SYNTHESIZING SYSTEMATIC VARIATION AT BOUNDARIES BETWEEN VOWELS AND OBSTRUENTS

Sebastian Heid and Sarah Hawkins
Department of Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, U.K.

ABSTRACT

This work assesses whether natural-sounding excitation near segment boundaries enhances the intelligibility of formant synthesis. Excitation type at fricative-vowel (FV) and vowel-fricative (VF) boundaries and durations of voicing in voiced stop closures are described for one male speaker of British English. Most VF boundaries have mixed aperiodic and periodic excitation, whereas most FV boundaries change abruptly from aperiodic to periodic excitation. Syllable stress, vowel height, and final/non-final position within the phrase influenced the incidence and duration of mixed excitation. Voicing in stop closures varied in well-understood ways. Synthesized phrases proved more intelligible in noise when excitation at fricative boundaries and in voiced stop closures was structurally appropriate. Implications for formant synthesis are discussed.

1. INTRODUCTION

This work is part of ProSynth [1, 2], a research program that addresses the premise that synthetic speech will sound more natural and be more intelligible, especially in adverse conditions, if it includes the systematic variation that is found in natural speech and reflects linguistic structure which is typically ignored in synthesis. This is important because, though synthetic speech may be as intelligible as natural speech in good listening conditions, it is much less intelligible in adverse conditions [3].

We have argued [4] that this fragility of synthetic speech is due to its unnatural quality. When humans speak, there is a tight relationship between movements of the vocal tract and properties of the emitted sound. Thus natural speech is *perceptually coherent* because it is *acoustically coherent:* its acoustic-phonetic fine detail reflects vocal tract behavior. When vocal-tract behavior systematically reflects linguistic structure, then that linguistic structure should be more transparent to the listener. It follows that if synthetic speech includes subtle but systematic phonetic variation, its intelligibility should rise. This seems to happen: when subtle but systematic changes in formant frequency that reflect the presence of an /r/ are introduced over several syllables of synthetic speech, intelligibility rises significantly [5, 6]. Some of these spectral changes are barely noticeable in good listening conditions; their impressionistic effect is best described as making the speech sound more coherent overall rather than as making phones obviously clearer. This impression of coherence may arise because the changes reflect the behavior of the tongue body in the general vicinity of an English /r/ [7]. The changes also, of course, spread information that there is an /r/ in the message over a longer stretch of the signal.

Another contribution towards perceptual coherence may come from the detailed waveshape. The waveform amplitude envelope provides useful perceptual information to listeners [8, 9]. The waveshape at segment boundaries seems especially likely to contribute perceptual coherence, and hence more natural-sounding and robust synthetic speech. For example, the abruptness of voicing offset reliably differentiates between voiced and voiceless fricatives at vowel-fricative boundaries [10], and in the vicinity of the boundaries between obstruents and non-obstruent articulations there are often regions of mixed periodic and aperiodic excitation as the vocal-tract opens or closes through a critical range of constriction areas. Acoustic patterns of these types may contribute to perceptual coherence by enhancing stream segregation, as for nonspeech sounds [11].

One reason for studying the distribution and perceptual importance of segment-boundary patterns is that these patterns arise naturally in concatenative synthesis, but must be produced at some computational cost in formant synthesis. Concatenative synthesis is immediately appealing because it seems to sound more natural than formant synthesis. This impression may arise partly because the acoustic fine-detail in the vicinity of segment boundaries contributes crucial perceptual coherence. If that is so, then formant synthesis would benefit from such fine detail.

The two properties chosen for this study are the excitation type at the boundaries between voiceless fricatives and sonorants, and the duration of voicing in the closure of voiced stops. Mixed excitation at fricative boundaries represents a difficult case for formant synthesizers, which typically produce abrupt changes between periodic and aperiodic sources. Voicing in the closure of stops is normally included in some contexts in formant synthesis systems, but was included in this study for two reasons. First, [12] have argued that low-amplitude, low-frequency periodicity is processed by the auditory system in a way that is compatible with the interpretation that it contributes to perceptual coherence through auditory streaming; in this sense, it is distinct from yet comparable with the class of fricative-sonorant boundaries. Second, whereas the literature seems silent on the incidence of mixed excitation in the vicinity of the boundaries between voiceless fricatives and sonorants, structurally-determined differences in voicing in the closure of voiced English stops are comparatively well understood.

The first step, then, is to identify the distribution of the patterns of interest in natural speech. The second step is to synthesize phrases containing the phoneme sequences of interest, and to assess their intelligibility, with and without the patterns of excitation type found in the natural speech.

2. ANALYSIS OF NATURAL SPEECH

The speech analysed is from the ProSynth database. It comprises single tokens of phrases, exemplifying a number of prosodic structures, from a male speaker of Southern British English. Because the database satisfies a number of other constraints, it is neither large nor balanced for the factors studied here, but it offers a good range of relevant structures.

2.1. Fricatives

Voiceless fricatives were divided into two sets according to the position of the boundary with the vowel: fricatives preceded vowels in the FV set, and followed them in the VF set. When a fricative fell between two vowels, it contributed to both groups. Further subdivisions were made according to structural and prosodic context, and vowel height, as outlined below.

Voiceless fricatives classed as having mixed excitation (or *overlap*) had at least 8 ms of simultaneous periodic and aperiodic excitation at the segment boundary. Thus some cases classed as simple excitation did have a short region of mixed excitation, but it was low amplitude and less than about one glottal period, which seems unlikely to be perceptually significant.

boundary	n	% simple	% mixed
FV	173	82	18
VF	158	27	72

Table 1. Percentage of fricatives with simple vs. mixed excitation types at the boundary with an adjacent vowel. n = total number.

Table 1 shows that the vast majority of FV boundaries had simple excitation types, whereas most VF boundaries had mixed types. Most FV boundaries were simple, but of the minority with mixed excitation, there was a much higher proportion of cases in unstressed compared with stressed contexts (Table 2; $\chi^2 = 10.31$, df = 1, $p < 0.01$). Stressed contexts have stress either preceding or following the fricative; unstressed contexts have unstressed vowels on both sides of the fricative. No other significant influences were observed for the mixed FV boundaries.

context	n	% simple	% mixed
stressed	75	92	8
unstressed	98	73	27

Table 2. Percentage of FV boundaries with simple vs. mixed excitation types, when the vowel is stressed or unstressed.

Amongst the minority of VF boundaries that had simple excitation, less than a quarter had onset fricatives while almost half had coda fricatives, but these trends are not significant ($p < 0.1$, $\chi^2 = 2.88$, df = 1).

Durations of overlap at the 32 mixed FV boundaries ranged between about 8 ms (by definition) and 18 ms, in a tight distribution with all but 5 cases less than 12 ms: the mean and median were both about 10 ms (s.d. 1.98 ms). The mean duration of overlap for all FV boundaries (simple + mixed) was of course much less since most of them were classed as simple.

The VF distribution was both wider and more skewed than the FV distribution. The mean duration of mixed excitation for the 114 mixed VF boundaries was 18 ms (s.d. 11 ms). Removing 6 outliers (range 38-78 ms) reduced the mean to 16 ms (s.d. 5.61 ms, median 18 ms, range 8-34 ms). The 6 outliers all had stressed low vowels preceding the fricative, which was the syllable coda; the 4 longest were phrase-final. (The fricative was in the coda in all 21 cases with low vowels.) The duration of mixed excitation in VFs in *un*stressed contexts with preceding low vowels fell within the typical range, even phrase-finally. Tables 3-5 summarize these points; contexts are separated to avoid small Ns. Mid and central vowels are omitted from vowel height analyses.

vowel at VF boundary	n	mean	s.d.
high: /i ɪ u ʊ/	75	15	5.8
low: /a ɑ ɒ ʌ/	21	27	21.3

Table 3. Durations of mixed excitation (ms) at VF boundaries classed as mixed (≥ 8 ms), for high and low vowel contexts.

VF	n	mean	s.d
non-final	93	16	6.9
final	21	25	20.2

Table 4. Durations of mixed excitation (ms) at VF boundaries classed as mixed for phrase non-final and final syllables.

	VOWEL HEIGHT					
	high: /i ɪ u ʊ/			low: /a ɑ ɒ ʌ/		
CONTEXT	n	mean	s.d.	n	mean	s.d.
stressed	64	15	5.7	15	31	23.4
unstressed	11	15	6.4	6	14	4.1

Table 5. Durations of mixed excitation (ms) at VF boundaries classed as mixed, according to vowel height and stress context.

In summary, FV boundaries are normally simple i.e. the transition from aperiodic to periodic excitation is abrupt. When there is mixed excitation, it is usually in unstressed contexts, and only about 10 ms long. In contrast, VF boundaries are normally mixed, and although the typical duration of mixed excitation is around 16-20 ms, it can be as long as 50-80 ms, especially in phrase-final syllables with stressed low vowels.

These observations seem to indicate that mixed excitation results mainly from asynchronies between glottal and upper-articulator movements, rather than aerodynamic factors stemming from differences in the time the articulators take to achieve a constriction area that produces frication. If gross articulator movement determined the patterns, we might expect more and/or longer regions of mixed excitation with slower rates of change in jaw height. Slower changes would be expected at FV boundaries and with high, unstressed vowels, because jaw velocity is typically lower for opening than closing, for unstressed than for stressed syllables, and for high rather than low vowels [e.g. 13]. Similarly, the cross-sectional area at the place of maximum vocal tract constriction changes more abruptly in VF compared with FV sequences [14]. This should produce more abrupt changes in excitation type at VF rather than FV boundaries, whereas we found the opposite. Obviously articulatory kinematics and direct aerodynamic consequences of oral constriction areas influence the data, but since they seem to predict the opposite pattern from our findings, the main determinant of the asymmetries in our data is probably asynchrony in coordinating glottal and upper- articulator movement.

2.2. Voiced stops

We measured the duration of periodicity in the closure of voiced stops, and the duration of the closure itself (to the burst). The patterns were as expected, so are only summarized. When the

preceding sound was voiceless, there was no periodicity in the closure; these are not discussed further. In the 212 closures preceded by a voiced sound, voicing was very slightly longer when the following sound was also voiced, but the standard deviations are quite big (39 ms, s.d. 18.7 ms, n = 154 vs. 34 ms, s.d. 15.2 ms, n = 58). Vowel height and place of articulation influenced the duration of closure voicing, with longer durations after low rather than high vowels, and bilabial > alveolar > velar stops; these patterns may reflect differences in oral cavity volume, for larger volumes allow the transglottal pressure-drop to last longer. When the following sound was voiced, syllable stress affected closure voicing: periodicity in a stop before an unstressed syllable was about 43 ms, regardless of whether the stop was an onset or a coda, and of whether the preceding syllable was stressed or unstressed. In contrast, mean closure voicing was only 34 ms when the stop was the onset of a stressed syllable. This absolute value is the same as that for stressed and unstressed codas in phrase-final position or followed by voiceless sounds. However, the perceptual impression will be different: whereas periodicity in stressed onset stops takes up only 40% of the closure, that in codas followed by voicelessness take up 70% of the closure. Other following voiced contexts have even longer (>80%) proportions of closure voicing.

3. PERCEPTUAL EXPERIMENT

3.1. Method

3.1.1. Material. 18 phrases from the database were copy-synthesized into HLsyn using PROCSY [15], 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: at the onset of a stressed syllable—*in a field*; unstressed onset—*it's surreal*; coda of an unstressed syllable—*to disrobe*; between unstressed syllables—*disappoint*; coda of a final stressed syllable—*on the roof, his riff , a myth, at a loss, to clash*; and both unstressed and stressed onsets—*fulfilled.* The other 8 items had voiced stops as the focus: in the coda of a final stressed 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 contexts—*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. From each right version, a "wrong" one was made be substituting a type or duration of excitation that was inappropriate for the context. Changes were systematic; no attempt was made to copy the exact details of the natural version of each phrase, as our 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. See Fig. 1. Right versions were expected to be more intelligible than wrong versions of fricatives.

Each stop had one of two types of wrong voicing: longer-than-normal voicing for *band* and *begged* (see Fig. 2) 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 6 were expected to be less intelligible in noise if naturalness and intelligibility co-varied.

Figure 1. Spectrograms of part of /IS/ in *disappoint*. Left: natural; mid: synthetic "right" version; right: synthetic "wrong" version.

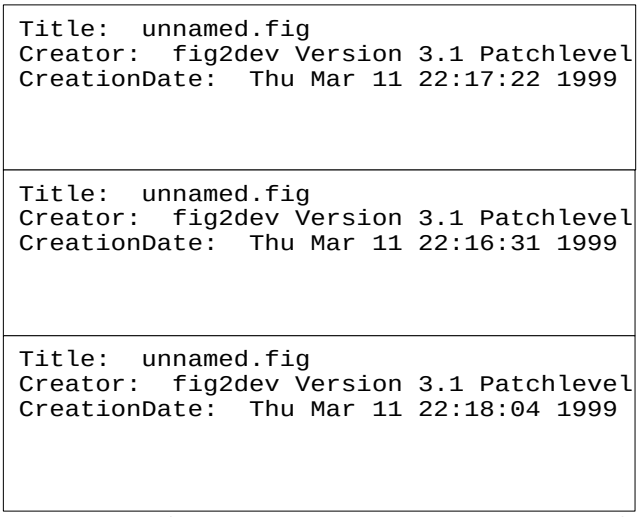

Figure 2. Waveforms showing the region around the closure of /b/ in *he begged*. Upper panel: natural speech; middle: "right" synthetic version; lower: hyper-speech synthetic version.

3.1.2. Subjects. The 22 subjects were Cambridge University students, all native speakers of British English with no known speech or hearing problems and less than 30 years old.

3.1.3. 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. They 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. They pressed a key to hear each item, and wrote down what they heard. Each listener 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.

3.2. Results

Responses were scored for number of phonemes correct. Wrong insertions in otherwise correct responses counted as errors. There were two analyses, one on all phonemes in the phrase, the other on just three—the manipulated phoneme and the 2 adjacent to it. Table 6 shows results for 16 phrases i.e. excluding the two hyper-speech phrases. Responses were

significantly better ($p < 0.02$) for the right versions in the 3-phone analysis, and achieved a significance level of 0.063 in the whole-phrase analysis.

context	version of phrase		t(21)	p (1-tail)
	"right"	"wrong"		
3 phones	69	61	2.35	0.015
entire phrase	72	68	1.59	0.063

Table 6. Percentage correctly identified phonemes in 16 phrases.

Responses to the hyper-speech words differed: 84% vs. 89% correct for normal vs. hyper-speech *begged*; 85% vs. 76% correct for normal vs. hyper-speech *band* (3-phone analysis). Hyper-speech *in the band* was often misheard as *in the van*. This lexical effect is an obvious consequence of enhanced periodicity in the /b/ closure of *band*.

4. DISCUSSION

We have shown for one speaker of Southern British English that linguistic structure influences the type of excitation at the boundaries between voiceless fricatives and vowels, as well as the duration of periodic excitation in the closures of voiced stops. Most FV boundaries are simple, whereas most VF boundaries are mixed. Within these broad patterns, syllable stress, vowel height, and final vs. non-final position within the phrase all influence the incidence and/or duration of mixed excitation. We interpret these data as indicating that the principal determinant of mixed excitation seems to be asynchrony in coordinating glottal and upper articulator movement. Timing relationships seem to be tighter at FV than VF boundaries, and there can be considerable latitude in the timing of VF boundaries when the fricative is a phrase-final coda.

Our findings for voiced stops were as expected, if one assumes that the main determinants of the duration of low-frequency periodicity in the closure interval are aerodynamic. One interesting pattern is that voicing in the closures of prestressed onset stops is short both in absolute terms and relative to the total duration of the closure.

We further 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 i.e. one edge of one phone or diphone, was manipulated in most of the phrases, and that there are relatively few data points, the significance levels achieved testify to the importance of synthesizing edges that are appropriate to the context. It is encouraging that differences were still fairly reliable in the whole-phrase analysis under these circumstances, since we would expect more response variability over the whole phrase.

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 presumably enhance its robustness in noise considerably. However, it is a truism that at times the speech style that is most appropriate to the situation is not necessarily the most natural one. The two instances of hyper-speech are a case in point. 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. But this manipulation risked making the /b/s sound like /v/s, effectively widening the neighborhood of *band* to include *van*. Since *in the van* is as likely as *in the band*, contextual cues could not help, and *band*'s intelligibility fell. *Begged*'s intelligibility may have risen because there were no obvious lexical competitors, and because we also enhanced the voicing in the syllable coda, thus making a more extreme hyper-speech style, and, perhaps crucially, a more consistent one. These issues need more work.

The perceptual data 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 with HLsyn, on which PROCSY is based.

REFERENCES

[11] Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press.

[4] Hawkins, S. 1995. Arguments for a non-segmental view of speech perception. *Proceedings of the XIIIth ICPhS*, 3, 18-15.

[5] Hawkins, S., and Slater, A. 1994. Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *Proceedings of ICSLP 94*, 1, 57-60.

[1] Hawkins, S., House, J., Huckvale, M., Local, J., and Ogden, R. 1998. ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Proceedings of the 5th International Conference on Spoken Language Processing.* 1707-1710.

[7] Kelly, J., and Local, J. 1989. *Doing Phonology*. Manchester, UK: Manchester University Press.

[12] Kingston, J., and Diehl, R.L. 1994. Phonetic knowledge. *Language* 70, 419-454.

[13] Ostry, D., and Munhall, K. 1985. Control of rate and duration of speech movements. *J. Acoustical Society of America* 77, 640-648.

[3] Pratt, R.L. (1986). On the intelligibility of synthetic speech. *Proceedings of the Institute of Acoustics* 8 (7), 183-192.

[15] PROCSY: http://kiri.ling.cam.ac.uk/procsy/

[2] ProSynth: http://www-users.york.ac.uk/~lang19/

[9] Rosen, S., and Howell, P. 1987. Auditory, articulatory, and learning explanations of categorical perception in speech. In Harnad, S., ed. *Categorical Perception: The Groundwork of Cognition*. Cambridge, UK: Cambridge University Press. 113-160.

[14] Scully, C., Grabe-Georges, E., and Castelli, E. 1992. Articulatory paths for some fricatives in connected speech. *Speech Communication* 11, 411-416.

[6] Tunley, A. 1999. *Coarticulatory influences of liquids on vowels in English.* Unpublished PhD dissertation. University of Cambridge, UK.

[8] van Tasell, D.J., Soli, S.D., Kirby, V.M. and Widin, G.P. 1987. Speech waveform envelope cues for consonant recognition. *J. Acoustical Society of America* 82, 1152-1161.

[10] Zue, V. 1988. *Speech Spectrogram Reading.* MIT (RLE).