ProSynth: An Integrated Prosodic Approach to Device-Independent, Natural-Sounding Speech Synthesis

Sarah Hawkins[*], *Jill House*[**], *Mark Huckvale*[**], *John Local*[***], *Richard Ogden*[***]

[*] University of Cambridge, [**] University College, London, [***] University of York

ABSTRACT

This paper outlines ProSynth, an approach to speech synthesis which takes a rich linguistic structure as central to the generation of natural-sounding speech. We start from the assumption that the speech signal is informationally rich, and that this acoustic richness reflects linguistic structural richness and underlies the percept of naturalness. Naturalness achieved by structural richness produces a perceptually robust signal intelligible in adverse listening conditions. ProSynth uses syntactic and phonological parses to model the fine acoustic-phonetic detail of real speech, segmentally, temporally and intonationally.

1. INTRODUCTION

ProSynth explores the viability of a phonological model that addresses phonetic weaknesses found in current concatenative and formant-based text-to-speech (TTS) systems, in which the speech often sounds unnatural because the rhythm, intonation and fine phonetic detail reflecting coarticulatory patterns are poor. Although intelligibility in quiet conditions may compare well with natural speech, it is seriously impaired under conditions of high cognitive load or noise.

Building on [1, 2, 3, 4], ProSynth integrates and extends existing knowledge to produce the core of a new model of computational phonology and phonetic interpretation which will deliver high-quality speech synthesis. Key objectives are: (1) demonstration of selected parts of a TTS system constructed on linguistically-motivated, declarative computational principles; (2) a system-independent description of the linguistic structures developed; (3) perceptual test results using criteria of naturalness and robustness. To initially test the viability of our approach, we use a set of representative linguistic structures applied to Southern British English. The three focal areas of research are intonation, morphological structure, and systematic segmental variation.

2. THE PHONOLOGICAL MODEL

Our declarative phonological structure makes extensive use of a prosodic hierarchy, with phonological information distributed across the structure. Phonetics is related to phonology via a one-step phonetic interpretation function which makes use of as much linguistic knowledge as necessary. Systematic phonetic variability is constrained by position in structure, not by a set of phonological rules. The basis of phonetic interpretation is not the segment, but phonological features at places in structure. We thus extend the principle successfully demonstrated in [3, 4], to larger phonological domains.

Systematic phonetic variability, as determined by phonological structure, includes more acoustic fine detail than is standardly implemented in synthetic speech, consistent with the view [1] that, to understand speech, listeners use all available sensory information in proportion to its actual and perceived reliability, and that systematic suballophonic acoustic variation provides essential acoustic coherence in the speech signal. Some acoustic fine detail affects only adjacent segments, while other aspects, termed resonance effects [5], may extend over longer temporal domains of up to several syllables. Listeners are sensitive to such variation in both natural [6] and synthetic speech [7, 8], in auditory and visual modalities [9], consistent with spreading activation models of speech perception.

2.1 The Prosodic Hierarchy

The phonological structure into which text is parsed has units at the following levels: syllable constituents (Onset, Rhyme, Nucleus, Coda); Syllable; Foot; Accent Group; Intonational Phrase. Linguistic contrast can occur at each level in the hierarchy.

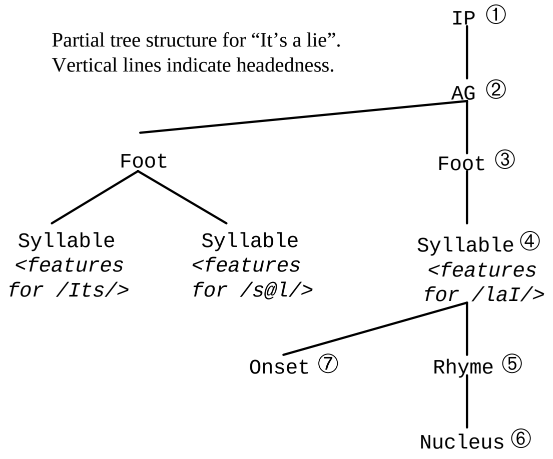
Partial tree structure for "It's a lie".
Vertical lines indicate headedness.
IP ①
AG ②
Foot
Foot ③
Syllable <features for /Its/>
Syllable <features for /s@l/>
Syllable ④ <features for /laI/>
Onset ⑦
Rhyme ⑤
Nucleus ⑥

Fig. 1. Partial tree structure of the utterance: "it's a lie". Indices (such as ❿) relate to the XML structure in Fig. 2.

Each smaller unit is dominated by a unit at the next highest level (Strict Layer Hypothesis [10]). This produces a linguistically well-motivated and computationally tractable hierarchy. 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. Structure-sharing is explicitly recognized through ambisyllabicity.

Although relevant to phonetic interpretation, particularly in terms of timing, the Phonological Word has no place in our strictly layered prosodic hierarchy. Word boundaries may not coincide with those of our prosodic constituents: some words contain several feet; some feet straddle word boundaries. Information about word breaks is available through links to the syntactic hierarchy, which can contribute as required to phonetic interpretation. Fig. 1 shows a partial parse of the phrase “It’s a lie” into the Prosodic Hierarchy.

2.2 Units of Structure and their Attributes

Input text is parsed to head-driven syntactic and phonological hierarchical structures. The phonological parse allots material to places in the prosodic hierarchy and is supplemented with links to the syntactic parse. The lexicon itself is in the form of a partially parsed representation. Phonetic interpretation may be sensitive to information at any level, so that it is possible to distinguish, for instance, a plosive in the onset of a weak foot-final syllable from an onset plosive in a weak foot-medial syllable.

Headedness: When a unit branches into sub-constituents, one of these constituents is its Head. If the leftmost constituent is the head, the constituent is said to be left-headed. If the rightmost, the structure is right-headed. Properties of a head are shared by the nodes it dominates [11]. Therefore a [+heavy] syllable has a [+heavy] rhyme; the syllable-level resonance features [±grave] and [±round] can also be shared by nodes they dominate: this is how coarticulation is modelled.

Phonological features: We use binary features, with each *attribute* having a *value*, where the *value* slot can also be filled by another attribute-value pair. To our set of conventional features we add the features [±rhotic], to allow us to mimic the long-domain resonance effects of /r/ [5, 8], and [±ambisyllabic] for ambisyllabic constituents (see below). Not all features are stated at the terminal nodes in the hierarchy: [±voice], for instance, is a property of the rhyme as a whole in order to model durational and resonance effects.

Syllables: The Syllable contains the constituents Onset and Rhyme. The rhyme branches into Nucleus and Coda. Nuclei, onsets and codas can all branch. The syllable is right-headed, the rhyme left-headed. Attributes of the syllable are [weight] (values heavy/light), and [strength] (values strong/weak): these are necessary for the correct assignment of temporal compression (§2.4).

Ambisyllabicity: Constituents which are shared between syllables are marked [+ambisyllabic]. Ambisyllabicity makes it easier to model coarticulation [4] and is an essential piece of knowledge in the overlaying of syllables to produce polysyllabic utterances. It is also used to predict properties such as plosive aspiration in intervocalic clusters (§2.4).

Feet: All syllables are organised into Feet, which are primarily rhythmic units. The foot is left-headed, with a [+strong] syllable at its head, and includes any [-strong] syllables to the right. Types of feet can be differentiated using attributes of [strength] and [headedness]. Any phrase-initial, weak syllables are grouped into a weak, headless foot. A syllable with the values [+head, +strong] is stressed.

Accent Groups (AG): An accented syllable is a stressed syllable associated with a pitch accent; an AG is a unit of intonation initiated by such a syllable, and incorporating any following unaccented syllables. The head of the AG is the leftmost strong, headed foot within it. A weak foot is also a weak, headless AG. AG attributes include [headedness], pitch accent specifications, and positional information within the IP.

Intonational Phrase (IP): The IP, the domain of a well-formed, coherent intonation contour, contains one or more AGs; minimally it must include a strong AG. The rightmost AG—traditionally the intonational nucleus—is the head of the IP. It is the largest prosodic domain recognised in the current implementation of our model.

2.3 Segmental information

The temporal extent of systematic spectral variation due to coarticulatory processes is modelled using two intersecting principles. One reflects how much a given allophone blocks the influence of neighbouring sounds, and is like coarticulation resistance [12]. The other principle reflects resonance effects, or how far coarticulatory effects spread. The extent of resonance effects depends on a range of factors including syllabic weight, stress, accent, and position in the foot, vowel height, and featural properties of other segments in the domain of potential influence. For example, intervening bilabials let lingual resonance effects spread to more distant syllables, whereas other lingual consonants may block their spread; similarly, resonance effects usually spread through unstressed but not stressed syllables.

2.4 Temporal information

Timing relations in ProSynth are handled primarily in terms of (1) temporal compression and (2) syllable overlap. Like spectral detail, temporal effects are treated as an aspect of the phonetic interpretation of phonological representations. Linguistic information necessary for temporal interpretation includes a grammar of syllable and word joins, using ambisyllabicity and an appropriate feature system. Such details as formant transition times, and inherent durational differences between close and open vowels, are handled in the statements of phonetic exponency pertaining to each bundle of features at a given place in structure.

A model of temporal compression allows the statement of relationships between syllables at different places in metrical structure [3], 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 can be related to the metrical structure as a whole. The primary timing unit is the syllable.

Syllable overlap: By overlaying syllables to varying degrees (making reference to ambisyllabicity), it is possible to lengthen or shorten intervocalic consonants systematically. There are morphologically bound differences which can be modelled in this way, provided that the phonological structure is sensitive to them. For instance, the Latinate prefix *in-* is fully overlaid with the stem to which it attaches, giving a short nasal in *innocuous*, while the Germanic prefix *un-* is not overlaid to the same degree, giving a long nasal in *unknowing*. Differences in aspiration in pairs like *mistake* and *mis-take* can likewise be treated as differences in phonological structure and consequent differences in the temporal interpretation of those structures.

2.5 Intonational information

There is a dimension of paradigmatic choice in modelling intonation: the pitch pattern used is not predictable from structure but is determined by discourse factors. The pattern for an IP depends on the pitch accents assigned to AGs, and on boundary tones associated with the edges of domains. The interpretation of the selected pitch contour in terms of f0 is, like other phonetic parameters, structure-dependent. Precise alignment of contour turning-points is constrained by the properties of units at lower levels in the hierarchy.

3. IMPLEMENTATION

We have so far (1) recorded and begun analysis of a speech database and (2) implemented our phonological representations using XML.

3.1 Design and Construction of a Database

Analysis for modelling has begun on a database of recorded speech, produced by a single male speaker of Southern British English. The database has been designed to exemplify a subset of possible structures. Currently we are looking at IPs of up to two AGs, themselves containing one or two feet of up to three syllables, and using a consistent falling intonation pattern. Even these limited structures show systematic variability in the alignment of f0 and the timing of different feet. Database sentences include prosodic domains differing in structure and length, and segmental sequences that differ in the extent to which intervening segments block the spread of coarticulatory effects. The perceptual salience of measured acoustic-phonetic regularities is assessed, and those that prove to be used by listeners are incorporated into the prosodic hierarchy.

3.2 Linguistic Representation and Processing

For linguistic representation and processing, we have formatted our computational structures using the extended mark-up language XML [13]. XML provides a powerful and computationally tractable representation for our hierarchical structures. It is also an upcoming internet standard and one supported by available toolkits such as the Edinburgh Language Technology Group toolkit LT-XML [14].

Currently we are using XML to represent: (1) lexicon, including the parts of speech and word pronunciation data; (2) utterance audio file information, including speaker name, utterance identifier, file name; (3) utterance word sequence, including time alignment information and cross references into the syntactic and prosodic hierarchies; (4) utterance parse, including detailed word tag, phrase structure and syntactic functions; (5) utterance prosodic structure, including phonetic features derived from the signal.

We use 'hyperlinks' within XML to indicate structural relationships between the syntactic and prosodic hierarchies and word-sequence within an utterance. This allows us, for example, to identify a syllable contained within a particular word or positioned at a particular place within a grammatical phrase. The links also allow us to identify the timing of a word from a phonetic alignment with a signal. Fig. 2 shows a partial XML representation of the parsed utterance, "It's a lie", whose tree structure representation is shown in Fig. 1.

```
<IP ➓ START="0.2206" STOP="0.9727">
 <AG ❡ START="0.2206" STOP="0.9727">…
  <FOOT ① START="0.5011" STOP="0.9727">
   <SYL   FPOS="1" RFPOS="1" RWPOS="1"
      START="0.5011" STOP="0.9727"
      STRENGTH="STRONG" WEIGHT="HEAVY" WPOS="1"
      WREF="WORD3">
    <ONSET   START="0.5011" STOP="0.6615"
      STRENGTH="WEAK">
     <CNS AMBI="N" CNSCMP="N" CNSGRV="N" CNT="Y"
       FXGRD="52.4" FXMID="115.6" NAS="N"
       RHO="N" SON="Y" START="0.5011"
       STOP="0.6615" STR="N" VOCGRV="N"
       VOCHEIGHT="CLOSE" VOCRND="N"
       VOI="Y">l</CNS></ONSET>
    <RHYME  CHECKED="N" START="0.6516"
       STOP="0.9727" STRENGTH="WEAK" VOI="N"
       WEIGHT="HEAVY">
     <NUC ★ CHECKED="N" LONG="Y" START="0.6516"
       STOP="0.9727" STRENGTH="WEAK" VOI="N"
       WEIGHT="HEAVY">
      <VOC FXGRD="-160.6" FXMID="106.0" GRV="Y"
       HEIGHT="OPEN" RND="N" START="0.6516"
       STOP="0.8620">a</VOC>
      <VOC FXGRD="-105.3" FXMID="95.4" GRV="N"
       HEIGHT="CLOSE" RND="N" START="0.8620"
        STOP="0.9727">I</VOC></NUC>
    </RHYME>
   </SYL>
  </FOOT>
 </AG>
</IP>
```

Fig 2. Partial XML representation of utterance: "it's a lie".

To annotate an existing audio file with XML annotations requires the following steps: (1) create a basic XML description of the audio data in the file; (2) add in a word level transcription; (3) update with parts of speech and pronunciations to word; (4) copy over prosodic structures from lexicon; (5) align prosodic structure with automatically-derived phone labels on audio file; (6) transfer parameters of modelled fundamental frequency into XML structure.

Our database of XML annotated files can be searched to find structures matching a specific pattern so that analysis can be made of timing, f0 patterns and ultimately segmental realisations in context. To provide the required flexibility of pattern-matching across the syntactic and prosodic hierarchies, we have developed our own pattern-matching system. For example, the following pattern

```
UTT
.WORDSEQ
..WORD(ID=$1) /the/
.IP
..AG
...FOOT
....SYL(WREF=$1)
.....*RHYME
....SYL
.....ONSET
......CNS /j/
```

searches and reports the rhyme in the word "the" before a syllable containing a /j/ in its onset. The indented structure reflects the pattern of the annotation hierarchy. The pattern-matching language will be extended to express the kind of declarative linguistic knowledge about timing, fundamental frequency form and segmental realisation in context required by our synthesis system.

4. FUTURE WORK

Work is in progress [15] to automatically copy-synthesize database items into parameters for HLsyn, a Klatt-like formant synthesizer that synthesizes obstruents by means of pseudo-articulatory parameters. This method allows for easy production of utterances whose parameters can then be edited. Utterances can be altered to either conform to rules of the model, or to break such rules, thus allowing the perceptual salience of particular aspects of phonological structure to be assessed. Tests will assess speech intelligibility when listeners have competing tasks involving combinations of auditory vs. nonauditory modalities, and linguistic vs. nonlinguistic behaviour.

A statistical model based on our hypotheses about relevant phonological factors for temporal interpretation will be constructed from the database, leading to a fuller non-segmental model of temporal compression. Temporal, intonational and segmental details will be stated as the phonetic exponents of the phonological structure.

5. REFERENCES

1. Hawkins, S. "Arguments for a nonsegmental view of speech perception." *Proc. ICPhS XIII*, Stockholm. Vol. 3, 18-25, 1995.
2. House, J. & Hawkins, S., "An integrated phonological-phonetic model for text-to-speech synthesis", *Proc. ICPhS XIII*, Stockholm, Vol. 2, 326-329, 1995.
3. Local, J.K. & Ogden R. "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, New York. 109-122, 1997.
4. Local, J.K. "Modelling assimilation in a non-segmental rule-free phonology." In G J Docherty & D R Ladd (eds): *Papers in Laboratory Phonology II*. Cambridge: CUP, 190-223, 1992.
5. Kelly, J. & Local, J. *Doing Phonology*. Manchester: University Press, 1989.
6. Hawkins, S., & Nguyen, N. "Effects on word recognition of syllable-onset cues to syllable-coda voicing", *LabPhon VI*, York, 2-4 July 1998.
7. Hawkins, S. & Slater, A. "Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech." *ICSLP* 94, 1: 57-60, 1994.
8. Tunley, A. "Metrical influences on /r/-colouring in English", *LabPhon VI*, York, 2-4 July 1998.
9. Fixmer, E. and Hawkins, S. "The influence of quality of information on the McGurk effect." Presented at Australian Workshop on Auditory-Visual Speech Processing, 1998.
10. Selkirk, E. O., *Phonology and Syntax*, MIT Press, Cambridge MA, 1984.
11. Broe, M. "A unification-based approach to Prosodic Analysis." *Edinburgh Working Papers in Cognitive Science* 7, 27-44, 1991.
12. Bladon, R.A.W. & Al-Bamerni, A. "Coarticulation resistance in English /l/." *J. Phon* 4: 137-150, 1976.
13. http://www.w3.org/TR/1998/REC-xml-19980210
14. http://www.ltg.ed.ac.uk/
15. Heid, S. & Hawkins, S. "Automatic parameter-estimation for high-quality formant synthesis using HLSyn." Presented *at 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.