ASSESSMENT OF NATURALNESS IN THE PROSYNTH SPEECH SYNTHESIS PROJECT

Sarah Hawkins[1], Sebastian Heid[1], Jill House[2], Mark Huckvale[2]

Abstract

This paper describes the approach to the assessment of the naturalness of synthetic speech taken within the ProSynth collaborative speech synthesis project. The view expressed is that an important aspect of naturalness is ease of understanding, and the consequences are that this leads to a means for the evaluation of scientific hypotheses through perceptual tests. The premise within ProSynth is that listeners' processing of synthetic speech will be faster and more accurate when the signal includes phonetic fine detail that systematically varies with the linguistic structure. Four perceptual experiments are outlined which demonstrate both the application of the approach and the effectiveness of the basic principle.

1. Introduction: What is Naturalness?

What would it mean to say that synthetic speech was 'natural' sounding? A naïve answer would be that such a signal 'could have been produced' by a human being; a more stringent answer that the signal 'is indistinguishable' from human speech; an operational answer that the speech is 'typical' of human production of that sentence.

None of these is satisfactory from a scientific point of view: a signal that could have been produced by a human being might still be unintelligible; a signal that is recognisably a machine may still be perfectly acceptable to listeners; a signal that is close to the mean on multiple parameters may be rather dull.

The problem is that 'naturalness' is a multifaceted characteristic, and it doesn't make sense to measure or assess our synthetic speech systems along a single dimension. Consider a rating scale experiment in which listeners are asked to rank samples for naturalness: listeners will be attending to a number of characteristics of the speech, among them: voice quality, prosody, intelligibility, coarticulatory coherence, or presence of acoustic processing artifacts. Since it is probable that listeners are most affected by voice quality, this will affect final scores most. However the problems of synthetic speech are more wide ranging than voice quality.

Within the ProSynth collaboration (Hawkins *et al*, 1998; Ogden *et al*, 2000), we have taken the view that the most significant deficiency of contemporary synthetic speech is that it is *much more difficult to understand* than natural human speech; and that this is the case even when the words themselves are relatively intelligible when heard individually. Evidence for this comes from studies which have found that the intelligibility of synthetic speech declines more strongly than human speech in adverse listening conditions (Pratt, 1986; Duffy & Pisoni, 1992), as well as from all our own experience. The cause of this increase in cognitive load can be attributable to a range of poorly modelled phenomena in synthetic speech: in the unpredictable rhythm, in the mismatch between prosody and information structure, in the failure to model the precise context-sensitive realisation of elements. In our view then, 'natural-sounding' synthetic speech is speech that is as *easy to understand* as human speech.

[1] Linguistics, University of Cambridge
[2] Phonetics & Linguistics, University College London

The significance of taking such a view is that it makes naturalness open to scientific investigation. We can design perceptual experiments to compare synthetic against human, or more usefully synthetic with properties A against synthetic with properties B. In this way our experiments allow us to choose between scientific hypotheses, to rank the importance of deficiencies, and to estimate the size of the remaining gap.

This paper gives brief details of some perceptual experiments conducted within the ProSynth collaboration, and uses these to support this approach to improving synthetic speech. In the following sections we outline the ProSynth framework and then discuss how the experiments were constructed and what they showed.

2. The ProSynth Framework

ProSynth is an integrated prosodic (i.e. structure based) approach to speech synthesis: the interactions between grammatical, prosodic and segmental parameters in speech production are captured through a single, highly-structured, computationally-tractable linguistic formalism (Huckvale, 1999). The design is influenced by work that shows it is possible to combine, within a declarative framework, phonological with phonetic knowledge in a process of phonetic interpretation (e.g. Local & Ogden, 1997); it is also influenced by recent phonetic research that shows that speech is rich in non-phonemic information which contributes to robustness (e.g. Hawkins & Slater, 1994); and also by proposals to integrate intonational, rhythmical, and segmental effects (House & Hawkins, 1995). A more detailed justification can be found in Ogden *et al* (2000).

Current work in the ProSynth framework has been on a limited range of phenomena for one accent of British English. We have modelled systematic variation in timing, intonation and some segmental realisation effects for relatively short declarative sentences. For further information about the speech we have analysed or for information about available software, see our web pages [http://synth.phon.ucl.ac.uk/prosynth/].

3. Experimental Design

How can we go about designing experiments to test naturalness which at the same time provide diagnostic information about which approaches are most useful? In general we need to contrast *performance* on some task with two versions of the synthesized speech. Version 1 is the version under test, the one with the new algorithm say, while version 2 is the control. For example, version 1 might have syllable durations sensitive to the position of the syllable in the metrical foot, while version 2 has average durations insensitive to position. This might be called a test of RIGHT vs. AVERAGE: it is a test of whether sensitivity to particular contexts is worth incorporating. An alternative is to make version 2 arise from predictions from the model for the wrong structure: for example swapping the syllable durations for first and second position in the foot. This might be called a test of RIGHT vs. WRONG: it is a test of whether the model is correct.

What kind of perceptual measure is useful for such tasks? Measures need to be *simple*: listeners need to be able to understand what is required of them; and measures need to be *appropriate*: they need to tap sensitivity to the feature in question. Thus measures of word or phoneme intelligibility may be appropriate for aspects of segmental quality or timing, but possibly inappropriate for intonation where a deeper measure of comprehension is required.

Experiment 1. Intonation experiment

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

Preliminary f0 modelling on neutral, declarative, discourse-final utterances in the ProSynth database showed a statistically significant difference in alignment of the f0 contour dependent on the type of foot. The f0 turning points of an H*L pitch accent occur consistently later in an accented syllable when it is part of a disyllabic rather than a monosyllabic foot. This rightward shift is not obviously dependent on the internal structure of the accented syllable, since it was observed across a wide range of structures (House, Dankovicová & Huckvale, 1999). It was hypothesized that phrases would be judged more natural when the f0 alignment was appropriate for the foot structure: in other words, that the f0 alignment is perceptually salient. If the hypothesis were supported, then the implication would be that synthesis of intonation should take account of foot structure.

The stimuli were 32 pairs of utterances, each with a final, monosyllabic foot, for example *the terrain*; *he was mad*; *it's a lie*. Segmental durations within each MBROLA-synthesized stimulus matched those of the original utterance in the database. Before the final foot, f0 was also sampled from the original utterance; in the final foot, values were specified at the turning points of the f0 template, with linear interpolation between them.

The two members of each pair were identical except for the alignment of two f0 turning points which marked the beginning and the end of the most steeply falling portion of the contour. In right items, the f0 turning points were appropriate for the monosyllabic final foot. In wrong items, they were modified to follow the pattern appropriate for the same syllable in a disyllabic final foot. For the three examples above, the respective utterances with disyllabic feet were *it was **rain**ing*; *for a **mad**man*; *they were **ly**ing*. The precise f0 manipulation was sensitive to the properties of the onset and coda in the accented syllable.

The 32 pairs of phrases were randomized in ten blocks of 32. There were 10 subjects. They were told that the members of a pair differed only in melody (for 4 nonphoneticians) or intonation (for 6 phoneticians), and that they should focus only on that and ignore all other properties. The subjects were instructed to press one of two buttons, depending on whether they judged the first or the second member of a pair to sound more natural in the sense of neutral, factual, cool and normal, yet without abnormally low emotion. Even if the particular words in some utterances meant that a livelier, more emphatic or more excited pronunciation sounded more appealing than a more neutral one, they were told to choose the more neutrally-spoken item.

Overall, 78% of the responses favoured right items, and 22% wrong items. A paired *t*-test comparing mean responses for each S confirmed that right items were preferred significantly more often than chance (78% vs. 50%; $t(9) = 5.71$, $p < 0.0002$). Unsurprisingly, phoneticians were more consistent in their choices than nonphoneticians, but the preference for right rather than wrong intonation patterns is significantly better than chance for each subgroup. For phoneticians, 86% preferences for right items ($t(5) = 6.09$, $p < 0.0009$); for nonphoneticians, 66% preferences for right items ($t(3) = 4.01$, $p = 0.014$).

What kind of testing procedure is required? Most synthetic speech is already highly intelligible when the utterances are short and heard in good listening conditions, so that intelligibility testing will have to take place in additive noise to avoid ceiling effects. When testing speech that is readily comprehensible, it may be necessary to give listeners a simultaneous competing task to perform.

Experiment 1 (see text box) is perhaps indicative of the standard approach and shows its weaknesses. Listeners asked to rate intonation contours for naturalness were found to require a complex definition: "neutral, cool, factual without abnormally low emotion"; and phoneticians gave different results to non-phoneticians. This suggests that naturalness judgments are sensitive to the way in which the task is presented. Direct questions about naturalness suffer from the problems discussed above: listeners are influenced by a range of characteristics of the signal, and end up being relatively insensitive to any small changes actually being tested.

The next two experiments exemplify the alternative approach: to look directly at the communicative efficiency of the signal: both use intelligibility in noise.

Experiment 2. Segmental detail experiment

This experiment assessed whether natural-sounding excitation near segment boundaries enhances the intelligibility of formant synthesis. Observations from the ProSynth database showed systematic variation in the incidence of (a) mixed periodic and aperiodic excitation at boundaries between vowels and voiceless fricatives, and (b) the duration of periodicity in the closures of voiced stops (Heid & Hawkins, 1999). In brief, most vowel-fricative (VF) boundaries have mixed aperiodic and periodic excitation, whereas most fricative-vowel (FV) boundaries change abruptly from aperiodic to periodic excitation. Syllable stress, vowel height, and final/non-final position within the phrase influence the incidence and duration of mixed excitation. Similarly, the duration and proportion of voicing in the closures of phonologically voiced stops depend systematically on vowel height, place of articulation, stress context and the syllabic position of the stop. It was predicted that synthesized phrases would be more intelligible in noise when they conformed to the natural patterns of excitation for the particular structural context, because they would add both to the signal's perceptual coherence and to its informativeness about linguistic structure.

18 phrases from the database were copy-synthesized into a formant synthesizer, HLsyn (Bickley et al., 1997), using PROCSY (Heid & Hawkins, 1998), and hand-edited to a good standard of intelligibility, as judged by a number of independent listeners who did not serve in the perceptual tests. In 10 phrases, the sound of interest was a voiceless fricative in a number of structural contexts; in the other 8 it was a voiced stop. The sound of interest was synthesized with the "right" type of excitation pattern at its boundaries. From each right version, a "wrong" one was made by substituting at just one boundary a type or duration of excitation that was inappropriate for the structural context. For fricatives between sonorants, either the VF or the FV boundary was manipulated (e.g. VF in *his r<u>iff</u>*; FV in *in a <u>fie</u>ld*). For stops, voicing during closure was manipulated (e.g. in *the <u>d</u>elay*).

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. Subjects pressed a key to hear each item, and wrote down what they heard. Each subject heard each phrase once: half the phrases in the right version, half wrong. 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.

Responses were scored for number of phonemes correct on three phonemes: the manipulated one and the 2 adjacent to it. Insertions of spurious elements in otherwise correct responses counted as errors. Responses were significantly better for the right versions using a one-tailed paired t-test (69% vs. 61%, $t(21) = 2.35$, $p = 0.015$).

In Experiment 2 (see text box) listeners were asked to identify short synthetic phrases in noise but were marked for their accuracy on just three phonemes. The excitation used in the realisation of these phonemes was the parameter under test. The appropriate excitation pattern for the context was contrasted with the inappropriate one. The results showed a small but significant intelligibility gain.

In Experiment 3 (see text box) listeners were asked to identify short synthetic phrases in noise and were marked for overall phoneme accuracy. The contrast here was on the predicted duration of syllable rhyme for the first syllable in the last foot. The predicted durations were compared with those predicted for the same segments in a different structure. The results showed a small but significant intelligibility gain.

Apart from validating the approach to naturalness testing presented in this paper, these two experiments show that small but *structurally appropriate* changes to the realisation of synthetic speech facilitate processing. This is a tenet of the ProSynth approach: contextually sensitive variation in the realisation of phonological structures is exploited by listeners - it provides additional information as to their identity. Synthetic speech lacking these changes in context is more difficult for listeners to process since these cues are absent or contradictory. These particular results show small improvements in intelligibility, but we expect there to be a number of such systematic patterns that could be exploited; and when integrated together these changes will add up to a significant advance.

Experiment 3. Rhyme duration experiment

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

Twelve different linguistic structures were chosen, each with two exemplars, making a total of 24 pairs of phrases in all. Each phrase was synthesized with an approximation to its natural f0 in the ProSynth database. A "right" and a "wrong" version of each phrase was produced as follows. In the right version, segmental durations of the first part of the phrase, up to and including the onset of the first syllable of the last foot, were copied from the natural utterance. Segmental durations for the rest of the final foot were those predicted by the ProSynth model for the particular linguistic structure. The wrong version of each phrase was made by exchanging the ProSynth-predicted segment durations of the strong rhyme between the two phrases in each pair, where those segments were identical. So the durations for *ob* in *he's a robber* were replaced by the durations for *ob* in *to rob them* and vice versa. In cases where the segment strings of two STRONG rhymes were not identical (as in /ɛt/ and /ɛlt/, or /aɪn/ and /aɪnd/), the durations of the same or similar segments were exchanged. So durations for /ɛt/ in *to get them* were replaced by the durations of /ɛ/ and /t/ in *to belt them* and vice versa; durations for /aɪn/ in *to mine it* were replaced by durations of /aɪn/ in *to remind us* and vice versa.

In absolute terms, the mean difference between the right and wrong ProSynth-predicted durations is 22 ms for both nucleus and coda. In relation to normal speech synthesis standards, the test is thus of subtle rather than gross rhythmic effects.

Since the manipulations affect the rhythm of the whole phrase, intelligibility was assessed by scoring phonemes correct for each entire phrase. Responses were about 4% better for the right versions (79% vs. 75%). This improvement, though small, is strongly significant in a one-tailed paired t-test on the mean right vs. wrong scores for each subject: $t(24) = 3.13$, $p = 0.0023$. Even with long-short vowel difference data excluded, the correct rhythm engendered better phoneme intelligibility, suggesting that even rather subtle temporal patterns enhance intelligibility when modelled systematically.

Experiment 4. Intonation by speed of comprehension

This pilot experiment was conducted to explore the potential of a comprehension test to evaluate intonation. Since the design needs refinement, the experiment is described only briefly here. Listeners read a story, then decided whether answers to questions about the story were true or false, and responded accordingly by pressing the appropriate one of two buttons. The number of correct responses and reaction time (RT) were measured. Questions appeared one at a time on a computer screen; answers to the questions were the same stimuli used in Experiment 1. Each of the 32 phrases appeared once as a true answer to a question, and once as a false answer. Each of 36 Ss heard each answer only once: either as a true or a false answer (to questions about different stories), and with either the right or the wrong intonation pattern (for the answer to the same question about the same story).

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

Returning to the problem of how to assess intonation demonstrated by Experiment 1, we have been piloting an alternative approach that taps the ability of a listener to comprehend an utterance directly. In Experiment 4 (see text box) listeners were timed in their responses to true/false questions after having heard a short story. The questions were synthesised with intonation contours right or wrong for the segmental structure of the nuclear syllable. Looking at error rate and speed of response showed a slight advantage for the appropriate contour, but the results were not significant. This work needs to be developed further, either by increasing the control over materials and testing (to

reduce response variance) or by adding a competing cognitive task (to increase the sensitivity of the listeners to the differences).

4. Conclusions

In this paper we have presented a view that an important aspect of the naturalness of synthetic speech is how easy it is to understand. This approach to naturalness is important because it opens the possibility of scientific evaluation of competing hypotheses using perceptual tests. Through the use of examples conducted within the ProSynth project, the paper has shown how such tests might be constructed, and has highlighted both strengths and weaknesses. Tests that attempt to tap 'naturalness' directly are very sensitive to the instructions given to subjects. Tests using reaction time have a lot of response variation that make statistical analysis difficult. On the other hand, intelligibility testing in noise has shown listener sensitivity to small but structurally sensitive changes in phonetic realisation. There may be many possible ways in which systematic variation with linguistic structure may be observed in the signal. To achieve complete naturalness by modelling them all may take a long time, but it seems that each one could make a useful contribution.

Acknowledgements

The authors of this paper thank all the ProSynth project collaborators. This project was supported by the U.K. Engineering and Physical Sciences Research Council.

References

Bickley, C.A., Stevens, K.N., & Williams, D.R. (1997). A framework for synthesis of segments based on pseudo-articulatory parameters. In *Progress in Speech Synthesis* (van Santen J., Sproat, R., Olive, J., Hirschberg, J., eds) Springer, New York, pp. 211-220.

Duffy, S.A., & Pisoni, D.B.(1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35, pp. 351-389.

Hawkins, S., & Slater, A. (1994). Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech. *Proceedings of the International Conference on Speech and Language Processing*, pp. 57-60.

Hawkins, S., House, J., Huckvale, M., Local, J., Ogden, R., (1998), "ProSynth: an integrated prosodic approach to device-independent natural-sounding speech synthesis", *Proc. Int. Conf. Spoken Language Processing*, Sydney.

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

Heid, S. & Hawkins, S. (1999). Synthesizing systematic variation at boundaries between vowels and obstruents. *Proceedings of the XIVth International Congress of Phonetic Sciences* (Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., and Bailey, A.C. eds.), 1, pp 511-514. University of California, Berkeley, CA.

House, J., Dankovicova, J. & Huckvale, M. (1999). Intonation modelling in ProSynth: an integrated prosodic approach to speech synthesis. *Proceedings of the XIVth International Congress of Phonetic Sciences* (Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., and Bailey, A.C. eds.), 3, pp. 2343-2346. University of California, Berkeley, CA.

House. J., & Hawkins, S., (1995). An integrated phonological-phonetic model for text-to-speech synthesis. In *Proceedings XIII International Congress of Phonetic Sciences* (Elenius, K., Branderud, P., eds) Stockholm, Sweden, pp. 326-329.

Huckvale, M., (1999). Representation and processing of linguistic structures for an all-prosodic synthesis system using XML. *Proceedings EuroSpeech-99*, Budapest, Hungary.

Local, J., & Ogden, R., (1997). A model of timing for nonsegmental phonological structure. *In Progress in Speech Synthesis* (van Santen, J., Sproat, R., Olive, J., Hirschberg, J., eds) Springer, New Yo rk, pp. 109-122.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovicova, J., & Heid, S., (2000) ProSynth: An Integrated Prosodic Approach to Device-Independent, Natural-Sounding Speech Synthesis, *Computer Speech and Language*, in press.

Pratt, R. (1986). On the intelligibility of synthetic speech. *Proceedings Institute of Acoustics*, 8 pp. 183-192.