Naturalness in a High Level Synthetic Speech System

Mark Tatham
Eric Lewis

Reproduced from: Tatham, M. and Lewis, E. (1995) Naturalness in a high level synthetic speech system. In J. Pardo (ed.), *Proceedings of Eurospeech '95,* 1815-1818. Madrid: European Speech Communication Association

ABSTRACT

Naturalness in synthetic speech is to a large extent determined by how well the system models the variability found in human speech. Good models of variability are now emerging, and this paper describes how variability of several different types is incorporated into SPRUCE — a high level text-to-speech synthesis system. The synthesiser is carefully engineered according to the requirements of a recent computational model of speech production. The resulting voice output illustrates the usefulness of well motivated theory in speech synthesiser design.

1. INTRODUCTION

During the research and development of SPRUCE – our high level text-to-speech synthesis system — we have been investigating the characterisation of 'natura1ness' in synthesis systems. In this paper we consider one contributory factor — variability — and examine the model of variability which is implemented in SPRUCE. After discussing three kinds of variability we focus on the concepts of monitoring and supervision.

2. THE SPRUCE ENGINE

SPRUCE[1] is a modular text-to-speech synthesis engine. It inputs plain orthographic text and produces output files suitable for driving several different low level synthesisers. It can, for example, satisfactorily drive several low-level synthesisers, including the JSRU[2] and DECTalk[3] formant synthesisers, the CNET diphone based PSOLA concatenated waveform synthesiser[4] as well as the IBM[5] wavelet based synthesiser. The system works slightly differently depending on which synthesiser it is driving; and in the case, for example, of the JSRU low-level formant synthesiser it relies on an inventory of stored words and syllables modelled from normalised parametrically analysed human speech. The acoustic parameters of the word and syllable models are those needed for driving the low-level synthesiser.

High level SPRUCE is dictionary based. The system incorporates a lexical knowledge base consisting of phonetic, phonological, syntactic and semantic models of words in the input text based on our own adaptations of contemporary linguistic theory. These adaptations are designed specifically to extend the domain of linguistic theory into application driven modelling.

The existence of a dictionary allows systematic substitution in the input text of matching word models retrieved from the dictionary. The models are such that fast and efficient parsing is possible — a prerequisite for assigning good prosodics. The negative side of dictionary based systems is that they perform best in limited domain situations to keep the dictionary of manageable size — though the SPRUCE engine provides for at least 100,000 words. At the time of writing the system incorporates a 30,000 word English dictionary and runs in virtually real time, with no speed penalty accruing from the use of a large dictionary. Provision is made for dealing with words not modelling in the dictionary in the traditional ways associated with systems like DECTalk and the high level JSRU text-to-speech synthesiser.

3. VARIABILITY IN SPEECH

Speech production theories differ in the way they model variability, and in ordinary speech variability seems much greater than is usually observed under experimental conditions.[6] This is particularly true in the case of speech in a dialogue environment — a major application area of speech technology. The vast majority of data collection up till now has been from *normalised* utterances which are quite different from the real-life utterances of ordinary conversation.[7] There is general agreement on three major types:

- **phonological variability** — the systematic substitution of surface alternates for underlying phonological objects. For example: the use in English of palatal /1_j/ in some phonological environments, and velar /l_w/ in others, despite the fact that underlying these two alternates we need only one /L/ for distinguishing between morphemes. These alternations were noted in traditional phonology[8] and in more recent classical generative linear phonologies.[9] Phonological variability is *intended* by the speaker of the language — it is language specific.
- **coarticulatory variability** — the systematic occurrence of variants of underlying objects in particular phonetic contexts. For example: inter-nasal nasalisation of vowels — a nasalised vowel in the word *none.* There is an extensive literature on the subject of coarticulation.[10] Coarticulatory variability is *not* intended by the speaker, and arises mainly from mechanical, physiological and aerodynamic constraints within the articulatory system.
- **random variability** — the apparently unsystematic occurrence of variability in repetitions of speech which would be classified as *same* under the above two variability categories. For example: the repetition of any speech segment by a single speaker or by another speaker produces waveforms which are not identical. The investigative literature here is sparse, though the phenomenon of random variability is regularly noted. Random variability is *not* intended by the speaker.

4. VARIABILITY IN SYNTHESIS

There is need for an explicit framework for modelling variability in speech synthesis,[11] but all high level synthesis systems already incorporate phonological and coarticulatory variability in some form. Sometimes it is unclear whether to assign a particular effect to the phonological rules or to the coarticulation rules,[12] but this not considered in depth in most text-to-speech systems. In SPRUCE, however, how to handle speech production variability is critical. The reason for this is that SPRUCE is intended to capture naturalness to enable more effective use in dialogue systems. This is accomplished in the system's modularity and the use of a particular speech production theory.

Random variability is usually neglected except in systems which rely on stored representations of human speech. So, for example, the JSRU and DECTalk text-to-speech systems always reproduce a particular allophone identically in a given phonological or phonetic context, whereas in a pcm diphone-based system, in the CNET concatenated diphone waveform system, the IBM wavelet system, or in SPRUCE parallel formant or concatenated waveform modes, some aspects of random variability are retained. The reason for this is that the JSRU and DECTalk systems provide only static models of allophones with no possible provision for dynamic variability, whereas the other systems incorporate dynamic modelling of their segments (of whatever size). This contributes significantly to perceived naturalness.

The traces of random variability captured in these systems are differences in the realisation of allophones, which are phonologically identical, because their inventory representations are dynamic and have been taken from different samples of human speech. For example, in a syllable-based system the model of initial [t] in *tack* — [tæk] — will be different from that in *tap* — [tæp]. This would not be true for an allophone-based system; here they would be identical.

A syllable-based system does not, however, capture the variability which occurs when a human being pronounces the same syllable on different occasions — the system contains only one model of the complete syllable. In contrast, a system which derives data from a large database of connected human speech will maximise variability because it will have access to several versions of a particular sequence of segments, and if choice is not constrained for some other reason it can randomly select any version. To some extent any version will be within the range of acceptability.

5. COARTICULATORY VARIABILITY

Coarticulatory variability is usually modelled as a set of constraints introduced by aerodynamic, mechanical, or other a-linguistic effects. Even in recent speech production theory, such as that of which Articulatory Phonology[13] is a part, this coarticulation (or coproduction) is ascribed to sub-linguistic systems and forms no part of the phonological description of the language. However, this theory fails to account for regularly occurring phonetic phenomena which cannot be simply explained in terms of such effects. For example, if the inter -nasal nasalisation of vowels is truly a-linguistic it would not vary in degree from language to language or dialect to dialect. Yet we see that in English dialects there are systematic differences, as between southern British and most American accents, with American English permitting markedly more nasalisation than British English. Recently the suggestion has been made[14, 15] that Articulatory Phonology linked with the Task Dynamic Model[16] can be adapted to account for these phenomena.

Articulatory Phonology aims to provide a unified phonological and phonetic theory by integrating the modelling of *planning* and *execution.* In attempting this, though, it falls short of explaining the kind of data exemplified in the British/American English example. By introducing a *supervisory monitor* to oversee the way low-level processes are executed, we can adequately explain the difficult data — and go on to reproduce it in synthesis. The supervisory monitor controls the extent to which coarticulatory constraints are permitted to apply.

6. MONITORING AND SUPERVISING COARTICULA TORY VARIABILITY IN SYNTHESIS

In the standard version of SPRUCE coarticulatory variability *within the syllable* is incorporated as an intrinsic property of the syllable-size models used for generating the output signal. However, inter-syllabic coarticulatory variability has to be calculated. For this we use a non-linear interpolation technique in which the transition ‘shape’ varies depending on

- which allophones fall on either side of the boundary, and
- whether or not syllables on either side of the boundary carry primary stress.

The results are satisfactory since it seems that listeners are most sensitive to errors within the syllable, and these are automatically guaranteed not to occur because the smallest segment model *is* the syllable. We are achieving good results at syllable boundaries where listener sensitivity is less critical.

Monitoring and supervision of coarticulatory variability are novel concepts in speech synthesis. One area in which we model these processes is the control of precision of articulation. Articulatory precision varies[17, 18] depending on the speaker’s prediction of the probability of listener error — the greater the predicted error probability the greater an attempt is made at articulatory precision. Precision is modelled as supervised motor control constraints on coarticulation; control is applied according to production instructions which tie in with the theory’s linguistic orientation.

7. MODELLING ARTICULATORY PRECISION

In SPRUCE we monitor the requirement for precision of articulation in certain specific contexts. Let us take an example of how we do this. The lexical models contained within the dictionary embody phonological representations — three instances of such representations might be: *cat* /kæt/, *cad* /kæd/, *dog* /dog/. If a pair of words of similar syntactic category occur within a single sentence and the phonological models of those words differs in only one segment, then the precision of articulation of the second of the pair is increased. So, in the sentence *The cad chose a cat for a pet* we find the pair *cad* and *cat* of similar syntactic category and differing in only one phonological segment — /d/ *vs.* /t/. Where such a collocation does *not* occur the /t/ of *cat* would be subject to a rule of imprecision which would provide a phonetic model of the syllable ending in an unreleased [t] if the word is followed by a word beginning with a consonant. But on this occasion, where increased precision is needed, the syllable model called includes a released [t] — the perceptual effect being an increase in number and degree of contrastive parameters between *cad* and *cat.*

In human speech this phenomenon is regularly observed and it is hypothesised that it is there to minimise predicted perceptual confusion. When listening to synthetic speech it follows that the effect is expected, and that if it does not occur the result will at best be a perceived local dip in naturalness and at worst possible perceptual confusion.

There are several parameters describing precision within SPRUCE's word and syllable models and there is a comprehensive set of rules which trigger different degrees of precision in particular contexts.

8. OTHER VARIABILITY TO MODEL

In SPRUCE this concept of variable modelling for linguistic elements within a speech synthesis system extends beyond the variants just described, which are by definition intrinsically determined (factors within the sentence text trigger the effects). We have also incorporated extrinsic triggering of a range of similar effects for improved naturalness.

An example of this is the need from time to time to introduce contrastive emphasis on a particular word while modelling particular speaking styles.[19] A recent development is to incorporate pragmatic features into the linguistic model. The phonetic effects are accounted for within the theory of pragmatic phonetics.[20, 7]

One of the many ways in which this effect might be accomplished involves supervising the lengthening of the syllabic nucleus, modifying the fundamental frequency during this part of the syllable and adjusting formant frequencies to negate the calculated effects of coarticulation between segments within the syllable.

These and other effects determined both intrinsically and extrinsically to the input text itself make for a continuously varying acoustic output for SPRUCE, and it would be rare for two sentences to ever follow precisely the same output patterning. The result is a considerable improvement in perceived naturalness because much of the anticipated variability of human speech is now replicated in the synthetic signal.

7. CONCLUSION

In this paper we describe how three kinds of variability in speech production are modelled in speech production theory and carried over into high level speech synthesis. We note that by introducing dynamic models of word and syllable acoustics and by careful lexical modelling on several different levels, we are able to supervise the production of the output signal in such a way that all three kinds of variability are systematically introduced. In this way we are beginning to model important properties of human speech within the SPRUCE high level speech synthesis system. These are precisely the properties of speech which contribute to its humanness and therefore to the naturalness of the simulation.

REFERENCES

[1] E. Lewis and M.AA Tatham. 'A generic front end for text-to-speech synthesis systems.' *Proceedings of Eurospeech'93.*1993, pp. 913-916

[2] J.N. Holmes. *Speech Synthesis and Recognition.* Van Nostrand Reinhold, Wokingham (U.K.). 1988

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

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

[5] RA Sharman. 'Concatenative speech synthesis using sub-phoneme segments.' In *Proceedings of the Institute of Acoustics* 16. Institute of Acoustics, St. Albans (UK). 1994,pp.367-374

[6] O. Engstrand. 'Systemacity of phonetic variation in natural discourse.' *Speech Communication* 11. 1992, pp. 337-346

[7] K. Morton and M.A.A. Tatham. 'Pragmatic effects in speech synthesis.' *Proceedings of Eurospeech* '95. Madrid. 1995, this set of volumes

[8] A.C. Gimson. *An Introduction to the Pronunciation of English.* Edward Arnold, London. 1984

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

[10] C. Fowler. 'Coarticulation and theories of extrinsic timing.' *Journal of Phonetics* 8. 1980, pp. 113-133

[11] R Carlson. 'Synthesis: modelling variability and constraints.' *Speech Communication* 11. 1992, pp. 159166

[12] K. Morton and M.AA Tatham. 'Devoicing, aspiration, and nasality — cases of universal misunderstanding?' *Occasional Papers* 23. Linguistics Dept., Essex University (U.K.). 1980, pp. 90-103

[13] C.P. Browman, and L. Goldstein. 'Articulatory phonology: an overview.' *Phonetica* 49. 1992, pp. 155-180

[14] M.AA Tatham. 'The supervision of speech production — an issue in speech theory.' In *Proceedings of the Institute of Acoustics* 16. Institute of Acoustics, St. Albans (U.K.). 1994, pp. 171-181

[15] M.AA Tatham. 'The supervision of speech production.' In C. Sorin, J. Mariani, H. Meloni and J. Schoentgen (eds.) Levels in Speech Communication: Relations and Interactions. Elsevier, Amsterdam 1995, pp. 114-126

[16] E. Saltzman. 'Task dynamic coordination of the speech articulators: a preliminary model.' In H. Heuer and C. Fromm (eds.) *Generation and Modulation of Action Patterns.* Springer-Verlag, Berlin. 1986, pp. 129-144

[17] M.AA Tatham and K. Morton. 'Precision.' *Occasional Papers* 23. Linguistics Dept., Essex University (U.K.). 1980, pp. 107-116

[18] K. Morton. 'Cognitive phonetics — some of the evidence.' In R Channon and L. Shockey (eds.) *In Honor of Ilse Lehiste.* Foris, Dordrecht. 1986, pp. 191-194

[19] B. Granstrom. 'The use of speech synthesis in exploring different speaking styles.' *Speech Communicationl1.* 1992,pp.347-355

[20] K. Morton. 'Pragmatic phonetics.' In W.A. Ainsworth (ed.) *Advances* in *Speech, Hearing and Language Processing* 2. JAI Press, London. 1992, pp. 17-53