PROCSY: A HYBRID APPROACH TO HIGH-QUALITY FORMANT SYNTHESIS USING HLSYN *

Sebastian Heid and Sarah Hawkins

Phonetics Laboratory
Department of Linguistics
Sidgwick Avenue
Cambridge CB3 9DA
U.K.

May 1999

Abstract

PROCSY is a hybrid method of automatically producing natural-sounding formant-based synthetic speech from an existing speech signal by using copy-synthesis and estimated articulatory trajectories as input to the HLsyn^{TM} synthesizer. The purpose is to allow controlled manipulation of selected acoustic parameters. Parameters for HLsyn are derived from prosodically parsed and labelled speech files in two ways. Broadly, vowels and approximants are copy-synthesized from the acoustic signal, while obstruents and nasals are synthesized by rule: articulatory trajectories and constriction areas are estimated from the segment label and duration, together with attributes such as syllable stress where relevant, and converted into HL parameter values. HLsyn combines information from both sources to calculate parameter values for a Klatt-type synthesizer. Strengths of the method are (i) simple HLsyn input captures acoustically complex obstruents, and (ii) HLsyn parameters automatically produce complex acoustic properties that accompany consonantal closures, especially at segment boundaries. These properties are hard to synthesize and thus typically absent in formant TTS, yet they provide some of the systematic variability we hypothesize contributes to robust, natural-sounding synthesis. Potential applications are discussed.

*Expanded version of the paper presented at the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, Jenolan Caves, Australia, Nov. 1998. The present version is submitted for publication.

1 INTRODUCTION

The work described here is part of ProSynth (Hawkins, House, Local, and Ogden (1998), *http://synth.phon.ucl.ac.uk/prosynth/*), a research program to develop a linguistically-informed, device-independent text-to-speech (TTS) system. Hence this work reflects some of ProSynth's specific requirements, but it also has general applications.

ProSynth's motivating hypothesis is that the intelligibility of synthetic speech under adverse listening conditions will only approach that of natural speech when the synthesizer reproduces the fine acoustic-phonetic detail that reflects the systematic variation of natural speech. This systematic variation can reflect subtle acoustic consequences of vocal-tract behaviour, and the detailed linguistic structure of an utterance.

Our position, developed in more detail in Local (1992), Local and Ogden (1997), Hawkins (1995), is partly based on the finding that even when formant-based synthetic speech is about as intelligible as natural speech in good listening conditions, it is much less intelligible in noise: natural speech is about 15% less intelligible at 0 dB s/n than in quiet, whereas synthetic speech can drop by 35%-50% (Pratt (1986)). We outline our argument here in order to explain why PROCSY depends on HLsyn rather than on standard formant synthesis.

We conjecture that the fragility of synthetic speech in noise is related to its unnatural quality. The tight relationship between vocal-tract behaviour and the properties of the emitted sounds make natural speech acoustically coherent and hence perceptually coherent, by which we mean that its acoustic-phonetic fine detail reflects vocal tract behaviour and identifies the signal as coming from one person. This fine detail is found in all aspects of speech, e.g. in correlations between glottal waveshape and upper articulator behaviour, especially at abrupt segment boundaries; in the amplitude envelope governing perception of rhythm and of 'integration' between stop bursts and following vowels; and in long- and short-domain coarticulatory effects on formant frequencies. Effects of these types contribute to signal variability, but they do so systematically, adding information rather than noise.

Some aspects of perceptual coherence, such as local coarticulation of formant frequencies across segment boundaries, are fundamental to basic intelligibility, and TTS systems include them. Others, generally absent from TTS systems, provide naturalness that makes real speech easier to understand in adverse listening conditions. For example in some contexts, "long-domain resonance effects" (Kelly and Local (1989)) due to a particular consonant provide weak but consistent acoustic cues to the consonant's phonemic identity over several syllables. When rule-based synthetic speech includes such long-domain coarticulatory variation due to consonants such as /r/ and /l/, phone identification for real and nonsense words in noise can improve by around 15% (Hawkins and Slater (1994), Tunley (1999)). Interestingly, it is not just the phoneme whose resonance effects are modelled that sounds more intelligible, but all of the affected phones, presumably because the whole sequence reflects the patterns of the naturally spoken utterance better. Similarily, amplitude envelope and excitation at segment boundaries contribute to the signal's acoustic coherence and improves its intelligibility (e.g. Heid and Hawkins (1999)). These and other sources of evidence encourage our position that, to understand speech, listeners use all available sensory information in proportion to its reliability (cf. Warren and Marslen-Wilson (1987), Marslen-Wilson and Warren (1994), Fixmer and Hawkins (1998)). It follows that at the presence of such "redundant" long-domain and local information in any type of synthetic speech should significantly improve its naturalness and intelligibility in adverse listening conditions.

One aim of ProSynth, then, is to develop a device-independent control structure that automatically produces relevant long-domain coarticulatory effects, as well as more local spectral variation that contributes perceptual coherence and/or information about linguistic structure. However, little is known about which of these effects help speech understanding (but cf. Tunley (1999)), and ideally ProSynth will include only those that do.

The immediate use of PROCSY is thus to allow perceptual tests of particular hypotheses using natural-sounding synthetic speech. Formant synthesis is necessary because the effects in question require precise control in the spectral domain, which it is impractical to try to achieve using PSOLA-manipulated concatenated natural speech. Moreover, though concatenated natural speech may include appropriate fine phonetic detail at segment boundaries, the contribution of this type of detail to the signal's robustness is one of the things we want to investigate and that too cannot be easily manipulated by PSOLA-type techniques. On the other hand, standard formant synthesis sounds unnatural and cannot be done quickly. Extracting parameter values by copy-synthesis can speed the process up, but copy-synthesizing obstruents is notoriously difficult, and still leaves the problem of unnatural-sounding segment boundaries. We attempt to circumvent these disadvantages by copy-synthesizing utterances with known attributes, of which the principle ones are segmental labels and durations, using the HLsyn synthesizer.

HLsyn is uniquely suited to this purpose because its small set of quasi-articulatory parameters can be used to simply model the complex acoustic consequences of constricting the vocal tract enough to produce obstruent and nasal consonants, while formant frequencies for vowels and approximants can be synthesized in the standard way. Although developed initially for perceptual evaluation, PROCSY ultimately should be usable as one of the acoustic front ends for the final ProSynth system.

PROCSY as described here is still work in progress and differs significantly from the system described at the Jenolan Workshop. It has mainly changed in three respects. a) The rules and the processing program have

PARAMETER	DESCRIPTION
ag	area of glottal opening (mm^2)
al	area of lip constriction (mm^2)
ab	area of tongue blade constriction (mm^2)
an	area of velopharyngeal port (mm^2)
ap	area of posterior glottal chink (mm^2)
f0	fundamental frequency (deciHz)
f1	frequency of 1st formant (Hz)
f2	frequency of 2nd formant (Hz)
f3	frequency of 3rd formant (Hz)
f4	frequency of 4th formant (Hz)
dc	delta compliance of the walls of the vocal tract (percent)
ue	rate of active change in vocal tract volume (cm^3/s)
ps	subglottal pressure (cm H_2O)

Table 1: The parameters of HLsyn

been separated; the separate rule file makes the rules easier to work with than they were when coded directly in the program. (b) The segmental and timing information are now provided by an XML file that contains not only the sequence of segments but also the phonological structure of the whole utterance, which can therefore be used to specify the context for the application of a rule. This change went hand-in-hand with the third change, c) that whereas the rules in the original system were mainly context independent, they are now heavily context- and structure-dependent and will continue to be more so as we increase our understanding of the underlying structural influences on systematic phonetic variation, and their perceptual importance.

2 Outline of HLsyn

HLsyn is a quasi-articulatory high-level front end to SenSyn, a Klatt-type cascade-parallel formant synthesizer (Sensimetrics Corporation, Bickley, Stevens, and Williams (1997)). A small set of parameters (listed in Table 1) allows the user to synthesize an utterance in a mixture of acoustic, aerodynamic and quasi-articulatory terms. The articulatory parameters control the excitation source and aspects of spectral shape. They trigger the type of excitation by controlling cross-sectional areas at the glottis and in the oral cavity. Spectral consequences of changing HL parameter values include, for example, automatically introducing pole-zero pairs when the velopharyngeal port is modelled as open (with frequencies that are partly functions of the specified oral constrictions), and, optionally, intrinsic modifications of formant frequencies and f0 due to tongue height and aerodynamic factors. These complex interactions generate the intricate acoustic details which occur in natural speech at the margins between consonantal and vocalic segments, and which are both difficult and immensely time-consuming to produce by hand in formant synthesis.

For example the parameter **ab** (the cross-sectional area at the point of maximum constriction in the oral cavity made by the tongue blade) specifies when a constriction is small enough to change the sound source from periodicity (voice), which feeds into the cascade branch of the synthesizer, to aperiodic friction that feeds into the parallel branch. But the changes at such a boundary not only involve changes to the excitation source; there are also gradual changes of open quotient, and hence spectral tilt and formant amplitudes and bandwidths, which further affect the output spectrum. These properties, of course, are also affected by interactions of **ab** with other parameter values, such as formant frequencies and subglottal pressure. To set the necessary acoustic parameter values by hand is tedious and error prone, and indeed few individuals have enough knowledge to do it successfully. HLsyn exploits the fact that these changes are all related and can be modelled from acoustic theory, as discussed in Stevens (1999), so that changes to only one parameter, such as **ab**, can lead to natural-sounding effects by introducing a host of changes to spectral details. Thus HLsyn can be viewed as a constrained system which models basic physical and physiological principles to provide realistic-sounding speech.

3 Outline of PROCSY

PROCSY is on the one hand a pre-processor for HLsyn that interprets rules written in a simple rule syntax, and on the other hand it is a set of rules specifying how the parameters should be set for certain nodes in particular linguistic structural contexts. Unlike the previous version of the system, which had only segments as input, the system now relies on a prosodic tree providing a structured phonological representation of an utterance. This tree, provided by the XML file described in Section 4, allows PROCSY to make use of contextual information from all nodes in the prosodic hierarchy.

The separation of rule file and program was motivated partly to cope with the growing complexity of the rules, and also to make it faster to test new rules individually or to compare different versions of rules by having different rule files. The option of testing rules individually is helpful when PROCSY is used as a research tool. The option of having a number of independent rule files allows us to have different ones for different accents, speech styles or ultimately languages, without changing the program.

Figure 1 outlines the process of synthesizing an utterance with PROCSY. A signal is automatically labelled phonemically, hand-marked for some additional information such as the time of the transient due to the release of

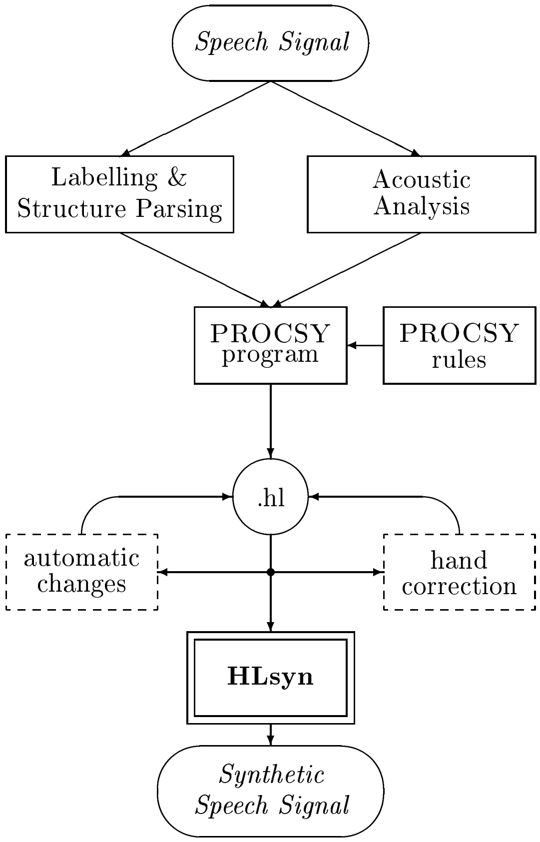

Figure 1: Outline of PROCSY. The dashed boxes represent optional processing stages.

stop closures and then hand-checked. The version of PROCSY described in Heid and Hawkins (1998) used labels that included a great deal of detailed information, for example about multiple stop bursts or the duration of mixed excitation. The current version works mainly with phonemic labels in an effort to produce by rule those aspects of the detailed labels that are systematic. In principle, either phonemic or more detailed labels could be used, or a mixture of both.

The signal is also analyzed acoustically using *xwaves*' automatic formant tracker. This performs an lpc-analysis (49-ms cos^4 window, 70% preemphasis) every 5 ms followed by dynamic programming post-processing to provide smoothed contours of formant frequencies, together with bandwidths, f0, energy of the signal, spectral slope and the probability of voicing. These two sources of information, the labeled attributes and the automatically tracked parameters, are used to carry out the pure copy-synthesis part of PROCSY. The rules that are used to set the quasi-articulatory parameters of HLsyn are based on the phonological structure discussed in the next section.

The PROCSY program is written in Python and the rule file is written in a rule syntax that we developed to suit the particular properties of ProSynth's phonological structure and the HLsyn-parameters. The output of the PROCSY program is an HLsyn input file with the extension "*.hl*". As Figure 1 shows, this *.hl* file is read into HLsyn and used to generate the synthetic speech. It is in ASCII and can easily be accessed by additional programs or edited by hand. Some hand-editing is desirable when the copy synthesis is unsatisfactory in some way, either because the rules are incomplete or because the automatic formant extraction was faulty. The first problem becomes less frequent as PROCSY's rules improve, but the second one is likely to stay. Because the HL parameters are relatively transparent in terms of vocal-tract behaviour yet can have profound acoustic consequences, it is usually rather easy to find the cause of the problem and quickly correct it. So even when the automatic procedures fail to provide the desired quality of synthetic speech, this system still saves time. For the utterance *What if you ride* as in Figure 2, it takes less than a minute for PROCSY to generate the *.hl* file from the *xwaves* and XML data files, and for HLsyn to synthesize the waveform und display its spectrogram. We estimate it would take at least half an hour for a reasonably expert HLsyn user to achieve a comparable quality by hand from the same information.

Figure 2 shows an example of the performance of PROCSY: the topmost spectrogram (a) shows a natural utterance "What if you ride", and the middle one (b) shows the purely automatically-generated version. The latter is flawed because the formants in the /w/ at the beginning of the utterance were wrongly tracked by *xwaves* and their values were then rejected by the constraints of the rules. (The formants which are excited in (b) are default values which are set at the beginning of the utterance.) The bottom spectrogram (c) shows the effect of some hand-editing. All

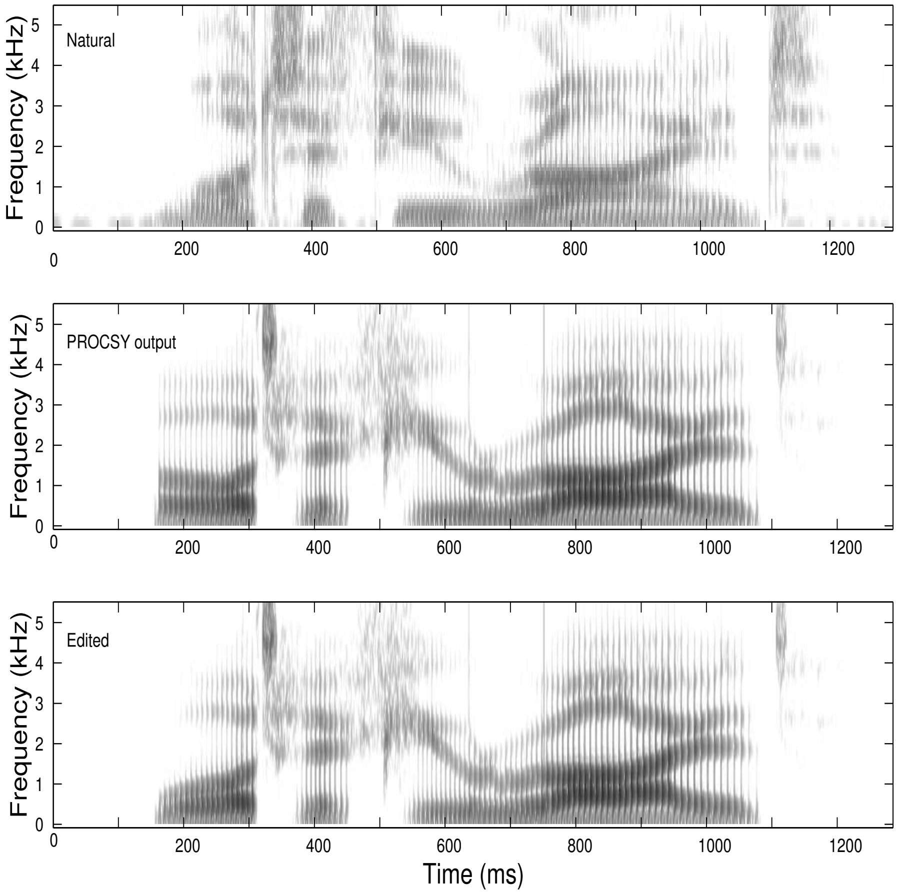

Figure 2: Upper panel: naturally-spoken *What if you ride?*. Middle panel: automatically-generated copy synthesized version of the natural utterance. Lower panel: Same as the middle panel, except that the F1 and F2 frequencies have been hand-edited at 162 and 215 ms.

that was done was the insertion of four values: F1 and F2 at the beginning and the end of the /w/.

4 The phonological structure

A fundamental principle of ProSynth research is commitment to a prosodic phonological structure that brings together different levels of linguistic description. The centrality of the prosodic structure stems from the theoretical position that much of the systematic variability in, say, segmental timing, not only reflects linguistic structure, but, crucially, provides essential perceptual cues to that structure. While such details may not significantly affect the intelligibility of short stretches of speech in good listening conditions, they are almost certainly important in longer passages and adverse conditions: they provide an additional type of informational richness to that discussed in terms of perceptual coherence of the auditory signal in Section 1.

The phonological structure has a prosodic hierarchy with the following levels: Intonational Phrase, Accent Group, Foot, Syllable. The syllable has the standard constituents of Onset, Rhyme, Nucleus, and Coda. These elements are structured to obey the Strict Layer Hypthesis (Selkirk (1984)) and principles of headedness, as described in Hawkins, House, Local, and Ogden (1998).

PROCSY's computational structure is declarative, like ProSynth's. That is, information cannot be deleted or modified. Thus the structure determines the processing of the utterance, in that the rule processing starts at the root node of the tree and steps through the structure down to the leaf nodes. In this way nodes in the hierarchy are always processed in the order in which they influence each other. The order of processing can thus play a role in the rule execution, but only by reflecting the order of the nodes in the structure and thereby the structural dependencies. In contrast, rule-ordering at a specific node plays no role at all: rules for a given node can be thought of as executed simultaneously. This framework makes impossible the rule ordering problems that are endemic in traditional phonological rule systems.

The declarative structure means that PROCSY can set HLsyn parameter values to operate over any domain specified in the prosodic hierarchy, by attaching the parameter value to the appropriate node in the tree. At present, PROCSY makes relatively little use of these higher nodes in the hierarchy, partly because one major use is in assigning the f0 contour, which in PROCSY is copied from the acoustic signal, and partly because we are still developing the knowledge that will allow higher nodes to be used. However part of ProSynth research is aimed at finding out more about how higher-order information affects segmental realisation, and as this is accomplished, it can be used in the current framework immediately.

One example of the type of use PROCSY does currently make of the higher nodes is in assigning values for subglottal pressure (**ps**). Subglottal pressure is controlled via variables that are set at the Intonational Phrase and Foot levels and then used to set **ps** at syllable nucleus level. This generates a time-varying subglottal pressure contour with local maxima on accented syllables and slow "decay" in between. This pattern roughly implements the findings of Lieberman (1967).

Another example of how the higher-level structure is currently used is in specifying that rules should only apply to a node if certain contextual requirements are met. For example, a given rule might only apply when the syllable that governs the current node contains an Onset whose segments include the features Sonorant and Voiced (i.e. an approximant).

The more we know about contextual dependencies, the more detail can be included in the rule files, and the better the output synthetic speech. The better the synthesis, the wider the range of dependencies that can be tested, since we assume that many of the more subtle dependencies have negligible perceptual consequences unless the overall quality is already fairly good.

5 Rules

5.1 General principles

To apply the rules, the nodes are analyzed in terms of their component phonological features. The features associated with a node determine the articulatory settings. For example a node corresponding to a /z/ has the features **CNT = Y, SON = N, VOI = Y, CNSGRV = N, CNSCMP = N, NAS = N, STR = Y, RHO = N**. The features **SON=N** (non-sonorant) **CNT=Y** (continuant) and **VOI=Y** (voiced) specify the sound as a voiced fricative. **CNSGRV=N** (non-grave consonant) and **CNSCMP = N** (non compact consonant) specify the sound as an alveolar. When the rules interpret these features, they make a constriction by the tongue blade which evokes frication noise and they produce a glottal constriction which allows for voicing under these conditions. The formant frequencies specify the place of articulation of the /z/ as alveolar rather than dental, palatoalveolar or retroflex, which results in HLsyn specifying an appropriate value for each filter of the parallel branch of the synthesizer, thus achieving the right spectral shape without having to evoke the parallel branch or introducing necessary spectral zeros explicitly in the rules.

The time of onset and offset of obstruent and nasal segments is specified by timing information attached to the relevant node in the XML file. The rules specify the precise locations of articulatory constrictions necessary to produce acoustic segments of the right durations, along with the duration and rate of change of transitions between constrictions.

The challenge of this conceptually simple approach lies in the systematic, structurally-determined variation in the natural signal that must be copied to the resulting synthetic signal. The HLsyn manual offers recommended parameter values for synthesizing American English consonants between stressed vowels and in a few clusters. There are few or no recommendations for unstressed environments, and even if there were, they would not all be suitable for Southern British English, since American and British English treat many unstressed syllables quite differently. The approach we have taken is to start as simply as possible, only adding more complex rules when it is clear that they are necessary. In the first version of PROCSY (Heid and Hawkins (1998)), the rules were all context-free, with only the timing information for individual segments determining parameter coutours. We tried to produce various types of systematic variability, for example in excitation type at segment boundaries and in degree and extent of coarticulation of nasality, by using a form of underspecification: control points for the quasi-articulatory parameters were set at certain landmarks and slowly interpolated between them. Although this very simple, underspecified version worked surprisingly well for some utterances, it quickly became clear that more elaborate modelling was needed in most cases. The current version therefore includes more information about the phonological structure, and the rules are heavily context-sensitive. The rest of this section discusses some of the present rules and their guiding principles.

5.2 Voicing

Voicing depends mainly on the parameter **ag**, which specifies the glottal opening in mm^2. When the oral cavity is unconstricted, values of **ag** of around 4.0 mm^2 produce modal voice, while high values (> 12.0 mm^2) result in breathy noise or friction if there is an appropriate supraglottal constriction. The range between 4.0 and 12.0 mm^2 results in increasing breathiness and greater spectral tilt with less overall energy, as is often found at the edges of voiced segments when glottal vibrations are just starting or stopping. When the oral cavity is constricted enough to produce friction high values of **ag** result in high oral flow consequently in high amplitude friction, while values of **ag** around 4-8 mm^2 add perdiodicity to the aperiodic excitation in the vicinity of the oral constriction.

Voicing can be influenced by two other HLsyn parameters: **ue**, the rate of active change in vocal tract volume, and **dc** the change in compliance of the vocal tract walls. At present we use these parameters only to generate voicing in the closure of voiced plosives in certain contexts.

Table 2 shows an example of a rule that specifies two different settings of **ag** at the boundary between a voiced segment and a voiceless segment.

```
if SEG+1:VOI is N and              ; if next segment is voiceless
   (VOC≫ or VOI is Y) then         and current seg is a vowel
                                   or a voiced consonant, then

   if SEG+1:CNT is N then          ; if next seg is –continuant (=stop), then

      set ag at stop-10 to 4.0 fi  ; set ag to 4.0 (modal)
                                   10 ms before end of current seg
                                   end if

   if SEG+1:CNT is Y then          ; if next seg is +continuant (=fricative), then

      set ag at stop-40 to 4.0;    ; set ag to 4.0 (modal)
                                   40 ms before end of current seg

      set ag at stop+40 to 20.0;;  ; set ag to 20.0 (wide open)
                                   40 ms after end of current seg
                                   end rule
```

Table 2: Example of a rule to control voicing at the boundary between a voiced segment and a voiceless stop or fricative. See text for explanation.

SEG+1 functions as a pointer to the features of next segment (similarily **SEG-1** or **SEG+5** could be used to test the features of the previous segment or the fifth next segment respectively). The expression **VOC≫** means apply this rule only when in a vowel node. The expression **VOC≫ or VOI is Y** is necessary because voice is a feature only specified for consonants.

The glottal area **ag** is set to 4.0 (modal voicing) 10 milliseconds before the end of the voiced segment if the next segment is a voiceless plosive, or 40 ms before the end of the voiced segment if the next segment is a voiceless fricative.

In the later case, **ag** reaches its local maximum (20 mm^2) 40 ms into the fricative. (Maximum **ag** for a voiceless stop is set in a separate rule that applies to plosives and their bursts.) Because HLsyn interpolates between set control points, these timing differences generate the differences in quality of voicing offset to be expected before fricatives and plosives.

5.3 Obstruent consonants

5.3.1 Duration and spectral shape

As noted in Sections 1 and 5.1, obstruents are hard to copy-synthesize in conventional formant synthesis, but can be done by rule relatively straightforwardly in HLsyn by specifying a constriction for a particular articulator, together with some formant frequency information. The latter is necessary for all obstruents, but is particularly important in distinguishing different places of articulation made by the same articulator. For example, alveolar and postalveolar fricatives (/s z/ vs. /ʃ ʒ/) are both made by constrictions of the tongue blade, **ab**.

The only information that is copied from the original signal for fricatives is the times when frication starts and stops; for stops, it is the onset of closure, the time of the release burst, and the time of the onset of periodicity following the release (i.e. VOT) if appropriate. The features at the segment nodes determine which HL parameter is used to form the constriction: the lips (**al**) for /p b f v/, the tongue blade (**ab**) for /t d θ ð s z ʃ ʒ/. For velars (only /k g/ in English), the tongue body position is indirectly specified via a low F1. The cross-sectional areas of the oral and glottal constrictions together determine the excitation type, and the spectral shape is determined by rules within HLsyn that are triggered by particular combinations of formant frequencies and oral constriction parameters. Thus when **ab** is close to 0.0, indicating a complete occlusion made by the tongue blade, the formant frequencies determine whether F4 and F5 excite the spectrum (e.g. for /t/ or /s/) or F3 and F4 (e.g. for /ʃ/). When all these parameters are set appropriately, the result is remarkably natural obstruents at little computational cost.

Automatically-tracked formant information, especially in fricatives and the aspiration periods of stops, can be used to achieve the appropriate spectral shape, but it proved hard to devise a general algorithm that retains formant values that conform to HLsyn's constraints and discards those that do not, in all possible contexts. Consequently, automatically-tracked formant frequencies during obstruents are not used in PROCSY. Instead, formant frequencies are set by rule in the central region of fricatives and at plosive bursts.

For most plosives, two main parts have to be generated: an oral closure and a burst, possibly followed by aspiration or aspiration plus frication. Thus the rules model the closure of a particular articulator, its rapid release, and a transition appropriate for the following sound. For bilabial and alveolar plosives, the closure is produced by setting the cross-sectional area for the appropriate oral articulator to 0.0 mm^2; for velar plosives, F1 is set to 180 Hz. During the closure, HLsyn models a pressure increase in the oral cavity. For voiceless plosives, the rules normally enhance this increase by opening the glottis wide — say to 30 mm^2. For voiced plosives in appropriate contexts, the rate of rise in oral pressure can be decreased in order to allow voicing to continue some way into the closure; this is done by manipulating **dc** and **ue** as described in Section 5.1. Rapid release of the closure in the oral articulator parameter causes HLsyn to generate a burst with the right spectral amplitude and shape. After the burst, **ag** stays at or returns to 4.0 mm^2 if a vowel follows the stop, or it stays open if voiceless sounds follow. Fricatives are produced along similar principles except that rates of change of cross-sectional areas are typically somewhat different from those for plosives.

These default rules for obstruent consonants produce good output for

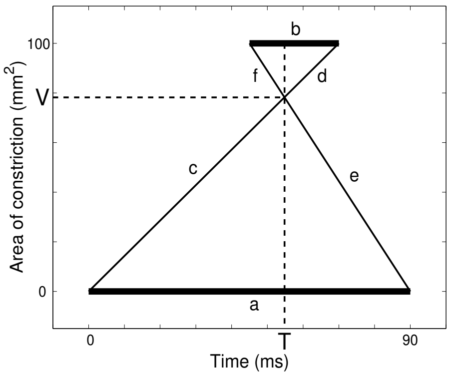

Figure 3: Schema for the calculation of control points when a short vowel or a sequence of vowel with approximant mean that the default transitions of an area parameter overlap. See text for further explanation.

many contexts, but some obstruents need different rules to produce more contextually-appropriate patterns of friction and/or aspiration.

One such "special case" is when a vowel adjacent to an obstruent is so short that there is not enough time to perform the transition in the standard way (cf. *pip*, *miss*). This situation is most common when a lax and/or unstressed vowel lies between two obstruents. Our general solution is to preserve the standard slope of the transition, in order to produce the same acoustic effects at the segment boundaries, but to reduce the maximum cross-sectional area of the oral articulatory parameter during the vowel. When the obstruents are homorganic, as in *pip*, PROCSY calculates the requisite degree of undershoot in the oral articulator as it opens into the vowel and then closes to make the second obstruent, as illustrated in Figure 3. The line marked **a** represents the actual time interval of the vowel. The slopes of the lines **cd** and **ef** represent the desired rate of change of the constriction area, e.g. **al** or **ab**. Thus the line **b**, which is parallel to **a** at a value of 100 mm^2, the ideal maximum cross-sectional area, represents the time interval by which the vowel is too short to accommodate the two standard transitions. It is easily shown that **a**:**b** = **c**:**d** = **e**:**f**. Since values associated with **a** and **b** are known, then we can calculate the time (T) at which the articulator reaches its maximum opening area, and the value (V) of that area.

5.3.2 Mixed excitation

In our database, the duration of mixed periodic and aperiodic excitation at vowel-fricative boundaries is greater than that at fricative-vowel boundaries. HLsyn automatically produces this difference to some extend, but

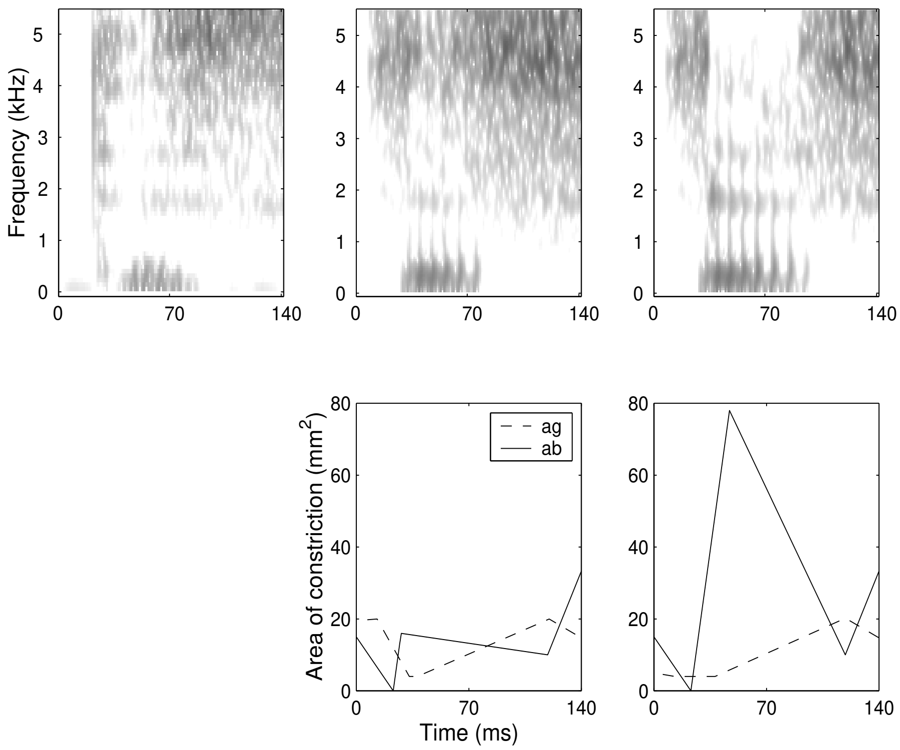

Figure 4: Upper row: spectrograms of *dis* in *disappoint*: left = naturally spoken; middle = copy-synthesized version with large degree of mixed excitation; right = copy-synthesized version with small degree of mixed excitation. Parameter values for **ab** and **ag** used to generate each synthetic version are shown immediately below their respective spectrograms.

we have found that syllable stress, vowel height, and final/non-final position within the phrase influence the incidence and duration of mixed excitation. We produce different durations of mixed excitation by changing the slope of the transition of the oral articulator parameter. Figure 4 shows spectrograms of *dis* from the word *disappoint*. In the top row, a naturally-spoken token from our database is in the leftmost panel; the other two panels show HLsyn-synthesized versions: one with the context-sensitive slope of **ab** (middle panel), and the other with the standard slope (rightmost panel). The two panels in the bottom row show the relevant parameter values, **ab** and **ag**. The context-sensitive slope, which is shallower and has a maximum of only about 16 mm^2, produces a longer period of mixed excitation that resembles that of the natural syllable. In contrast, the standard slope, which achieves a maximum area of 78 mm^2, produces very little overlap.

5.3.3 Stop bursts and aspiration

Section 5.1 describes how the time of occurrence and spectral shape of the plosive bursts are achieved. This section discusses how to achieve different qualities of excitation after the plosive is released. In natural speech, voice onset time (VOT, the duration between release and the onset of periodicity) varies with the place of articulaton of the plosive, vowel quality, stress, and the overall rate of speech. Other properties such as the amplitude of frication and aspiration can also vary with these factors. In PROCSY copy-synthesis, these complexities do not cause difficulties because VOT itself is specified in the XML file and HLsyn automatically produces many of the other context-dependent properties as long as PROCSY produces the right parameter values, such as increased **ps** for stressed syllables. In the standard case, the correct VOT is produced simply by setting **ag** to 4 mm^2 at or shortly after the onset of periodicity following the release of the oral constriction. Normally therefore, **ag** in voiceless stops decreases linearly from 30 mm^2, its value at the time of release, to 4 mm^2 at the onset of periodicity for the following voiced sound. Voiced stops behave similarly but usually have a smaller glottal opening at release.

As the above discussion of mixed excitation at vowel-fricative boundaries suggests, different qualities of excitation can be produced by varying the rate at which **ag** changes. If **ag** stays wide open for some time after the burst and then approaches its modal setting relatively rapidly, the result includes higher-amplitude noise excitation during the VOT interval, and a relatively abrupt start to modal voicing. Conversely, very shallow slopes in **ag** can result in extended periods of breathy voice at voicing onset. These properties can be varied to suit different structural contexts, and individual speaker characteristics.

Another example of a contextually-dependent variant in the quality of excitation during the VOT is extensive high-amplitude frication, rather

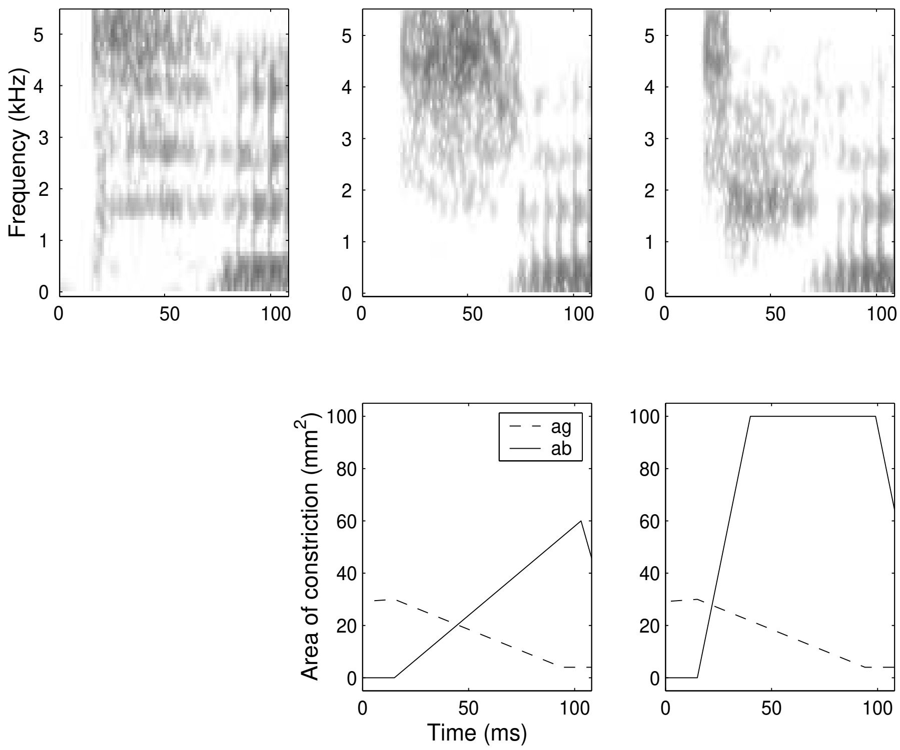

Figure 5: Upper row: spectrograms of *to* in *to deride*: left = naturally spoken; middle = copy-synthesized version with friction after the burst; right = copy-synthesized version mainly with aspiration after the burst. Parameter values for **ab** and **ag** used to generate each synthetic version are shown immediately below their respective spectrograms.

than aspiration. This is common for Southern British English alveolar stops in the context of unstressed high vowels or schwa. It is less common in many other accents of English, including American, although a similar type of sound is found in so-called voiceless vowels in Japanese. We produce this type of sound by making the slope of the oral articulator very shallow. Figure 5 shows an example of this for the first syllable of *to deride*.

5.4 Nasality

Nasality is controlled in HLsyn via the parameter **an**, which represents the area of the velopharyngeal port in mm^2. When the port is open, the nasal and oral cavities are coupled, resulting in the introduction of extra pole-zero pairs and a shifting in the frequencies of oral formants. The details depend on the size of the velar opening, the shape of the oral cavity, and the volume of the nasal cavity. Non-low vowels are not usually nasalized in non-nasal contexts in British English, but one would expect them to be at least partly nasalized when adjacent to a nasal consonant.

5.4.1 Nasal stops

Place of articulation of nasal consonants is controlled by setting **al** or **ab** to 0 mm^2, or f1 to 180 Hz, for /m n ŋ/ respectively, following the same basic principles outlined in Section 5.2 for obstruents. Formant frequencies (other than f1 for /ŋ/) can normally be tracked automatically. During the oral occlusion, **an** is set wide open (40 mm^2). Transitions between nasals and obstruents demand a rapid movement of the velum that is simple to produce by rule. Transitions between nasal consonants and adjacent vowels or approximants require more care, as described in the next Section.

5.4.2 Coarticulation of nasality

To model the coarticulation of nasality into vowels and approximants that precede a nasal consonant, we start opening the velum (**an**) at a point one third into the duration of the vowel, or approximant-vowel or vowel-approximant cluster. Maximal opening (40mm^2) is achieved at the boundary of the nasal and continues throughout the nasal consonant. Thus the last two thirds of the vowel or approximant-plus-vowel cluster becomes increasingly nasalized. Perseverative coarticulation of nasality takes up a smaller proportion of any following vowel and/or approximant, at least if they are in stressed syllables. More work is needed on how far the coarticulation of nasality is affected by stress.

If there is a nasal on both sides of the nucleus (plus approximants), nasality continues right through them in the current model, but more work is needed here too, since the result is often too nasalized. Syllabic /l/ is treated in the same way as vowels in this respect: coarticulation of nasality can be anticipatory (*Pat'll make it*), persevarative (*Sam'll loose it*), or both (*Sam'll make it*).

Coarticulation of nasality is an example of rules which might be better represented at a speaker-dependent level in future systems: different speakers may show different kinds of coarticulatory strategies for effects like nasalization, so it would be reasonable to make such effects speaker-dependent.

5.5 Dealing with wrong formant frequency measurements

Since automatic formant measurements are never completely error free, one set of rules filters out presumed wrong candidates. These rules represent physical and physiologically-motivated constraints on possible formant configuration. Some are already implemented, but work is still in progress on others, e.g. in setting limits on the acceptable frequency range for a given formant in a given context. There are three types of rule. The first type reflects vocal-tract constraints. Formant frequencies that fall outside pre-defined limits are rejected, not only because they are probably wrong, but also because they can cause HLsyn to behave inappropriately, in that HLsyn interprets formant frequencies not only as parameter values for input to the actual synthesizer, SenSyn, but also as indicators of place of articulation. Certain wrong formant frequencies can therefore lead to serious errors. F1 is especially prone to producing such errors because, as noted in Section 5.3, the HLsyn rules use values of less than 200 Hz in F1 to calculate degree of velar closure. F1 frequencies below 200 Hz will thus effectively lead to closure of the tongue dorsum against the velum, an error which is obviously to be avoided during a vowel. As noted in Sections 5.3.1 and 5.3.3, formant frequencies affect the spectral shape of obstruents: they do this by determining the amplitudes of the poles of the parallel branch.

At present, expected ranges of formant frequencies are specified for each formant independently of the others. Similar rules could be written that use possible patterns of formant frequencies, for example the expected ranges of ratios of F1:F2, F2:F3 and so on. Measured ratios could be compared with expected ratio ranges for the particular segment, and more acceptable ones substituted if the measured values fell outside the expected range. We have not found this method worth implementing, however, partly because *xwaves*' formant tracker is rather good, and also because of the way we use PROCSY. So far our applications have typically produced small sets of utterances and each utterance is further manipulated to contrast with the original signal in some particular parameter. In this type of work it is probably advantageous to be confronted by the clear error that a completely missed formant produces than by the more subtle one that a 'reasonable substitution' might produce, since hand-editing is fast (see Figure 2 and accompanying text). The more complex substitution method might be valuable for applications that demand larger sets of utterances and in which reasonable but not always good quality is preferable to the higher quality that can be achieved using hand-correction.

The second way formant frequency errors are detected is by constraining formant movement. During a vowel, a change of 100 Hz in F1 over only 5 ms is unlikely, while one of 500 Hz is certainly an error.

Thirdly, context-dependent rules operate from knowledge about problems connected with specific sounds or groups of sounds. For example formant measurements are more likely to be wrong in glides than in vowels and similarly, measured frequencies are more likely to be wrong when two formants are closely spaced, as F1 and F2 can be in /o/, /ɑ/, and especially in British English /ɔ/.

Since HLsyn interpolates between input parameter values to generate the output signal, formant frequencies that are likely to be wrong can easily be omitted. The result is usually a smooth contour, since remaining formant values are normally sufficient to produce the right output. Some points, of course, are more crucial than others. If frequencies are unreliable over a crucial region, then hand-editing is required at present. In general, however, a good principle is "no input is better than wrong input".

6 APPLICATIONS

As outlined above, our motivation for developing this method is our need for carefully controlled speech stimuli to test the perceptual salience of coarticulatory effects and other fine phonetic details that we hypothesize result from the linguistic structure of an utterance, and that could increase the robustness of synthetic speech in adverse listening conditions.

Two contrasting strategies can be used to test the salience of specific acoustic properties, one constructive and the other destructive. With the constructive strategy, the properties of interest are inserted into the signal and gains in perceptual accuracy or speed are assessed. With the destructive strategy, the effect of removing or reducing the extent of putative perceptual cues is assessed. The HLsyn parametric descriptions are amenable to both approaches. For example, hand editing can introduce or destroy long-domain dependencies in the realization of liquids, whose contribution to acoustic-perceptual coherence is currently a central focus of research in ProSynth. Similarly, programs are in development which allow automatic systematic variation of the parameter structure, e.g. to alter timing relations between parts of the utterance or in the sequencing of changes in different parameters. Another program produces different types of interpolation between parameter control points, and different degrees of linearizition (stylization), so that we can assess how simple the final control structure of the output can be.

One experiment (Heid and Hawkins (1999)) assesses whether natural-sounding excitation near segment boundaries enhances the intelligibility of formant synthesis. Excitation type at fricative-vowel and vowel-fricative boundaries and the duration of voicing in voiced stop closures were shown to vary systematically with linguistic-phonetic structure for the male speaker of British English whose utterances comprise the ProSynth database. Some of these utterances were copy-synthesized using PROCSY, and the excitation type at selected segment boundaries was varied to either follow the patterns observed in the natural speech, or to violate them, so that each utterance was produced in two versions, a "right" and a "wrong" one. The "right" versions proved to be significantly more intelligible than the "wrong" versions in a listening test in which they were presented in cafeteria noise at an average s/n ratio of -4 dB relative to the maximum amplitude of the phrase.

Heid and Hawkins (1999) use the "constructive" method to assess the importance of local effects at segment boundaries, which are necessarily very short. An example of the "destructive" method is work that estimates the contribution to speech intelligibility of long-domain effects due to /r/ and /l/, such as those mentioned in Section 1. We use PROCSY to copy-synthesize utterance pairs like *What if you hide?* and *What if you ride?* We then "cross-splice" the HLsyn ASCII files to produce two new versions, each with inappropriate resonance effects: *ride* is preceded by *What if* from *What if you hide*, and *hide* is preceded by *What if* from *What if you ride*. Other splicing points could also be used. The same procedure can be used with the utterance *What if you lied?*, cross-splicing it with either or both of the other two. Parameters are automatically smoothed between control points across "splicing" points in the ASCII file. Perceptual tests can be run in adverse listening conditions: either in noise, or with high cognitive load, as when subjects are required to carry out another task simultaneously with listening and responding to the stimuli. The hypothesis is that the disruption of the natural long-domain resonances affects performance.

Another example of a potential application in speech perception is an extension of work by Hawkins and Nguyen (to appear), who have shown that in some circumstances listeners use the durational and spectral properties of an /l/ in the syllable onset to predict the voicing of an obstruent in the syllable coda. Hawkins and Nguyen used a speeded lexical-decision task, which brings with it certain disadvantages. However, at that time it seemed the best alternative, partly because it allowed the use of natural speech: standardly-produced synthetic speech was too poor in quality and too time-consuming to produce for the large number of stimuli required. Future work can use PROCSY, with significant gains in the degree of experimental control, at relatively little cost in terms of loss of quality of the stimuli.

7 Concluding remarks

HLsyn is an appealing platform for formant synthesis by rule because it allows the complex acoustic consequences of vocal tract dynamics at segment boundaries to be easily synthesized, and because its parameters can straightforwardly be used in order to translate linguistic structure into acoustic specifications. PROCSY's rules exploit the fact that the whole linguistic-phonological specification of an utterance is in the XML encoded phonological structure. Their guiding principle is to use formant measurements in the sonorant parts of an utterance, and to specify trajectories for the quasi-articulatory parameters from the phonological information in the XML file to generate the obstruents and nasals. This simple concept allows PROCSY to overcome drawbacks of either standard formant synthesis or PSOLA-based methods. Moreover, higher-level structural information is used to trigger the application of particular rules and also to directly control some parameters. PROCSY is currently used for fast copy-synthesis to produce stimuli for testing hypothesis about what type of acoustic variation matters in particular structural contexts. As more contextually-sensitive detail is added to the rules, however, it approaches a rule-driven TTS system.

8 Acknowledgements

This work is funded by EPSRC grant GR/L53069. We are grateful to John Local, Richard McGowan, and Kenneth Stevens for insightful comments.

References

Bickley, C. A., K. N. Stevens, and D. R. Williams (1997). A framework for synthesis of segments based on pseudoarticulatory parameters. In J. P. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*, pp. 211–220. New York: Springer.

Fixmer, E. and S. Hawkins (1998). The influence of quality of information on the McGurk effect. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson (Eds.), *Proceedings of the International Conferencs on Aditory-Visual Speech Processing*, Terrigal, Australia, pp. 27–32.

Hawkins, S. (1995). Arguments for a nonsegmental view of speech perception. In K. Elenius and P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, pp. 3:18–25. KTH and Stockholm University.

Hawkins, S., J. House, J. Local, and R. Ogden (1998). Prosynth: An integrated approach to device-independent, natural-sounding speech synthesis. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia.

Hawkins, S. and N. Nguyen (to appear). Effects on word recognition of syllable-onset cues to syllable-coda voicing. In J. Local, R. Ogden, and R. Temple (Eds.), *Proceedings of the Sixth Conference on Laboratory Phonology.* Cambridge: Cambridge University Press.

Hawkins, S. and A. Slater (1994). Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. In *Proceedings of the 3rd International Conference on Spoken Language Processing*, Volume 1, Yokohama, pp. 57–60.

Heid, S. and S. Hawkins (1998). PROCSY: A hybrid approach to high-quality formant synthesis. In *Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 219–224.

Heid, S. and S. Hawkins (1999). Synthesizing systematic variation at boundaries between vowels and obstruents. San Francisco. Paper to be presented at *the XIVth International Congress of Phonetic Sciences.*

Kelly, J. and J. Local (1989). *Doing Phonology.* Manchester University Press.

Lieberman, P. (1967). *Intonation, Perception, and Language.* Cambridge, MA: MIT Press.

Local, J. (1992). Modelling assimilation in non-segmental rule-free synthesis. In G. J. Docherty and D. R. Ladd (Eds.), *Papers in Laboratory Phonology II — Gesture, Segment, Prosody*, pp. 190–223. Cambridge University Press.

Local, J. and R. Ogden (1997). A model of timing for nonsegmental phonological structure. In J. P. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*, pp. 109–121. New York: Springer.

Marslen-Wilson, W. and P. Warren (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review 101*, 653–75.

Pratt, R. (1986). On the intelligibility of synthetic speech. *Proceedings of the Institute of Acoustics 8*, 183–92.

Selkirk, E. O. (1984). *Phonology and Syntax.* Cambridge, MA: MIT Press.

Stevens, K. N. (1999). *Acoustic Phonetics.* Cambridge, MA: MIT Press.

Tunley, A. (1999). *Coarticulatory Influences of Liquids on Vowels in English.* Unpublished Phd dissertation, Cambridge University.

Warren, P. and W. Marslen-Wilson (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics 41*, 262–75.