IMPROVING TEXT-TO-SPEECH SYNTHESIS

Mark Tatham and Eric Lewis***

*Department of Language and Linguistics, Essex University, Colchester, UK-CO4 3SQ

email: Mark.Tatham@essex.ac.uk

**Department of Computer Science, University of Bristol, Bristol, UK-BS8 1UB

email: Eric.lewis@bristol.ac.uk

ABSTRACT

Naturalness in human speech is dependent on a number of factors and the extent to which a text-to-speech synthesis system can account for these factors in its model will be a measure of its success in the marketplace.

As well as the obvious factors of rhythm and intonation there is the more difficult question of modelling the variability in human speech. This paper discusses how SPRUCE [1], a high-level text-to-speech synthesis system, incorporates several different types of variability.

1. INTRODUCTION

From the early stages of development of SPRUCE one of the most important concepts that the authors have tried to design into the system has been 'naturalness'. SPRUCE achieves this naturalness in the following ways:

- by modelling intonation accurately,
- by modelling rhythm accurately,
- by modelling variability.

SPRUCE prosody has been the subject of previous papers [2, 3] and SPRUCE rhythm will be the subject of a future paper. Here we discuss how variability can be modelled and how it can be incorporated into a text-to-speech synthesis system.

2. THE SPRUCE SYSTEM

SPRUCE is a modular text-to-speech synthesis system which, in its default configuration, inputs plain orthographic text and outputs an output file selected from a range of types suitable for driving several different low-level synthesisers. The system has previously been demonstrated[4] driving the JSRU[5] and DECTalk[6] formant synthesisers, the CNET diphone based PSOLA[7] concatenated waveform synthesiser, as well as the IBM[8] wavelet based synthesiser. In all cases SPRUCE relies on an inventory of stored words and syllables obtained from recorded human speech to build the correct output file. For formant synthesisers it is necessary to analyse the recorded speech samples to obtain the appropriate acoustic values driving the synthesiser parameters, while for waveform synthesisers it is necessary to extract and store the corresponding pitch synchronisation information in addition to the recorded speech.

The top level of SPRUCE is dictionary based. The system incorporates a lexical knowledge base in which each entry is matched to a set of linguistic models - phonetic, phonological, syntactic and semantic - based on our own adaptations of contempory linguistic theory. These models allow the inclusion of a fast and efficient parser for providing the necessary markers for imposing good prosodic contours upon the synthesised sentences.

Although a lexical knowledge base is necessarily single word oriented it is possible to identify, within each word model, parameters which are potentially subject to revision once the word is located in phrase or sentence context. Thus, as a simple example, the phonological model for the word '*word*' marks the voicing parameter towards the end as a candidate for reduced vocal cord vibration under certain contextual sentence conditions when the current symbolic representation is later substituted by a physical representation. Similarly, an entry like '*content*' which is ambiguous with respect to syntactic category has this fact marked in the syntactic model. This last example is interesting because the identification of alternate syntactic category markers for '*content*' (noun and adjective) links to a marker in the phonological model of the same word flagging alternate pathways through the model to permit the representation of two different word-stress patterns. There is nothing novel in the fact that a grammar has to describe such phenomena as phonological contextual effects or the relationship between syntactic category and stress pattern - what is novel in SPRUCE is that we are systematically marking words as candidates for such phenomena in the dictionary. Theorists will correctly argue that such an approach loses descriptive generality, but this loss buys a speed improvement in the system.

We do, however, list and use the associated rules later in the system - our markers enable them to be triggered directly without any need to spend time scanning all input text for only occasional rule application.

SPRUCE is a modular system, so it is pertinent to ask which module handles the variability of speech. The answer is that no one module can provide all the necessary information since variability needs to be modelled in several different ways. The relevant SPRUCE modules are essentially those dealing with segmental phonology (which we call simply the phonological module), supra-segmental phonology (which we call the prosodic module), and the low-level interface module. Variability is a contributor to all of these.

3. VARIABILITY

SPRUCE endeavours to take into account three well-defined types of variability, namely

1. phonological variability,
2. coarticulatory variability,
3. stochastic variability.

Phonological variability is variability *intended* by the speaker, and is typically demonstrated in English by the use of palatal [l] in some phonological environments and velar [l] in others, despite the fact that underlying these two alternates we need only one /l/ for distinguishing (on a phonological basis) between morphemes. This variability can be catered for in the phonological module by modifying the phonetic form of a word or syllable appropriately. [By **phonetic form** we mean a symbolic representation capable of interpretation as a prescription for eventual physical acoustic representation.] All text-to-speech systems incorporate phonological variability with good results.

Coarticulatory variability has been the subject of an extensive literature [9]. SPRUCE accounts for this variability in two ways. Since it effectively leads to changes in pronunciation according to context we automatically solve the problem of **intra-syllable coarticulation** by reason of the fact that SPRUCE is *syllable* based. The acoustic model of each syllable already incorporates the intra-syllable coarticulation identified and described in segmentally-based phonetic theories. **Inter-syllable coarticulation**, however, has to be catered for by the introduction of a dedicated set of rules to be applied in the low-level interface module. It is easy to see that coarticulation is a phenomenon which occurs at linguistic boundaries, and that the *size* of the linguistic unit which is chosen to form the basis of the synthetic speech is what determines how much coarticulatory modelling has to be incorporated. Allophone-based systems (such as DECTalk) need most coarticulatory modelling because allophones are the smallest linear units identified in linguistics. Most modern text-to-speech systems are capable of modelling coarticulatory variability well, although in some phonetic contexts pointing, we believe, (see below) to inadequacies in the theory of coarticulation rather than to poor engineering.

Using speech synthesis as a means of *testing* linguistic models is, for the theorist, a serious application of this technology. The fact that after some thirty years of attempts to render coarticulation satisfactorily there are still unexpected effects could legitimately lead the theorist to conclude that a theory of speech production which focuses on allophone-sized units may be inadequate. It is by no means certain that the best way of describing speech production involves the linear concatenation of allophones.

This long-standing debate has been revisited over the past decade [10], and, in the view of the authors, there may now be a sufficient body of evidence for us to reconsider our synthesis strategy. SPRUCE focuses on the syllable as the minimal linguistic unit, and, as mentioned above, thus avoids consideration of intra-syllabic coarticulation altogether. For the purposes of the application of some phonological rules we do, however, still represent our syllables in the lexical database in terms of concatenated allophones. We return later to modelling coarticulation.

Stochastic variability is the apparently unsystematic occurrence of variability in repetitions of speech occurring at different times; that is, the repetition of the *same* word or syllable, in the phonological sense, at different times produces slightly different waveforms. Two ways could be tried for introducing this type of variability:

1. in the light of no suitable model for the variability we might introduce a random jitter into various parameters of the acoustic model;
2. to use relatively large units of 'speech' for the acoustic models in the low-level synthesis system.

Having informally approached the problem of accommodating stochastic variability over a great deal of research experience we suspect that it is not, in fact, random but determined by factors in the mechanics and aerodynamics of speech production control which await adequate description. If this is true, then the non-randomness of stochastic variability is partially captured in any low-level synthesis system based on original human speech, *provided* that the inventory of samples contains multiple copies of each entry, and that these copies occurred in differing phonological or phonetic contexts in the original recordings. In practice one way of accommodating this is to have sample inventory entries of different linguistic lengths - that is, *stretches* of speech which do not necessarily correspond to any one pre-determined linguistic unit but which perhaps span several. Clearly the longer the stretches of speech used in creating the synthesised output the better will stochastic variability be captured.

Systematically incorporating phonological, coarticulatory and stochastic variability into speech generated by SPRUCE gives us a naturalness which we feel has not been bettered in a general purpose text-to-speech system. All systems improve, however, if

their *domains* (in the sense of discourse domains) are restricted. Thus a recently created SPRUCE system limited to the domain of **weather forecasting** improves on the general purpose version in terms of overall naturalness and in consistency of naturalness - the naturalness error rate (that is, percentage of words sounding unacceptably less than natural) falls.

4. NATURALNESS FACTORS

We have, thus far, identified several areas which separately and together contribute to the overall naturalness rating of a synthesis system. In summary, these are:

1. domain size in which the synthesiser is to function,
2. completeness of the resident lexical database,
3. incorporation of phonological variability (at the symbolic level, reflecting language-imposed constraints on the linear co-occurrence of phonological units),
4. incorporation of coarticulatory variability (at the physical phonetic level, reflecting motor and acoustic constraints on the linear co-occurrence of allophones),
5. incorporation of stochastic variability (at the acoustic level, reflecting constraints on the consistency of speech production in general),
6. size of linguistic unit modelled at the low, physical, level.

But there is another area which needs identifying and incorporating into any synthesis system purporting to focus on naturalness. This area is best described in extensions to the basic theory of coarticulation which phoneticians have developed over the last few years [11], because although there have been attempts to model speech production and perception without adhering to the traditional idea of phoneme or allophone sized units as fundamental to speech [10], the idea that speech is made up of a linear concatenation of small, permutable units is still dominant.

The theory of coarticulation is essentially a patch to account for the fact that when such units are articulated under the variable constraint of **time**, distortions or degradations occur in what would otherwise be an ideal articulatory realisation of the abstract requirements computed at the phonological level. This view of speech as being a somehow imperfect signal downgraded from some abstract ideal is directly traceable to the general theory of language prevalent in the 60s and 70s (see [12] for a typical overview of phonology using this approach).

Classical coarticulation theory models all such distortions as deriving from motor, mechanical (usually inertial), aerodynamic or acoustic constraints, all of which are properties of the physical periphery of speech production. As such the constraints and their effects are a-linguistic and are of only marginal interest to the theory of language.

Extensions to this theory were developed in the 70s and 80s [13] to account for observations of a type of variability not listed above - a *systematic* variation in coarticulation which could not be explained in terms of the properties of the speech production system itself [14]. Two classic examples will serve to illustrate *two* dimensions to this new category of variation:

1. The phonetic nasalisation of the phonologically non-nasal vowel in a word like *man* in English is attributed in coarticulation theory to a time-governed inertial effect on velar control. That is, movement of the mass of the velum is necessarily temporally constrained to result in incomplete velar closure during the vowel, thus permitting nasal airflow and, acoustically, nasal resonance. If this account were adequate in its assignment of the degradation of the oral vowel to a purely physical effect we would expect all dialects of English to exhibit the same degree of nasalisation on inter-nasal vowels, within the limits imposed by stochastic variability. The obvious differences between the various dialects, especially noticeable between general British and American pronunciations, means that this explanation fails completely to explain the observable facts.

2. The range of coarticulatory variability in the pronunciation of the initial consonants in English *sell* and *shell* is very much narrower (at the motor, articulatory and acoustic levels) than the variation observed in the pronunciation of [s] in most dialects of Spanish and Greek. But if coarticulation were purely physically dominated, why would not the variability be the same for these three languages?

In all cases adduced so far of this variation of coarticulatory variation, the explanation has a linguistic origin. In the nasalisation example the interpretation of the oral vowel is allowed more variability in American English than in British English - a dialectal or stylistic (linguistic) constraint. In the [s] example the wider variability in Spanish and Greek is allowed because there is no [sh] sound in those languages to cause a perceptual morphemic clash under conditions of wide variability. The theory of Cognitive Phonetics [15] was developed to account for this linguistically-governed constraint on coarticulation. Later versions of the theory, applied to non-segmental developments in speech production and perception refer to linguistically dominated supervision of gestures within the dynamic speech scenario [16, 17].

SPRUCE adopts two approaches to modelling cognitive phonetic constraints on coarticulation. One is to mark candidate violations of standard coarticulatory theory in the lexical database, signalling constraints on later selection from the available acoustic models contained in the low-level inventory of physical units. The other is to make certain that the acoustic models in the inventory properly reflect the cognitive phonetic constraints of the dialect or language being synthesised. We believe that this approach accounts for some of the additional naturalness of SPRUCE.

5. CONCLUSION

This paper claims that **naturalness** in synthetic speech is essentially the successful rendering of variability in the final acoustic signal, once we get beyond the obvious factors such as limiting the domain of discourse within which the system is to operate. In SPRUCE we identify and treat distinctly several sources of variability in human speech, adhering carefully to contemporary speech production theory. We believe that this approach renders transparent the *sources* of naturalness, and at the same time enables us to manipulate what we feel to be an important interplay between the various types of variability.

Further information regarding SPRUCE can be obtained on the Internet at the sites http://www.cs.bris.ac.uk/~eric/, and http://speech/essex.ac.uk/speech/

6. REFERENCES

1. Lewis, E. and Tatham, M.A.A. "SPRUCE - A New Text to-Speech Synthesis System," *Proc. 2nd European Conference on Speech Communication and Technology, Vol. 3, pp. 1235-1238, 1991.*

2. Tatham, M.A.A. and Lewis, E. "Prosodics in a Syllable-based Text-to-Speech Synthesis System," *Proc. ICSLP, pp. 1179-1182, 1992.*

3. Tatham, M.A.A. and Lewis, E. "Prosodic Assignment in SPRUCE Text-to-Speech Synthesis", *Proc. Inst. Acoustics, UK, Vol. 14, Pt. 6, pp.447-454, 1992.*

4. Lewis, E. and Tatham, M.A.A. "A Generic Front-End for Text-to-Speech Synthesis Systems", *Proc. 3rd European Conference on Speech Communication and Technology, Vol. 2, pp. 913-916, 1993.*

5. Holmes, J.N. *Speech Synthesis and Recognition,* Van Nostrand Reinhold, Wokingham(UK), 1988.

6. Allen, J, Hunnicutt, M.S. and Klatt, D. *From Text to Speech: The MITalk System,* Cambridge University Press, Cambridge, 1987.

7. Moulines, E. and Charpentier, F. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication, Vol. 8, pp. 453-467, 1990.*

8. Sharman, R. "Concatenative speech synthesis using sub-phone segments," *Proc. Inst. Acoustics(UK), Vol. 16, pp. 367-374, 1994.*

9. Fowler, C. "Coarticulation and theories of extrinsic timing," *Journal of Phonetics Vol. 8, pp. 113-133, 1980.*

10. Browman, C.P. and Goldstein, L. "Articulatory phonology: an overview," *Phonetica, Vol. 49, pp. 155-180, 1992.*

11. Ohman, S. "Numerical model of coarticulation," *Journal of the Acoustical Society of America, Vol. 41, pp.310-328, 1967.*

12. Chomsky, N. and Halle, M. *The Sound Pattern of English,* Harper and Row, New York, 1968.

13. Tatham, M.A.A. " Some problems in phonetic theory", in H. and P. Hollien (eds.), *Amsterdam Studies in the Theory and History of Linguistic Science IV: Current Issues in Linguistic Theory, Vol. 9 - Current Issues in the Phonetic Sciences, pp. 93-106,* John Benjamins B.V., Amsterdam, 1979.

14. Morton, K. "Cognitive phonetics - some of the data," in *In Honor of Ilse Lehiste. R. Channon and L. Shockey (eds.). pp.191-194,* Foris Publications, Dordrecht, 1987.

15. Tatham, M.A.A. "Cognitive phonetics," in W.A. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing, Vol.1, pp.193-218,* JAI Press, London, 1990.

16. Tatham, M.A.A. "The supervision of speech production," in C. Sorin et. al. (eds.) *Levels in Speech ommunication pp.114-126,* Elsevier, Amsterdam, 1995

17. Tatham, M.A.A. "Dynamic Articulatory Phonology and the supervision of speech production,". *Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Vol.1, pp. 58-61, 1995*