecified, a value is obtained for each feature through spreading. In YorkTalk, spreading of *voc* features is defined in the phrase structure rules on a case-by-case basis. In IPOX, this is done by means of templates, such that *voc* features appear to spread from *strong* nodes (where they

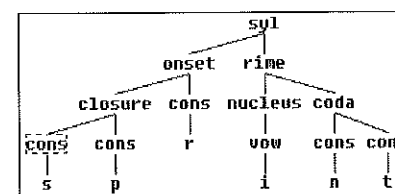


FIGURE 8.1. Syllable structure of /sprint/.

originate) to *weak* nodes.⁴ As an example, consider onset /s/ in *sprint*, analyzed as shown in figure 8.1.

Having no inherent specification for *voc* features, /s/ (or, rather, the onset constituent that directly dominates /s/) must obtain values for these features through spreading. Since /r/ is unspecified for the feature *rnd*, /s/ obtains the value [-rnd] in accordance with /i/, just as it would in *spit*. However, because prevocalic /r/ is dark [OGC93, p. 216], /r/ is specified [+grv], which is shared with /s/. With respect to this feature, /s/ in *sprint* is more like /s/ in *spot*. As a result, the *voc* features associated with /s/ in *sprint*, *spit*, and *spot* are all different. Phonetic interpretation is sensitive to these differences, which are reflected in the output waveforms as subtly but appropriately different frication spectra associated with /s/. (See Demo 1 on the CD-ROM.)

8.2.2 Phonetic Interpretation 1: Temporal Interpretation

At the next stage, the metrical-prosodic representation is assigned a temporal interpretation. As in YorkTalk, constituents are allowed to overlap with one another in various ways. Unlike YorkTalk, temporal interpretation in IPOX is determined by the compiler, and the user is prevented from incorporating ad hoc temporal interpretation constraints. All the developer can and must do is provide statements for durations assigned to constituents, which can be treated as terminals as far as phonetic interpretation is concerned, and the extent of overlap between adjacent nodes in the metrical-prosodic representation.

These statements are interpreted differently depending on whether a constituent occurs within a syllable. Across syllables, a *left-to-right strategy* is used, in which the duration of a constituent is the sum of the durations of each of the individual syllables within that constituent minus the amounts of overlap specified at the boundaries between those syllables. Within syllables, a *co-production strategy* is used, in which the duration of a constituent equals the duration assigned to the prosodic head of that constituent, and weak nodes are overlaid on their strong sister nodes at the left or right edge, depending on whether the configuration is *weak-strong* or *strong-weak*, respectively.

⁴In actuality, we are dealing with nondirectional feature sharing. However, if a feature is inherently specified on a vowel, and is shared with a consonant, then it “appears” to have been spread from the vowel onto the consonant.

When the *coproduction strategy* is used, the following conventions are observed:

1. If no duration is specified for a weak node, the duration is equal to the amount of overlap between the weak node and its strong sister node.
2. Alternatively, if no overlap is specified between a weak and a strong node, the amount of overlap equals the duration of the weak node.
3. A negative overlap quantity specifies the amount of nonoverlap (i.e., separation) between a weak node and its strong sister node.

As an example of a duration rule, the following statement assigns a duration of 680 ms to a nucleus marked as [+long] (a long vowel or a diphthong):

```
nucleus: [+long] => 680.
```

In the phonology, long vowels and diphthongs are modeled as branching nuclei, dominating both a vowel and an off-glide constituent (e.g., /iy/ for a front high long vowel; /ay/ for a front low-to-high diphthong). The above statement, however, effectively makes the nucleus a terminal element of the structure as far as phonetic interpretation is concerned. In other words, the phonetic interpretation of long vowels and diphthongs is holistic rather than compositional, even though they are assigned internal structure in the phonological analysis.

As an example of a rule for overlap, the following statement fixes the overlap between onset and rime at 200 ms:

```
onset + rime => 200.
```

Because the metrical relation between onset and rime is *weak-strong*, the onset is overlaid on the rime for the first 200 ms.

The various possibilities offered by the coproduction model are exemplified by the temporal interpretation of the syllable *sprint*, shown in the screen clip in figure 8.2.

As can be seen, consonants in the onset are temporally interpreted rather differently from coda consonants. Specifically, onset constituents do not have inherent durations: their durations follow from statements about nonoverlap between sister constituents within an onset. Coda constituents do have inherent durations, as well as statements about nonoverlap between sister constituents within a coda.

The phonetic exponents of a constituent are calculated relative to the start time and duration of the constituent. This does not mean, however, that these exponents are confined to the space allocated by temporal interpretation. Thus, durations of constituents cannot be equated with concatenative segment durations, nor can

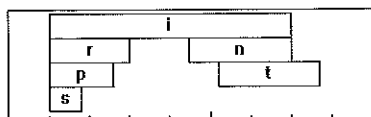


FIGURE 8.2. Temporal interpretation of /sprint/.

the boundaries in figure 8.2 be understood as boundaries between segments in the orthodox sense. On the other hand, the relation between perceptual segment durations and temporal interpretation is not totally arbitrary in actual practice, even if it is in principle. For example, the “distance” between onset and coda in a syllable is usually a fairly good measure of the perceptual duration of the vowel with which they are coproduced (see chapter 9 of this volume).

8.2.3 Phonetic Interpretation 2: Parametric Interpretation

Temporal interpretation is only one aspect of phonetic interpretation. In addition, values for the synthesizer parameters must also be determined in order to provide a complete phonetic interpretation of the abstract phonological structures we initially computed. Parametric phonetic interpretation is done by evaluation of a set of phonetic exponency rules, which have the following general form:

```
Category:Features => Exponents.
```

These rules are evaluated for each node in the metrical-prosodic representation. The nodes of the tree are visited in a *top-down* fashion. In this way, holistic properties of a constituent are computed first, to be worked out in finer detail by lower-level constituents. As an example, consider the generation of F_0 contours for metrical feet within a single phrase. The phonetic exponency of a foot for this parameter is a simple linear high-to-low fall. At lower levels of prosodic structure, individual consonants and vowels contribute small details to the complete F_0 specification (e.g., so-called consonantal perturbations), similar to the F_0 algorithm briefly discussed in [PB88, pp. 176–177].

Across syllables, the order of interpretation is *left-to-right*. The motivation for this is that in connected speech each syllable needs to be “glued” to the previous one (see also subsection 8.3.1). Within syllables, however, the order of interpretation is *head-first*. In this way, weak nodes may adapt their parameters to those of their strong sister nodes because the parameters of the strong sister node (the head) are already known. The phonetic motivation for this is that consonants coarticulate with vowels, either directly or indirectly through other consonants, but not vice versa.

The results of evaluating phonetic exponency rules are added to the *phonetic exponency database*. Each entry in the database has the following general form:

```
<Parameter, Time1, Time2, Value1, Value2>
```

As an example, consider the following exponency rule for the second formant F_2 for fricatives in onset position:

```
cons: [-coda, -son, +cvt] =>
A = END,
B = F2_END,
C = F2_VAL,
D = F2_LOCUS,
E = F2_COART,
F = f2(0.2*A),
```

$$G = f2(A+(B*A)),$$

$$f2(0.2*A, 0.5*A, 0.9*A, A, A+(B*A)) = (F, C, C, D+E*(G-D), G).$$

In this rule, the feature specification $cons: [-coda, -son, +cnt]$ specifies a fricative consonant is dominated by an onset constituent. The $[-coda]$ feature is inherited from the onset node dominating $cons$. The letters A to G are variables for times and parameter values, which are local to the rule. The built-in macro END returns the duration of the current constituent. F2_END, F2_VAL, F2_LOCUS, and F2_COART are calls to user-defined lookup tables; the current constituent is passed as an argument to the call. $f2(0.2*A)$ and $f2(A+(B*A))$ query the phonetic exponency database for a value for F2 at a particular point in time (at 20 percent of the duration and at slightly more than 100 percent of the duration, respectively); if such a value cannot be inferred from the database, the system returns a default value. $f2(0.2*A, 0.5*A, 0.9*A, A, A+(B*A))$ specifies the points in time between which values for F2 are to be interpolated, and $= (F, C, C, D+E*(G-D), G)$ specifies the corresponding values for F2. The term $D+E*(G-D)$ implements a locus equation of the form $Locus + Coart * (Vowel - Locus)$ (see [AHK87, pp. 113–115]).

The above rule implements the general structure of F2-exponents for fricatives in onset position. Lookup table entries are used to fill in the specifics of different types of fricatives in various contexts. For example, in the case of /s/ the lookup table for F2_VAL returns a value of 1700 or 1400, depending on whether the voc features define a spread or a rounded context, respectively.

By way of illustration, figure 8.3 shows the application of the above rule in different contexts. Figure 8.4 shows the formant structure and voicing for the syllable *sprint*.

8.2.4 Parameter Generation and Synthesis

If figure 8.4 seems messy, this is because all phonetic exponents of a structure are shown simultaneously. However, the phonetic exponency database is more than just a mixed bag of parameter tracks, as it also takes into account the way in which some constituents are overlaid on other constituents. Phonetic exponency in IPOX, as in YorkTalk, is very much like paint applied to a canvas in layers, obscuring

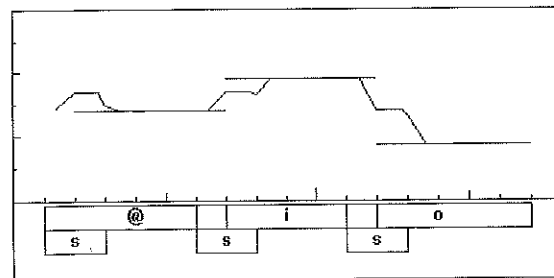


FIGURE 8.3. Phonetic exponency of /s/ for F2.

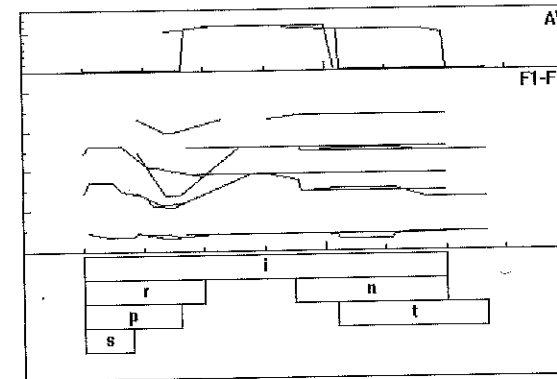


FIGURE 8.4. Phonetic interpretation of *sprint*.

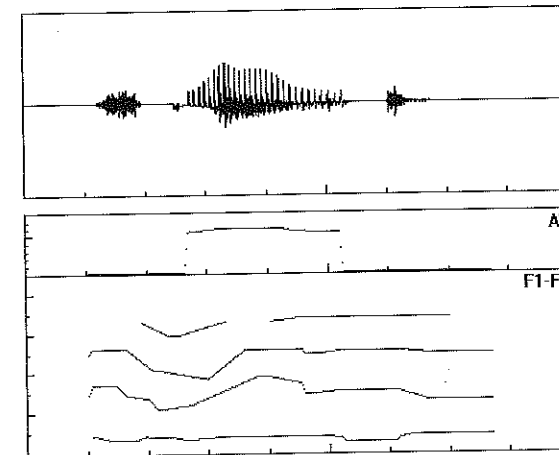


FIGURE 8.5. Parameter generation and synthesis of *sprint*.

earlier layers where it is applied and blending in with the background at the edges. Thus, it is possible to take the top-most value for a parameter at arbitrary points in time (such as every 5 ms), and use this value as input to the synthesizer. If the exponency rules have done their job well, the result is a continuous set of parameter tracks, as shown in figure 8.5, which also shows the waveform generated by the Klatt synthesizer.

Note that in our system it is not necessary to control each synthesis parameter at every point in time. For example, in figures 8.4 and 8.5 it can be seen that the fourth formant F4 is actively controlled only for onset /r/ and the coda nasal, and simply left at its default value (not shown) elsewhere. In fact, the default in IPOX is to do nothing. For this reason, and because of the declarative nature of the system, each exponency rule substantiates an independent empirical claim about the phonetic realization of a piece of phonological structure.

In a similar fashion, the lookup tables used by the exponency rules are sparse tables: a distinction made in one case (e.g., front versus back, spread versus round) need not be made in all cases. Thus, it is feasible to supply phonetic detail as seems necessary, or make sweeping generalizations as seems appropriate.

Also, the use of a specific synthesizer restricts the exponency rules to a specific set of parameters and the range of admissible values for each parameter. The system as a whole, however, is not dependent on a particular synthesizer configuration, and could easily be adapted to any synthesizer that accepts a parameter file as input, including an articulatory synthesizer or hybrid approach between acoustic and articulation-based synthesis (see, e.g., chapter 16 of this volume).

8.3 Polysyllabic Words

Phonetic interpretation in IPOX and YorkTalk is *compositional*, that is, the interpretation of a constituent is a function of the interpretation of its parts and the way in which they are combined. For example, the phonetic interpretation of a syllable consists of the phonetic interpretation of the rime, combined with the phonetic interpretation of the onset that is overlaid on the rime. Thus, within a syllable the mode of combination is head-first, as dictated by metrical structure, using the coproduction model of temporal interpretation. Across syllables, the mode of combination is slightly different (left-to-right, such that the onset is overlaid on the coda of the previous syllable). However, this does not change the fact that phonetic interpretation is compositional across syllables as well. That is, the phonetic interpretation of a polysyllabic utterance is simply the combination of any number of syllables, each of which is no different from its isolation form. On the face of it, this point of view seems problematic in that it raises the questions of how to deal with the special properties of intervocalic consonant clusters as well as how to handle vowel reduction and speech rhythm.

If we are to maintain that phonetic interpretation is compositional, the answer to these two questions must lie in the way in which syllables are combined to form utterances. The next subsections discuss our solutions to these problems.

8.3.1 Ambisyllabicity

One of the problems raised by compositionality is how to make sure that a single intervocalic consonant is properly coarticulated with its neighboring vowels. For example, in a word such as *bottle* /bot@l/ the /t/ should be interpreted as a coda with respect to the preceeding vowel, and as an onset with respect to the following vowel. Thus, the intervocalic /t/ must be parsed twice, once as the coda of the first syllable, once as the onset of the second syllable. In other words, the /t/ must be made *ambisyllabic*. This way we derive the coarticulation of /t/ with the preceding vowel /o/, just as in /bot/, as well as with the following vowel /@/, just as in /t@l/. However, without doing anything else, the resulting speech is simply the

concatenation of two syllables, separated by a short pause. This type of juncture is appropriate if the two syllables belong to different prosodic phrases, but not word-internally. To avoid this, the onset /t/ is overlaid on the coda /t/ in temporal interpretation just enough to create the effect of a single intervocalic /t/.

More generally, different amounts of overlap are used in different contexts, depending on the nature of the intervocalic cluster as well as the prosodic configuration in which it appears [Col95]. In the case of two identical stops, different possibilities for overlap between syllables can be visualized as follows (see Demo 2 on the CD-ROM):

- very short closure

```
--Vowel-|-Closing-|---Closure---|-Release-|---
                    -Closure-|-Release-|-Vowel--
```

- normal intervocalic closure

```
--Vowel-|-Closing-|---Closure---|-Release-|---
                    -Closure-|-Release-|-Vowel--
```

- long closure

```
--Vowel-|-Closing-|---Closure---|-Release-|---
                    -Closure-|-Release-|-Vowel--
```

In IPOX and YorkTalk, ambisyllabicity is not restricted to single intervocalic consonants (as in most phonological theories). Within (Latin parts of) words, intervocalic clusters are parsed with *maximal ambisyllabicity*. Compare: /win[t]@r/, /a[sp]rin/ and /si[st]@m/ (see Demo 3 on the CD-ROM).

By parsing the bracketed clusters as ambisyllabic, we derive the fact that in each of these cases the first syllable is heavy, and that the /t/ in /wint@r/ is aspirated, whereas in /sist@m/ it is not aspirated, as it occurs in the onset cluster /st/.

Various kinds of assimilation occurring in intervocalic clusters must be dealt with as well. We will not discuss assimilation in any detail here, but briefly sketch a general approach (for a detailed analysis see [Loc92]).

A declarative system does not allow feature-changing rules, as would be employed in traditional generative analyses of assimilation phenomena. However, it is possible to use underspecification in the analysis component, combined with feature-filling rules to further instantiate underspecified structures. As an example, consider the following rule for determining the amount of overlap between syllables with an ambisyllabic voiceless stop (in this rule, the value of *cns* is a complex feature structure defining consonantal place of articulation):

```
coda:[-voi, -son, -cnt, cns=A] +
onset:[-voi, -son, -cnt, cns=A] => ...
```

If both coda and onset are fully specified for *cns* features, this rule merely checks for equality, and assigns the appropriate amount of overlap (indicated by . . .). However, if either the coda or the onset is underspecified for *cns* features, these features become shared, effectively assimilating the coda to the following onset or vice versa. Also, the cluster receives the temporal interpretation of a single ambisyllabic consonant.

In actual practice, though, cases of full assimilation are fairly rare, and a slightly more subtle approach is needed, in which the feature structures of assimilated consonants are similar but not identical, and in which the temporal interpretation of an assimilated cluster is slightly different from ambisyllabic consonants. Also, it will be necessary to use different assimilation rules for different morphophonological and prosodic contexts.

8.3.2 Vowel Reduction and Speech Rhythm

Connecting syllables as discussed in the previous subsection helps to smooth the boundaries between them. It does not, however, produce a very satisfying result if all syllables in a word or utterance have the same durations as their isolated monosyllabic counterparts. Therefore, to model the rhythm of connected speech, syllables that appear in a polysyllabic utterance are assigned a “compression factor,” depending on the position of a syllable in a metrical foot, and the internal makeup of a syllable.

This compression factor is taken into account during temporal interpretation, such that various constituents within a syllable are compressed accordingly. In parametric interpretation, a distinction is made between absolute and relative timing. For example, fricatives and sonorants are relatively timed (i.e., all references to points in time are expressed as proportions of the duration of a constituent), but stop consonants use absolute timing. Because of this, a stop consonant in a compressed syllable occupies a larger proportion of the syllable than in a syllable that is not compressed. In other words, compression is not linear.

Also, rules for overlap between constituents take into account compression with one major exception: Overlap between onset and rime is always 200 ms (this is an arbitrary constant), even if constituents within the onset are compressed. Thus, the notional start of the vowel is always located 200 ms after the start of the syllable. As a consequence, vowels are more sensitive to compression than their neighboring consonants.

The effect of syllable compression is illustrated in figure 8.6 for the second syllable of *bottle*: the left panel shows the temporal and phonetic interpretation of this syllable without compression; the right panel shows the same syllable compressed to 62 percent of its duration.

Because compression is not linear, the two pictures are strikingly different. In the compressed version, the onset overlaps a much greater portion of the rime, and the vowel is almost totally eclipsed. Also, as a result of compression, the formant transitions are qualitatively different. This can be seen rather clearly in the transitions of the second formant F2. In both versions the coda /l/ is coarticulated

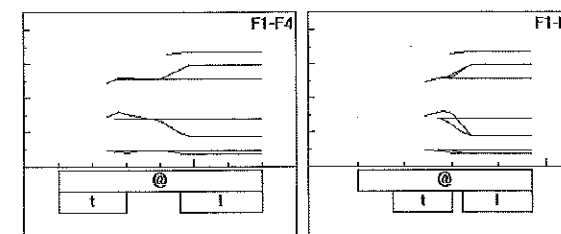


FIGURE 8.6. /t@l/ without (left) and with (right) compression.

with the vowel. However, in the compressed version the onset /t/ is coarticulated with the coda /l/ and only indirectly with the vowel. Perceptually, this creates the effect of a *syllabic sonorant* (see Demo 4 on the CD-ROM).

It is important to note that although the vowel appears to have been deleted, its presence in the background is still notable in the formant transitions, which would have been different if we had selected another vowel. Thus, the question is raised as to whether an analysis of vowel reduction/deletion in terms of syllable compression can be generalized to full vowels as well. We have preliminary evidence that this is the case.

As an example, consider the unstressed prefix *sup* in words such as *suppose* and *support*. In received pronunciations of these words, it appears that the vowel has been deleted. However, we have synthesized successful imitations of these words, using the full vowel /ʌ/ (as in *but*). If the first syllable in *suppose* is compressed to 60 percent, then the vowel obtains a slightly reduced quality. With compression to 52 percent, the vowel is eclipsed, resulting in the pronunciation *s'ppose* (see Demo 5 on the CD-ROM).

A segmental analysis of vowel elision would seem to predict that a version of the word *support* in which the first vowel is deleted is phonetically similar, if not identical, to the word *sport*. By contrast, an analysis in terms of compression predicts subtle, but notable differences in temporal as well as spectral structure. Specifically, it is correctly predicted that /p/ is aspirated in *s'pport*, but not in *sport* (see Demo 6 on the CD-ROM).

Even more challenging are the apparent changes in vowel quality in a stem such as *photograph* when it appears in a “stress-shifting” Latinate derivation. In isolation, the main stress is on the first syllable, and the vowel in the second syllable is reduced. In *photography*, however, the main stress is on the second syllable, and the vowels in the first and third syllables undergo reduction. Finally,

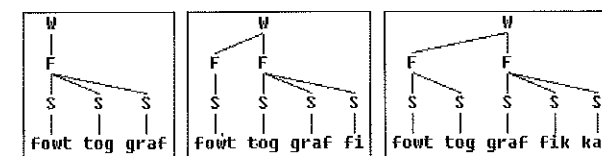


FIGURE 8.7. *Photograph*, *photography* and *photographical*.

in *photographical* the main stress is on the third syllable, and reduced vowels are found in the second and final syllable.⁵ In a segmental system, we would need to posit alternations with schwa for each vowel in the stem *photograph*. In a declarative system like IPOX, we could use a feature such as [+reduced] to trigger a special set of phonetic exponency rules. However, often this does not appear to be necessary. We have successfully synthesized these three words using full vowels (/ow/ as in *blow*, /o/ as in *pot*, and /a/ as in *sad*) in analysis as well as phonetic interpretation. The prosodic analysis trees assigned to these words by IPOX are shown in figure 8.7. By varying syllable compression in accordance with the metrical-prosodic structure, we obtain the expected alternations between full and reduced vowels (see Demo 7 on the CD-ROM). Also, the reduced vowels are appropriately different in quality depending on the “underlying” vowel, which we think is an additional advantage with respect to an analysis using alternation with schwa.

The evidence presented above suggests that in our system vowel reduction is the natural consequence of variations in speech rhythm that are independently motivated. In order to further substantiate this claim, however, we need to examine alternations between full and reduced vowels more systematically.

8.4 Connected Speech

As mentioned in section 8.1, the current version of IPOX lacks interfaces between morphosyntactic structure on the one hand, and metrical-prosodic structure on the other. In this section we briefly discuss the kind of interface we have in mind, illustrated with an analysis of the sentence “*This is an adaptable system.*” Also, we discuss how, with a few adaptations to the prosodic grammar, we have been able to generate this sentence, despite the lack of a syntax-prosody interface.

The main problem posed by this sentence is a severe mismatch between morphosyntactic and prosodic structures. This is illustrated in the metrical represen-

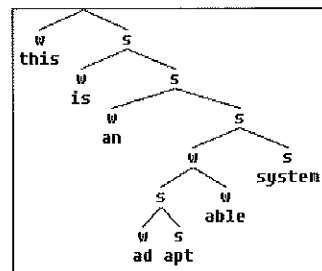


FIGURE 8.8. Morphosyntactic structure.

⁵Note that in the novel, but well-formed derivation *photographicality* we would find a full vowel for the affix *-al*.

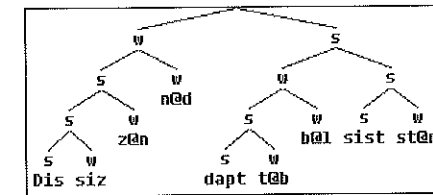


FIGURE 8.9. Prosodic structure (metrical tree).

tations in figures 8.8 and 8.9 (in these screen clips, category labels are not shown, and metrical structure is shown by labeling nodes with w(eak) or s(trong)). The structure in figure 8.8 is obtained by parsing orthographic input using simple IPOX grammars for English syntactic and morphological structure, in which phrase structure rules are annotated with metrical structure. The structure in figure 8.9 is obtained by parsing phoneme-based input using a prosodic grammar for English. (The internal structure of syllables is not shown.)

Although the two structures are radically different, they are systematically related, and a general solution would need to take both into account. In generative phonology, this relation is usually described in procedural terms: unstressed function words and stray initial syllables initially remain “unparsed,” and are “adjoined” at a later stage to the preceding foot. In declarative theory, such a solution is not available, so the two structures must be produced in parallel. In our current example, the structure in figure 8.9 is arrived at by requiring that heavy syllables such as /dapt/ and /sist/ appear as the head of a foot, whereas light syllables are weak nodes of a foot (except phrase-initially). Such a restriction, however, is warranted only in the Latin part of the lexicon [Col94], and should not be generalized to phrase-level prosodic structure. However, setting the grammar up this way allowed us to experiment with the generation of short sentences.

Figure 8.10 shows the prosodic structure of our sentence again, this time in the “headed” format of figures 8.1 and 8.8. Figure 8.11 shows the generated formant structure and F_0 , as well as the waveform generated from this structure (see Demo 8 on the CD-ROM).

This experiment illustrates that generating connected speech is not qualitatively different from generating a two-syllable word, although for a longer utterance we would need to worry about prosodic phrasing as well as be able to supply a more varied intonation. In the present setup, intervocalic clusters are ambisyllabic as much as possible, and each metrical foot receives the same linear falling F_0 . In future work, we envisage incorporation of a more sophisticated treatment of F_0

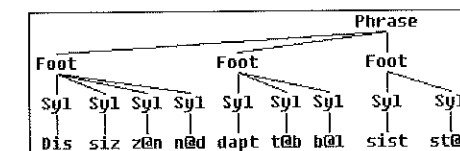


FIGURE 8.10. Prosodic structure (headed tree).

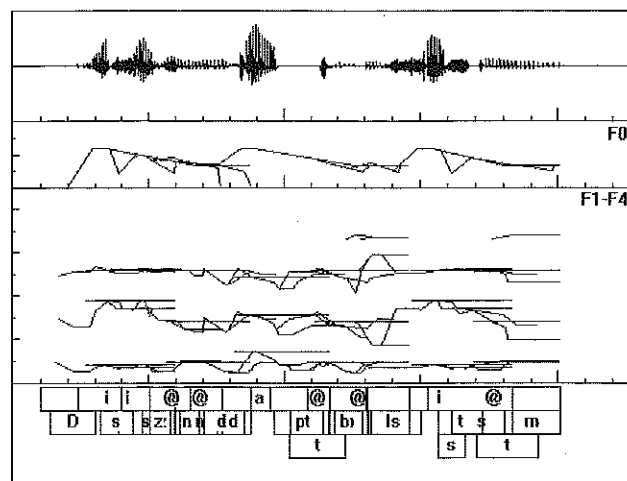


FIGURE 8.11. Phonetic interpretation.

generation along the lines of [Pie81], whose autosegmental approach to intonation is close in spirit to the methods employed in IPOX.

8.5 Summary

We have described an architecture for analyzing a sentence syntactically, morphologically and prosodically, and computing phonetic parameters for the Klatt synthesizer from such an analysis. A number of phenomena that are unrelated in a more conventional system based on rewrite rules, such as coarticulation, unstressed vowel shortening, centralization of unstressed vowels, syllabic sonorant formation, and elision of unstressed vowels before syllabic sonorants, are modeled in IPOX as natural concomitants of the all-prosodic view of phonological structure and phonetic interpretation. In the future we shall improve the phonetic quality of our English and Dutch phonetic parameters, as well as address the prosody-syntax interface and the phonetics of intonation more thoroughly.

An inspection copy of the IPOX system is available on the World Wide Web from one of the following locations:

<ftp://chico.phon.ox.ac.uk/pub/ipox/ipox.html>
<http://www.tue.nl/ipo/people/adirksen/ipox/ipox.html>

The software and documentation are freely available for evaluation and non-profit research purposes only. The authors reserve the right to withdraw or alter the terms for access to this resource in the future. Copyright of the software and documentation is reserved ©1994, 1995 by Arthur Dirksen/IPO and John Coleman/OUPL.

Acknowledgments: This chapter has benefited from comments by two anonymous reviewers. The research of Arthur Dirksen has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

REFERENCES

- [AHK87] J. Allen, M. S. Hunnicut, and D. Klatt. *From Text to Speech: The MITALK System*. Cambridge University Press, Cambridge, 1987.
- [Col92] J. S. Coleman. "Synthesis-by-rule" without segments or rewrite rules. In *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T. R. Sawallis, eds. Elsevier, Amsterdam, 211–224, 1992.
- [Col94] J. S. Coleman. Polysyllabic words in the YorkTalk synthesis system. In *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, P. A. Keating, ed. Cambridge University Press, Cambridge, 293–324, 1994.
- [Col95] J. S. Coleman. Synthesis of connected speech. To appear in *Work in Progress* No. 7. Speech Research Laboratory, University of Reading, 1–12, 1995.
- [Dir93] A. Dirksen. Phonological parsing. In *Computational Linguistics in the Netherlands: Papers from the Third CLIN meeting*, W. Sijtsma and O. Zweekhorst, eds. Tilburg University, Netherlands, 27–38, 1993.
- [DQ93] A. Dirksen and H. Quené. Prosodic analysis: the next generation. In *Analysis and Synthesis of Speech: Strategic Research Towards High-Quality Text-to-Speech Generation*, V. J. van Heuven and L. C. W. Pols, eds. Mouton de Gruyter, Berlin, 131–144, 1993.
- [Loc92] J. K. Local. Modelling assimilation in nonsegmental, rule-free synthesis. *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, G. J. Docherty and D. R. Ladd, eds. Cambridge University Press, Cambridge, 190–223, 1992.
- [OGC93] J. P. Olive, A. Greenwood and J. Coleman. *Acoustics of American English Speech: A Dynamic Approach*. Springer-Verlag, New York, 1993.
- [Pie81] J. B. Pierrehumbert. Synthesizing intonation. *J. Acoust. Soc. Amer.* 70(4):985–995, 1981.
- [PB88] J. B. Pierrehumbert and M. E. Beckman. *Japanese Tone Structure*. MIT Press, Cambridge, MA, 1988.

Appendix: Audio Demos

Demo 1 - Coarticulation:

Spreading of vocalic place features in the phonology is reflected in phonetic interpretation by subtle differences in frication spectra associated with /s/: spit, spot, sprint.

Demo 2 - Syllable overlap:

Three versions of /bot@l/ bottle, with ambisyllabic /t/, generated with different amounts of syllable overlap:

Demo 3 - Ambisyllabicity:

Intervocalic clusters are parsed with maximal ambisyllabicity. In the following words, the bracketed clusters are ambisyllabic: /win[t]@t/, winter; /si[st]@m/, system; /a[sp]rin/, aspirin. Note that /t/ is aspirated in winter, but not in system.

Demo 4 - Syllable compression I:

Again, three versions of /bot@l/ bottle, this time with different amounts of compression for the first and second syllable: Note that when the second syllable /t@l/ is compressed to 62 percent, the vowel is almost fully eclipsed, creating the impression of a syllabic sonorant.

Demo 5 - Syllable compression II:

Two versions of /spowz/ suppose, with different amounts of compression for the unstressed prefix /sp̄/.

Demo 6 - Syllable compression III:

A segmental analysis of vowel elision would seem to predict that "s'pport" is phonetically identical to "sport." Our analysis in terms of syllable compression correctly predicts subtle (and less subtle) differences: Note that /p/ is aspirated in "s'pport" but not in "sport."

Demo 7 - Syllable compression IV:

The three words below have been synthesized using full vowels (/ow/ as in blow, /o/ as in pot, and /a/ as in sad) in analysis as well as phonetic interpretation. By varying syllable compression in accordance with metrical-prosodic structure, we obtain the expected alternations between full and reduced vowels: /fowtograf/, "photograph"; /fowtografi/, "photography"; /fowtografikal/, "photographical."

Demo 8 - Connected speech:

Our first attempt at generation of a full sentence with IPOX.
/Disiz@n@dapt@b@lsist@m/, "This is an adaptable system."

9

A Model of Timing for Nonsegmental Phonological Structure

John Local
Richard Ogden

ABSTRACT Usually the problem of timing in speech synthesis is construed as the search for appropriate algorithms for altering durations of speech units under various conditions (e.g., stressed versus unstressed syllables, final versus non-final position, nature of surrounding segments). This chapter proposes a model of phonological representation and phonetic interpretation based on Firthian prosodic analysis [Fir57], which is instantiated in the YorkTalk speech generation system. In this model timing is treated as part of phonetic interpretation and not as an integral part of phonological representation. This leads us to explore the possibility that speech rhythm is the product of relationships between abstract constituents of linguistic structure of which there is no *single* optimal distinguished unit.

9.1 Introduction

One of the enduring problems in achieving natural-sounding synthetic speech is that of getting the rhythm right. Usually this problem is construed as the search for appropriate algorithms for altering durations of segments under various contextual conditions (e.g., initially versus finally in the word or phrase, in stressed versus unstressed syllables). Van Santen [van92] identifies a number of different approaches employed to control timing in synthesis applications and provides an overview of their relative strengths and weaknesses. He refers to these as (i) sequential rule systems [Kla87]; (ii) lookup tables; (iii) binary (phone) classification trees [Ril92]; and (iv) equation techniques, based directly on duration models [CUB73]. All the approaches he discusses 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. Van Santen's own, sophisticated approach to modeling duration (employing sums-of-products models) has produced improved results in segmental timing for the AT&T synthesis system [van94]. However, as he acknowledges, units other than the segment are required to make synthetic speech sound natural.

In large part, the pervasive problem of rhythm in synthesis can be seen to arise from the adherence by researchers to representations which are based on concatenated strings of consonant and vowel segments that allocate those segments