

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.

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 HLsynTM synthesizer (Sensimetrics Corporation). The purpose is to allow controlled manipulation of selected acoustic parameters. Parameters for HLsyn are derived from 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, 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.

1 INTRODUCTION

The work described here is part of ProSynth [12], [3], 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 position, developed in more detail in [5], [2], 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% [8]. We outline our argument here in order to explain why PROCSY depends on HLsyn rather than 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: its acoustic-phonetic fine detail reflects vocal tract behaviour and identifies the signal as coming from one place. 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 systematically, adding information rather than noise.

Some aspects of acoustic coherence 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 conditions. So, when rule-based synthetic speech includes 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% [5], [10]. In some contexts, these long-domain resonance effects [6] provide weak but consistent acoustic cues to phoneme identity over several syllables. If we are correct in assuming that, to understand speech, listeners use all available sensory information in proportion to its actual and perceived reliability (cf. [11], [7], [1]), then providing such long-domain 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. However, little is known about which resonance effects help speech understanding, and ideally ProSynth will include only those that do.

The immediate use of PROCSY is thus to allow perceptual tests with natural-sounding synthetic speech. Formant synthesis is necessary because the effects in question require precise control in the spectral domain, whereas PSOLA-manipulated concatenated natural speech allows easy f0 and timing manipulation, but is impractical for

ag	area of glottis
al	area of lip constriction
ab	area of tongue blade constriction
an	area of nasal opening
ue	rate of active change in vocal tract volume
f0	fundamental frequency
f1	frequency of 1st formant
f2	frequency of 2nd formant
f3	frequency of 3rd formant
f4	frequency of 4th formant
ps	subglottal pressure
dc	delta compliance of the walls of the vocal tract
ap	area of posterior glottal chink

Table 1: The parameters of HLsyn

spectral manipulation. 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 using utterances with known segmental labels and durations, and the HLsyn synthesizer. Ultimately, although developed here for perceptual evaluation, PROCSY should be usable as one of the acoustic front ends for the final ProSynth system.

2 OUTLINE OF HLSYN

HLsyn is a quasi-articulatory high-level front end to SenSyn, a Klatt-type cascade-parallel formant synthesizer. A small set of parameters (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 intrinsic modifications of formant frequencies and f_0 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 put in by hand in formant synthesis.

HLsyn can thus be viewed as a constrained system which models basic physical and physiological principles to provide realistic-sounding speech. Rule-based generation of obstruents is relatively straightforward, while acoustic parameters allow formants to be produced by standard copy

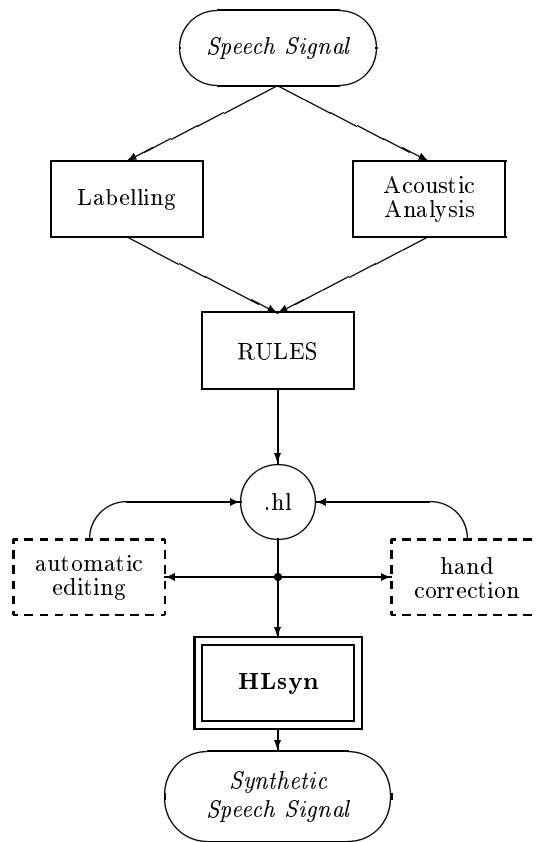


Figure 1: Outline of PROCSY

synthesis. Interpolation between control points potentially allows rules and parameter values to be underspecified.

3 OUTLINE OF PROCSY

Figure 1 outlines the structure of PROCSY. Each signal is labelled in fairly fine phonetic detail (Section 4), and acoustically analyzed for formant frequencies, f_0 , energy and probability of voicing. The labels and automatically-measured acoustic data feed into the rules, which generate the HLsyn parameter values stored in a .hl file. This file is read into HLsyn and used to generate the synthetic speech. It is in ASCII and can easily be accessed by additional programs (Section 7) or edited by hand.

4 LABELLING

The utterance labels provide the symbolic input to the rule section. At present, we use only phonetic segment labels — English phonemic SAMPA symbols with added phonetic details where necessary. Eventually we will include higher-level phonological categories so that segment labels will be effectively context-dependent. These categories, described in [3], include syllabic, morphological, and prosodic information. The labelling system is hierarchical: it indicates preferred segmentation points, and enables access to sim-

ple phonemic information, or to more detailed information, as desired. Labels with more detailed information are subsumed by more general labels. The hierarchy is intended to allow the database of utterances to be searched for particular events. For example, a search for all the burst labels, will find the label burst-&-aspiration-&-voice in Figure 2, as well as all other labels containing information regarding bursts.

The added phonetic details divide phones into subsegments: to distinguish traditional subunits such as the silence and burst-&-aspiration of stops; to mark extensive regions whose segmentation is uncertain, such as between approximants and glides; or to note aspects of the utterance which are not part of the standard description of that phoneme. Thus, excitation type is marked if it differs from the standard feature specification for the phoneme, and regions of mixed periodic and aperiodic excitation typical of the boundaries between obstruents and sonorants are marked if they are judged to be perceptually significant. Typically, monophthongal vowels take just the phoneme label, together with any non-standard excitation, while diphthongs are divided into two steady states and the transition. Stops and voiced fricatives often receive more detail.

The waveform in Figure 2 illustrates this labelling strategy for an intervocalic /b/. A stop has two main labels, closure and burst-&-aspiration. In this case, the closure contains periodicity which dies away about 20 ms before the release transient. Consequently, this phone contains a sub-label to distinguish the voiced part of the closure from the silent part. This example is expected in this context. Less standard instances of added detail include either residual periodicity or aspiration at the beginning of closure for a voiceless stop, and all possible combinations of periodic and aperiodic noise, as well as silence and transients, during the course of a voiced fricative. In Figure 2, the acoustic segment following the burst includes mixed aspiration and voicing, and the presence of both is reflected in the segment label used. Extensive ambiguous regions are sometimes found in formant trajectories between approximants and vowels, and between adjacent vowels. These are systematically marked wherever they seem to be great enough to significantly affect perception.

This detailed label information can be used to fine tune HLsyn's articulatory events to reproduce those regions of acoustic complexity and/or segmental ambiguity that are common in natural speech but rare in synthetic speech. Statistical analyses relating these detailed phonetic events to the prosodic and syllabic context are planned, so that context-sensitive regularities can be reflected in the output of the linguistic control structure. As yet, the value of these additional labels has not been systematically tested. If some prove unnecessary, the hierarchical labelling system means they can be ignored. However, we anticipate that some will contribute to naturalness and intelligibility.

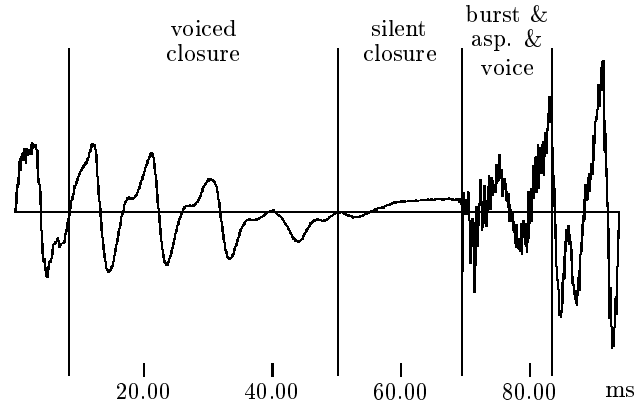


Figure 2: Waveform showing three parts of a stop consonant.

5 ACOUSTIC ANALYSIS

The acoustic analysis uses *xwaves'* automatic formant tracker. This performs an lpc-analysis every 5 ms with final dynamic programming post-processing to provide smoothed contours of formant frequencies and bandwidths, f_0 , energy of the signal, spectral slope and the probability of voicing. Currently, for compatibility with HLsyn, the signal is downsampled to 10 kHz and the first 4 formants are measured (49-ms \cos^4 window, 70% pre-emphasis). A shell script generates one file which contains all the acoustic information plus a time stamp calculated from the sampling frequency.

6 RULES

The rule section is implemented in a C program which takes the acoustic measurement file and the label file as input and then generates a .hl-file of HLsyn parameter values. The rules exploit the fact that the whole linguistic-phonological specification of the utterance is in the labels. Their guiding principle is to provide maximally general articulatory specifications, in the form of HLsyn parameter values. This simple concept produces good copy-synthesis because parameter values are partly derived from timing information attached to segment labels, and the inherent mechanisms of HLsyn take care of much of the generation of natural sounding speech.

To apply the rules, the labels are analyzed in terms of their component features. The features associated with a label determine the articulatory settings. For example the label /z/ produces a glottal constriction which allows for normal modal voicing, plus a constriction made by the tongue blade which evokes frication noise. The formant frequencies specify the place of articulation of the /z/, which results in appropriate values for all filters of the parallel branch of the synthesizer, thus achieving the right spectral shape without explicitly evoking the parallel branch or introducing spectral zeros. The timing information attached to the label specifies the onset and offset of the articulatory specification. The rules specify the precise location

of changes in articulatory constrictions within the acoustic segment, along with the duration and rate of change of transitions between constrictions. At present, these rules are very simple. Eventually, they are expected to be sensitive, for example, to syllabic stress and the relative tongue height of neighbouring segments.

The rest of this section exemplifies these principles. Examples are available from [13].

6.1 Voicing

Voicing depends mainly on the parameter **ag** which specifies the glottal opening in mm^2 . High values ($> 10.0 \text{ mm}^2$) for **ag** result in breathy noise, or friction if there is an appropriate supraglottal constriction. Values around 4.0 mm^2 produce modal voice. The range between 4.0 and 10.0 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. Hence to produce a modally-voiced sound **ag** is usually set to 4.0 mm^2 .

The interesting problem for the rules lies in how to produce the right excitation at the margin between voiced and voiceless sounds. For a vowel, modal voice is specified from just after the start to the end of the segment, because vowels are assumed to have modal voicing throughout. Glides, on the other hand, are not fully specified for **ag**. For them, **ag** is set to 4.0 mm^2 only at the end of the segment. Thus if a glide occurs between two vowels it will be modally voiced throughout; but if it occurs utterance-initially or after a voiceless fricative for which **ag** is set to values > 10.0 , then **ag** will fall throughout the glide to reach 4.0 mm^2 by the end, thereby generating the typical voicing pattern at the start of the glide.

6.2 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. This pattern can be achieved by specifying **an** only once for vowels, at their mid-points. If a vowel is preceded or followed by a nasal consonant, which is of course fully specified for nasality, the vowel will be partially nasalized, and it will be nasalized throughout its duration if it is surrounded by nasals. More complex patterns, such as greater anticipatory than perseverative coarticulation, can be achieved when syllabic and stress information is added. Incidentally, this is an example of a rule 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.

6.3 Obstruents

As already mentioned, 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 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**.

In PROCSY, the only information that is copied from the original signal is the time of occurrence of each part of the obstruent. The labels determine which HL parameter is chosen to form the constriction: the lips (**al**) for $/p b f v/$, the tongue blade for $/t d s z ʃ ʒ θ ð/$ (**ab**). For velars (only $/k g/$ in English), the tongue body position is indirectly specified via a low F1. Automatically-tracked formant information, especially in fricatives and aspiration periods of stops, can be used to some degree, but it proved hard to decide which formant values were useful and which should be regarded as errors. On the other hand, when formant information during the obstruent is ignored, the formant frequencies of the surrounding vowels usually provide enough information for HLsyn to get the place of articulation right. Consequently, automatically-tracked formant frequencies during obstruents are not used in PROCSY.

For most stop consonants, two main parts have to be generated: an oral closure and a burst. Thus the rules model the closure of a particular articulator, its rapid release, and a transition appropriate for the following sound. For bilabials and alveolars, the closure is produced by setting the cross-sectional area for the appropriate oral articulator to 0.0 mm^2 . During the closure, HLsyn models a pressure increase in the oral cavity; for voiceless consonants, the rules normally enhance the increase by opening the glottis more than 4.0 mm^2 . Rapid release of the closure in the oral articulator parameter causes HLsyn to generate a burst with the right spectral shape. After the burst, **ag** stays at or returns to 4.0 mm^2 if a vowel follows the stop, or it may stay open, possibly even widening, if voiceless sounds follow.

Thus voice quality in stops is controlled as follows: **ag** is specified at both the start and the end of the closure period, but only at the start of the burst. In consequence, as in the examples of nasality and voicing, what happens after the burst depends on the following context: in effect, interpolation between control points allows HLsyn to automatically find its way through the desired acoustic segments.

These default rules for stop consonants are not enough for non-standard instances. Complex excitation types are especially common at the borders of sonorant and (especially

voiceless) obstruents. As explained in Section 4, the labels indicate whether the original utterance contains significant ambiguous regions with more than one excitation source. Probably the most physiologically accurate way to model these regions is by modifying the trajectories of the upper articulators so that Hlsyn models an increase in oral pressure and a consequent rise in frication noise, possibly with concomitant reduction in the amplitude of vocal fold vibration. This method has been successfully used to model devoiced vowels in English [9]. We will take the same approach in the expectation that changes in the configuration of the upper articulators will affect formant frequencies as well as the type of excitation: these correlated changes in spectral shape and excitation type exactly reflect the principles that ProSynth researchers espouse as crucial to producing a robust synthetic signal.

We are investigating a number of other ways to produce extended regions of such mixed excitation. A short-cut is to leave the excitation type unspecified in these regions. This produces long transitions between the remaining control points for these parameters, so that both source types co-occur. One way to model unusually strong frication between a stop release burst and following vowel is to prolong the stop closure period, thus causing a rise in oral pressure, a higher-amplitude release burst, and in all probability higher-amplitude aperiodicity subsequent to the release. However, this is not an option for the current copy-synthesis, because segmental durations must mirror those of the original utterance. Instead, the pressure increase is created after the release either by retarding the rate of opening of the upper articulators as described above, and/or by increasing **ap**, which specifies the area of the posterior glottal chink. At present, we use **ap**, as suggested by Sensimetrics. Although this is not necessarily physiologically accurate, it is an easy way to provide different degrees of aperiodicity superimposed on a voiced segment.

Similar adjustments are at times needed at the start of obstruent closures to produce prominent although decaying voicing. A label indicates the temporal extent of such an ambiguous region; the system opens the glottis very slowly in that region, thereby producing a slowly decaying amplitude envelope in the waveform, rather than abrupt voicing offset. To produce the slow decay, **ag** must be $< 10.0 \text{ mm}^2$ at the end of the voiced part of the closure. Again, this method probably does not mirror the physiological facts, but the short-cut is simple and produces the right type of excitation, which is all that is necessary for our immediate purpose. However, in the longer term we intend to model these effects more realistically by controlling parameters such as upper articulator constriction area and the compliance of the vocal-tract walls, **dc**.

Those subtle effects which are not yet part of the rules can be produced very quickly by hand-editing the **.hl** file parameters.

6.4 Dealing with wrong formant frequencies

Since automatic formant measurements are never completely error free, a second 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 thresholds. There are three types of rule. The first type reflects vocal-tract constraints. Constraints on possible vocal-tract shapes limit the patterns that the first four formant frequencies can take: e.g. frequencies of 200, 400, 4000 and 4100 Hz are obviously impossible. Formant frequencies that fall outside pre-defined limits are rejected, not only because they are probably wrong, but also because they cause Hlsyn to behave inappropriately, because Hlsyn interprets formant frequencies not only as such but also as indicators of place of articulation. Wrong formant frequencies therefore lead to the assumption of places of articulation which are not sensible. F1 is especially prone to producing such errors because, as noted in Section 6.3, 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, for instance.

The second way formant frequency errors are detected is by constraining formant movement. At least during a modally-voiced vowel, a change of 100 Hz in F1 over only 5 ms is unlikely, while one of 200 Hz is certainly an error.

Thirdly, context-dependent rules operate from information implicit in the labels. 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/, /a/, or /ɔ/.

Since Hlsyn interpolates between input parameter values to generate the output signal, suspicious formant frequencies can easily be omitted. The remaining formant values are normally sufficient to produce the right output, although of course some points 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”.

6.5 Post-processing

Once the parameters are in an **.hl**-file they are easily edited. For the purposes of ProSynth, the **.hl**-file serves as input for automatic editing programs that allow us to produce modified versions of the same utterance to be used in perceptual tests (see below). Additionally, it is in the **.hl**-file that the Hlsyn parameters are edited by hand if the signal is unsatisfactory in some way. 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: we estimate that PROCSY is about 6 times faster than manual use of HLsyn.

7 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 prosodic structure of an utterance, and 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 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 linearization (stylization), so that we can assess how simple the final control structure of the output can be.

An example of a potential application in speech perception is an extension of work by [4], 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 being 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.

REFERENCES

[1] Fixmer, E. and Hawkins, S. The influence of quality of information on the McGurk effect. Presented at AVSP98 (Satellite of *ICSLP 98*), 1998.

[2] Hawkins, S. Arguments for a nonsegmental view of speech perception. In Elenius, K. and Branderud, P.

(eds.), *Proc. ICPHS XIII*, 3:18–25. KTH and Stockholm Univ., 1995.

[3] Hawkins, S., House, J., Huckvale, M., Local, J., and Ogden, R. Prosynth: An integrated approach to device-independent, natural-sounding speech synthesis. Presented at *Proc. ICSLP 98*, 1998.

[4] Hawkins, S. and Nguyen, N. Effects on word recognition of syllable-onset cues to syllable-coda voicing. Presented at *LabPhon VI*, 1998.

[5] Hawkins, S. and Slater, A. Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *Proc. ICSLP 94*, 57–60, Yokohama, 1994.

[6] Kelly J. and Local, J. *Doing Phonology*. Manchester University Press, 1989.

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

[8] Pratt, R.L. On the intelligibility of synthetic speech. *Proc. Inst. Acoustics*, 8:183–92, 1986.

[9] Rodgers, J. *Vowel devoicing in English*. PhD thesis, Dept. of Linguistics, University of Cambridge, 1998.

[10] Tunley, A. Metrical influences on /r/-colouring in English. Presented at *LabPhon VI*, 1998.

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

[12] <http://synth.phon.ucl.ac.uk/prosynth/>

[13] <http://kiri.ling.cam.ac.uk/procsy/>