Proceedings of the Institute of Acoustics

THE JSRU TEXT-TO-SPEECH SYSTEM AND SOME THOUGHTS ABOUT ITS PRONUNCIATION TASK

E Lewis(1), R Sampson(2)

(1) Department of Computer Science, University of Bristol, Bristol.
(2) Department of French, University of Bristol, Bristol.

1. INTRODUCTION

Research into speech synthesis has been particularly active over the last thirty to forty years with one of the prime objectives being the production of a text-to-speech system capable of accurate reproduction of the human voice. Development of such systems has reached the stage where commercial text-to-speech synthesisers are available but none can yet claim to provide truly natural sounding speech. In 1986, Logan, Pisoni and Green[1] evaluated eight text-to-speech systems for segmental intelligibilty and their results indicated that the expected rate of perceptual errors for the best performing systems varied from about three to twelve times that for natural speech. In terms of "naturalness", however, the authors are of the opinion that even the best systems are not yet acceptable for reproducing large amounts of natural sounding speech. None of the systems assessed by Logan et al was British but a British system, subsequently referred to throughout this paper as the JSRU system, has been in existence since 1964. It was developed at the Joint Speech Research Unit at GCHQ, Cheltenham, and was first described by Holmes, Mattingly, and Shearme[2]. Subsequent development of the system has been described in publications by Holmes, Wright, Yates and Judd[3] and Edward[4],[5].

In 1985 the JSRU was moved to the RSRE, Malvern to become the Speech Research Unit, and further development of the system by the SRU ceased until very recently. The JSRU system has, however, been made available to other organisations and research groups such as GEC, British Telecom, PA Technology and various university departments, and these groups have continued to develop the software, although no commercial system based on the JSRU software has been released yet. In 1989 a Speech Research Unit Research Group was formed to provide a forum for researchers interested in improving both the hardware and software of the JSRU system. This group has formulated some proposals for further development of the JSRU system which include the creation of a much larger lexicon, the improvement of the interaction between the pronunciation rules, the lexicon and the affix tables, the creation of a parsing module and the improvement of the intonation algorithm. This paper describes the activities of the authors in improving the pronunciation sub-system.

2. THE JSRU SYSTEM

In 1985 the JSRU text-to-speech system consisted of some 250 procedures, comprising 500K bytes of code written in the real-time language RTL/2 and distributed between 40 files. The documentation concerning the distribution and communication of the procedures between the files was minimal and, although RTL/2 is a language with many good features, similar in some respects to Algol, it has been surpassed in terms of general usage by languages such as Pascal, Modula-2 and C and,

in particular, lacks the portability of these languages. Consequently, one of the authors[6] took the decision to rewrite the complete system in C. This version of the software has been available since 1990 and is essentially a line by line copy of the RTL/2 version, the intention being to reproduce exactly the JSRU system, including its errors. There is no doubt that the structure and content of the code could be improved considerably should one decide to completely redesign the software. The system has also been rewritten in Pascal by British Telecom.

The steps involved in converting text to speech in the JSRU system are shown in Figure 1 and described in outline below. The system consists essentially of five tasks which are as follows:

1. Conversion task. This task is essentially a pre-processing phase for converting unrestricted text to restricted text. Unrestricted text consists of such items as numbers, abbreviations, unpronounceable words, and especially acronyms which need to be spelled out(e.g. BBC), while restricted text consists of plain English text composed from alphabetic characters plus the punctuation characters comma, full-stop and question mark. This task first tries to retrieve the phonetic form of its input from the exceptions dictionary and, if successful, passes it directly to the next task. Otherwise, it converts its input to restricted text and checks that it is pronounceable.

2. Pronunciation Task. This task converts text to phonemes with associated stress values. It first looks in the dictionary, unless this has already been done by the conversion task, and if the word is not found then it removes the affixes one at a time, checking at each stage if the current stem is in the dictionary. If, after removing all the affixes, the root is not in the dictionary, then the task works out its pronunciation by rules and appends the pronunciation of the affixes. The stress application rules are subsequently applied to the whole word followed by some further pronunciation rules for reducing long vowels. The output from this task is the phonemic representation of the word with the start of syllables and stress patterns marked.

3. Allophonic task. This task builds a complete breath group from the words passed to it by the pronunciation task and then applies a set of rules for modifying the pronunciation of phonemes according to their context. The output from this task is a broad phonetic representation of each phrase or sentence.

4. Prosody task. This task adds intonation and timing to the phonetic text and expands each phoneme into one or more phonetic elements. The output consists of a list of these elements together with their corresponding pitch and duration, identified in Figure 1 by the term 'narrow phonetic representation'.

5. Lower Phonetic task. This task converts the output from the prosody task into control parameters for the hardware synthesiser. The parameters are calculated as a succession of frames, each of 10ms duration. Transitions between phonetic elements are calculated using information from the lower phonetic table for the target and boundary values for each element.

Input and output to the system can be made at most of the interfaces between the tasks. Input can be unrestricted text, restricted text, phonemic text, broad phonetic text or narrow phonetic 'text',

THE JSRU TEXT-TO-SPEECH SYSTEM

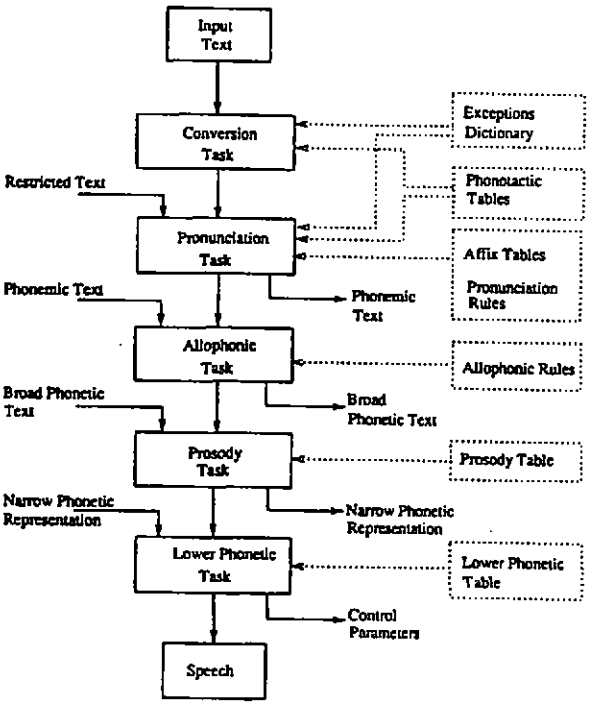

Figure 1: Main processing tasks of the JSRU system

while output can be phonemic text, broad phonetic text, narrow phonetic 'text', synthesiser control parameters or speech.

The system is capable of taking any text and converting it into speech, spelling out any words that are marked as such or that it decides are unpronounceable. Not surprisingly the system does not always assign the correct pronunciation and stress. Errors can be compensated for to a certain extent by false spelling or by including the phonetic form as part of the original text. However since the objective must be to provide a system which requires no special input markers this is not a very satisfactory solution to the problem. The longer term answer is to modify the existing system to provide correct processing of the input.

3. FURTHER DEVELOPMENTS

The Holmes synthesiser has been shown to be capable of producing extremely good quality synthetic speech provided the control parameters are suitably chosen[7,8]. The problem, therefore, is

to produce the correct model for deriving these parameters. Arising from the SRURG's meetings last year the following areas have been identified as ones which should prove beneficial for making improvements.

1. At the pronunciation level the rules clearly cannot cope with all situations and the current dictionary is not large enough to cope with the exceptions. The interaction of the dictionary, affix tables and pronunciation rules needs to be revised with the object of incorporating a much larger lexicon. Some work along these lines has already been completed by PA Technology and is available to academic users for research purposes.

2. In order to provide a proper lexical and sentential stress pattern it is essential to have syntactic information. The current format of the dictionary could be modified to provide such information and a syntax module could be added to the system between the pronunciation and prosodic tasks to provide a syntactic parse. The appropriate stress information would then be passed to the prosody module.

3. Ideally some semantic processing should be incorporated into the system in order to provide a more natural sounding intonation contour. This could also be used to incorporate some of the results of Morton's[9] work on including mood in synthetic speech.

4. The allophonic table currently provided is extremely small and needs to be enlarged to cater for the many allophonic variants that can occur in English. Again some work to this effect has already been carried out at PA Technology.

5. At the prosody and lower phonetic levels more flexibility is required in order to provide more detailed loudness and fundamental frequency contours. In addition the tables could be modified to allow asymmetric transitions between phonetic elements. Such improvements have been implemented at GEC by Holmes and Pearce[10].

Although this paper is concerned with the JSRU speech synthesis software it is, perhaps, worth mentioning that the design of the hardware synthesiser could almost certainly be improved by utilising current technology more fully. One hardware improvement which has been made by GEC is to increase the bandwidth of the synthesiser substantially so that it can better synthesise female voices.

4. PRONUNCIATION SUB-SYSTEM

In the organisation of the pronunciation task, a basic issue to be addressed is the extent to which the assignment of pronunciations to words should be achieved by the application of rules or by means of a look-up dictionary. The original JSRU version relied principally on the former alternative. The result was an extensive set of phonological rules, completed by a small dictionary for exceptions which contained about 100 entries only.

Increasingly, however, the desirability of changing this balance has become apparent. The number of exceptional items has been found to be much larger than anticipated, and experience has

shown that the devising of further, more intricate pronunciation rules to try to handle them is unlikely to be successful. Given this and the availability of improved systems of data storage and retrieval, the implication might be that phonological rules could be largely dispensed with and that the pronunciation of words might instead be determined primarily by direct reference to a comprehensive dictionary suitably compiled. However, such a strategy is not entirely appropriate for two reasons. On the one hand, the scale of any such dictionary would be enormous and require considerable storage. The Shorter Oxford Dictionary, for instance, contains approximately 160,000 entries, and these are *lexemes*. The number of actual *word forms* (e.g. *propose, proposes, proposing, proposed* are derived from the single lexeme *propose*) would doubtless run to half a million or more. On the other hand, the problem of new lexical creations presents itself. Without a set of fairly detailed pronunciation rules, pronounceable forms such as *gazzamania* and *post-thatcherism* would be merely spelt out. A balanced arrangement is thus required, the proposed pattern being that the phonological rules would handle regular and productive patterns whilst a substantial dictionary would cover irregular forms as well as forms displaying minority regularity. The key question is what principles to adopt in determining where the frontier should lie between the use of phonological rules and the dictionary in pronunciation assignment.

Two test cases might be discussed in this context, the first concerning affixation. Few English words consisting of one morph contain more than two syllables, and indeed the great majority are monosyllabic. These, therefore, are not really problematic for stress assignment, and given their fairly limited number and the fact that new lexical creations typically do not involve monomorphic forms, the phonemic representation of such forms can readily be determined by pronunciation rules backed by dictionary entries for exceptional items. Words composed of two or more morphs are, however, problematic and it is this class of word which represents the really productive area of the lexicon for neologisms.

To handle the pronunciation of words derived by affixation, the JSRU system contains an affix component, comprising 46 prefixes and 39 suffixes each specified as to its pronunciation and to its significance for stress assignment. As in the MITalk[11] system, words are subjected to affix-stripping before the application of pronunciation rules. The high frequency of affixed word forms, especially in texts of a more technical nature, indicates the importance of having a maximally exhaustive affix component to ensure the specification of correct pronunciations for derived forms. This the JSRU does not have (nor, incidentally does DECtalk, probably the best commercial system available, which produces pronunciations such as *évanesce, phónometry*). Fudge[12] provides an invaluable bank of data to draw upon for expanding the set of affixes, but further improvements are also possible. These rely often on relaxing the strict theoretical linguistic criteria usually adopted for morph recognition and taking on the perspective of the "naïve native speaker" for whom awareness of the identity of morphs may at times be somewhat nebulous. Thus, included amongst suffixes could be:

1. quasi-suffixal sequences like *-een, -oo, -esce*, as in *halloween, kangaroo, deliquesce* (which Fudge does recognise), since these have the important property of attracting primary stress to themselves.

2. "learned" suffixal elements, typically taken from Greek, such as *-metry, -ology, -osis* (many but not all of which are recognised by Fudge). These have absolutely consistent pronunciations and stress-assignment characteristics, and they form a relatively closed set, albeit a substantial one.

3. systematic use of unanalysed complex suffixes such as *-mental, -ologist, -scopic*, whose pronunciation and stress implications the native speaker of English recognises at first sight. Removing these from the application of cyclical stress rules thus actually seems to reflect normal language use, and at a practical level would unclutter the rules concerned with stress assignment. (Interestingly, the JSRU system seems to recognise this possibility by including amongst the suffixes recognised *-ic, -al, -ical, -alic*).

The enrichment of the affix component that results from including the suffix or pseudo-suffix types indicated will permit correct pronunciations to be assigned automatically to the complex forms being generated constantly in the most dynamic part of the English lexicon, namely words of polymorphic structure.

The other test case concerns the phonological ramifications arising from the "adde" (i.e. add -E) rule. This applies when certain suffixes are recognised and has the effect of adding a final -E to the stem sequence after affix stripping. For this rule, the total set of 39 suffixes is divided into two subsets, with just 10 suffixes being identified as not triggering the addition of final -E after their removal. Amongst the latter set are *-ic, -ical* and *-ify* whilst in the larger subset appear such suffixes as *-ed, -er* and *-ing*. The purpose behind the identification of this twin set of suffixes is primarily to be able to determine the correct phonemic representation for the stressed vowel of the stem. Thus, in *logic, magical* and *gratify*, the presence of a short stressed vowel argues against the addition of a final -E after suffix removal, since a final -E automatically causes the lengthening of a preceding stressed vowel if just one consonant precedes it (as in *doge, rage*). On the other hand lengthening is found in *paged, wider* and *raving*, and hence the addition of final -E is required after suffix stripping in these forms.

However, the original rule has been complicated by the problem of how to handle the palatalisation of "c" and "g" when these appear stem final. Looking in particular at words with stem final "g", we find that these give rise to considerable complexities. In forms such as *paged* and *waging*, a correct pronunciation is assigned following the addition of final -E, but, in the JSRU rule, *logic* and *magical* preserve a velar value for "g" since any addition of final -E after suffix stripping would yield unwanted stressed vowel lengthening. Further problems come with words such as *singer, finger* and *ginger*. Here a suffix *-er* is recognised, but the addition of a final -E is specifically blocked so that all the pronunciations assigned rhyme with *singer*. Thus, *logic, magical, finger* and *ginger* have all to be included in the dictionary.

An alternative approach might be to develop the existing "adde" rule further by including forms whose stems end in *-ng* or reassigning the suffixes *-ic, -ical and -ify* to the larger set of suffixes whose removal triggers the addition of a stem final -E. However, this complication of an already complex rule would itself result in new entries being needed in the dictionary. Neither solution seems to be

entirely satisfactory.

A rather simpler approach might be adopted instead. The "adde" rule falls down primarily because it is tackling not one, but two, phonological problems simultaneously, those of assigning appropriate vowel length in suffixal forms (e.g. *raged* versus *magic*) and the palatalising of stem final "c" and "g". The former it handles successfully. The latter problem is of a different type and requires separate treatment. If the situation with forms such as *singer, finger* and *ginger* in particular is considered, the decision as to what forms to treat as regular and deal with by rule and what forms to view as exceptional and include in the dictionary is determined apparently on the basis of the relative frequency of occurrence of the items within running text. The pattern seen in *singer* and *ringing* is thus chosen as regular.

However, another criterion exists which is perhaps preferable here. This ties in with phonological productivity and reflects the native speaker's intuitions. When presented with a novel English word form ending with the sequence *-nger*, we may enquire how a normal speaker would pronounce it. Usually, the outcome would probably be [nd͡ʒ], since the palatalisation of "g" before "i", "e" or "y" is, of course, a general process in English unless other factors intervene. This fairly straightforward realisation is, however, obscured by the fact that a limited set of established, high-frequency words do not follow this pattern.

The conclusion, therefore, is that it is items such as *singer, banger* and *banged* which should properly be included in the dictionary, since they represent a non-productive pattern of pronunciation, for all the familiarity of such forms and, more generally, it may be assumed that all forms with stem final "c" and "g" preceding suffixes such as *-ic, -ical, -ed* and -y should participate in a general palatalising pronunciation rule.

5. CONCLUSIONS

In order to obtain good quality speech from the JSRU text-to-speech system some of the algorithms for converting orthographic text to synthesiser control parameters need to be substantially revised. One of the tasks comprising the JSRU text-to-speech system which has been identified by the SRURG for further development is the pronunciation task involving the interaction of the exceptions dictionary, affix tables, phonotactic tables and pronunciation rules. From examining two components of this task, affix stripping and the "adde" rule, it emerges that considerable improvements can be obtained by the application of relatively simple principles since this approach helps to clarify the scope of the dictionary and the pronunciation rules, and the way in which they might interact more efficiently.

6. REFERENCES

[1] J S LOGAN, D B PISONI & B G GREENE, 'Measuring the Segmental Intelligibility of Synthetic Speech: Results from Eight Text-to-Speech Systems', *Behav. Res. Meth. Instr. Comp.*, 18 (2) p100 (1986)

THE JSRU TEXT-TO-SPEECH SYSTEM

[2] J N HOLMES, I G MATTINGLY & J N SHEARME, 'Speech Synthesis by Rule', *Lang. Speech*, 7 p127 (1964)
[3] J N HOLMES, R D WRIGHT, J W YATES & M WJUDD, 'Extension of the JSRU Speech Synthesis by Rule System', *9th Int. Cong. Acoust., Madrid*, (1977)
[4] J A EDWARD, 'Pronunciation Rules for English Text', *JSRU Res. Rep. No. 1014*, Speech Research Unit, RSRE, MAlvern (1982)
[5] J A EDWARD, 'Rules for Synthesising Prosodic Features of Speech', *JSRU Res. Rep. No. 1015*, Speech Research Unit, RSRE, Malvern (1982)
[6] E LEWIS, 'A C Implementation of the JSRU Text-to-Speech System', *Rep. No. TR-89-15*, Comp. Sci. Dept, Bristol University (1989)
[7] J N HOLMES, 'The Influence of the Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesiser', *IEEE Trans. Audio Electroacoust*, AU-21 p298-305 (1973)
[8] W J HOLMES, 'Copy Synthesis of Female Speech Using the JSRU Parallel Formant Synthesiser', *Proc. Eur. Conf. Speech Comm. Tech* p513 (1989)
[9] K MORTON, 'Pragmatic Phonetics', *Advances in Speech, Hearing and Language Processing*, ed. W A Ainsworth, JAI Press London (*forthcoming*)
[10] W J HOLMES & J B PEARCE, 'Automatic Parameter Derivation for Synthesis-by-Rule Allophone Models', *Proc Inst. Acoust.*, 12 (10) p491 (1990)
[11] J ALLEN, M S HUNNICUTT & D KLATT, 'From Text to Speech: The MITalk System', CUP (1987)
[12] E FUDGE, 'English Word Stress', London: Allen & Unwin (1984)