SPREAD OF CV AND V-TO-V COARTICULATION IN BRITISH ENGLISH: IMPLICATIONS FOR THE INTELLIGIBILITY OF SYNTHETIC SPEECH

Sarah Hawkins and Andrew Slater

Department of Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK

ABSTRACT

Formant-based synthetic speech is less robust in noise than natural speech and is often criticised as "robot-like". One reason may be the failure to model systematic spectral variation that is not crucial to phoneme identification. This study investigates how some aspects of consonantal context and stress contribute to this variation. CV sequences were embedded in carrier phrases to give quasi-meaningful English sentences. Vowels were /a u i ə/ or /ʌ u i ə/; /z/ or /ɪ/ followed /u/, other consonants were either all /b/ or all /d/. Stress was on the second syllable (Set 1), or the first and third (Set 2). F2 and F3 frequencies were lower when consonants were /r/ and /b/ rather than /z/ and /d/, but "r-lowering" was not significant in Set 2, presumably because vowel quality, and hence tongue position, were more constrained by stress. R-lowering may spread to syllables which are not adjacent to /r/, typically across unstressed vowels and labial consonants. The measured differences were usually audible. Modelled in synthetic speech, they can increase the phonemic intelligibility of the speech in noise by about 15%.

1. INTRODUCTION

1.1. Contribution of naturalness & acoustic coherence to the robustness of synthetic speech

Formant-based synthetic speech can be nearly as intelligible as natural speech in good listening conditions, but it sounds less natural, and in noise it is much less intelligible. Natural speech is about 15% less intelligible at 0 dB SNR than in quiet, whereas synthetic speech drops by 35%-50% [7]. We conjecture that the fragility of synthetic speech in noise is related to its unnaturalness. When humans speak, there is a tight relationship between the movements of the vocal tract and properties of the emitted sound. Thus natural speech is acoustically coherent: its acoustic-phonetic fine detail reflects vocal tract behaviour. This fine detail is found in all aspects of speech, e.g. in the periodic glottal source spectrum and the amplitude envelope, especially at abrupt segment boundaries, and in coarticulatory effects on formant frequencies. Effects of these types contribute to the variability of the signal. But variation from these sources is systematic, lawful variation, and adds information to the signal. We suggest that this lawful variation contributes acoustic coherence to the signal; and that while some aspects of acoustic coherence are essential to achieve basic intelligibility, other aspects provide the naturalness that makes real speech easier to understand under difficult listening conditions.

There is evidence from at least three fields of enquiry that synthetic speech will be easier to understand if it is maximally acoustically coherent (or natural): from auditory psychophysics; from theories of human speech perception; and from the popularity of concatenated natural speech over formant-based synthesis in speech applications.

Experiments show that when sounds have certain temporal and spectral relationships to one another, humans group them into coherent patterns. This is called auditory streaming [1]. Whether a time-varying signal (like speech) is heard as coming from a unitary source depends on tight spectral and temporal relationships between its various events. Appropriate relationships are heard as a coherent pattern from a single source; inappropriate relationships are heard as several streams, or else unrelated acoustic events.

Severe incoherence is now rare in synthetic speech, but subtle forms still persist: some sequences sound much more acceptable than others that are just as intelligible. Evidence from auditory streaming strongly suggests that to produce robust synthetic speech, we must pay attention to the fine acoustic detail: to variation that has typically been ignored as not essential to the identification of phonemes.

Theories of speech perception, most notably the motor theory [5] and direct realism [3], provide the second strand of evidence. One need not espouse such theories in their entirety to acknowledge the importance of their central tenet: that the listener's knowledge of the relationship between vocal tract behaviour and sound profoundly influences his or her interpretation of the signal. Presumably this knowledge includes lawfully varying fine acoustic detail caused by the vocal tract system.

Thirdly, the phonetic quality of formant-based synthetic speech is not much better than it was a decade ago, and, largely in consequence, many applications use concatenated natural speech. In formant synthesis, strict phonemic intelligibility (of isolated CV, VC, and VCV syllables) seems to reach a ceiling of performance above which it is difficult to make significant improvements. Yet our impression is that the appeal of concatenated speech is mainly that it sounds more natural. This is not coincidental. If our arguments are right, concatenated speech has at least two built-in advantages over formant synthesis: it contains all the short-term lawful variation (especially at segment boundaries) of natural speech, and, if done well, it contains
at least some of the longer-term lawful variation of natural speech. Typically, formant synthesis mimics only some of these relationships - mainly those basic to phoneme identification, speech rhythm, and intonation.

To summarise, we suggest that natural speech is robust in noise because it is acoustically coherent, due to the acoustic-phonetic fine detail that reflects vocal tract behaviour. The fragility of most synthetic speech implies that it lacks this acoustic coherence. As well as sounding less natural, it is less redundant and harder to attend to and track in difficult listening conditions. We hope to improve both the naturalness and robustness of our synthesis by modelling the lawful variation found in natural speech.

1.2. Lawful variation in vowel quality

Many types of fine detail contribute to lawful variation. In this paper we focus on spectral variation in vowels induced by consonants. Our theoretical motivation is twofold. First, variation in vowel formant frequencies has both short-term and longer-term aspects (e.g. CV transitions and vowel-to-vowel coarticulation respectively) and is presumably central to achieving a coherent signal from a unitary sound source. Second, vowel formant frequencies reflect complex articulatory interdependencies, and so promise insights into the structural description of phonological, articulatory, and perceptual units. A practical motivation is to improve vowel quality in Infonova British English text-to-speech and develop the system to optimise the production of such variation. Southern British English is particularly good to work on because of its tendency to centralize vowels and the consequent gradations in quality of the same phoneme in different contexts.

We are especially interested in consonantal effects on vowels that extend beyond the temporal domain of local coarticulation. These we term consonantal resonance effects because they normally arise when the oral cavity assumes a shape throughout most of a syllabic nucleus that is due to a nearby consonant. At least four such resonance effects can be identified for English: palatality or coronality (not distinguished at this stage), rhoticity, nasality, and labiality. We consider palatality and rhoticity here because they have received less attention than nasality and labiality.

2. EXPERIMENT 1: ANALYSIS OF NATURAL SPEECH

2.1. Assumptions and hypotheses

We chose /z/ to exemplify palatality/coronality, and /r/ for rhoticity. Sentences containing sequences likely to allow or to block the spread of these lingual resonance effects were constructed. We assumed that the effects can spread if the tongue is relatively free: when surrounding consonants are not lingual (e.g. /b/), and when vowels are unstressed or weakly stressed. Conversely, we assumed lingual resonance effects will not spread if the tongue is constrained by a conflicting articulation, as when nearby consonants are lingual but have a conflicting configuration (e.g. /s/ surrounding /r/), or when vowels are stressed.

These assumptions led to four hypotheses.

1. Resonance effects due to /z/ will raise F2 frequencies in surrounding vowels, due partly to the location of the constriction and partly to the tapering of the back cavity immediately behind the constriction.
2. A rhotic resonance effect will lower F2 frequencies in surrounding vowels due to retroflexing the tongue tip and/or bunching the tongue body.
3. F2 frequencies will be lower in /b/ than /d/ contexts, due to the effects of bilabial vs alveolar occlusion.
4. Resonance effects of /z/ and /r/ will spread to nonadjacent unstressed vowels in /b/ contexts, but /d/ contexts will block the spread of effects due to /r/, and be neutral towards or enhance effects due to /z/.

Resonance effects of /z/ or /r/ were /a:/ and /a/ because it varied a lot within context in our pilot work; /a/ because it often varies a little, perhaps due to the degree of palatal contact with its, yet its quality does vary when it is subject to /l/ vs /r/ resonance effects e.g. Henry vs. Henley.

2.2. Method

Table 1 shows two sets of CV sequences differing mainly in terms of stress. Vowels were /a/ or /a:/ or /a/ /i/ /a:/; /z/ or /r/ followed by /a:/; other consonants were either all /b/ or all /d/: Primary stress was on either the second syllable (Set 1), or the first and third (Set 2). These sequences were embedded in quasi-meaningful sentences listed in Table 2. Fully comparable meaningful sentences with all factors constrained on both sides of the /z/ and /r/ do not exist, but sentences with constraints on only one side are common. For example, They have a boring book vs. They have adoring parents.

Table 1. Sequences used in Experiment 1.

<table>
<thead>
<tr>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ba'buizibab/</td>
<td>/ba'bu'zi:ibab/</td>
</tr>
<tr>
<td>/ba'bu:ribab/</td>
<td>/ba'bu'ri:bab/</td>
</tr>
<tr>
<td>/bo'duizibab/</td>
<td>/bo'du'zi:ibab/</td>
</tr>
<tr>
<td>/bo'du:ribab/</td>
<td>/bo'du'ri:ribab/</td>
</tr>
</tbody>
</table>

Table 2. Sequences in quasi-meaningful sentences.

<table>
<thead>
<tr>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grab a booby baboon today.</td>
<td>Speak to Bubboo Ziba before Monday</td>
</tr>
<tr>
<td>Grab a boorly baboon today.</td>
<td>Speak to Bubboo Ribba before Monday</td>
</tr>
<tr>
<td>Grab a doozy doublloon today.</td>
<td>Speak to Buddoo Rida before Monday</td>
</tr>
<tr>
<td>Grab a doory doubloon today.</td>
<td>Speak to Buddoo Zida before Monday</td>
</tr>
</tbody>
</table>

The sentences were randomised in blocks to give 10 tokens of each. The resultant list was read by 5 speakers of British English, 4 men and 1 woman, aged 23-45 years, all phonetically-trained fluent readers. They were told to think of "Bubboo Ribba" etc. as people and practised the material before recording onto DAT in a sound-treated room.

The speech was digitized at 16 kHz using Silicon Graphics Indigo A/D converters. The first three formant frequencies were measured from 30 ms Burg lpc spectra at the midpoint of each of the four vowels in each critical sequence (18 poles for the female voice, 20 for the males).
2.3. Results

Only F2 data are presented, since F1 and F3 showed smaller, less consistent effects. Figure 1 shows the mean frequencies of F2 at the midpoints of each of the four vowels in Set 1 (1a) and Set 2 (1b). The overall shape of the curves reflects the pattern of F2 frequencies for these vowels. We are interested in the vertical differences within each vowel, because these reflect the effects of the different consonants. ANOVAs were done on each vowel separately (r/z x b/d context x subjects). Differences reported are significant at p = 0.05 or better.

Fig. 1. Mean frequencies of F2 for 5 speakers at the midpoints of vowels in (a) Set 1 and (b) Set 2. Solid lines, /|/ contexts; dashed lines, /d/ contexts; filled squares, X; crosses, X.

(a) Set 1
kHz 2.5

(b) Set 2
kHz 2.5

Both sets show the same pattern in 3 respects. (1) F2 is lower in /|/ contexts than /d/ contexts throughout. (2) In each set, the last vowel varies more with b/d context than the first, and the first /|/ in Set 1 varies more than the /|/ in Set 2. This suggests (a) more carryover than anticipatory coarticulation (b/d fall after the first /|/ or /d/ but before the last /|/; other contexts are the same); (b) unstressed /|/ seems to be more susceptible than stressed /|/ to b/d effects. (3) Variation due to b/d context is greater for /|/ than for /d/; though differences are significant within both vowels. This suggests that /|/ is relatively stable, but can vary in quality when particular articulations dominate.

syllabic structure. Contrary to our expectations, /|/ and /d/ appear to dominate these syllables, rather than /z/ and /r/.

Set 1 differs from Set 2 in that /r/ engenders significantly lower F2 frequencies than /z/ in all four vowels in Set 1 (.baboon..), whereas /z/ and /r/ had almost no effect on the midpoint formant frequencies in Set 2 (.Bubboo..). Although our predictions about the blocking effects of stressed vowels are consistent with this finding, the difference between the two sets is better explained in terms of allomorphic differences conditioned by the position of /z/ or /r/ in the word. In Set 1, /r/ and /z/ are word-medial, whereas they are word-initial in Set 2. Many speakers of British English lower and back the allophone of /u/ from [u] to something closer to [u] or [o]. This large change in the quality of /u/ then conditions changes in vowels in the other syllables. This explanation is confirmed by additional data from our female speaker: when she used [u] instead of [u] in boory and doory, there was no difference between the /u/s in /z/ and /r/ contexts.

2.4. Discussion

As expected, /|/ and /d/ condition differences in mid-vowel F2 frequencies, with the biggest differences in following vowels, presumably due to carryover articulation. Stress and vowel quality (/u/ vs /i/) affect the size of these differences. The expected effects of /z/ were only indirect: word-medial /r/ preceded by /u/ conditioned a lower and more backed allophone that in turn affected vowels in nonadjacent syllables. We believe that this trans-syllabic spread of r-lowering to the schw of a nonadjacent syllable may be best viewed as vowel-to-vowel coarticulation, albeit conditioned by a consonant somewhat remote from the schw. Our main interest is in assessing the effects of measured differences on the robustness of synthetic speech, space prevents further discussion.

3. EXPT. 2: INFLUENCE OF RESONANCE EFFECTS ON INTELLIGIBILITY OF SYNTHESIS

3.1. Method

Similar sentences to those used in Set 1 of Expt. 1 were synthesized using the default male voice of the Infovox British English system (Version 3.1). Set 2 was omitted since it was not interestingly different from Set 1. The more natural [u] allophone of /u/ was used for boory and doory. Two further word pairs, daddy leaves and dandy leaves, were added for interest. Local [6] notes differences in Tyneside English word-final /l/, conditioned by the rhythm of the preceding accented syllable. Our data for Southern British English show small differences in these words and surrounding central vowels. Unlike Expt. 1, the noncritical parts of each sentence were all slightly different, as shown in Table 3; critical words are in bold.

Two versions of each sentence were synthesized. The Rule Form was generated by the Infovox rules, with small changes to segment durations and f0 if necessary. The Experimental Form was the same as the rule form, except that the frequencies of F1, F2 and F3 in the critical 4 vowels were modified in accordance with the findings of Expt. 1. Where formant changes were inconsistent between
speakers, the pattern for one male subject was followed.

Table 3. Sentences used in Experiment 2.

<table>
<thead>
<tr>
<th>Grab a zooey baboon today</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snob a boorey baboon on Tuesday</td>
</tr>
<tr>
<td>Nab a douzeey dubbleoon tomorrow</td>
</tr>
<tr>
<td>Rob a doorey dubbleoon on Thursday</td>
</tr>
<tr>
<td>Her daddye leaves the theatre early</td>
</tr>
<tr>
<td>The dandy leaves the train at Royston.</td>
</tr>
</tbody>
</table>

Two tapes were made, each with half the critical sentences such that each sequence of words occurred only once on each tape. The critical sentences were randomised with 6 filler sentences and preceded by 10 practice sentences, all generated by rule. The synthetic sentences and some randomly varying (cafeteria) noise were digitized at 16 kHz, mixed at 10 dB s/n (RMS) and recorded on DAT with 10s ISI. The same noise was used for both tapes. Two groups of 30 subjects, all students with no phonetic training, each heard one tape, free-field in a sound-treated room. Subjects were told that some of the words were nonsense. They had a list of the sentences with the critical word(s) omitted, and filled in the blank(s).

The data were scored for the number of phonemes correct in the critical word or words. The maximum score for each word was the number of its phonemes, 1 point was deducted for each false insertion. The minimum score per word was 0. Raw scores were converted to percentage of phonemes correct per item for subsequent analysis.

3.2. Results and Discussion

The average percentage of phonemes correct for the critical words are given in Table 4. The experimental forms were significantly more intelligible than the rule forms (F(1,356) = 10.5, p < 0.001 in ANOVA (Group x Version x Subjects, collapsed over stimulus word). There were no other significant effects.

Table 4. Percentage of phonemes correct in critical words of listening test

<table>
<thead>
<tr>
<th>Rule</th>
<th>Experimental</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>boozy</td>
<td>61</td>
<td>77</td>
</tr>
<tr>
<td>boorey</td>
<td>81</td>
<td>93</td>
</tr>
<tr>
<td>douzeey</td>
<td>54</td>
<td>71</td>
</tr>
<tr>
<td>doorey</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>daddye leaves</td>
<td>98</td>
<td>97</td>
</tr>
<tr>
<td>dandy leaves</td>
<td>81</td>
<td>86</td>
</tr>
</tbody>
</table>

Of the 4 forms studied in Expt. 1, all except doorey were much more intelligible in the experimental version than in the rule-generated version. The synthesis of doorey was difficult: neither the experimental nor the rule form sounded good, and they were almost indistinguishable. This contrasts with the other 3 words, whose experimental forms all sounded better than their rule forms (markedly for douzeey, rather little for boorey). A better-sounding version of doorey was made, but was not tested since it lowered F2 in /a/ so much that the relationship in our natural speech data between F2 in douzeey, doorey, boozy and boorey was violated.

Daddy leaves and dandy leaves were worked on for a very short time and the spectral changes made were small. Presumably a ceiling effect stopped daddy leaves from improving. Dandy leaves improved acceptably, given the time spent on it and the small spectral changes made.

4. CONCLUSIONS

Consonant-conditioned variation in vowel quality throughout a syllable seems well worth incorporating into synthetic speech. In our data, it can increase the phonemic intelligibility of speech in moderate degrees of randomly fluctuating noise by about 15%, even for open-response to nonsense words. Although we have not tested the drop in intelligibility between quiet and noise, this improvement in performance could close the gap in intelligibility between natural and synthetic speech shown with the less stringent DRT [7]. These data suggest a role for the syllable and foot in the segmental-phonetic structure of text-to-speech, and have theoretical implications for speech perception. They support a parallel-processing model in which probability values associated with phonetic features are continuously updated from the incoming signal, and in which relatively weak evidence for phonetic features that is consistent over time can be as effective as clear, rapid acoustic events in determining phonemic identity [4,9]. Such models are compatible with the view that pattern-matching techniques like HMMs are relevant to speech perception by humans.

[Supported in part by Infovox AB, Sweden.]

REFERENCES