Units of Representation for Speech Synthesis

Mark Tatham

Reproduced from: Tatham, M. (1991) Units of representation for speech synthesis. In *Proceedings of the Institute of Acoustics*, R. Lawrence (ed.) Vol. 13 Part 2 (1991). St Albans: Institute of Acoustics

---

0. INTRODUCTION

Until relatively recently there was little choice concerning units of representation for speech synthesis: the field has been dominated by the use of either an allophonic representation or a diphone representation for building blocks, which, when suitably concatenated make up the basis for creating a waveform simulating the acoustic signal produced when a human being speaks.

It now seems, however, that we have left behind the novelty stage in the development of synthesis, and can now begin to distinguish distinct areas of usage for synthetic speech, and experiment with different approaches — perhaps for different purposes. This paper reports on findings from preliminary work conducted as part of the **SPRUCE** Project (Tatham 1990) in the Advanced Speech Technology Laboratory at Essex University and collaborative with the Department of Computer Science at Bristol University (Eric Lewis, co-investigator). The experiment was designed to investigate the relative merits of different approaches, particularly in the choice of units of representation.

The uses for synthetic speech which were considered ranged from the creation of predetermined sentences, through the output for limited restricted domain dialogue systems, to full-blown interactive systems where it might be impossible to predict in advance anything of what the device might have to speak. The need for all such systems is gaining momentum; so it seems appropriate to investigate the various strategies from a relatively practical perspective: a synthesis system need not be able to do more than it has to in its dedicated environment. Put quite simply: which units of representation work best in which environments?

The chosen method for creating the waveform was what is known as formant synthesis. This is the most widespread method, and, within certain relatively unimportant constraints and under the right conditions, is capable of very high quality output almost indistinguishable from human speech (Holmes 1985). The synthesiser chosen was the **JSRU** model (Holmes 1988) in the implementation engineered by Loughborough Sound images Ltd. The device itself exceeds, in its output quality, our very best efforts in producing an input to drive it which can match its theoretical design capabilities, given that, in the work reported here, all inputs produced were done so by rule.

1. THE USES OF SYNTHETIC SPEECH

There is no need here to rehearse in detail the various uses to which synthetic speech can be put. The most obvious ones are those involving announcements of some kind or other where the message is fixed *('Fasten your seat belts ')*, variable in a limited *way (‘The number you require is.. two.. six.. two.. five.. nine.. o’)*, quite variable *(‘Because of expected drizzle with intermittent heavy rain, driving conditions on the M11 will probably be difficult till after teatime’)*. In dialogue systems we can imagine very restricted limited domains (telephone directory inquiries) and less limited domains (airline booking systems). In some circumstances the need may be for totally unpredictable speech content (spoken e-mail, talking books).

The more restricted of these domains — certainly fixed message announcements — can be handled by ordinary recordings or coded recordings, but that is not at issue: the investigation here is concerned with how to do all of these things using formant synthesis

rather than whether or not we should so do. What the investigation required therefore was examples of input to the formant synthesiser based on choice of various units of representation.

2. THE FORMANT SYNTHESISER

The device used was the LSI implementation of the **JSRU** formant synthesiser. The model is parametrically oriented, the parameters being twelve in number: frequencies and amplitudes of the first three formants, amplitude of the fourth formant (fixed in frequency), frequency of the nasal formant, low frequency amplitude (a parameter used in conjunction with first formant amplitude to determine nasal formant amplitude), degree of variable mix of periodic and aperiodic excitation, the mark/space ratio of the glottal opening for periodic excitation and the fundamental frequency of the periodic excitation. The arrangement is described in detail in Holmes (1985).

The input consists of a so-called parameter file comprising a sequence of frames of parameter values, where a frame contains a value for each parameter to be spoken for 10 ms. Thus one second's worth of speech requires a parameter file of 100 frames each containing 12 values. In the work described here the mark/space ratio of the glottal opening and the nasal frequency were fixed at appropriate values for the voice being synthesised.

3. THE SYNTHESIS PROGRAM

The synthesis program used is part of the **SPRUCE** Project. Its task is to produce a parameter file for driving the synthesiser according to a plain text input. The system is dictionary based to avoid difficult orthography-to-phoneme conversion and to permit storage of certain syntactic, semantic and phonological information. The synthesis strategy is the correct concatenation of representations of the various chosen units to form a parameter file for output to the synthesiser. According to the program's interpretation of the input text, and in conjunction with information retrieved from the dictionary a sequence of units is selected from an inventory. The selected units are conjoined and combined with prosodic information derived in parallel before creating the output parametric file which is then taken to the synthesiser. For the purposes of the work reported here the **SPRUCE** synthesis program is arranged to provide inventories of different sized units of representation, and the concatenation rules are in parallel sets appropriate for conjoining units of differing types.

4. THE UNITS OF REPRESENTATION

The units of representation chosen for comparison were: sentence, phrase, word, syllable and allophone. These are of course linguistic units — that is, they have a meaningful place in linguistics and are therefore theoretically motivated. The possible use of diphones or other units is not discussed here because such units do not have the same linguistic motivation.

Put simply, the task was to create a sentence by concatenating either sentences (not a very difficult task, since the entire sentence was already represented!), or phrases, or words, or syllables or allophones where each sentence created would consist only of concatenations of similar units: that is, all words or all syllables, for example. Since the text input to the system was always the same orthographic representation the dictionary search always resulted in the same output and the calculated prosodics were the same for all sentences, no matter what the units of representation. This meant that the variables were restricted to

- the accuracy of the representation of the units,
- the correctness of the rules for conjoining the units and
- the correctness of the algorithm for fitting the prosodic contours to the conjoined units.

5. THE INVENTORY OF UNITS

Several different sub-inventories were prepared in this work, each holding different sized units of representation (allophones, syllables, words, phrases, sentences), and arranged in

parallel such that a given input text sentence could be synthesised by using a single sentence, a sequence of phrases, a sequence of words, a sequence of syllables or a sequence of allophones.

6. ALLOPHONES

In most text-to-speech systems the units which form the building blocks for creating sentence parameter files are allophones (though they are sometimes called phonemes). In fact, though, they are not allophones but rather special abstract representations of what have been ***called extrinsic allophones*** — that is, an allophone derived *phonologically* rather than phonetically. Any one entry in the inventory can be thought of as a single frame specifying synthesiser parameters. Accompanying this representation is a value for its duration (usually shown as a single number). Prior to conjoining therefore an allophone is rather like a single frame repeated for the number of frames typical of that entry in running speech and showing no variation of value for any parameter throughout that duration. Various conjoining algorithms are in use; all are designed to smooth any abrupt changes in value for parameters from unit to unit in the concatenation.

The representation of allophone units used in the work reported here is quite different. It takes the form not of a single frame but of the number of frames that this allophone needs for a normalised representation, and where the values in each frame have been derived from sequential 10 ms samples of an entire allophone in real speech. Another way of saying this is that the inventory stores a complete running allophone from real speech, together will all running variability which might have been present in the real speech. Coarticulatory effects, since they are specific to a particular phonetic context, are not present. These representations are derived by parametric analysis of appropriate samples of natural speech.

Parametric analysis is the direct analysis (in this case on a 10 ms sampling basis) of the natural speech waveform. ASTL's parametric analysis procedure is very accurate, being partly automatic and partly interactive with the researcher, and designed to be directly complementary to the **JSRU** synthesiser design.

7. OTHER UNITS

The other units stored in the alternative parallel inventories consisted of whole sentences, phrases and words. As with the allophone representations, these were also derived from natural speech by our parametric analysis procedure to preserve the variability found in speech waveforms. In the case of these units coarticulatory effects were, of course, preserved within the unit, though they were taken out at the unit boundaries. Clearly then, all coarticulatory effects were preserved in the case of sentence units, whereas none were preserved in the case of allophone units.

8. THE DATA

The data was prepared to make up the five inventories in the synthesis program. To give a single example from the many actually used, this consisted of

1. 'how is it different' — among others
2. 'how' 'is it' 'different' — among others
3. 'how' 'is' 'it' 'different' — among others
4. 'how' 'is' 'it' 'diff' 'rent' — among others
5. 'h' 'au' 'i' 'z' 't' 'd' 'f' 'r' 'a' 'n' — among others

[**JSRU** phonetic transcription]

These representation began as recordings of a human speaker.

1. was a recording of a complete sentence;
2. isolated phrases;
3. individual words;

4. individual syllables;
5. individual phonetic segments.

For 2. to 5. the items were placed a neutral frame by the speaker. These frames ensured that there were no coarticulatory effects at the boundaries — or at least ensured that the coarticulatory effects were consistent and not sentence-contextual. All items were recorded in random order, not in the order they were later to be assembled or in the order shown above.

The recordings were then parametrically analysed into files for the inventory, consisting of a set of 10 ms samples with values for 12 parameters in each. In the inventory the fundamental frequency of the original was set throughout to a single value — that is, a monotone.

The **SPRUCE** program took, in this example, as its textual input the sentence *'How is it different'*, performed a syntactic analysis on this and assigned a preliminary intonation contour. Word stress was derived from the **SPRUCE** dictionary, and sentence stress was assigned in accordance with the syntactic analysis and merged with the intonation contour.

In session 1 the complete sentence representation was retrieved from the inventory, the calculated intonation and stress contours assigned and the resulting file was taken to the synthesiser (Fig. 1). In session 2 the system retrieved from the phrase inventory the appropriate phrases, conjoined them by rule and assigned the prosodics before output (Fig. 2). Sessions 3, 4 and 5 were similar: words, syllables and allophones were retrieved and conjoined in the correct order, again by rule. The conjoining rules were different in each case: the most complex algorithm being used in the case of allophones and the least complex (so-called 'straight line joining' of each parameter for each adjacent segment) in the case of phrases (Figs. 3-5 respectively).

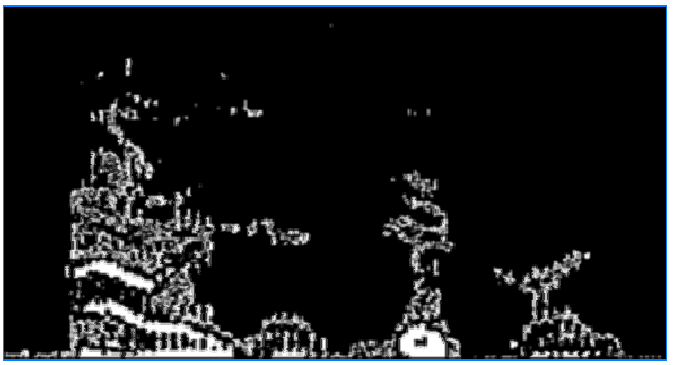

Fig. 1 *'How is it different?'* — resynthesised from a sentence sized unit.

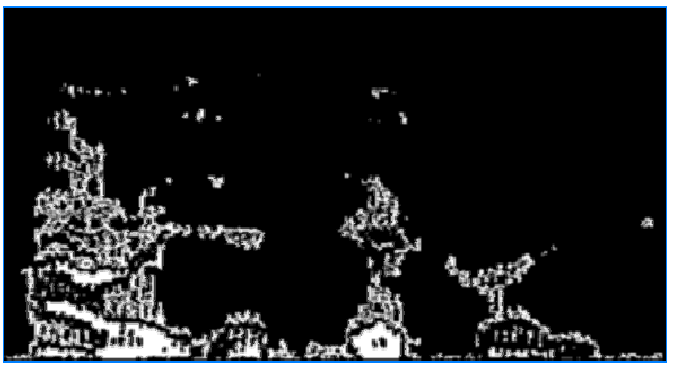

Fig. 2 '*How is it different?'* — synthesised from phrase sized units.

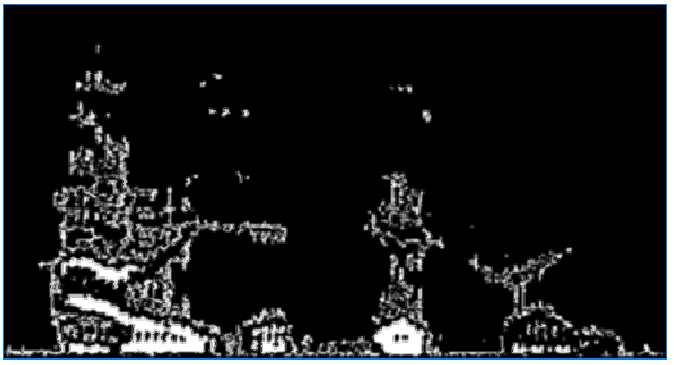

Fig. 3 '*How is it different?'* — synthesised from word sized units.

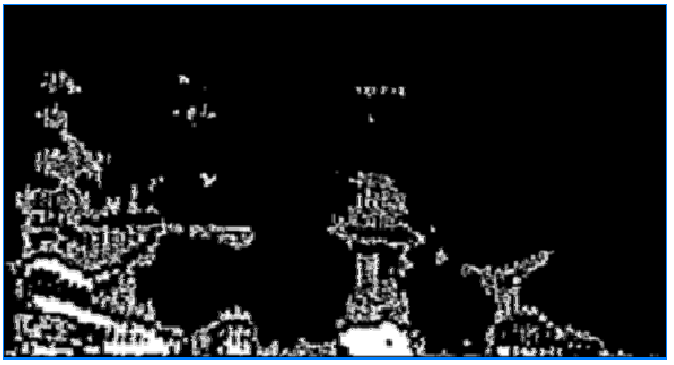

Fig. 4 '*How is it different?'* — synthesised from syllable sized units.

Fig. 5 '*How is it different"* — synthesised from allophone sized units.

Fig. 6 Some of the syllables used — 'how' — 'is' — 'it' — 'diff' — 'rent'.

Fig. 6 shows spectrograms of some of the syllables as entered in the inventory, but put through the synthesiser with no further processing. The effects of the conjoining rules can be seen by comparing Fig. 4 with Fig. 6. The time scale of Figs. 1-5 is 1s, and of Fig.6 1.8s. The frequency scale ranges from 0Hz to 5kHz.

9. RESULTS

The results were in general not surprising. They showed that synthetic speech could be generated automatically from various sizes of linguistically motivated units. The best rendering was, of course, the resynthesis of the complete sentence; the least satisfactory was the sentence recreated from conjoined allophones.

What was of interest, however, was the direct correlation between size of unit of representation and naturalness of the synthetic output when the amount of processing is taken into account. The more processing the less natural, but the more versatile. The same syllables, or allophones, for example could have been combined into different sentences, whereas the whole sentence was, of course, just that — a single unique sentence.

What was fairly novel was the synthetic output made from conjoining representations of allophones resynthesised from human speech. The conjoining algorithm used here was an adaptation, and simplification, of the **JSRU** algorithm (Holmes, Mattingly and Shearme 1964), but errors seemed to be masked by the subjective effect of a dramatic increase in naturalness compared with the usual method of representing allophones (discussed above).

10. CONCLUSION

There is a trade off between the size of the unit of representation in text-to-speech synthesis and the versatility of the system. A system that uses sentences must, of course, have as many sentences units inventory as are going to be needed in the use the system is to be put. Allophonically based systems, on the other hand, can in principle speak any sentence. Systems based on words and syllables require much larger inventories than the allophone systems, but once again are in principle capable of speaking any sentence.

Naturalness of output correlates with size of unit — the longer the unit the more natural. This is not surprising since the longer the unit is the more of the known and unknown features which contribute to naturalness are preserved. The more conjoining is necessary the less natural the speech sounds.

The novel feature of the work reported here was the use of resynthesised allophones in the minimal unit system. The resultant speech was markedly more natural than most systems I have heard and probably produced less listening fatigue.

The oral paper presented at the Meeting of which this printed version appears in the Proceedings demonstrates recordings of samples of the system's output. Unfortunately this paper cannot do that, but the spectrograms do give some idea of the similarity between the outputs based on the different inventories of units used.

REFERENCES

Holmes, J. N. (1979) Synthesis of natural sounding speech using a formant synthesizer, in *Frontiers of Speech Communication Research* (B. Lindblom and S. Ohman, eds.). London: Academic Press, pp. 275-285

Holmes, J. N. (1985) A parallel formant synthesizer for machine voice output, in *Computer Speech Processing* (F. Fallside and W. A. Woods, eds.). London: Prentice Hall International, pp. 163-187

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

Holmes, J. N., Mattingly, I. G. and Shearme, J. N. (1964) Speech synthesis by rule. *Language and Speech* 7, pp. 127-143

Tatham, M. A. A. (1990) Preliminaries to a new text-to-speech synthesis system. *Proceedings of the Institute of Acoustics*, Vol.12: Part 10, pp. 233-240