ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis

Richard Ogden,† Sarah Hawkins,‡ Jill House,§ Mark Huckvale,§ John Local,† Paul Carter,† Jana Dankovičová§ and Sebastian Heid‡

†University of York, ‡University of Cambridge, §University College, London

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 acoustic richness of the speech signal reflects linguistic structural richness and underlies the percept of naturalness. Naturalness achieved by paying attention to systematic phonetic detail in the spectral, temporal and intonational domains produces a perceptually robust signal that is intelligible in adverse listening conditions. ProSynth uses syntactic and phonological parses to model the fine acoustic–phonetic detail of real speech. We present examples of our approach to modelling systematic segmental, temporal and intonational detail and show how all are integrated in the prosodic structure. Preliminary tests to evaluate the effects of modelling systematic fine spectral detail, timing, and intonation suggest that the approach increases intelligibility and naturalness.

© 2000 Academic Press

1. Introduction

Speech synthesized by rule has yet to make a significant impact as an output channel for information systems, despite continued engineering advances in text-to-speech (TTS) systems. A recurrent complaint is the perceived “unnatural” quality of the synthetic speech: that the speech does not sound as if it could have been produced by a human speaker. Such problems persist despite improvements in textual analysis, pronunciation and signal generation. For example: although the use of a large corpus of recorded speech for polyphone concatenation has produced signals with sections with a highly natural voice quality, utterances still exhibit disfluencies, broken rhythm and lack of coherence. Contemporary synthetic speech still suffers from unexpressive and often inappropriate prosody, and from poorly modelled coarticulation. These failings arise from the poverty of the linguistic representation underlying the utterance to be produced, as well as a fundamental lack of attention to the systematic fine detail in human production—fine detail that listeners expect and also utilize when listening in noise.
Segmental intelligibility data illustrate the scale of the problem. When heard in noise, most synthetic speech loses intelligibility much faster than natural speech. For example, Pratt (1986, see also Pols, 1989) used the Diagnostic Rhyme Test (isolated words or syllables), to compare the intelligibility of natural speech and nine TTS systems in quiet and at 0 dB s/n ratio. The nine systems included three DECTalk voices, Call Text 5010 v1.2, Infovox provisional British English, Prose 2000 v1.2, Texas Instruments TI-Speech, Namal Type & Talk, and Computer Concepts. The natural speech was about 15% less intelligible at 0 dB s/n ratio than in quiet, whereas even the DECTalk voices were 20–30% less intelligible in the noise, despite the fact that, in quiet conditions, the best DECTalk voice and the natural voice had almost identical intelligibility scores. Allowing for floor effects, the typical drop of the other synthetic systems was even greater—about 35–50%.

Although it is tempting to dismiss Pratt’s result on the grounds that current state-of-the-art concatenative synthesizers should be less disadvantaged in noise than formant synthesizers from the mid-1980s, there seem to be no supporting data, and there are some indications to the contrary. For example, Logan, Greene and Pisoni (1989) used the Modified Rhyme Test to compare the intelligibility (in quiet) of 10 TTS systems, including one Linear Predictive Coding (LPC)-resynthesis and one concatenative system. The results were remarkably similar to Pratt’s, with the better formant synthesizers vastly outperforming the other two. More recently, the results of the 1998 ESCA/COCOSDA evaluation exercise (van Santen et al., 1998) indicated that the crucial factor determining intelligibility was not the type of synthesis, but the quality of the linguistic model that drove the system. One way to interpret these results is that the intelligibility of a TTS system tends to reflect the amount and quality of effort put into it. Formant synthesizers often sound less attractive, but are just as capable of successfully transmitting a message. In other words, concatenated natural speech avoids problems related solely to voice quality and local segment boundaries, but suffers just as much as formant synthesis from poor models of timing, intonation, and systematic variation in segmental quality that is dependent on word and rhythmical structure. Even when the grammatical analysis is right, one string of words can sound good, while another with the same grammatical pattern does not. When an utterance sounds wrong (i.e. unnatural), it is hard to understand, regardless of the type of synthesizer (cf. Duffy & Pisoni, 1992). Moreover, the cumulative effect of a succession of trivial errors can be devastating, making it crucially important to use the right linguistic model when utterances are long.

ProSynth is an integrated prosodic (i.e. structure-based) approach to speech synthesis. At its core is a phonological model which allows for structurally important distinctions to be made, even when the phonetic effect of these distinctions is subtle. The phonological model in ProSynth draws together insights from current theories of phonology, and makes it easier to model phonetic and perceptual effects. Recent research in computational phonology (e.g. Bird, 1995) combines highly structured linguistic representations (more technically, signs) with a declarative, computationally tractable formalism. Recent research in phonetics (e.g. Simpson, 1992; Hawkins & Slater, 1994; Manuel, 1995; Zsiga, 1995) shows that speech is rich in non-phonemic information which contributes to its naturalness and robustness. Other work (Local, 1992, 1993, 1995a,b; Ogden, 1992; Local & Ogden, 1997) has shown that it is possible to combine phonological with phonetic knowledge by means of a process known as phonetic interpretation: the assignment of phonetic parameters to pieces of phonological structure. All these strands of work have contributed to the phonological model which ProSynth uses. By mimicking as far as possible the systematic spectral, temporal and intonational details which are observable in natural speech, we aim to improve the intelligibility of synthetic speech.
This paper has the following structure. Section 2 outlines the motivation for the ProSynth model. Section 3 describes the linguistic model used to represent the information necessary for modelling the kinds of phonetic effects described in Section 2. Section 4 sets out how the model described in Section 3 is implemented, and Section 5 demonstrates how segmental, temporal and intonational detail are modelled. Section 6 presents preliminary tests to evaluate the effects of incorporating this kind of detail.

2. Motivation

That there are interdependencies between grammatical, prosodic and segmental parameters is well known to phoneticians and to everyone who has synthesized speech. When these components are developed for synthesis in separate modules, the apparent convenience is offset by the need to capture the interdependencies, which often leads to problems of rule ordering and rule proliferation to correct effects of earlier rules. In our view, much of the robustness of natural speech is lost by neglecting systematic detail, a neglect that results partly from an inappropriate emphasis on phoneme strings rather than on linguistic structure. Fine phonetic detail, also called systematic or lawful variation (or variability, cf. Elman & McClelland, 1986), contributes to making the time-varying speech signal an effective communicative medium because it reflects multidimensional properties of both vocal-tract dynamics and linguistic structure.

Accordingly, ProSynth models more phonetic detail than is standard in synthetic speech. Such detail includes secondary resonance effects, timing and rhythm, and f0 alignment. The aim is to create a signal that sounds natural because it seems to come from a single talker and provides rich phonetic information about the linguistic structure of the utterance. The well-known “redundancy” of the speech signal, whereby a phone can be signalled by a number of more-or-less co-occurring acoustic properties, contributes some of this detail, but in our view, other less well-documented properties are just as important. As implied above, they can be roughly divided into two groups: those that make the speech signal sound as if it comes from a single talker, and those that reflect linguistic structure for a given accent.

A speech signal sounds as if it comes from a single talker when it is perceptually coherent, meaning that its properties reflect details of vocal-tract dynamics. To be heard as speech, time-varying acoustic properties must bear the right relationships to one another. When they do, the perceptual system groups them together into an internally coherent auditory stream (Bregman, 1990) or more abstract entity (cf. Remez, Rubin, Berns, Pardo & Lang, 1994; Remez, Fellowes, Pisoni, Goh & Rubin, 1998). A wide range of properties seems to contribute to perceptual coherence. The influence of some, such as patterns of formant frequencies, is widely acknowledged (cf. Remez, Rubin, Pisoni & Carrel, 1981). Others are known to be important but are not always well understood; examples are the amplitude envelope which governs some segmental distinctions (cf. Rosen & Howell, 1987) and also perceptions of rhythm and of “integration” between stop bursts and following vowels (van Tasell, Soli, Kirby & Widin, 1987); and correlations between the mode of glottal excitation and the behaviour of the upper articulators, especially at abrupt segment boundaries (Gobl, 1988; Pierrehumbert & Talkin, 1992; Ní Chasaide & Gobl, 1993).

A speech signal will not sound as if the talker is using a consistent accent and style of speech unless all the systematic phonetic details are right. This requires producing often small distinctions that reflect different combinations of linguistic properties. As an example, take the words mistakes and mistimes, whose spectrograms are shown at the left-hand side of Figure 1. The beginnings of these two words are phonetically different in a number of ways,
Figure 1. Left: spectrograms of the words *mistimes* (top) and *mistakes* (bottom) spoken by a British English woman in the sentence *I’d be surprised if Tess ___ it* with main stress on *Tess*. Right: syllabic structures of each word.

even though the first four phonemes are the same. The /t/ of *mistimes* is aspirated and has a longer closure, whereas the one in *mistakes* is not aspirated and has a shorter closure. The /s/ of *mistimes* is shorter, and its /m/ and /t/ are longer, which is heard as a rhythmic difference: the first syllable of *mistimes* has a heavier beat than that of *mistakes*.

These phonetic differences arise because the morphological structure of the words differs: *mistimes* contains the morphemes *mis*+*time*, which each have a separate meaning; and the meaning of *mistimes* is straightforwardly related to the meaning of each of the two morphemes. However, the meaning of *mistakes* is not obviously related to the meaning of its constituent morphemes, and the word might best be regarded as monomorphemic. This morphological difference is reflected phonologically in the syllable structure, as shown on the right of Figure 1. In *mistimes*, /s/ is the coda of syllable 1, and /t/ is the onset of syllable 2. Conversely, the /s/ and /t/ in *mistakes* belong to both syllables and form both the coda of syllable 1 and the onset of syllable 2. In an onset /st/, the /t/ is always unaspirated (cf. *step*, *stop*, *start*). The durational differences in the /m/ and the /t/ arise because the morphologically-conditioned differences in syllable structure result in *mist* being a rhythmically heavy syllable whereas *mis* is rhythmically light, while both syllables are metrically weak (i.e. unstressed). Thus the morphological differences between the words are reflected in structural phonological differences; and these in turn have implications for the phonetic detail of the utterances, despite the segmental similarities between the words.

Some types of systematic fine detail may contribute to both perceptual coherence and information about linguistic structure. So-called resonance effects (Kelly & Local, 1989) provide one example. Resonance effects associated with /t/, for example, typically manifest acoustically as lowered formant frequencies, and can spread over several syllables, but the factors
that determine whether and how far they will spread include syllable stress, the number of consonants in the onset of the syllable, vowel quality, and the number of syllables in the foot (Tunley, 1999).

On the one hand, including this type of fine phonetic detail (or systematic variation) in synthetic speech makes it sound more natural in a subtle way that is hard to describe in phonetic terms but seems to make the signal “fit together” better—in other words, it seems to make it more coherent. On the other hand, the fact that the temporal extent of rhotic resonance effects depends on linguistic structure means not only that cues to the identity of a single phoneme can be distributed across a number of acoustic segments (sometimes several syllables), but also that aspects of the linguistic structure of the affected syllable(s) can be subtly signalled.

Listeners can use distributed acoustic information to identify naturally-spoken words (Warren & Marslen-Wilson, 1987; Hawkins & Nguyen, to appear), and when such information is included in synthetic speech it can increase phoneme intelligibility in noise by 10–15% or more (Hawkins & Slater, 1994; Tunley, 1999). Both classical and recent experiments (Repp, 1982; Strange, 1989; Duffy & Pisoni, 1992; Pisoni, 1997; Kwong & Stevens, 1999) suggest that most systematically varying properties will enhance perception in at least some circumstances. Natural-sounding, systematic variation of this type may be especially influential in adverse listening conditions or when the cognitive load is high.

In summary, ProSynth is based on the philosophy that natural speech is robust because it contains many phonetic details at the spectral, temporal and intonational levels. These details vary systematically to form a perceptually coherent whole and are the product of the phonetic interpretation of a rich linguistic structure. In ProSynth, we attempt to model declaratively the richness of both linguistic structure and of the acoustic–phonetic signal which results from its interpretation (Pierrehumbert, 1990). The next sections set out how the phonological model is organized, and how we interpret it phonetically.

3. ProSynth: a linguistic model

ProSynth uses a phonological model which encodes phonological information in a hierarchical fashion using structures based on attribute-value pairs. Each phonological unit occurs in a complete metrical context. This context is a prosodic hierarchy with phonological contrasts available at all levels, as described in Section 3.1. The complex interacting levels of rules present in traditional layered systems are replaced in ProSynth by a one-step phonetic interpretation function operating on the entire context, which makes rule ordering unnecessary. Instead of the relatively poor structure and complex, interacting rules used in many older synthesis systems, in common with several other current synthesis systems (e.g. Bell Labs’ Approach, Sproat, 1997; The Festival Speech Synthesis System, Taylor, Black & Caley, 1998), ProSynth uses a rich structure and applies simple rules of phonetic interpretation which are highly structure bound. Systematic phonetic variation is thus constrained by position in structure. The basis of phonetic interpretation is not the segment, but phonological features at places in structure. We thus extend the principles successfully demonstrated in Local (1993) and Local and Ogden (1997) to a wider variety of phonological domains and phonetic details. The details of the units of structure and their attributes are set out in Section 3.2.
3.1. The prosodic hierarchy

The phonological structure is organized as a prosodic hierarchy, with phonological information distributed across the structure. The knowledge is formally represented as a kind of tree structure. Trees are commonly used for phonological representation.

The hierarchy has units at the following levels: syllable constituents (onset, rhyme, nucleus, coda); syllable; foot; accent group (AG); intonational phrase (IP). The prosodic hierarchy, building on House and Hawkins (1995) and Local and Ogden (1997) is a head-driven (Pollard & Sag, 1994) and strictly layered structure. Each unit is dominated by a unit at the next highest level (Selkirk, 1984, the Strict Layer Hypothesis). This produces a linguistically well-motivated and computationally tractable hierarchy which accords with the representational requirements of its implementation in XML (Section 4.3). 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.

Figure 2 shows a partial phonological structure for the phrase Come with a bloom. Note that phonological information is spread around the structure. For example, the feature [voice] is treated as a property of the rhyme as a whole, and not of just one of the terminal nodes headed by the rhyme. Timing information is also included: in Figure 2, the [start] of the IP is the same as the [start] of the onset of the first syllable of the utterance, and the [end] of the IP is the same as the [end] of the coda of the last syllable, as indicate by the tags 1 and 2. The value for [ambisyllabic] is shown for two consonants: note that for the [ambisyllabic:+] consonant /D/, the terminal node is immediately dominated by two other nodes.

There is no separate level of phonological word within the hierarchy. Such a unit does not sit happily in a strictly layered structure, because the boundaries of prosodic constituents like AG and Foot can occur in the middle of a lexical item. Conversely, word boundaries can occur in the middle of a Foot or AG. For example, in the phrase maths department there are two feet: [maths de-], and [-partment]. The second begins in the middle of a word, and the first contains a word boundary.

The computational representation of the prosodic structure allows us to get round this problem: word-level and syntactic-level information is hyper-linked (i.e. a feature Word Reference (WREF) on each syllable gives the ID of the word node in a syntax tree which contains it) into the prosodic hierarchy. 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 word-final syllable from an onset plosive in a weak word-medial syllable. In this way, lexical boundaries and the grammatical categories of words can be used to inform phonetic interpretation.

3.2. Units of structure and their attributes

Input text is parsed into both a syntactic and a phonological structure. The phonological parse allot 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. This section describes in more detail the units of structure—in particular supra-syllabic constituents—and their attributes.

Phonological features. Features are represented as <attribute, value> pairs.1 To the set of conventional features are added the features [rhotic:±], to allow us to mimic the long-

---

1 In <attribute, value> pairs, where the value is non-Boolean, such as [weight: heavy/light], we abbreviate this to eg. [light] where it is clearer to do so in the text.
domain resonance effects of /t/ (Kelly & Local, 1989; Tunley, 1999), and [ambisyllabic:±] for ambisyllabic constituents (see later this section). Phonological <attribute, value> pairs are distributed around the entire prosodic hierarchy rather than at just the terminal nodes (or even associated to just terminal nodes), as in many phonological theories. [voice:±], for instance, is a property of the rhyme as a whole in order to model durational and resonance effects, but this does not preclude differentiating sonorants and obstruents in the coda for the purposes of intonation modelling. Attributes at any level in the hierarchy may be accessed for use in phonetic interpretation.

**Headedness.** When a unit branches into sub-constituents, one of these constituents is its head. If the leftmost sub-constituent is the head, the unit is said to be left-headed. If the rightmost sub-constituent is the head, the unit is right-headed. AGs and feet are left-headed. Properties of a head are shared by the nodes it dominates (Broe, 1991; Ogden, 1999). 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 some aspects of coarticulation are modelled.

The feature [head:±] is used to mark headedness. A constituent with the feature [head:+] is the head of the superordinate constituent it belongs to. In Figure 2, headedness is indicated by vertical lines, as opposed to slanting ones. Phonetic interpretation proceeds head-first and is therefore determined in a structurally principled fashion without resort to extrinsic ordering.
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 head of the IP is the rightmost AG—traditionally the intonational nucleus. The IP is the largest prosodic domain recognized in the current implementation of the ProSynth model. The attributes of IP are (1) position in discourse, (2) speech act function, and (3) focus. (1) and (2) together determine pitch range f0 scaling and boundary tones; and (2) determines pitch accent type, whereas (3) determines intonational nucleus placement, using information from the syntax or the lexicon as a default when other discourse information is unknown.

Accent groups (AG). AGs are units of intonation. They immediately dominate one or more feet. The head of the AG is the leftmost [heavy] foot, and is associated with an intonational pitch accent. AG attributes include [weight: heavy/light], number of component feet, position within the IP and pitch accent specifications. Only [heavy] AGs can have pitch accents assigned to them. When an IP begins with one or more unaccented syllables, we maintain the strictly layered structure by analysing them as constituting a [light] or “degenerate” AG, which in turn contains a [light] foot. Degenerate AGs have no head, cannot carry pitch accents, and can only occur as the first AG in an IP.

Feet. All syllables are organized into feet, which are units of rhythm. Types of feet are differentiated using attributes of [weight: heavy/light], [strength: strong/weak], [head:+] and number of component syllables. Feet with the attribute [head:+] act as domains for the realization of pitch accents (see above). The attribute [weight] distinguishes between fully-formed ([heavy]) and degenerate ([light]) feet. A degenerate foot cannot act as a site for rhythmic stress because it is also [weak]. Only [strong] feet are associated with a rhythmically stressed position. The leftmost syllable within a foot acts as its head, so the syllable at the head of a [strong] foot, itself [strong], is stressed. However, [strong] syllables may occur inside [weak] feet; for example, the fourth syllable known in the phrase in the well-known maths department is [strong], but is dominated by a rhythmically [weak] foot.

This example illustrates how an AG may contain more than one foot: maths department, a compound noun, is a single AG, but contains two feet, [maths de-] and [-partment] (see also Figure 3). Similarly, when the discourse context requires an early intonation focus within the IP, AGs containing two or more feet are expected: for example, in a discourse situation where the maths department had already been mentioned, the sentence He joined the maths department would typically have an AG stretching from joined to the end of department, giving three feet in all ([joined the], [maths de-] and [-partment]). Automatic correct assignment of AG structure will eventually require access to information coded at lexical or discourse levels.

Syllables. The syllable contains the constituents onset and rhyme. The rhyme branches into nucleus and coda. Nuclei, onset and codas can all branch. Onset and codas contain consonants, while nuclei contain vowels. Both onsets and codas contain vocalic features which are inherited from the nucleus, which is the head of the syllable. This allows for the accurate modelling of coarticulation (Coleman, 1992; Local, 1992; Ogden, 1992).

Syllables are right-headed, rhymes left-headed. Attributes of the syllable include [weight: heavy/light], and [strength: strong/weak]: these are necessary for the correct assignment of temporal compression (Section 5.2). Foot-initial syllables are strong.

Weight is defined with regard to the sub-constituents of the rhyme. A syllable is [heavy] if its nucleus attribute [length] has the value [long] (in segmental terms, if it contains a long...
vowel or a diphthong). A syllable is also [heavy] if its coda has more than one constituent, as in /rent/, /task/, /taks/. Other syllables are [light]. In polymorphemic syllables such as cat+s, the weight of the syllable is determined according to the stem, and the suffix is treated as a syllable appendix.

There is not a direct relationship between syllable strength and syllable weight. Strong syllables need not be heavy. In loving, /l/ has a [short] nucleus, and the coda has only one constituent (corresponding to /v/), yet it is the strong syllable in the foot. Similarly weak syllables need not be light. In department, the final syllable has a branching coda (i.e. more than one constituent) and therefore is [heavy] but [weak]. ProSynth does not use extrametricality: all phonological material must be dominated by an appropriate node in structure.

Figure 3 illustrates the partial metrical structure for the syllable, foot, AG and IP nodes for the phrase in the well-known maths department, along with low-level syntactic tags.

<table>
<thead>
<tr>
<th>AG</th>
<th>IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>[POS: 0]</td>
<td>[POS: 1]</td>
</tr>
<tr>
<td>[feet: 1]</td>
<td>[feet: 2]</td>
</tr>
<tr>
<td>[light]</td>
<td>[heavy]</td>
</tr>
<tr>
<td>[head:-]</td>
<td>[head:-]</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>[head:-]</td>
<td>[head:-]</td>
</tr>
<tr>
<td>[weak]</td>
<td>[strong]</td>
</tr>
<tr>
<td>[light]</td>
<td>[heavy]</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>[head:-]</td>
<td>[head:-]</td>
</tr>
<tr>
<td>[weak]</td>
<td>[weak]</td>
</tr>
<tr>
<td>[light]</td>
<td>[light]</td>
</tr>
<tr>
<td>in</td>
<td>the</td>
</tr>
<tr>
<td>Prep.</td>
<td>Det.</td>
</tr>
<tr>
<td>AG</td>
<td>IP</td>
</tr>
<tr>
<td>[POS: 2]</td>
<td>[POS: 2]</td>
</tr>
<tr>
<td>[feet: 2]</td>
<td>[feet: 2]</td>
</tr>
<tr>
<td>[heavy]</td>
<td>[heavy]</td>
</tr>
<tr>
<td>[head:+]</td>
<td>[head:-]</td>
</tr>
<tr>
<td>[strong]</td>
<td>[strong]</td>
</tr>
<tr>
<td>[heavy]</td>
<td>[heavy]</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>[head:+]</td>
<td>[head:+]</td>
</tr>
<tr>
<td>[strong]</td>
<td>[strong]</td>
</tr>
<tr>
<td>[weak]</td>
<td>[weak]</td>
</tr>
<tr>
<td>[light]</td>
<td>[light]</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>[head:+]</td>
<td>[head:-]</td>
</tr>
<tr>
<td>[strong]</td>
<td>[strong]</td>
</tr>
<tr>
<td>[weak]</td>
<td>[weak]</td>
</tr>
<tr>
<td>[light]</td>
<td>[light]</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>[head:+]</td>
<td>[head:-]</td>
</tr>
<tr>
<td>[weak]</td>
<td>[weak]</td>
</tr>
<tr>
<td>[light]</td>
<td>[light]</td>
</tr>
<tr>
<td>in</td>
<td>the</td>
</tr>
<tr>
<td>Prep.</td>
<td>Det.</td>
</tr>
<tr>
<td>known</td>
<td>maths</td>
</tr>
<tr>
<td>N(N)</td>
<td>N</td>
</tr>
</tbody>
</table>

Ambisyllabicity. Ambisyllabicity means that a consonant can simultaneously belong to two adjacent syllables. Formally, ambisyllabicity is represented as re-entrant nodes at the terminal level: i.e. a consonant may simultaneously be ultimately dominated by two syllable nodes by being in the coda of one syllable and in the onset of the next. Constituents which are shared between syllables are marked [ambisyllabic:+]. Ambisyllabicity makes it easier to model coarticulation and is an essential piece of knowledge in the correct temporal relations between adjacent syllables. It is also used to predict spectral properties such as plosive aspiration in intervocalic clusters (see also Section 2).

Constituents are [ambisyllabic:+] wherever this does not result in a breach of syllable structure constraints. Loving comprises the two syllables, /l/v/ and /v/ls/, since /v/ is both a legitimate coda for the first syllable, and a legitimate onset for the second. Love- less has no ambisyllabicity, since /zl/ is neither a legitimate onset nor a legitimate coda.
Clusters may be entirely ambisyllabic, as in risky (/rɪsk/+/ski/), where /sk/ is a legitimate coda and onset cluster; partially ambisyllabic (i.e. one consonant is [ambisyllabic:+], and one is [ambisyllabic:−], as in selfish /sɛlf/+/fɪʃ/), or not ambisyllabic as in risk them (/rɪsk/+/θem/).

4. Implementation

This section describes how the phonological model described in the previous section has been implemented computationally. It describes the database used for the spectral, temporal and intonational modelling, the architecture and the use of the Extensible Markup Language (XML) to represent linguistic structure.

4.1. Database

Analysis for modelling is based on a speech database of over 470 utterances, designed specifically so as to provide coverage of the targeted linguistic structures, and recorded by a single male speaker of Southern British English (SBE). Database speech files have been exhaustively labelled to identify segmental and prosodic constituent boundaries, using hand-correction of an automated procedure. Fundamental frequency contours, calculated from a simultaneously recorded Laryngograph signal, can be displayed time-aligned with constituent boundaries.

The database comprises a subset of possible linguistic structures, each with a number of exemplars which together offer a wide range of the types of systematic variation of interest. Each utterance consists of one IP, and up to two AGs. The foot types within the AG vary according to the weight of the head syllable, the number and type of consonants in the onset and rhyme, the syllabic affiliation of intervocalic consonants, and vowel length. There are also phrases containing segments whose secondary resonance is expected to spread, and some which are expected to block the spreading of such effects.

The database thus provides us with material for analysis of the spectral, temporal and intonational phenomena we aim to synthesize.

4.2. Architecture

ProSynth uses an open computational architecture for synthesis. There is a clear separation between the computational engine and the computational representations of data and knowledge. The overall architecture is shown in Figure 4.

Text marked for the type and placement of intonational accents is input to the system, and a pronunciation lexicon is used to compose the strictly layered metrical structure for each intonational phrase in turn. The overall utterance is represented as a hierarchy, as already described Section 3. The composed structure is then modified according to rules for phonetic interpretation stored as external scripts. Rules apply when they match specific sub-structures in the data and modify node attributes in a reversible way that preserves the declarative aspect of the rules (Section 4.4).

The interpreted structure is converted to a parametric form depending on the signal generation method. The phonetic descriptions and timing can be used to select diphones and express their durations and pitch contours for output with the MBROLA system (Dutoit, 1997). The phonetic details can also be used to augment copy-synthesis parameters for the HLSyn quasi-articulatory formant synthesizer (Heid & Hawkins, 1998; in press). The timings and pitch
information have also been used to manipulate the prosody of natural speech using PSOLA (Hamon, Moulines & Charpentier, 1989).

4.3. Linguistic representation and modelling

XML is an extremely simple dialect of SGML (Standard Generalized Markup Language), the goal of which is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML is a standard proposed by the World Wide Web Consortium for industry-specific mark-up for a number of applications, such as: vendor-neutral data exchange, media-independent publishing, collaborative authoring, the processing of documents by intelligent agents and other metadata applications (see www.w3.org/XML/).

XML is used as the external data representation for the phonological structures in ProSynth. The features of XML which make it ideal for this application are: storage of hierarchical information expressed in nodes with attributes; a standard text-based format suitable for networking; a strict and formal syntax; facilities for the expression of linkage between parts of the structure; and readily-available software support.

In the ProSynth system, the input word sequence marked with prosodic phrasing and accent types is automatically converted to an XML representation of the hierarchical phonological structure. Conversion from marked text to XML is performed completely automatically using a pronunciation dictionary and a phonological parser. This representation is successively refined and expanded in the stages of phonetic interpretation. Within each stage, declarative rules matching parts of the structure make changes to the features stored on the nodes. Finally, special purpose code translates the XML encoded structure into parameter tables for signal generation.

XML is used to encode the following in ProSynth:

Word sequences. The text input to the synthesis system needs to be marked up in a number of ways. Importantly, it is assumed that the division into prosodic phrases and the assignment of accent types to those phrases has already been performed. This information is added to the text using a simple mark-up of IPs and AGs.
Lexical pronunciations. The lexicon maps word forms to syllable sequences. Each possible pronunciation of a word form has its own entry comprising: SYLSEQ (i.e. syllable sequence), SYL, ONSET, RHYME, NUC, ACODA (material in inflexional morphemes), CODA, VOC and CNS nodes (for vocalic and consonantal information, respectively). Information present in the input mark-up, possibly derived from syntactic analysis, selects the appropriate pronunciation for each word form.

Prosodic structure. Each composed utterance comprising a single intonational phrase is stored in a hierarchy that uses: UTT, WORDSEQ, WORD; and IP, AG, FOOT, SYL, ONSET, RHYME, NUC, CODA, ACODA, VOC and CNS nodes. Syllables are cross-linked to the word nodes using linking attributes. This allows for phonetic interpretation rules to be sensitive to the grammatical function of a word as well as to the position of the syllable in the word.

Database annotation. The ProSynth database has been manually annotated for the location of segmental boundaries and pitch accents. A prosodic structure, together with timing information derived from the labelling procedure, has been constructed for each phrase. This annotation is stored in XML using the same format as for synthesis. Tools for searching this database help us in generating knowledge for interpretation.

As described in Section 3.2, ambisyllabicity is a particular case of re-entrancy in a tree. Since XML rigidly enforces a strict hierarchy of components it is necessary to duplicate and link nodes in order to represent ambisyllabicity in XML.

An extract of a prosodic structure expressed in XML is shown in Figure 5, taken from the phrase *It was risky*. The representation was generated for one of the synthetic files used for perceptual testing. (In the XML representations, Y/N are used in place of the +/− used elsewhere in the text.)

4.4. Knowledge representation

Knowledge for phonetic interpretation is expressed in a declarative form that operates on the prosodic structure. This means firstly that the knowledge is expressed as unordered rules, and secondly that it operates solely by manipulating the attributes on the XML-encoded phonological structure. To encode such knowledge, a representational language called ProXML was developed in which it is easy to express the hierarchical contexts which drive processing and to make the appropriate changes to attributes. The ProXML language is read by an interpreter PRX written in C which takes XML on its input and produces XML on its output. ProXML is a very simple language modelled on both C and Cascading style Sheets (see www.phon.ucl.ac.uk/project/prosynth.htm for more information). A ProXML script consists of functions which are named after each element type in the XML file (each node type) and which are triggered by the presence of a node of that type in the input. When a function is called to process a node, a context is supplied centred on that node so that reference to parent, child and sibling nodes is easy to express.

Figure 6 shows a simple example of a ProXML script to adjust syllable durations for strong syllables in a disyllabic word whose second and final syllable is weak. The compression factors in the script are computed from regression tree data (Section 5.2) taken from the ProSynth database and represents a generalization about temporal relations within the database. If the first syllable is heavy, the rule is dependent on the length of the vowel. In this example, the DUR attribute on SYL nodes is set as a function of the phonological attributes.
Figure 5. Partial XML representation of the utterance: *It was risky*. Note how ambisyllabicity is represented: AMBI = Y for the feature description of /sk/; and /sk/ is in both the Coda of one syllable and the Onset of the next.
Figure 6. Example ProXML script, which modifies syllable durations dependent on the syllable-level and nucleus-level attributes.

found on that node and on others in the hierarchy. Note that the rules modify the duration attribute (\( * = \) means “scale existing value of DUR”, where the starting point is equivalent to Klatt’s “inherent duration”) rather than set it to a specific value. In this way, the declarative aspect of the rule is maintained.

5. Modelling

5.1. Spectral detail

5.1.1. Segmental identity

Whichever type of synthesis output system is used, the immediate input comes from the XML file. As shown in Figure 4, concatenative synthesis currently uses the MBROLA system, with sound segments chosen in the standard way from the MBROLA inventory for British English. Formant synthesis uses HLsyn driven by PROCSY, which is part copy-synthesizer from labelled speech files, and part rule-driven from information in the XML file (Heid & Hawkins, 1998; in press). Most formant trajectories for vowels and approximants are copy-synthesized, while obstruent consonants and some other sounds are produced by rule.

5.1.2. Fine-tuning spectral shape

In concatenative synthesis, the task of fine-tuning spectral shape is achieved by selecting appropriate units. ProSynth as yet makes no attempt to improve upon the standard MBROLA unit selection, but ultimately our work should have applications in unit selection in as much as it should increase our understanding of how factors such as long-domain resonance effects and grammatical dependencies influence spectral variation. The work described in the rest of Section 5.1 was done using formant synthesis using PROCSY and HLsyn.

When the constriction areas and other vocal-tract properties represented by the higher-level parameters are set to appropriate values, HLsyn automatically produces much local fine-tuning of spectral shape. Consequently, in comparison with standard formant synthesizers, it is relatively straightforward to produce complex acoustic changes at segment boundaries that closely mimic those of natural speech. Most notably, HLsyn produces natural-sounding, perceptually robust transitions between adjacent segments that differ in excitation type, such as the transition between vowels and voiced or voiceless stops or fricatives. This attribute of HLsyn means that some of the immediate appeal of concatenative synthesis—natural-sounding, perceptually robust transitions between adjacent segments, together with a pleasant voice quality—is also available in formant synthesis at little computational cost.
Although these types of acoustic fine-detail are relatively easily achievable using HLsyn, they have to be programmed to occur in only the right contexts. PROCSY provides the rules that do this. Some of the systematic variation is programmed by reference to the structure of the prosodic hierarchy, and some in the traditional way by reference to linear segmental context. Examples of prosodically-dependent rules include stress-dependent variations in the waveform amplitude envelope, and stress-dependent differences in excitation type in certain CVC sequences. For example, close or central vowels preceded and followed by alveolar obstruents involve much more voiceless frication or aspiration when they are unstressed as in the underlined syllables of today, disappoint, and attitude than when they are stressed, as in turtle, tiddler and titter. Examples of rules that rely mainly on local segmental context include coarticulation of nasality and the amount of voicing in the closure of voiced stops. These sorts of properties, though in need of more work, are reasonably well understood and most are relatively straightforward to implement to a satisfactory standard.

More challenging, because it is more subtle and less well understood, is the temporal extent of perceptually salient long-domain coarticulatory processes such as the resonance effects discussed in Section 2. For example, in SBE, /r/-colouring varies with vowel height and the number of consonants in the syllable onset and spreads for at least two syllables on either side of the conditioning consonant, as long as those syllables are unstressed and especially if they are in feet of three or more syllables (Tunley, 1999). Thus, whereas strong /r/-colouring might be expected to be found throughout a phrase like The tapestry bikini, it would be expected to be weak and confined only to bad and rap in a phrase like The bad rap artist (in a non-rhotic accent), which has open vowels and stressed syllables near the /r/. Work by West (1999) is broadly supportive of these observations.

We are still investigating what limits the spread of rhotic resonance effects. For example, when an /r/ occurs in a context that is susceptible to /r/-colouring, such as the last syllable of tapestry, is the resonance effect blocked by the next stressed syllable, or can it spread through into unstressed syllables of the adjacent foot? Preliminary evidence suggests that it can. Equally, however, some segmental articulations reduce the magnitude of resonance effects even in adjacent unstressed syllables. The way that resonance effects are modelled in ProSynth will depend to a large extent on how these stress and segmental interdependencies pattern. If rhotic resonance effects can pass through stressed syllables into adjacent feet, then the feature [rhotic] might have to be modelled as an attribute of the foot. However, some of the control of [rhotic] resonance must be determined at the level of the segmental feature, since its acoustic realization must also take account of the segmental context. One question is whether it might be possible to model the distinction between resonance effects governed by foot-level nodes vs. by segmental- or featural-level nodes in terms of how the particular instance of the effect is achieved—whether by manoeuvres of the tongue body, the tongue tip, the lips, or by combinations of all three.

5.2. Temporal modelling

One of the goals of temporal modelling is to model English rhythms accurately. The ProSynth timing model is foot based (Ogden, Local & Carter, 1999), and for any given syllable takes into account (1) its strength, (2) its weight, (3) its place in the foot, and (4) the strength and weight of adjacent syllables. Information about word boundaries is also available, allowing e.g. word-finality to influence the temporal interpretation of any syllable.

Abercrombie (1964) describes two rhythms which are important for disyllabic words in the variety of English being modelled: (1) short–long: happy, funny, city, (2) equal–equal:
Table I. Syllable durations in relation to weight of the first syllable in disyllabic, utterance-final feet

<table>
<thead>
<tr>
<th>Weight of 1st syll</th>
<th>Duration of 1st syll (ms)</th>
<th>Duration of 2nd syll (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy</td>
<td>381</td>
<td>330</td>
</tr>
<tr>
<td>Light</td>
<td>276</td>
<td>268</td>
</tr>
</tbody>
</table>

*hamper, funding, seedy.* The words with short–long rhythm have a light first syllable, while the words with equal–equal rhythm have a heavy first syllable. The vowels in the second syllables in the two sets are durationally different. For all items in the database that have utterance-final, disyllabic feet and short vowels in the first syllable, the duration of both the first and the second syllable is sensitive to the weight of the first syllable (Table I). The duration of a second syllable after a heavy first syllable is 23% greater than after a light first syllable.2

As well as durational differences there are also qualitative differences in the second-syllable vowels. The second-syllable vowels of words with short–long rhythm are diphthongized, while in words with equal–equal rhythm there are monophthongal vowels. The implication of these results is that when the second syllable of words like these is phonetically interpreted, it is necessary to have information available about the strength and weight of the preceding syllable. Similar, but more complex, statements must also be made for longer feet.

As well as rhythmic properties, there are “segmental” durational effects which relate to smaller stretches of speech but which (perhaps paradoxically) reflect higher levels of linguistic organization. For example, Fougéron and Keating (1997), and Keating, Cho, Fougéron and Hsu (to appear) have shown that the duration of various segment types is sensitive to at least three levels of structure in the prosodic hierarchy. Such observations provide further evidence that the accurate modelling of durations depends on having a rich phonological structure that phonetic interpretation accesses. In other words, temporal phonetic interpretation is reliant on the informational richness which is encoded in the phonological structure.

The temporal interpretation model is based on a CART (Classification and Regression Tree) analysis of the database, taking into account the phonological features in the prosodic hierarchy. CART analysis is succinctly described by van Santen (1994):

> [. . .] CART-based methods [. . .] construct a tree by making binary splits on factors so as to minimize the variance of the durations in the two corresponding subsets. [. . .]
> When a CART tree encounters a feature vector not observed in the training database, it can still find a path in the tree that, up to some point, matches the new feature bundle.

This means that if nothing in the database matches the required pattern exactly, then a near approximation will be found.

The labelled waveforms of the database and their XML-parsed description files are searched according to relevant feature information (e.g. syllable weight and strength), and a CART model is used to generalize across these data and generate duration statistics for feature bundles at given places in the phonological structure. The resulting duration model can be used to drive MBROLA diphone synthesis, since it predicts the durations of acoustic segments.

2The duration of the first syllable includes ambisyllabic consonants; so in a word like *city*, the /t/ is counted as belonging to both the first and the second syllables. This explains why the first syllable of words with short–long rhythm, nevertheless, is longer in duration than the second syllable.
The analysis model works top–down—that is, it factors out first the effects of IP, then of AG, and so on, down the tree to the features at the terminal level. This reflects the assumption that the IP, AG, foot and syllable are all levels of timing, and that details of lower-level differences (such as segment type) can be overlaid on details of higher-level differences (such as syllable weight and strength; the strength and weight of an adjacent syllable; etc.). The top–down model also has the effect of constraining search spaces. For instance, nuclei in [light] syllables do not split by [long:±], since no light syllable can be [long:+]; therefore in a [light] syllable, the model does not attempt to sub-divide the data by [long:±]. The resulting timing model is such that each node in the hierarchy has a multiplicative compression factor associated with it. An example of this has already been provided in Figure 6. The fact that it is a multiplicative model means that the order in which the statements of temporal interpretation are applied is irrelevant. It also makes the model compositional.

As an example, consider the interpretation of /p/ in happy. In order to interpret the /p/ accurately, the model refers to (at least) the following pieces of information:

- /p/ is located in a rhyme whose nucleus contains a short open vowel
- /p/ is [ambisyllabic:+] and is in the coda of a [strong], [light] syllable and in the onset of a [weak] syllable

Each of these facts—along with other, higher-level ones—affects the temporal interpretation of the /p/ in happy.

This method of timing assumes that segment durations, as measured from the database, are in fact what a duration model must replicate. However, another way to look at the speech signal is to consider segments as an artefact of the temporal overlaying of phonetic parameters. This view of timing has been explored in earlier work, such as Coleman (1992), Local (1992) Ogden (1992) and Local and Ogden (1997). According to this model, higher-level constituents in the hierarchy are compressed, and their daughter nodes are compressed in the same way. The temporal interpretation of ambisyllabicity is the degree of overlap that exists between syllables, so an intervocalic consonant (typically ambisyllabic) has duration properties inherited from both the syllables it is in.

The temporal consequences of ambisyllabicity can be modelled by overlaying Syllable\(_n\) on to Syllable\(_{n-1}\) thus setting Syllable\(_n\)’s start point to be before the end of Syllable\(_{n-1}\). By overlaying syllables to varying degrees and making reference to ambisyllabicity, it is possible to lengthen or shorten intervocalic consonants systematically. There are morphologically related differences which can be modelled in this way, provided that the phonological structure is sensitive to them; the mistakes and mistimes example discussed in Section 2 is one such instance. As another example, the Latinate prefix in- is fully overlaid with the stem to which it attaches and is [ambisyllabic:+], giving a short nasal in innocuous, while the roughly synonymous Germanic prefix un- is not overlaid to the same degree and is [ambisyllabic:−], giving a long nasal in unknowing. Future work will focus on integrating the segment-based and the more syllable-based approaches in the model.

5.3. Intonational modelling

Intonational modelling in ProSynth relies on the assumption, common to most theories of intonation, that the highly variable f0 contours encountered in natural speech can be analysed into component parts and classified according to a finite set of possible pitch melodies, which need to be defined phonologically. An IP is the domain for a well-formed intonation contour, which is made up of one or more pitch accents, selected from a language-specific inventory,
between initial and final boundary tones, which can be low or high. There is, therefore, a
dimension of paradigmatic choice in modelling intonation: the overall pitch pattern selected
for an IP is not itself predictable from structure but is determined by discourse factors and the
speaker’s communicative intent. Similarly, the location within the phrase of the pitch accents
themselves, notably the final pitch accent—traditionally the intonation nucleus—is partially
dependent on the relative importance assigned by the speaker to different parts of the text.

In ProSynth it is the attributes of the IP (Section 3.2) which are seen as determining the
composition of the intonation contour in terms of component AGs, the selection of pitch
accents assigned to each AG and the choice of boundary tone. The intonation pattern for an
IP is thus composed of the pitch accents assigned to AGs, and of boundary tones associated
with the edges of the IP domain. The final pitch accent, and thus the nucleus, is assigned to the
final AG within the IP; this rightmost AG is the head of the IP. The ProSynth model currently
relies on manual editing to divide the text into AGs and to assign IP and AG attributes; a
long-term objective is to develop rules to automate these procedures.

Within each [heavy] AG there will be at least one [heavy, strong] foot, in which the first
syllable is stressed (Section 3.2). If the AG has been assigned a pitch accent such as H* L,
in ToBI-style notation (Silverman et al., 1992), then this pattern will be realized over its left-
most [strong] foot. In the example: in the well-known maths department, whose hierarchical
structure is illustrated in Figure 3, the final (nuclear) AG contains two strong feet; the first of
these is the head of the AG and the domain for the nuclear pitch contour, and the syllable at
its head is “accented”, in the sense that it is the stressed syllable normally associated with the
starred tone component of the pitch accent (H* in the present example). If the IP attributes
of this phrase show it to be discourse-final and declarative, it will typically be assigned a
low initial boundary tone [interpolated through the initial degenerate AG (AG0)], a relatively
high accent in AG1, a falling nuclear pitch movement in AG2, and a low final boundary tone:
[\%L H* H*L L\%].

Figure 7 shows the f0 contour produced by the database speaker for the phrase for a
madman. The Laryngograph-derived contour is shown time-aligned with the hierarchical
structure of the text, consisting of one degenerate AG (for a) followed by one fully-formed
AG (madman), itself containing a single foot. A phonological representation of this contour
would be [%L H* H*L L\%].

The interpretation of a specified pitch contour in terms of f0 is, like other phonetic pa-
rameters, structure dependent. Precise alignment of contour turning-points is constrained by
the properties of units at lower levels in the hierarchy. A key objective of the initial analysis
of the ProSynth database was to establish the most important structural factors influencing
alignment of a given pitch accent. To this end, items in the database were all recorded with a
pattern specified as declarative, discourse-final and “neutral” in emotional content. This has
yielded a consistent “low fall” (H*L) type of accent across the varied structures. Analysis of
the relationship of the resultant contours to the structural hierarchy has enabled us to model
the mapping of contour to text, and to devise preliminary structure-dependent f0 interpreta-
tion rules. The ProSynth approach, with its emphasis on the contribution to alignment made
by onset and rhyme constituents with different properties, has much in common with the
work reported in van Santen and Hirschberg (1994) and van Santen and Möbius (1997).

In the ProSynth model, described in more detail in House, Dankovičová and Huckvale
(1999), the H*L nuclear pitch accent is defined in terms of a template based on a sequence of
f0 contour turning-points. For the falling (H*L) pitch accent, three crucial contour turning-
points were identified: Peak ONset (PON), Peak OFFset (POF) and Level ONset (LON).
The use of both PON and POF reflects the observation that the “peak” associated with H*
accents is often manifested as more of a plateau, with its own duration, than as a single peak; PON and POF represent the start and end of such a plateau, with POF therefore denoting the beginning of the f0 fall. LON occurs at the end of the fall, and is the point from which the low tone spreads till the end of voicing in the AG.

The alignment of the turning-points was automatically derived from the Laryngograph recording used to calculate the f0 trace, and checked using informal listening tests to ensure that there was perceptual equivalence between natural f0 contours and those constructed by linking the target points identified. The following procedure was used to locate PON, POF and LON automatically. First, the absolute peak of the f0 contour in each phrase-final AG was found. Then, the value corresponding to 4% below the peak value was calculated, to delimit a range of frequencies which would be perceived at an approximately equal pitch by listeners (Rosen & Fourcin, 1986). This delimiting value defines the points in time preceding and following the absolute f0 peak that are designated PON (Peak ONset) and POF (Peak OFFset), respectively, as schematized in Figure 8. To determine the location in time of LON, the mean f0 value for the final 50 ms in the phrase was calculated. This f0 value served as the baseline for the overall range of the contour, with the absolute peak f0 as the maximum. LON was identified as the earliest point after the peak at which the f0 contour achieved the value equivalent to only 25% of this overall range. The alignment of PON, POF and LON could then be clearly related to the key syllable constituents, as labelled by the hand-corrected automatic segmentation.

Statistical analysis of the database suggests that the timing of all these points varies systematically with aspects of the structure of the foot where the pitch accent is realized. A particularly important factor is the length of the foot in terms of the number of component syllables. When alignment of the turning-points was measured in relation to the accented syllable within the foot, a significant shift to the right was observed for contours realized over a two-syllable foot compared with those in a monosyllabic foot. This is consistent with findings reported elsewhere (Bruce, 1990; Silverman & Pierrehumbert, 1990; Prieto, van Santen & Hirschberg, 1995). The rightward shift was not itself dependent on the internal structure of the accented syllable, since it was observed across feet containing a wide range of syllable structures (House et al., 1999). However, a re-analysis of the alignment, taking the complete
foot rather than the accented syllable as the relevant temporal domain, removed much of the variability that derived from foot structure. The preliminary rules therefore treat the foot as the primary domain for template realization (see also the use of “accent groups” as a domain in Möbius, 1995; van Santen & Möbius, 1997); they also specify the turning-points as attributes of the leftmost foot (the head of the AG) within the AG. If analysis of further data supports the use of the foot as a primary domain for modelling f0 it will be a considerable advantage in simplifying rules for synthesis. The analysis is currently being extended to a wider range of foot structures, including three-syllable feet.

Characteristics of the onset and rhyme of the accented syllable also have significant effects on f0 alignment. The effects of onset type, coda type and of interactions between the two are reported in House et al. (1999). Findings about the relationship between the peak (PON and/or POF) and the accented syllable which depend on syllable structure are consistent with those reported elsewhere (e.g. van Santen & Hirschberg, 1994; Rietveld & Gussenhoven, 1995, see also House & Wichmann, 1996; Wichmann & House, 1999, for a summary of the within-category alignment literature).

It is claimed that successful modelling of the speaker’s observed f0 values, and their integration with the models of timing and spectral properties that we are simultaneously deriving from the database, enhances the coherence of the synthesized output. At least some of the systematic variability which derives from structures is now explicitly modelled.

6. Perceptual evaluation

6.1. Introduction and general methods

Four preliminary perceptual experiments establish that the principles being implemented do increase the intelligibility and/or naturalness of synthetic speech. Experiments 1 and 2 tested the intelligibility in noise of segmental detail and timing, respectively. Experiments 3 and 4 tested aspects of intonation: experiment 3 assessed naturalness, or more properly, neutralness;
experiment 4 measured reaction time to answer questions about read stories. This section describes properties common to two or more experiments, while Sections 6.2–6.5 report the individual experiments. The four experiments are discussed together in Section 6.6

6.1.1. Materials
The experimental material consisted of phrases in the ProSynth database. Experiments 1, 2 and 3 used different sets of phrases; experiments 3 and 4 used the same set. In all four experiments, the phrases were synthesized in two forms, with and without the experimental manipulation of interest. Experiment 1 tested the influence of appropriate vs. inappropriate excitation at the boundary between vowels and obstruents. Experiment 2 compared the intelligibility of phoneme sequences with appropriate and inappropriate durations for the particular foot and syllable structure: inappropriate segmental durations were taken from the same phoneme sequences in different foot and syllable structures. Experiments 3 and 4 compared intonation, specifically the alignment of turning points in the f0 contour relative to the accentuated syllable of the last foot. Experiment 1 used formant synthesis while Experiments 2, 3, and 4 used MBROLA synthesis.

6.1.2. Subjects
A total of 53 listeners (Ss) was used, drawn from students and staff at the University of Cambridge. All were native speakers of Southern British English, or, while native to another part of Britain, had lived in the South of England for at least 3 years. None reported speech or hearing problems; 15 had phonetic training, varying from expert to undergraduate only. Most (37) were aged between 18 and 24 years; three were aged around 50, and the rest were less then 34 years old. Almost half (25) served in more than one experiment, but only two served in all four. Ss who served in more than one experiment completed them in a single session, except for the two who served in all experiments, for whom experiment 4 predated experiment 3 by 5 months. Order of presentation of the experiments within a session was balanced across subjects, except that, for design reasons, experiment 4 always preceded experiment 3, and was never the first or last experiment in a session.

6.1.3. Procedure
Synthesized utterances were presented to listeners over high-quality headphones, using a Tucker-Davis DD1 D-to-A system from a PC computer, and a comfortable listening level. Listeners were tested individually in a sound-treated room. Where stimuli were presented in noise (experiments 1 and 2) signal-to-noise ratios were chosen to produce error rates that would avoid ceiling and floor effects, as determined in pilot tests.

6.2. Experiment 1
Segmental detail: excitation type at vowel–obstruent boundaries

6.2.1. Introduction
This experiment assessed whether natural-sounding excitation near segment boundaries enhances the intelligibility of formant synthesis. Observations from the ProSynth database showed systematic variation in the incidence of (a) mixed periodic and aperiodic excitation at boundaries between vowels and voiceless fricatives, and (b) the duration of periodicity in the closures of voiced stops (Heid & Hawkins, 1999). In brief, most vowel–fricative (VF)
boundaries have mixed aperiodic and periodic excitation, whereas most fricative–vowel (FV) boundaries change abruptly from aperiodic to periodic excitation. Syllable stress, vowel height, and final/non-final position within the phrase influence the incidence and duration of mixed excitation. Similarly, the duration and proportion of voicing in the closures of phonologically voiced stops depend systematically on vowel height, place of articulation, stress context and the syllabic role of the stop. It was predicted that synthesized phrases would be more intelligible in noise when they conformed to the natural patterns of excitation for the particular structural context, because they would add both to the signal’s perceptual coherence and to its informativeness about linguistic structure. Superficially uncontroversial, it was not in fact obvious that this prediction would be supported because the differences involved are small: the mean duration of mixed periodic and aperiodic excitation at a VF boundary in the ProSynth database is only 18 ms, for example.

6.2.2. Material

Eighteen phrases from the database were copy-synthesized into formant synthesizer, HLsyn, using PROCSY (Heid & Hawkins, 1998, in press) at 11 025 kHz SR, and hand-edited to a good standard of intelligibility, as judged by a number of listeners. In 10 phrases, the sound of interest was a voiceless fricative, with structural contexts defined as follows: at the onset of a stressed syllable—*in a field*; unstressed onset—*it’s surreal*; unstressed coda before a stressed syllable—*to disrobe*; coda between unstressed syllables—*disappoint*; coda of a final stressed syllable—*on the roof*; his *riff*, a *myth*, at a *loss*; to *clash*; at both unstressed and stressed onsets—*fulfilled*. The other eight items had voiced stops as the focus: in the coda of a final stress syllable—*it’s mislaid*, he’s a *rogue*, he was *robbed*; stressed onset—*in the band*; unstressed onset—*the delay*, to *be wronged*; unstressed and final post-stress context—*to deride*; and in the onset and coda of a stressed syllable—*he begged*.

The sound of interest was synthesized with the “right” type of excitation pattern at its boundaries. From each right version, a “wrong” one was made by substituting at just one boundary a type or duration of excitation that was inappropriate for the structural context. In each case, the boundary manipulated is that of the structural context described above. For fricatives between sonorants, either the VF or the FV boundary was manipulated. For stops, only voicing during closure was manipulated. Changes were systematic; no attempt was made to copy the exact details of the natural version of each phrase, as the aim was to test the perceptual salience of the type of change, with a view to incorporating it in a synthesis-by-rule system.

At FV boundaries, the right version had simple excitation (an abrupt transition between aperiodic and periodic excitation), and the wrong version had mixed periodic and aperiodic excitation. VF boundaries had the opposite pattern: wrong versions had no mixed excitation. Each stop had one of two types of wrong voicing: longer-than-normal voicing for *band* and *begged* whose onset stops normally have a short proportion of voicing in the closure; and unnaturally short voicing in the closures of the other six words. The wrong versions of *band* and *begged* were classed as hyper-speech and expected to be more intelligible than the right versions. The other six were expected to be less intelligible in noise if naturalness and intelligibility co-varied.

6.2.3. Subjects

There were 22 Ss.
6.2.4. Procedure

The 18 experimental items were mixed with randomly-varying cafeteria noise at an average s/n ratio of +4 dB relative to the maximum amplitude of the phrase. Ss pressed a key to hear each item, and wrote down that they heard. Each S heard each phrase once: half the phrases in the right version, half wrong or hyper-speech. The order of items was randomized for each listener separately, and, because the noise was variable, it too was randomized separately for each listener. Five practice items preceded each test.

6.2.5. Results

Responses were scored for number of phonemes correct on three phonemes: the manipulated one and the two adjacent to it. Wrong insertions in otherwise correct responses counted as errors. Excluding the two hyper-speech phrases, responses were significantly better for the right versions using a one-tailed paired t-test (69% vs. 61%, \( t(21) = 2.35, p = 0.015 \)).

Responses to the two hyper-speech words were informative: an increase in intelligibility for normal vs. hyper-speech begged (84% vs. 89% correct); but a decrease (85% vs. 76% correct) for normal vs. hyper-speech band. Hyper-speech in the band was often misheard as in the van. Implications of this difference for the use of synthetic hyper-speech are developed below.

6.3. Experiment 2

Timing: duration of rhymes in strong syllables in phrase-final feet

6.3.1. Introduction

This initial test of ProSynth’s hypotheses about temporal structure and its relation to prosodic structure assessed whether listeners’ ability to understand synthetic speech is influenced by rhythmic effects that depend on the [weight] and [length] of the rhyme, and whether or not their codas are [ambisyllabic]. Rhyme [length] was included in the materials since it affects not only the duration of the nucleus but also that of the coda. Timing is known to be crucial to intelligibility; the issue here is whether the small, structurally-sensitive temporal differences which ProSynth predicts for syllable rhymes produces gains in intelligibility. Accordingly, pairs of phrases were synthesized which were identical except for the duration of the first rhyme of the final foot. Durations in the rhyme rather than the whole syllable were manipulated because the rhyme is the domain over which syllable weight operates.

6.3.2. Material

Twelve different linguistic structures were chosen, each with two exemplars, making a total of 24 pairs of phrases in all, as listed in Table II. Each phrase was synthesized with an approximation to its natural f0 in the ProSynth database. A “right” and a “wrong” version of each phrase was produced as follows. In the right version, segmental durations of the first part of the phrase, up to and including the onset of the first syllable of the last foot, were copied from the natural utterance. Segmental durations for the rest of the final foot were those predicted by the ProSynth model for the particular linguistic structure. The wrong version of each phrase was made by exchanging the ProSynth-predicted segment durations of the strong rhyme between the two phrases in each pair, where those segments were identical. So the durations for ob in he’s a robber were replaced by the durations for ob in to rob them and vice versa. In cases where the segment strings of two [strong] rhymes were not identical (as in /st/ and
Table II. Pairs of phrases used in experiment 2, showing structural similarities (Common attributes) and differences (Contrast). In “wrong” stimuli, the durations of the underlined part of each phrase in the column headed Phrase 1 were exchanged with those of the underlined part in the same row of the Phrase 2 column, and vice versa. Phrases used in more than one comparison are asterisked.

<table>
<thead>
<tr>
<th>Structural properties</th>
<th>Common attributes</th>
<th>Contrast</th>
<th>Stimulus pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>light</td>
<td>short</td>
<td>ambi vs. nonambi</td>
<td>he’s a robber to be loving to rob them</td>
</tr>
<tr>
<td>heavy</td>
<td>short</td>
<td>ambi vs. nonambi</td>
<td>*they were resting *it was risky to arrest them</td>
</tr>
<tr>
<td>heavy</td>
<td>long</td>
<td>ambi vs. nonambi</td>
<td>for a climber to roof it to climb them</td>
</tr>
<tr>
<td>heavy</td>
<td>short</td>
<td>ambi vs. partlyambi</td>
<td>*in a hospice *it was risky we dropped it</td>
</tr>
<tr>
<td>heavy</td>
<td>long</td>
<td>ambi vs. partlyambi</td>
<td>to mine it with a needle to remind us for a fielder</td>
</tr>
<tr>
<td>heavy</td>
<td>short</td>
<td>partly ambi vs. nonambi</td>
<td>they banned it they banned me it was selfless</td>
</tr>
<tr>
<td>heavy</td>
<td>long</td>
<td>partly ambi vs. nonambi</td>
<td>she returned it *we recalled it</td>
</tr>
<tr>
<td>ambisyllabic</td>
<td>short</td>
<td>light vs. heavy</td>
<td>in a letter to stop it *they were resting *in a hospice</td>
</tr>
<tr>
<td>non-ambisyllabic</td>
<td>short</td>
<td>light vs. heavy</td>
<td>to stop them we dropped them to belt them</td>
</tr>
<tr>
<td>heavy</td>
<td>ambi-syllabic</td>
<td>short vs. long</td>
<td>*in a hospice *it was risky for roasting very thirsty</td>
</tr>
<tr>
<td>heavy</td>
<td>partly ambisyllabic</td>
<td>short vs. long</td>
<td>it’s a splinter we repelled it they were dainty</td>
</tr>
<tr>
<td>heavy</td>
<td>non-ambisyllabic</td>
<td>short vs. long</td>
<td>to want them with a gauntlet it was needless</td>
</tr>
</tbody>
</table>

/εt/, or /ain/ and /aind/), the durations of the same or similar segments were exchanged. So durations for /εt/ in to get them were replaced by the durations of /ε/ and /t/ in to belt them and vice versa; durations for /am/ in to mine it were replaced by durations of /am/ in to remind us and vice versa. In Table II, the segments whose durations were exchanged are underlined.

Table III shows the durational differences between right and wrong stimuli, and between each of these and the durations produced by Klatt’s (1979) rules. For each of the nucleus, coda and rhyme there are three rows, representing: (1) the difference between the right and wrong ProSynth durations; (2) the difference between the right ProSynth durations and those of Klatt; (3) the difference between the wrong ProSynth durations and those of Klatt. The leftmost two columns of numbers show the mean and standard deviation, respectively, of the actual (signed) differences, while the rightmost two show the mean of the absolute values of differences. For the signed values, the mean durations of the ProSynth-predicted nuclei are very slightly shorter than those predicted by Klatt, while the codas and consequently the
Table III. Mean differences in duration between nuclei, codas and rhymes exchanges in experiment 2, excluding six pairs in which the contrast was between long and short nuclei. Right–wrong: difference between “right” and “wrong” nucleus (or coda, or rhyme) durations in ProSynth-synthesized phrases. Right-Klatt: difference between nucleus/coda/rhyme in right ProSynth phrases and the output of Klatt’s (1979) rules. Wrong-Klatt: difference between nucleus/coda/rhyme in wrong ProSynth phrases and Klatt’s (1979) rules. Mean difference: calculated from the actual differences between pairs. Mean absolute difference: calculated from the absolute values of the differences between pairs

<table>
<thead>
<tr>
<th>Constituent</th>
<th>Comparison</th>
<th>Mean differences (ms)</th>
<th>s.d.</th>
<th>Mean absolute difference (ms)</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td>right–wrong</td>
<td>-1</td>
<td>31</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>right-Klatt</td>
<td>-5</td>
<td>37</td>
<td>27</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>wrong-Klatt</td>
<td>-4</td>
<td>38</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>Coda</td>
<td>right–wrong</td>
<td>0</td>
<td>30</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>right-Klatt</td>
<td>16</td>
<td>38</td>
<td>34</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>wrong-Klatt</td>
<td>16</td>
<td>34</td>
<td>31</td>
<td>21</td>
</tr>
<tr>
<td>Rhyme</td>
<td>right–wrong</td>
<td>-1</td>
<td>42</td>
<td>35</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>right-Klatt</td>
<td>13</td>
<td>44</td>
<td>33</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>wrong-Klatt</td>
<td>14</td>
<td>44</td>
<td>34</td>
<td>31</td>
</tr>
</tbody>
</table>

Rhymes are slightly longer than those predicted by Klatt. In absolute terms, the mean difference between the right and wrong ProSynth-predicted durations is 22 ms for both nucleus and coda. This value is rather smaller than the differences between Klatt’s durations and either of ProSynth’s, which are similar to one another. That is, for both nucleus and coda, absolute mean differences for right-Klatt and wrong-Klatt are similar to one another, and rather larger than the right–wrong differences. These figures suggest that the so-called wrong durations are only slightly inappropriate compared with the right ProSynth-predicted durations; and that neither the right nor the wrong durations are very different from Klatt’s durations. In relation to normal speech synthesis standards, the test is thus of subtle rather than gross rhythmic effects.

6.3.3. Subjects

Data are presented for 25 Ss. Data from one further S were excluded because she misunderstood the instructions.

6.3.4. Procedure

The 96 stimuli (12 structures × 2 exemplars × 2 pair members × 2 right/wrong versions) were mixed with randomly-varying cafeteria noise and presented to two groups of listeners, following the procedure described for experiment 1 (Section 6.2.4). Differences from experiment 1 were that signal-to-noise ratio was +5 dB, and six practice items preceded each test.

6.3.5. Results

Since the manipulations affect the rhythm of the whole phrase, intelligibility was assessed by scoring phonemes correct for each entire phrase. Responses were about 4% better for the right
versions (79% vs. 75%). This improvement, though small, is strongly significant in a one-tailed paired \( t \)-test on the mean right vs. wrong scores for each subject: \( t(24) = 3.13, p = 0.0023 \). To provide a more stringent test, a second \( t \)-test was done which omitted the items in which the manipulation involved an exchange of a long with a short vowel (see Table II). Although such long–short exchanges affect coda duration, their primary effects are in the nucleus, and are large relative to the differences between members of the other pairs; they might therefore be expected to have a disproportionate influence on the results. This proved not to be the case. When the long/short exchanges are omitted, the mean difference between right and wrong versions drops by only 0.3% (from 3.711 to 3.405), and remains strongly significant \( (t(24) = 2.71, p = 0.0061) \). Thus, even with the long-short data excluded, the correct rhythm engendered better phoneme intelligibility, suggesting that even rather subtle temporal patterns enhance intelligibility when modelled systematically.

### 6.4. Experiment 3

#### Intonation: \( f_0 \) alignment (naturalness/neutralness)

#### 6.4.1. Introduction

This experiment investigated the effect of varying the alignment of an intonation contour on listeners’ judgements of the naturalness of an utterance as neutral, declarative and discourse-final. The alignment shift tested is an example of systematic structural variation determining the realization of a single pitch accent pattern in a given context.

Preliminary \( f_0 \) modelling on neutral, declarative, discourse-final utterances in the ProSynth database showed a statistically significant difference in alignment of the \( f_0 \) contour dependent on the type of foot. The \( f_0 \) turning points of an H*L pitch accent occur consistently later in an accented syllable when it is part of a disyllabic rather than a monosyllabic foot. This rightward shift is not obviously dependent on the internal structure of the accented syllable, since it was observed across a wide range of structures (House et al., 1999). It was hypothesized that phrases would be judged more natural when the \( f_0 \) alignment was appropriate for the foot structure: in other words, that the \( f_0 \) alignment is perceptually salient. If the hypothesis were supported, then the implication would be that synthesis of intonation should take account of foot structure.

#### 6.4.2. Material

The stimuli were 32 pairs of utterances, each with a final, monosyllabic foot, for example the terrain; he was mad; it’s a lie. Segmental durations within each MBROLA-synthesized stimulus matches those of the original utterance in the database. Before the final foot, \( f_0 \) was also sampled from the original utterance: at every 12% of the segment duration for a vowel, and at every 25% of the segment duration for a consonant. In the final foot, values were specified only at the turning points of the \( f_0 \) template, with linear interpolation between them.

The two members of each pair were identical except for the alignment of POF (Peal Off-set) and LON (Level ONset), which was either “right” or “wrong”. In right items, POF and LON alignment was appropriate for the monosyllabic final foot. In wrong items, it was modified to follow the pattern appropriate for the same syllable in a disyllabic final foot. For the three examples above, the respective utterances with disyllabic feet were it was raining; for a madman; they were lying. The particular \( f_0 \) manipulation was one of four types, depending on the properties of the onset and coda in the accented syllable, as summarized in Table IV.
TABLE IV. Types of utterance pairs used in experiments 3 and 4. V-initial, V-mid, V-final: start, midpoint, and, end, respectively, of the periodic segment corresponding to the vowel. Absent: turning-point was not present in original monosyllabic utterance. Various: miscellaneous onsets; if a Group (iv) onset is sonorant, the coda is voiced [whereas Group (iii) onsets have voiceless codas]

<table>
<thead>
<tr>
<th>Utterance pairs</th>
<th>Onset type</th>
<th>Coda type</th>
<th>F0 alignment: right (monosyllabic model)</th>
<th>F0 alignment: wrong (disyllabic model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group (i) 9 pairs</td>
<td>voiced (+/−sonorant)</td>
<td>single sonorant; or sonorant + obstruent cluster</td>
<td>V-initial</td>
<td>V-final</td>
</tr>
<tr>
<td>Group (ii) 5 pairs</td>
<td>voiceless</td>
<td>single sonorant; or sonorant + obstruent cluster</td>
<td>V-initial</td>
<td>V-final</td>
</tr>
<tr>
<td>Group (iii) 6 pairs</td>
<td>sonorant</td>
<td>all voiceless obstruent(s) or no coda</td>
<td>V-initial</td>
<td>absent</td>
</tr>
<tr>
<td>Group (iv) 12 pairs</td>
<td>various (+/−voice)</td>
<td>all obstruent(s)</td>
<td>V-initial</td>
<td>Vid-mid</td>
</tr>
</tbody>
</table>

6.4.3. Subjects

There were 11 subjects in all. Statistics are reported for 10, six of them phoneticians. Results for the remaining S, hereafter S11, an expert phonetician, are commented on separately since he was given different instructions.

6.4.4. Procedure

The 32 pairs of phrases were randomized in 10 blocks of 32, with the “right” item coming first on five occurrences and second on the other five. A separate randomization was made for each S. The interval between the two members of a pair was 200 ms; intertrial (response) interval was 2.5 s, with a pause after every twelfth trial, when the listener could rest. The next trial started when the listener terminated the pause by pressing a key on a computer keyboard. The 320 test trials were preceded by 12 practice items, which were randomly chosen but the same for all Ss. All Ss were told that the members of a pair differed only in melody (for non-phoneticians) or intonation (for phoneticians), and that they should focus only on that and ignore all other properties. Ten of the Ss were instructed to press one of two buttons, depending on whether they judged the first or the second member of a pair to sound more natural in the sense of neutral, factual, cool and normal, yet without abnormally low emotion; even if the particular words in some utterances meant that a livelier, more emphatic or more excited pronunciation sounded more appealing than a more neutral one, they should choose the more neutrally-spoken item. S11 was simply asked to choose the version he preferred.

6.4.5. Results

For the main group of 10 Ss, 78% of the responses favoured right items, and 22% wrong items. A paired t-test comparing mean responses for each S confirmed that right items were preferred significantly more often than chance (78% vs. 50%; t(9) = 5.71, p < 0.0002). Unsurprisingly, phoneticians were more consistent in their choices than non-phoneticians, but the preference for right rather than wrong intonation patterns is significantly better than
chance for each subgroup. For phoneticians, 86% preference for right items \( t(5) = 6.09, p < 0.0009 \); for non-phoneticians, 66% preferences for right items \( t(3) = 4.01, p = 0.014 \). S11, asked to choose his more preferred version, was equally consistent in the opposite direction, with only 11% preferences for right items. This difference underlines the fact that the instructions given in intonation experiments can strongly influence Ss’ judgements.

6.5. Experiment 4

Intonation: f0 alignment (speed of comprehension)

6.5.1. Introduction

Naturalness judgements are both less informative and less central to the main purpose of ProSynth research than assessments of whether a given intonation pattern facilitates listeners’ understanding of the message. However, standard intelligibility tests are also unsuitable in that phoneme error rate seems too insensitive a measure of the subtleties that intonation can convey. It is well known that given the same words and context, a number of different intonation patterns can convey the same message, and yet both the exact words and their context affect the interpretation of a given intonation pattern. We are therefore developing a test that assesses the influence of f0 manipulations on listeners’ comprehension of utterances heard in context. The more appropriate the intonation is to the context, the more easily listeners should be able to process the information, and this should be reflected in lower error rates and/or faster reaction times in answering questions.

A pilot experiment has been conducted to explore the potential of such a comprehension test to evaluate intonation. Since the design needs refinement, the experiment is described only briefly here. Listeners read a story, then decided whether answers to questions about the story were true or false, and responded accordingly by pressing the appropriate one of two buttons. The number of correct responses and reaction time (RT) were measured. Questions appeared one at a time on a computer screen; answer to the questions were the same stimuli used in experiment 3. Each of the 32 phrases appeared once as a true answer to a question, and once as a false answer. Each of 36 Ss heard each answer only one: either as a true or a false answer (to questions about different stories), and with either the right or the wrong intonation pattern (for the answer to the same question about the same story).

The results are promising in that right intonation patterns produced faster RTs than wrong ones for many utterances, but the difference was not statistically significant overall. Moreover, since RTs to true answers were significantly faster than to false answers, as expected, the design seems sufficiently sensitive to warrant more work, as discussed below.

6.6. Discussion of perceptual evaluations

Experiments 1–3 demonstrate the perceptual importance of correctly modelling acoustic–phonetic fine detail in synthesis for spectral, temporal and intonational phenomena: they confirm that listeners make use of variation that systematically reflects prosodic structure.

Experiment 1 showed that phoneme identification is better when the pattern of excitation at segment boundaries is appropriate for the structural context. Considering that only one acoustic boundary was manipulated in most of the phrases used in experiment 1, and that there are relatively few data points, the significance levels achieved testify to the importance of synthesizing segment boundaries that are appropriate to the phonological context.

The data of experiment 1 do not distinguish between whether the right versions are more intelligible because the manipulations enhance the acoustic and perceptual coherence of the
signal at the boundary, or because they provide information about linguistic structure. The two possibilities are not mutually exclusive in any case. The data do suggest, however, that one reason for the appeal of diphone synthesis is not just that segment boundaries sound more natural, but that their naturalness may make them easier to understand, at least in noise. It thus seems worth incorporating fine phonetic detail at segment boundaries into formant synthesis. It is relatively easy to produce these details using HLsyn.

Experiment 2 showed that a structurally-sensitive temporal model leads to significantly improved intelligibility in synthetic speech. As rhythm depends on relative timing throughout the whole phrase, the results reported are for the intelligibility of the whole phrase, including parts which were the same in both members of a stimulus pair. However, this decision to score intelligibility over the whole phrase when only one syllable rhyme was changed could reduce differences due to localized durational manipulations. Furthermore, the MBROLA diphone synthesis, with its limited set of diphones, does not produce the segmental effects predicted by ProSynth. For example, MBROLA produces aspirated plosives in words like roast\[\text{h}\]ing where the ProSynth model predicts non-aspiration. Such weaknesses are bound to lower the phoneme recognition rate either because the rhythm is disrupted or because inappropriate allophones may lead listeners to segment the signal into words incorrectly. We can thus conclude that quite small rhythmic differences will strongly influence intelligibility in noise when they reflect systematic structural differences.

Experiment 3 showed that structurally appropriate f0 alignment significantly affected the way listeners interpreted the context of an utterance. There are a number of ways to model these differences in f0 alignment. Rules can change the alignment relative to the accented syllable, or they can apply at the level of the foot. ProSynth uses the foot.

The f0 manipulations described in Section 6.4.2 did not produce wrong stimuli which sounded phonetically unnatural; rather, they sounded like plausible utterances spoken in a different context from the originals on which the right versions had been modelled. The low final boundary tone was retained in both members of each stimulus pair; the frequency value of the f0 peak was kept constant, but its alignment later in the accented syllable meant that in wrong versions a longer proportion of the accented vowel was produced with the pitch of the peak itself, producing a plateau with essentially the same f0 value. The percept which resulted from the alignment shift was of a falling tone realized over a wider pitch range: a “high” fall. High falls are associated with more emotional contexts, or with emphasis or contrast, and they do not show the lowering of pitch range that often marks discourse-final position. The data show that the difference in the two patterns is perceptually salient. However, there is no sharp category boundary between high and low falls: both may be phonologized as H∗L. It thus seems reasonable to conclude that synthetic intonation contours must reflect foot structure if a consistent affect is to be conveyed over feet of different lengths.

In addition to their generally positive confirmation of the value of ProSynth’s approach, these perceptual data highlight a number of issues in the design of perceptual evaluation tests. By far the most challenging is the evaluation of intonation (cf. Sonntag & Portele, 1998). There are obvious advantages in evaluating the influence of intonation pattern by assessing its influence on performance of an independent task. However, experiment 4 demonstrates some of the practical difficulties in doing so. It is possible that the differences in intonation pattern modelled were in general too subtle to produce a significant difference in experiment 4. However, the test was not ideal in a number of ways, and the data suggest that it is worth refining the present design. If the intonational effects are subtle, they may show up better when the segmental and timing quality is better than in the MBROLA stimuli used. To reduce irrelevant response variance, we will use stimuli that are more homogeneous in difficulty and
that impose a smaller cognitive load than in the present task, so that their truth value is easier to evaluate. Furthermore, Ss will be forced to respond fast. Faster responses should not only reduce RT variances but also produce more errors, which would allow analyses on errors as well as RTs. The advantages of assessing intonation indirectly justify refinement of this experimental technique.

The fact that there is a wide range of natural-sounding possibilities for the realization of intonation patterns underlines the need to provide a clear context in perceptual evaluations. Some inconsistencies in the data may be due to subjects using different notions of appropriate context. Experiment 3’s S11, asked simply which stimulus he preferred, with no context provided, usually chose the wrong members of each pair. Out of context, some wrong members sounded more lively and interesting, while others (e.g. in the act) invite a non-neutral intonation because they are rarely spoken as isolated phrases. Such a result confirms that if naturalness judgements are collected without controlling the assumed context of utterance we risk acquiring uninterpretable data. The issue of context may also be relevant to the inconclusive results of experiment 4. For some stimuli, the predicted RT differences support the hypothesis that a “discourse-final” intonation speeds utterance processing. However, for some of the answers that subjects evaluated as true or false, a more contrastive type of intonation could have been more helpful.

The two instances of hyper-speech in experiment 1 further demonstrate that at times the speech style that is most appropriate to the situation is not necessarily the most natural one. By increasing the duration of closure voicing in stressed onset stops, we imitated what people do to enhance intelligibility in adverse conditions such as noise or telephone bandwidths. However, this manipulation risked making the /b/s sound like /v/s, effectively widening the lexical neighbourhood of band to include van. Since in the van is as predictable a phrase as in the band, contextual cues could not help in these short phrases, and band’s intelligibility fell. Begged’s intelligibility may have risen because there were no obvious lexical competitors, and because the voicing in the syllable coda was also enhanced, thus making a more extreme hyper-speech style, and, perhaps crucially, a more consistent one. These issues need more work.

In summary, these experiments provide a first step towards assessing the validity of the theory that underpins ProSynth. In themselves they do not extend our knowledge of linguistic structure, but they do suggest that there will be gains in intelligibility or naturalness whenever detail that reflects linguistic structure is added to the signal. Presumably, the more systematic fine detail that is added, the more robust the signal. Thus, if local changes in excitation type at segment boundaries enhance intelligibility significantly, then systematic attention to boundary details throughout the whole of a synthetic utterance will considerably enhance its robustness in noise. Similarly, if finer temporal or f0 detail in the accented rhyme of a phrase-final foot enhances intelligibility and/or acceptability, then detailed modelling of the entire phrasal rhythm and intonation should provide further gains. Future experiments will assess the combined influence of accurately modelled, structurally-sensitive spectral, temporal and intonational fine detail.

7. Conclusion

This paper describes a model of speech synthesis that assumes that there is systematic phonetic detail in natural speech which in turn signals details of linguistic structures to listeners. We have shown that if synthetic speech models at least some of these details, it is more ro-
ProSynth: an integrated prosodic approach to speech synthesis

bust in adverse listening conditions, even (as in the case of the experiments with MBROLA synthesis) when the quality of the synthesis is not very high.

The ProSynth model assumes that speech is informationally rich both phonologically and phonetically. The linguistic representation used to generate synthetic speech contains a great deal of structural information, encoding hierarchical linguistic relations. Phonological information is distributed across the structure, making it possible to give a phonetic interpretation to structures. This linguistic structure is thus made audible in the spoken stream of speech. Systematic acoustic variation in spectral shape, timing, rhythm and intonation all contribute to making speech a coherent whole; and these details can be used by listeners as subtle but useful cues to decoding the speech signal successfully. Perceptual experiments suggest that modelling linguistic structure and its associated systematic phonetic detail in the ways we have suggested makes significant differences to the intelligibility and perceived naturalness of synthetic speech in poor listening conditions.

This work is supported by EPSRC grant numbers GR/L53069 (Cambridge), GR/L51829 (York) and GR/L52109 (UCL). We thank Daniel Hirst for discussions about the evaluation of intonation.

References


ProSynth: an integrated prosodic approach to speech synthesis

ProSynth: an integrated prosodic approach to speech synthesis

Horwood, Chichester.
Wichmann, A. & House, J. (1999). Discourse constraints on peak timing in English: experimental evidence. In...

(Received 2 November 1999 and accepted for publication 21 February 2000)