A KEYVOWEL APPROACH TO THE SYNTHESIS OF REGIONAL ACCENTS OF ENGLISH

Briony Williams and Stephen Isard
Centre for Speech Technology Research
University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, Scotland, UK.
Tel. +44 131 650 2790, FAX: +44 131 650 6351, E-mail: briony@cstr.ed.ac.uk

ABSTRACT

Most English text-to-speech synthesisers offer one of only two accents: General American or RP. Developing a new accent is laborious, since it is not possible to choose one accent as a base form and systematically translate to others. We use the approach of Wells ([1]), categorising vowels in terms of abstract keywords that encode classes of words. Thus it is unnecessary to use a phonemic transcription in either the development or the execution of a synthesiser. The "keyvowel" system can be used throughout the synthesis system, avoiding the need to make accent-specific changes manually. The same linguistic resources can be re-used for each new accent. More fundamentally, the keyvowel system functions as a meta-accent that subsumes vowel-related information in all accents of English.

1. THE NEED FOR REGIONAL ACCENTS IN ENGLISH SPEECH SYNTHESIS

A language may include several *accents*, differing not in their syntactic rules (as for *dialects*), but merely in the pronunciation rules. There are many accents of English, especially within the British Isles, but speech synthesisers have generally offered only General American or RP English. Most accents are mutually intelligible, but many users of synthesisers might prefer an accent closer to the one they are familiar with. This is especially true of vocally-impaired users, since the voice becomes their persona. The well-known British physicist Steven Hawking began to use a synthesiser with an American accent, since that was the only kind available at the time. After a long period of using it, he now has no wish to change to a British synthesiser, since "[he] would feel [he] had become a different person" [2]. This comment illustrates how fundamental the synthesiser's accent is to the user's self-perception.

There is a need for more variety in the number of accents offered, not only for disabled people, but also for publicity and presentation. A Scottish bank offering synthetic speech telephone services would probably prefer a Scottish accent to an English one. In addition, the availability of accents would add variety and interest to consumer products that use synthesised speech.

2. REGIONAL ACCENTS AND SPEECH SYNTHESIS: THE PROBLEM

Given that the synthesis of different accents is desirable, the next question is to decide on the most effective method. Various factors must be considered when selecting a method for synthesis.

2.1. Rule-based versus concatenative synthesis

The question of different accents will differ in its impact on rule-based speech synthesis and on concatenative speech synthesis. In the case of the former, preparation of a new accent will require detailed acoustic-phonetic knowledge of the accent, as well as preparation of an accent-specific phonetic lexicon and letter-to-sound (LTS) rules, and detailed phonological knowledge. The detailed acoustic knowledge will require much basic research into the acoustic characteristics of the accent before synthesis can even be attempted.

In the case of concatenative synthesis, this detailed knowledge of acoustic characteristics is not necessary. The resources needed for each accent are: the phoneset (phoneme inventory), the pronunciation lexicon, LTS rules, and a textual representation of a database of recorded speech. Even where the units are derived from a large database of *continuous* speech, this textual transcript of the database would still be required.

2.2. Types of linguistic variation between accents

For concatenative synthesis using existing methods, each accent requires a new phoneset and lexicon, as well as recordings and transcriptions of an accent-specific speech database. These are non-trivial tasks. If two accents differed only in the phonetic realisation of the same phonological system, there would be no difficulty, as the same phoneset, lexicon and text could be used. Accents can differ more fundamentally than this, however, in the following ways (from [1]):

2.2.1. Differences in phonotactic distribution

Two accents use the same phonological system, but the phonemes occur in different syllabic contexts. For example, both RP and Scottish English have the /r/

phoneme. In the latter it appears in any consonantal position in the syllable, but in RP it appears only before the vowel (i.e. in the onset) and not in the coda.

2.2.2. Differences in the phonemic system

Two accents differ in the number or identity of phonemes: e.g., RP contains two low unrounded vowels, /æ/ and /ɑ/, while Scottish English has only one, /a/.

2.2.3. Differences in lexical distribution of phonemes

Two accents may differ only in the phonemes selected for particular words. Even where two accents use the same phoneme system (unlike 2.2.2), and the same phonotactic distribution in syllables (unlike 2.2.1), the phonemes do not always appear in the same words. For example, a typical northern English accent and RP both contain /u/ and /ʊ/, with identical *syllabic* distribution but different *lexical* distribution: the northern accent has /u/ and RP has /ʊ/ in "hook", "look".

2.3. Methods of encoding linguistic variation

Since accents can differ in so many ways, existing methods of concatenative synthesis might use one of two approaches to develop a new accent of English:

2.3.1. "Brute force" approach

Develop an entirely new lexicon and set of LTS rules for each accent. This entails much time and effort, as well as detailed phonological and lexical knowledge. If the addition of a new accent is seen as desirable but not essential, then in commercial terms this approach may be judged not cost-effective.

2.3.2. "Base accent" approach

To simplify the process, develop a dictionary and set of LTS rules in a base accent (perhaps RP), and characterise each new accent's dictionary and LTS rules in terms of differences from this accent.

Although apparently easier, under the second approach any accent chosen as a base accent will at some point fail to show a distinction that occurs in some other accent. There seems to be no single accent containing all possible phonemes and distinctions of English accents. For example, RP English differentiates certain vowels that are not distinguished in Scottish English (eg. /ʊ/ and /u/) but lacks another distinction made in some Scottish accents (between the vowels of "tied" and "tide"). Whichever accent is chosen for the master lexicon, there will be some loss of information from the point of view of other accents, and so a simple translation from an existing accent is not possible.

3 SOLUTION: A KEYVOWEL SYSTEM

3.1. Wells' keyword system for English

Wells ([1]) elaborates a system for classifying the vowel phonemes of English allowing for variations across accents. Instead of stating that the word "pool" contains the vowel [u] in RP and the vowel [ü] in a Scottish accent, he states that it contains the GOOSE vowel, an abstract unit defined in terms of a class of words (eg. *loop, group, move, duke, sleuth*) rather than in terms of a specific pronunciation. The GOOSE vowel is later phonetically defined separately for RP and Scottish. Other keywords are KIT, THOUGHT and CLOTH, with a total of 27 vowel keywords. The string CLOTH (etc.) is treated as a symbol representing a wide range of actual vowel phonemes in various accents. In any given accent, it is possible for two or more keyword classes to be realised using the same vowel phoneme (for example, in near-RP accents, CLOTH and LOT words use the same vowel phoneme /ɒ/, but in General American the word classes use /ɔ/ and /a/ respectively).

3.2. Goodbye to phonemic transcription

This system avoids the need to re-specify all vowel phonemes for a different accent. If all vowels (in the lexicon and LTS rules) are specified in terms of keywords (and hence "keyvowels"), then exactly the same lexicon can be used for all accents. Given the use of a concatenative synthesis system, there is not even any need for a set of realisation rules giving the phonemes for that accent. The same text representation of isolated words can be used for all accents, and it is not necessary to research detailed acoustic-phonetic knowledge of the vowels of the different accents.

The important point is that this cuts out altogether the use of phonemic transcription. In text-to-speech synthesis, there are two stages in generating speech:

a) From orthographic form to phonemic transcription.
b) Phonemic transcription to sequence of speech units.

Using conventional methods of concatenative synthesis, both stages require extensive re-engineering when developing a new accent. The "keyvowel" method has two significant advantages over conventional methods:

3.2.1. Single stage during synthesis

There is only *one* stage. The system converts from the orthographic form directly to speech units specified in terms of keyvowels, with no intermediate phonemic transcription. Instead of grapheme-to-phoneme rules, there will be a set of grapheme-to-keyvowel rules, for use in the rare cases where an input word is not found in

the dictionary. The recorded database of speech units is specified in terms of the keyvowels and so can be accessed directly using them.

3.2.2. Maximal re-use of linguistic resources

Re-engineering this single stage for a new accent requires no modification of the linguistic resources used by the system, merely the processing of a new voice. The recording subject is given a script of "real words" and hence automatically provides the appropriate realisation of each keyvowel in the given accent.

4 KEYVOWEL-BASED DICTIONARY

A draft keyvowel dictionary has been produced, with 47781 entries. Each entry in this dictionary has three parts: index number, (lower-case) orthographic form), and pronunciation string. The vowel symbols in the pronunciation string represent keyvowels rather than actual phonemes of any particular accent.

4.1. Raw materials

Wells ([1]) defines each keyword in terms of words having vowel phoneme *x* in RP and vowel phoneme *y* in General American (GenAm). Therefore it is necessary to compare the pronunciations of words in both an RP lexicon and a GenAm lexicon in order to classify each entry in terms of keyword. The machine-readable CMU pronouncing dictionary of American English was used as the source for GenAm, while the BEEP pronouncing dictionary was used as the source for RP. The CMU dictionary is available on the World Wide Web at http://www.speech.cs.cmu.edu/cgi-bin/cmudict, and the (compressed) BEEP dictionary is available from ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries

4.2. Preparation

4.2.1. Wordlist preparation and initial pronunciations

From the two original dictionaries, a list was derived of all (lower-cased) orthographic strings common to both, making allowance for the correspondence of the 25 CMU final "-or" words with BEEP "-our" words. The resulting list of 47781 words was used to derive the pronunciation strings from each dictionary. Subsequent processing focussed on removing errors from this material, and on preparing it for use in deriving a keyvowel dictionary.

4.2.2. Processing of pronunciation strings

Several errors in the pronunciation strings required correction. Primary stress was missing in many words, while others had more than one primary stress. In the BEEP dictionary, most instances of secondary stress preceded primary stress, but in 329 cases the order was reversed and required correction. It was decided to edit the stress order in the CMU dictionary to adapt it to the BEEP order. Secondary stress is held to differ from tertiary stress in that only the former may receive primary stress under conditions of backshifting of stress: tertiary-stressed vowels remain unreduced but never receive primary stress.. The processing situation is summarised in Table 1 for pronunciations derived from the BEEP dictionary, and in Table 2 for CMU entries.

Type of case	Number	Editing method
No primary stress: monosyllabic words	7668	Automatic
No primary stress: polysyllabic words	5453	Manual
More than one primary stress	183	Manual
Secondary stress follows primary	329	Automatic

Table 1: Processing of BEEP pronunciations.

Type of case	Number	Editing method
No primary stress	20	Manual
More than one primary stress	391	Manual
Secondary stress follows primary	7211	Automatic

Table 2: Processing of CMU pronunciations.

A syllabification routine was written and applied. Syllable-dependent errors were corrected, as follows.

Schwa vowels were inserted manually to avoid syllabic consonants (eg. in *little, cotton*) in 4068 BEEP entries.

In the case of 1275 BEEP entries, the centring diphthongs /ɪə/, /uə/ were manually altered to disyllabic /ɪ . ə/, /ʊ . ə/, where these corresponded to two underlying syllables (as reflected in the CMU forms).

It was found that 590 BEEP forms displayed a postvocalic /r/, which was deleted automatically since RP is non-rhotic. On the other hand, 9673 CMU entries showed a missing postvocalic /r/ after the vowel symbol "er0", and these were added automatically.

Finally, 423 cases of /ɔ/ in the BEEP strings (shown by the symbol "ao") were edited by hand to the new symbol "oa". These cases corresponded to the FORCE vowel, as determined by the list in [1] (and derivatives of those words). These words are not distinguished in BEEP but must be differentiated in some other accents.

4.2.3. Harmonising segment numbers

The final preparatory stage ensured that, for each entry, the BEEP and CMU pronunciations contained the same number of segments, disregarding the systematic variation of postvocalic /ɹ/ (missing in BEEP) and postalveolar stressed /j/) (missing in CMU). This was needed for the automatic derivation of keyvowel forms. The harmonising of segment numbers entailed hand-editing 2128 pronunciation strings.

In some accents, words such as *perpetuate, appreciable*, are pronounced with the /tjʊ/ or /sɪ/ suggested by the orthography. In most accents these segments undergo "Yod Coalescence" ([1]: 3.3.3) to become /tʃʊ/ and /ʃɪ/ respectively. Since many entries in the dictionaries were shown with Yod Coalescence, some editing to the BEEP entries was necessary to restore underlying /tj/ and /s/, to allow for the accents that retain them. These cases are included in the 2128 cases referred to above.

4.3. Output dictionary

4.3.1. Symbol pairing

A program was written that read the BEEP and CMU pronunciation string for each entry, and output a bipartite symbol for each segment. The output symbol consisted of the BEEP symbol, followed by a colon, followed by the CMU symbol. In the case of *systematic* variation (postvocalic /ɹ/ and postalveolar stressed /j/) dummy symbols were used where necessary. It was this symbol pairing program that necessitated the preceding harmonisation of segment numbers.

4.3.2. Keyvowel strings

Rules were written and executed that inspected each bipartite symbol and output the appropriate "keyvowel"-level symbol. Consonant output symbols included two special symbols: "rr" for postvocalic /ɹ/, and "yy" for postalveolar /j/. Vowel output symbols were based on the keyword classes in [1]. For example, input bipartite symbol "oh1:aa1" (i.e. primary-stressed RP /ɒ/ and GenAm /ɑ/) became output symbol "oh1" (the LOT vowel, primary-stressed), while input symbol "oh1:ao1" (primary-stressed RP /ɒ/ and GenAm /ɔ/) was output as "aoo1" (the CLOTH vowel, primary-stressed). The output strings, when added to the corresponding orthographic strings, formed the keyvowel dictionary.

5. USE OF THE KEYVOWEL DICTIONARY

The keyvowel dictionary forms the vital resource for subsequent accent-specific linguistic resources that can be used in a text-to-speech synthesiser, as follows.

5.1. Accent-specific speech database texts

When developing a new accent for a speech synthesiser using concatenative synthesis, it is necessary to derive a text that characterises the words contained in the recorded speech database. Conventionally, such a text is linked with the pronunciation in phonemic form. Using the keyvowel dictionary, however, this text will be indexed with the keyvowel form of each word, thus allowing direct access to the appropriate speech unit on the part of the system developer.

In addition, in the case where the recording subject reads isolated words from a script, the developer needs only to extract those blocks of words where the corresponding keyvowel is distinctive in the given accent. For example, in RP, CLOTH words contain the same vowel as LOT words, and so only one of these two sets of words needs to be recorded for that accent. This will save on development time.

5.2. During synthesis

During the process of synthesis, the system will access the speech units in terms of the keyvowels by which they are coded, rather than by particular phonemes. This allows for a direct data pathway between dictionary and speech unit, as explained in 3.2.1 above.

5.3. Future work

Future application of the keyvowel dictionary will probably begin with the development of a text-to-speech synthesiser for Scottish English. It is hoped to develop synthesis in several different accents of English.

7. REFERENCES

[1] J.C. Wells, "Accents of English" (3 volumes). Cambridge University Press, Cambridge, 1982.

[2] R. Matthews, "Master of the Universe", *"CAM: University of Cambridge Alumni Magazine"*, Autumn Term 1995.