All-Prosodic Synthesis Architecture

Arthur Dirksen
Institute for Perception Research/IPO
adirksen@prl.philips.nl

John Coleman
Oxford University Phonetics Laboratory/OUPL
John.Coleman@Phonetics.Oxford.ac.uk

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 (Coleman 1992, 1994; Local 1992), but differs in that it uses a rule compiler (Dirksen 1993), which allows a clean separation of linguistic statements and computational execution, as well as a more concise statement of various kinds of generalizations.

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 dealt with 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. 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 non-heads, 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.

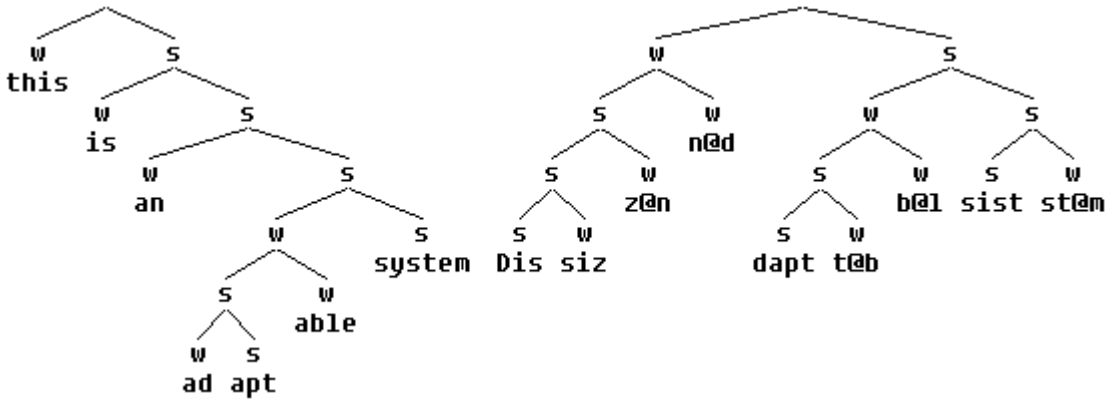
The architecture of IPOX is rather similar to that of YorkTalk (Coleman 1992, 1994; Local 1992), on which it is based, but is different in a number of respects:

1. YorkTalk representations are implemented as arbitrary Prolog terms, for example: *syllable(Head, onset(Closure, Glide), rime(Nucleus, Coda))*. In IPOX, metrical structure is made explicit in the representation (Dirksen and Quené 1993). This makes it possible to define general algorithms to process these structures in various ways.
2. The YorkTalk morphological and phonotactic parser is a Prolog DCG. 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 in the compiler.
3. IPOX includes a facility to state feature co-occurrence restrictions separately from the phrase structure rules (Dirksen 1993).

More generally, one can say that 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 into a single executable (which runs under Windows on a PC with an ordinary sound board, such as a Sound Blaster 16), using graphics to display analysis trees, phonetic parameters 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. Also, we have only just begun writing the grammars for English and Dutch, so speech output is still rather limited. Consequently, the focus in this paper is on how the IPOX/YorkTalk architecture may be used to generate connected speech rather than just isolated words, which until recently has been all that had been generated using this approach. To illustrate the approach, we shall discuss analysis and phonetic interpretation of a single utterance, the sentence "*This is an adaptable system.*"

2. Analysis

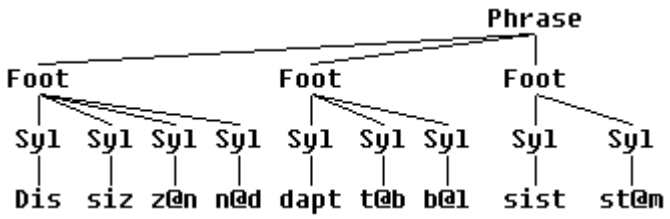
The main problem posed by this sentence is a severe mismatch between morphosyntactic and prosodic structures. This is illustrated in the metrical representations below. The one on the left is obtained by parsing orthographic input using simple IPOX grammars for English syntactic and morphological structure, in which phrase structure rules are annotated with metrical structure. The one on the right is obtained by parsing phoneme-based input using a prosodic grammar for English. (Internal structure of syllables is not shown. Ambisyllabic consonants are written out twice).



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 are produced in parallel. In our current example, the structure on the right is arrived at by requiring that heavy syllables such as /dapt/ and /sist/ only 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 Latinate part of the lexicon (cf. Coleman 1994), and should not be generalized to phrase-level prosodic structure.

3. Temporal interpretation

At the next stage, each node in the prosodic tree, which is repeated below in a slightly different format, is assigned a duration, and for each branching node the amount of overlap between its siblings is determined.



Subsyllabic constituents are assigned a duration on the basis of the feature structures associated with them in the analysis, as well as their prosodic environment. Durations of higher-level constituents are determined compositionally, taking overlap between constituents into account. Within the syllable, the duration of the head is assigned to the entire constituent (i.e. the duration of the syllable is equal to the duration of the nucleus). Here, "overlap" specifies the amount by which onset and coda are overlaid on the nucleus. The temporal structure of complex onsets and codas is determined in a similar way (Coleman 1994). Across syllables, however, the duration of a constituent is taken to be the sum of the durations of its siblings minus the amount of overlap between them.

This model of temporal interpretation (see also the paper by Local and Ogden in this volume) allows us to connect syllables in various ways, depending on their internal makeup as well as their prosodic environment. If no overlap (or a negative amount of overlap) is specified, syllables are simply concatenated. This type of juncture is appropriate between utterances and prosodic phrases, but not phrase-internally. For example, simply concatenating the syllables /sist/ and /t@m/ results in a pronunciation [sist' th@m], containing two stop bursts, two periods of aspiration and a short pause. In this case, it is necessary to make /st/ ambisyllabic, starting the second syllable somewhat before the end of the first syllable, such that the release of the unaspirated onset /t/ overwrites the release of the coda /t/, a type of juncture which is typical of ambisyllabic consonants. Other types of juncture are realized by specifying even more overlap ("flapping") or less overlap ("gemination").

Rhythmic properties of an utterance are accounted for by compressing the durations of syllables which appear in polysyllabic feet. For example, while a monosyllabic foot is assigned a compression factor of 1.0, the stressed syllable of a disyllabic foot might be assigned a compression of 0.9, and its weak sister one of 0.8. The rhythm of feet consisting of three or more syllables can be approximated by a compression pattern such as 0.85...0.65...0.75. (However, actual values for syllable compressions also depend on the internal structure of syllables).

4. Parametric interpretation.

After temporal interpretation, it is possible to determine absolute start and end times for each constituent, between which parametric interpretation rules are evaluated in top-down fashion, going left-to-right for constituents above the syllable level, head-first for subsyllabic constituents. In this way, holistic properties of prosodic phrases and metrical feet are computed first, to be worked out in finer detail by lower-level units. As an example, consider the generation of F0 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 (sometimes transcribed as H*+L). At lower levels of prosodic structure, individual consonants and vowels contribute small details to the complete F0 specification (e.g. so-called consonantal perturbations), similar to the F0 algorithm briefly discussed in Pierrehumbert and Beckman (1988:176-7).

The compression factors applied to weak syllables of a foot not only define the rhythm of an utterance, they also have an effect on vowel quality. Because onset and coda are overlaid on the nucleus, the latter determining the duration of the syllable, compressing a syllable reduces the vowel interval, possibly resulting in total elision. Thus, we can derive the effects of vowel reduction without altering the nucleus parameters. For example, we might expect /i/ in the second syllable of the first foot to have a somewhat reduced quality with respect to the first syllable. When the nucleus is evaluated, the difference is merely one of duration. However, when the onset and coda are overlaid on the nucleus, the onset-nucleus-coda transitions are often different for shorter vowels than for longer vowels, thus simulating the phenomenon of articulatory "undershoot". In a similar fashion, it is possible to derive the effects of syllabic sonorants such as /l/ in the final syllable of *adaptable* by providing the right compression factor.

5. Conclusion

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 modelled in IPOX as natural concomitants of the all-prosodic view of phonological structure and phonetic interrpretation. In the future we shall improve the phonetic quality of our English and Dutch phonetic parameters, as well as addressing the prosody-syntax interface more thoroughly.

References

Coleman, J.S. (1992) "Synthesis-by-rule" without segments or rewrite rules. In G. Bailly *et al.* (eds.), *Talking Machines: Theories, Models and Designs*. Amsterdam: Elsevier. 43-60.

Coleman, J.S. (1994) Polysyllabic words in the YorkTalk synthesis system. In P. A. Keating (ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*. Cambridge University Press. 293-324.

Dirksen, A. and H. Quené (1993) Prosodic analysis: the next generation. In V.J. van Heuven and L.C.W. Pols (eds.), *Analysis and Synthesis of Speech: Strategic Research towards High-Quality Text-to-Speech Generation*. Mouton de Gruyter. 131-144.

Dirksen, A. (1993) Phonological parsing. In W. Sijtsma and O. Zweekhorst (eds.), *Computational Linguistics in the Netherlands: Papers from the Third CLIN meeting*. Tilburg University. 27-38.

Local, J.K. (1992) Modelling assimilation in nonsegmental, rule-free synthesis. In G.J. Docherty and D.R. Ladd (eds.) *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press. 190-223.

Pierrehumbert, J.B. and M.E. Beckman (1988) *Japanese Tone Structure*. MIT Press.