



A TEXT-TO-SPEECH SYSTEM FOR BRITISH ENGLISH,
AND ISSUES OF DIALECT AND STYLE

Anthony Bladon*, Rolf Carlson**, Björn Granström**, Sheri
Hunnicuttt** and Inger Karlsson**

ABSTRACT

Although the concept of a multi-lingual text-to-speech system is a familiar one, rather little attention has been given to the question of the variety of each language that is synthesized. This question can be asked not only of national or dialectal varieties but of style differences within those varieties. We present a demonstration and discussion of a British English text-to-speech system. The issue of relatedness across dialects has been addressed in this system which incorporates British Received Pronunciation, in conjunction with a sister system which offers a General American pronunciation.

INTRODUCTION

We have earlier reported on the long term effort to develop high quality text-to-speech systems for several languages (ref 1,2). The approach taken has been to formulate the process in a coherent framework. One criterion was that linguists involved in creating, refining and maintaining the text-to-speech software should be able to work with constructs and conventions familiar to them without necessarily mastering conventional computer programming. Consequently distinctive features and phonemes are primes in our system; and the rule notation borrows heavily on that used in generative phonology, although it is expanded to easily handle continuous variables such as synthesizer parameters. Another goal was to streamline the transfer to a real-time system, which has the dual advantage of speeding up the testing of rules and of facilitating practical use in different applications.

The text-to-speech systems consist of a structure of rule components and various lexica. The rules are written in the same formalism throughout the system. A similar approach has since been followed by other researchers (ref 3-7). By contrast, some authors have employed different rule notations on different levels in the system. For example, Hertz has reported on a framework (ref 7), which is designed for easier handling of units of different sizes, such as phrases, words, syllables and phones, but this also implies a more complicated, less uniform notation. However in our system we formulate the whole text-to-speech process in a uniform framework and, in order to refer to different-level units, we attach appropriate features to our single stock of symbols. In this way everything from syntactic analysis to detailed sub-phonemic manipulations is handled in an analogous fashion.

* Infovox AB, Box 2503, S-171 02 Solna, Sweden and Phonetics Laboratory, University of Oxford, Oxford OX1 2JF, U.K.

** Department of Speech Communication and Music Acoustics Royal Institute of Technology Box 700 14, S-100 44 Stockholm, Sweden

ACCENT ISSUES

This emphasis upon internal homogeneity has been taken a stage further in our recent work. Two versions of an English text-to-speech system have been developed, one for American English (a General American accent) and one for British English (Received Pronunciation, RP). It was decided to develop these two rule-systems in tandem, with maximum mutual overlap. This decision had several important and perhaps controversial consequences. It meant, in particular, that both accents use a single set of phonemic symbols, despite their phonetic realizations sometimes being rather diverse. The exceptions lexicon and the corresponding reference corpora are as far as possible in a uniform phonemic transcription for both accents. Maximum overlap was applied also to such things as feature-assignments (e.g. which vowels are marked as -TENSE), the strategy for F0 control, and the rules for word-stress placement. At the allophonic level of course departures had to be introduced (e.g. in phonetic vowel qualities, both stressed and unstressed; /r/ and /l/ qualities; timing properties).

More generally, the implication of adopting maximum overlap was not that British was derived from American, nor vice versa, but that an artificial 'common base dialect' was created, from which both accents diverge. Consider the word 'butter'. Should its final syllable include an /r/ (or r-coloured vowel) as in American, or not, as in RP? We include the /r/ since the deletion rule for RP is exceptionless. But then, in the same word, should the intervocalic consonant be a plosive or a flap? We would opt for the plosive, since the plosive-to-flap rule is statable and since /d/ also flaps in American, the British plosive would not be recoverable from a flapped transcription. Consequently in 'butter', transcribed with /t/ and also with /*r/, we have created something of a hybrid.

Other consequences of our maximum overlap principle might look questionable at first sight. The initial vowels in 'Sirius, serious, Sears' have merged phonetically to /ir/ in American; but because their contrast is maintained in British, they would have to appear as contrastive in a shared lexicon. Even more inelegant is the attempt to achieve uniformity of transcription for the word classes represented phonetically in Table I, cf. (ref 8).

Table I Some classes of phonetic forms in (a conservative) American and in British RP pronunciations.

	American	British
rap	æ	æ
grass	-----	a
father	a	-----
hop	-----	ɒ
toss	ɔ	-----
broad	-----	ʌ

To handle this material, in effect, a separate 'phonemic' symbol has to be devised for each of the six classes exemplified by a keyword. Different vowel mergers then take place in the American and in the British rule-systems. From these and similar data one can generalize to say that, typically, a base dialect will reflect the dialect which does NOT have the phonemic mergers, splits, assimilations, elisions, contractions, stress reductions, and so on. Which accent that is, will vary from instance to instance, and hence the hybrid.

How then do we justify this monolithic approach? There are good reasons of several kinds. First, although perhaps least important in the applied context of this paper, there is some support from linguistic theory: language users who can comprehend two dialects (as is usually the case with American/British) appear to mediate this comprehension via cross-dialectal phonemic correspondences which are frequently quite violent to the phonetics and transcend large realizational differences. Second, and more practically, purchasers of text-to-speech systems and service engineers are not typically phoneticians and if they need to be exposed to a phonemic transcription (in a polylectal system) then better that it be just one. Making comparisons of system performance across accents is also greatly facilitated if the systems are as similar as possible. Finally, a monolithic system helps us to avoid getting too "locked in" to the two accents we have somewhat arbitrarily chosen: new accents, such as another rhotic one, Irish English for instance, could the more easily be derived.

STYLE ISSUES

Language variation is being incorporated into our system along a second, and rather ambitious dimension. Consider the pronunciations in Table II obtained from a small sample of speakers, but the issue is the principles involved, not the statistics:

Table II Some typical American and British pronunciations.

	American	British
natural, mutual	tʃ	tʃ
situation	tʃ	tʃ (tj)
statute, constituents	tʃ	tʃ tj
institute, constitute	t	tj (tʃ)
tuition	t	tj
tuna	t	tj

Many American speakers do not attest forms with a /j/ glide, after /t/ as in Table II and also after other alveolar consonants ('dune, new, lunar, assume'). In unstressed (note, not secondary stressed) conditions however, American affrication has gone a long way: and further than in British. But British is currently in a state of flux, with some of the vocabulary affricated to /tʃ/, some nearly so, some less so. More interestingly, British affrication appears to vary with "style": the more casual the speech, the more affrication tends to occur. This even spreads sometimes to the initially stressed words ('tuna, tune').

Our British text-to-speech implementation has been extended to

provide a "style variable", a user-set range of ten values. This device can be used, for example, to propagate more affrication with a "lower" style number. The area of the system in which we first explored this style variable was in fact that of the forty or so function words ('can, have, for, them' etc.) of British English whose pronunciation, though not their spelling, varies considerably with sentence context and style. As an example, the word 'can' in a phrase 'I can go' may have a large number of realizations, some of which may be just acoustically specifiable subtleties, but some at least of which can be rendered transcriptionally: [kæ̃n], [kən], [kəŋ], [kɿ], [ʔŋ]

It is probably reasonable to rank these forms (though in other cases it is often much more arguable) from left to right as graded from most formal to casual. They can therefore be synthesized with style variable values of say 9, 7, 5, 3 and 1 respectively.

To undertake this style ranking more widely through English phonetics is, in the present state of knowledge, rather an uncertain exercise. The normative data have hardly been collected at all. At the same time, there are two particularly good motives for pressing ahead. One is that, at present, the text-to-speech developer is faced with some uncomfortable decisions of simplification when specifying such a highly variable word as "can".

Another reason is a research issue. Suppose for illustrative purposes that the style level definitions for variants of 'can' are agreed, as above. Now if I decide to affricate my /t/ in 'intuition', where on the 'can' scale does this correspond to? Current research gives virtually no answers to such questions of style correspondence, overlap or clashes. It is worth remarking that the results would be of importance to speech recognition work also. Yet the text-to-speech system could be used, say interactively in an analysis by synthesis fashion, to elicit some of those answers. Just one of the benefits would be a more stylistically consistent text-to-speech output.

REFERENCES

1. Carlson, R., Granstrom, B., & Hunnicutt, S. (1982): "A multi-language text-to-speech module," Proc. IEEE-ICASSP.
2. Carlson, R. and Granstrom, B. (1975): "A phonetically oriented programming language for rule description of speech", in *Speech Communication*, Fant, G., Ed., Almquist & Wiksell, Stockholm.
3. Klatt, D. K. (1976): "Structure of a phonological rule component for a synthesis-by-rule program", IEEE Trans. ASSP-24
4. Hertz, S. R. (1982): "From text to speech with SRS", JASA, vol. 72 no. 4.
5. Kerkhoff, J. Wester, J. and Boves, L. (1984): "A compiler for implementing the linguistic phase of a text-to-speech conversion system", Proc. Inst. Phon. Cat. Univ. Nijmegen, vol. 8.
6. Holtse, P. and Olsen, A. (1985): "SPL: a speech synthesis programming language" Ann. Rep. Inst. Phon. Univ. Copenhagen. vol 19.
7. Hertz, S. R., Kadin, J. and Karplus, K. J. (1985): "The Delta rule development system for speech synthesis from text". Proc. IEEE-SIMSC.
8. Wells, J.C. (1982): *Accents of English*. Vol.I. Cambridge Univ Press.