AN INTEGRATED PHONOLOGICAL-PHONETIC MODEL FOR TEXT-TO-SPEECH SYNTHESIS

Jill House, Department of Phonetics and Linguistics, University College London
Sarah Hawkins, Department of Linguistics, University of Cambridge

ABSTRACT

We propose a model for text-to-speech synthesis (TTS) in which the units of phonological structure, correctly identified, determine phonetic interpretation in a non-arbitrary manner. A notion of dominance is used in the phonetic interpretation, which involves the interrelationship of all levels of structure in spectral, temporal and intonational domains.

1. PRELIMINARIES

1.1 Objectives

Our main aim is to maximise the intelligibility and naturalness of British English TTS using structures and procedures which are constrained in a principled, non-arbitrary way. Since a further objective is to design a model with multi-lingual potential, we must define structures which have quasi-universal applicability, while allowing that the detailed properties of the units identified, and of the processes they enter into, may be language-specific.

1.2 Theoretical considerations

Many current theories of phonology emphasise the discovery of non-arbitrary universal principles and language-specific parameter specifications; this represents a move away from an over-powerful re-write rule format in favour of rich structural representations. We too aim to identify a hierarchical structure whose components constitute all those units which enter into linguistic contrasts. In the phonetic interpretation we exploit a notion of **dominance** which involves the inter-relationship of all levels of structure in spectral, temporal and intonational domains.

Our model draws on the strengths of relational, structure-based systems such as YorkTalk [1]; an important difference is that we explicitly address phone-sized segments, within the structures of which they form part, in our phonetic interpretation. This allows us to make straightforward use of the segment-based transcriptions provided by the lexicon.

1.3 Technical considerations

The proposed model is adaptable in principle to both concatenative and formant synthesis systems, though details of the implementation would differ. Input to the model is assumed to be the output of a parser (e.g.[2]), which takes initial responsibility for building prosodic structure.

2. PROSODIC COMPONENT

2.1 The parser interface

The job of the parser is to derive a hierarchical prosodic structure using morphological and grammatical category information stored in a large pronunciation lexicon. Following [3], its grouping strategies are further motivated by principles of verb-balancing and verb-adjacency. To a large extent, the parser provides us with the required units of phonological structure, in tree format, with labelled nodes partially marked for prominence and boundary strength. Lexical look-up, supplemented where necessary by spelling-to-sound rules, supplies a segmental transcription, complete with lexical stress assignment to individual syllables, and an indication of the boundaries of phonological words.

The first task of the prosodic component is to check the well-formedness of the tree structure supplied by the parser. In practice this means using metrical principles to assign strong/weak values to nodes, supplementing the partial prominence orderings, and assigning segments to syllabic constituents. The completed structure is not subsequently altered.

2.2 The prosodic hierarchy

Following e.g. [4], we propose that utterances should be organised into constituent units arranged in a prosodic hierarchy. For us, the components of such a hierarchy include: *intonational phrase > pitch accent group > foot > syllable > syllabic constituents > skeletal positions (segments).* Units at the top of the hierarchy are made up of constituent units from the lower levels. One unit at each level acts as the dominant **head** constituent. The head typically constrains the phonetic interpretation of all constituents on the same level, and also of those units at lower levels which it directly dominates in the hierarchy. During the implementation, which proceeds top-down, information about the headedness of constituents is retained as an index that reflects the dominance hierarchy used in computing spectral, temporal and intonational parameters.

We recognise that intonational phrases are themselves loosely organised into "utterances", which in TTS typically correspond to sentence-length chunks of text, and that these chunks may be further organised into prosodic paragraphs. However, at levels above the intonational phrase the dominance hierarchy is not applicable.

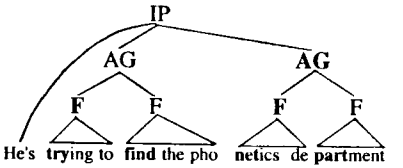
Constituents in the prosodic hierarchy have the following properties in English:
intonational phrase (IP): the domain for a well-formed intonation contour; all lower level constituents must be organised within an IP. The IP consists of one or more pitch accent groups, and the last of these constitutes the head.
pitch accent group (AG): this consists of an accented syllable (both stressed and pitch prominent), together with any unaccented syllables following it. The AG may contain several feet, of which the first (containing the accented syllable) is the head.
foot (F): like the pitch accent group, this unit is left-headed, consisting of an obligatory strong syllable (the head) and any weak syllables following it. For our purposes the foot is not bounded:

(1) $_{F}$[**prop**erties are de] $_{F}$[**creas**ing in] $_{F}$[**value**]

successive feet in (1) contain 5, 3 and 2 syllables respectively.

The organisation of IPs into AGs and feet is represented in (2). Constituents shown in **bold** are heads of the units of which they are daughters. Note that unstressed syllables at the beginning of an IP need not be organised into feet or pitch accent groups, but are dominated directly by the IP itself.

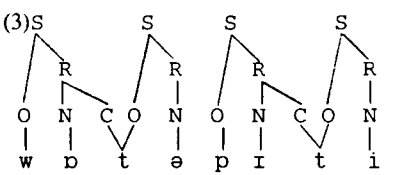
(2) $_{IP}$[He's $_{AG}$[$_{F}$[**try**ing to]$_{F}$[**find** the pho]]$_{AG}$[$_{F}$[**net**ics de]$_{F}$[**part**ment]]]



syllable (S): all segments are organised into syllabic constituents: obligatory **rhyme (R)** (head), and optional consonantal **onset (O)**.
syllabic constituents: the rhyme contains an obligatory **nucleus (N)** (head), normally filled by a vowel, and optional consonantal **coda (C)**. Onset, nucleus and coda may all branch. Dominance relationships between the segmental components of these constituents are expressed in terms of an index of coarticulation resistance (see 3.1).

Phrase-internally, consonants may be *ambisyllabic*, belonging simultaneously to the coda of one syllable and the onset of the next, proving phonotactic constraints are not violated.

(3)



Lexical stress further constrains ambisyllabicity word-internally: the /t/ in "PITy"(3), following a stressed vowel, is ambisyllabic, whereas the /t/ in "preTEND" would be assigned only to the onset of the second syllable. This reflects a different realisation of /t/ in such contexts in many varieties of English. In connected speech, ambisyllabicity applies more widely, to fill the empty onsets of vowel-initial words, regardless of stress: the /t/s in both "what a" (3) and "what ELSE" may be considered ambisyllabic.

2.3 The phonological word

One unit that is not included explicitly in our prosodic hierarchy is the **phonological word (PW)**. This important unit, created in the parser, involves modifying lexical representations by attaching weak,

monosyllabic function words as clitics to the last foot of a preceding word. Feet that are internal to the PW are prime sites for lenition processes (e.g. the reduction of "want to" to "wanna"). However, the phonological word is partly independent of other units in the hierarchy, since word boundaries and foot/AG boundaries need not coincide:

(3) This # im$_F$[PORtant] # $_F$[PRINciple] # $_F$[MUST be # re$_F$[MEMbered] #
(# = boundary of phonological word)

PW boundary information is passed on at the parser interface to ensure appropriate phonetic interpretation.

3. PHONETIC INTERPRETATION

We have three aims for the phonetic interpretation. First, the computed output values of each parameter should vary fairly continuously within a given range, so that the synthetic signal mimics the gradient parameter values of natural speech. Second, like the phonological structures, the phonetic principles we use should be language-independent; differences in parameter values reflect differences due to language, accent, and speech style. Third, for a given speech style, we seek a single control structure to account as far as possible for coarticulation, allophonic variation, timing/rhythm, and connected speech processes, in the belief that these are different facets of the same general process. For example, we claim that the syllable defines a unit of major importance for all these aspects; particular properties associated with the syllable will have different importance for, say, traditional coarticulation and timing. The work that will unify control of these four processes is not complete, so these are preliminary ideas.

For clarity in this short paper, we first discuss the control of coarticulation in its traditional sense, and then briefly indicate links with the other factors.

3.1 Coarticulation

Coarticulation is traditionally defined as the influence of one speech sound upon another, where a "sound" is a phone-sized unit. This definition typically includes obligatory effects due to aerodynamic and biomechanical properties of the vocal tract e.g. CV transitions, and certain non-obligatory processes that are often seen as easing articulatory demands, such as spread of lip rounding or nasalization. The definition usually excludes optional processes in which sounds mutually influence one another, such as assimilatory connected speech processes and controllable allophonic variation. With rare exceptions [5], explicit mention of timing is also excluded.

Most traditional coarticulatory effects for a given speech rate and style can be achieved through knowledge of the properties of the current syllable and of those adjacent to it. Within-syllable coarticulation accounts for CV and VC effects, while coarticulation across adjacent syllables (but not across a pause) accounts for V-to-V coarticulation and C ambisyllabicity.

We expect to achieve coarticulatory effects in a way similar to YorkTalk, by systematically computing parameter values in an order that reflects the domain of influence of each syllabic constituent, both within and between syllables. One difference from YorkTalk is that each terminal element (allophone) is associated with an index of coarticulation resistance [6] that affects the extent to which that phone influences the parameter values for the entire syllable. This coarticulation resistance index is incorporated into the dominance hierarchy.

3.2 The dominance hierarchy

While the domain of coarticulatory influence need be only as local as adjacent syllables, we need access to further information about position in the prosodic hierarchy and PW boundaries to compute the parameter values. The dominance hierarchy expresses the degree of influence that different factors have on syllabic constituents. Conceptually similar to [7]'s "acme articulations", it includes, but is broader than, "coarticulation resistance" [6]. Since the dominance hierarchy is expressed structurally, all information from the prosodic tree can contribute to it. Each node in the tree, from IP downwards, is associated with a weight, sensitive to the headedness of the constituent at each level. The weight associated with each terminal element is the product of all these weights on the structural components. For each relevant acoustic parameter in the syllable (e.g. aspiration amplitude, vowel duration, etc), this "terminal" weight is scaled appropriately and used in calculating the output value for that parameter. Other factors contributing to parameter specification include word or morphemic status, number of syllables, and position of the current syllable in the word.

The dominance hierarchy is intended to produce the gradient output that we aim for, and to be pivotal in bringing together coarticulation, allophonic variation, timing, and connected speech processes.

3.3 Allophonic variation, timing, and connected speech processes

Since we seek a general structure appropriate for different languages, accents, and speaking styles, we aim to standardise the control processes as far as possible, and to avoid rule proliferation. To this end, we aim to maximise the overlap in structural description and control parameters between coarticulation, allophonic variation, timing, and connected speech processes.

Since "coarticulatory processes" differ between languages, it seems that most aspects of coarticulation are under speaker control. We can thus conceptualise a continuum which extends from inevitable biomechanical and aerodynamic consequences (many traditional coarticulatory processes) at one end, to arbitrary, language-specific but highly controlled effects (many cases of allophonic variation) at the other. For example, CV transitions are evidence of coarticulation in the traditional sense, while differences in, say, the realisation of /u/ in the context of word-medial /r/ vs /z/ [8] represent allophonic variation of a more arbitrary type that can nonetheless be seen, and hence modelled, as coarticulatory in origin.

The structures relevant to (traditional) coarticulation and timing control are distinct in some ways but not in others. For example, the foot contributes directly to timing but not, we think, to coarticulation; the syllable, on the other hand, is a fundamental unit for both,, and desired coarticulatory effects can often be produced by appropriate adjustments to timing.

Similar patterns of convergence and difference exist for the other aspects of phonetic realisation. Assuming segments are defined in terms of structure, allophonic variation (e.g. aspiration amplitude and duration) largely reduces to gradient effects due to properties already defined by phonological structure, and hence represented in the dominance hierarchy, or to variations due to changes in speech rate or style (such as /t/ flapping). Some connected speech processes must be modelled for all fluent speech, but the degree and quality of these processes is style-dependent. Processes such as assimilations can be seen as having their origin in ease of articulation and hence in traditional coarticulation, but some cases may be analysed as involving structural differences.

4. SUMMARY

Our proposed model closely integrates a prosodic hierarchy and an acoustic dominance hierarchy, which together determine TTS parameters. They ensure a gradient output, properly constrained, which will increase naturalness and acceptability.

Partly supported by Telia Promotor Infovox AB

REFERENCES

[1] Local, J. (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.

[2] Youd, N.J. & A. Slater (1994), "Parsing for prosody", unpublished ms.

[3] Bachenko, J. & E. Fitzpatrick (1990), "A computational grammar of discourse-neutral prosodic phrasing in English", *Computational Linguistics* 16(3), 155-170.

[4] Nespor, M. & I. Vogel (1986), *Prosodic Phonology*, Dordrecht: Foris.

[5] Harris, K.S. & F. Bell-Berti (1981), "A temporal model of speech production", *Phonetica* 38, 9-20.

[6] Bladon, R.A.W. & A. Al-Bamerni (1976), "Coarticulation resistance in English /l/", *Journal of Phonetics* 4, 137-150.

[7] Kelly, J. (1989), " On the phonological relevance of some non-phonological elements", in T. Szende (ed.) *Proc. of the Speech Research '89 (Hungarian Papers in Phonetics* 21), 56-59.

[8] Hawkins, S. & A. Slater (1994), "Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech", *Proc. ICSLP 94*, 1, 57-60.