REPRESENTING THE ENVIRONMENTS FOR PHONOLOGICAL PROCESSES IN AN ACCENT-INDEPENDENT LEXICON FOR SYNTHESIS OF ENGLISH

Susan Fitt and Steve Isard

e-mail sue@cstr.ed.ac.uk, stepheni@cstr.ed.ac.uk
Centre for Speech Technology Research
University of Edinburgh
80 South Bridge
Edinburgh
UK

ABSTRACT

This paper reports on work developing an accent-independent lexicon for use in synthesising speech in English. Lexica which use phonemic transcriptions are only suitable for one accent, and developing a lexicon for a new accent is a long and laborious process. Potential solutions to this problem include the use of conversion rules to generate lexica of regional pronunciations from standard accents [1] and encoding of regional variation by means of keywords [2]. The latter proposal forms the basis of the current work.

However, even if we use a keyword system for lexical transcription there are a number of remaining theoretical and methodological problems if we are to synthesise and recognise accents to a high degree of accuracy; these problems are discussed in the following paper.

1. KEYWORD TRANSCRIPTION

Pronunciation lexicons for use in speech synthesis and recognition are readily available for General American and RP, but as use of speech technology grows more accents will be required. Developing lexicons for different accents of English is a long and potentially expensive process. Although conversion rules can be produced for semi-automatic accent generation [1], hand-checking is still required and the rules have to be rewritten for each new accent. A different solution is described in [2], based on Well' s keyword system [3]. Wells describes the vowels occurring in different accents in terms of keywords, so rather than saying that ' pool' contains the phoneme /u/ in RP and /ʉ/ in Scottish accents, he simply says that the word contains the GOOSE vowel.

1.1. Key-vowels and Key-consonants

Using Wells's keywords as the basis for producing a pronunciation lexicon, we might specify the symbol ' uu' as describing the GOOSE vowel, and transcribe ' pool' and other such words with this symbol; the ' uu' would be realised differently for Scottish English and for RP. Vowels are the main source of variation for British accents, and are the only sounds covered by Wells's keywords. However, as noted in [2], there is also variation amongst consonants, particularly post-vocalic /r/ (as in 'horse', RP versus Scottish English) and postalveolar /j/ (as in 'news', RP versus General American).

There are some additional consonants which are not covered in [2]. These are only used in a limited geographical area, such as /ɬ/ which is used in Wales and /x/ which is used in Scotland and Ireland. Both of these are used mainly for local words or names, such as /ɬ/ in the Welsh name 'Llewelyn', or /x/ in the Scottish word 'loch'. To some extent their use is predictable from the spelling and the pronunciation in other accents, but in a keyword system it is simpler to use key-consonants for these and encode them in the lexicon rather than try to produce them by rule.

1.2. Inclusion of Features in the Lexicon versus Derivation

At some point we need to draw the line between what we include in the lexicon and what we derive by rule, and to decide what, if anything, should be handled by exception lists. This topic will form the major part of this paper.

For keyword synthesis it might be assumed that all necessary information be marked by different symbols in the lexicon, but as we shall see there are details of pronunciation that an accent-independent lexicon needs to cope with, which would be better handled by rule. As a starting point, we might suggest that all phonological variation, such as the use of /ɬ/ versus /l/ in Welsh, be included in the lexicon, while all phonetic or allophonic variation, such as the use of dark and light /l/ in different accents, should be handled by accent-specific rules. Alternations which occur in only one or two words could be treated as exceptions.

2. PHONEMIC VARIATION ACROSS ACCENTS

Much of the variation of phonemic status in English accents can be covered by use of keyword symbols. As noted above, the primary source of variation is in the vowels, and use of keyword symbols for consonants covers still more regional variation. However, there are some problems in deciding

exactly what constitutes phonemic variation, which must be resolved if we are to produce accurate and consistent transcriptions for each accent. Furthermore, encoding all phonemic differences for various accents in the lexicon can lead to great complexity.

2.1. Phonemic versus Allophonic Variation

For most accents and most phones it is a simple matter to decide whether the difference between two phones is phonemic or allophonic. Phonemic differences are those represented by minimal pairs, such as 'hat' and 'hot' in most accents of English, or those with phonetically distinct sounds in complementary distribution, such as /h/ and /ŋ/. Allophonic differences are in complementary distribution but have phonetic similarity, such as the light [l] in RP 'look' and the dark [ɫ] in RP 'cool'. Typically speakers of the accent do not classify allophones of a particular phoneme as different sounds.

However, for some accents there are sounds whose phonemic status is borderline, and these pose a theoretical problem of classification. One example is long vowels in Scottish English. They are generally described in the literature as morphologically and phonologically conditioned, so while 'mood' [mʉd] and 'mooed' [mʉːd] may form an apparent minimal pair in terms of the phone string, the difference between the two is actually determined by the morphological structure, with the vowel in 'mooed' preceding a morpheme boundary. This would suggest that the two sounds are allophones, and if morphological information is included in the lexicon we can predict where long vowels will occur. However, for some speakers there are words such as 'leak' and 'leek' [3] which are minimal pairs but are not environmentally conditioned, suggesting that for these speakers the difference is phonemic. To make the situation still more complex, such speakers do not always agree on which words contain a difference of length.

One possible answer to such a dilemma is to record the accent of the majority of speakers, or of younger speakers if it is thought that the accent is in the process of change. Another solution, given that it appears to be impossible to encode the speech of all individuals in a keyword lexicon, is to aim for the simplest transcriptions; in the case of Scottish long vowels, this might mean ignoring the long/short distinction where it is not used by all speakers, and if this leads to non-phonemic status for the long vowels, they can be derived by rule. This does mean, though, that all the necessary environmental information must be contained in the dictionary. So, a word such as 'mooed', which contains two morphemes, must contain a morpheme boundary in the lexicon even though it is monosyllabic, for example (with + representing a morpheme boundary):

Word	**Keyword symbols**
mood	m * uu d
mooed	m * uu + d

2.2. Full and Reduced Vowels

An example of a particular problem in keyword synthesis is the use of the reduced vowels (schwa or /ɪ/), which varies greatly across accents. If the lexicon is to cover as many accents of English as possible, with a high degree of accuracy, this must be taken into account in the lexical transcriptions. For example (see [1]), Leeds English has full vowels in certain prefixes, with 'en' in 'entreat', 'envisage' and so on pronounced as /ɛn/ rather than /in/ as in RP. Cardiff English also uses full vowels in final closed syllables, with the pronunciation /ˈɛnd.lɛs/ for 'endless', rather than the schwa or /ɪ/ used in most other accents of English. Note that although these ' reduced vowels' can occur as variants of full vowels, in these examples their use in accents such as RP is obligatory; ' endless' iɛ́nd.ləs/ or /ˈɛnd.lɪs/ and the pronunciation /ˈɛnd.lɛs/ does not occur.

Since the reduced forms are more common, and the full forms can generally be derived by reference to the spelling, it is tempting to produce some of these variations by rule rather than hardwire them into the lexicon, although this would violate the principle that phonemic differences are encoded in the lexicon. Also, given the complexity of English spelling, accurate rules to determine the pronunciation from the spelling are not simple. This means that for every accent that contains full vowels where others use reduced vowels, extra keyword symbols must be created, giving us, for the examples in the accents above, three distinctions for /ɛ/-type vowels:

Word	**Keyword symbols**
entreat	E1 n . t r * ii t
endless	* E0 n d . l E2 s

One way of encoding such variation while maintaining a fairly readable lexicon is, as above, by use of numbers combined with the basic key-vowels. This type of encoding should allow us to use equivalent numbers for equivalent processes, for example with 0 always indicating an unreduced vowel, 1, 2 and 3 referring to reduced vowels in different accents, 4 being a vowel which is deleted or deletable in certain accents (as in 'secretary'), and so on.

3. PHONOLOGICAL PROCESSES

One area of difficulty in designing a system which contains all necessary information is caused by phonological processes, and some of these are discussed below.

3.1. Accent-Specific Allophones

One example of an accent-specific allophone is the flapping of /t/ in American English, in words such as 'city'; another is the realisation of /t/ as a glottal stop in many British accents. For instance, many British accents use a glottal stop in word-final position following a vowel, for example 'hot' [hɒʔ]. Others, such as Cockney, also use a glottal stop word-medially before an unstressed vowel, as in 'hotter', while in Edinburgh it may also occur before a stressed vowel [3]. We can deal with such pronunciations by:

i. Use of keyword symbols in the base lexicon to represent the different realisations of /t/. However, use of keyword symbols to represent allophones is inefficient, and this option is unrealistic for processes such as glottalisation of /t/, since different accents do this in different environments. Recording all the potential outcomes with different symbols in the lexicon unnecessarily increases its complexity, and includes information which is easily stated by rule, given appropriate keyword symbols, stress and syllabification.

ii. Use of keyword symbols for representation of the basic phonemes in the lexicon, with output phones chosen by the synthesiser. This alternative increases complexity in the synthesiser itself, since the synthesiser must now contain phonological rules for the different accents.

iii. Use of a meta-lexicon representing the phonemes, and compiled sub-lexica containing the output phones to be used. This option uses the same phonological rules as the second, but introduces an intermediate level of description; this may be advantageous for some applications.

Whether or not sub-lexica containing phone strings are explicitly generated, it should be noted that for concatenative synthesis, with segments recorded by a speaker, the allophonic variation must be taken into account when designing word sets for recording new accents. It would not be sufficient for the speaker to record the set of keywords as listed in Wells, even if this were extended to show all the key-consonants; instead we would need a set of keywords which included all the allophones of the accent. Where these are morphologically conditioned, as in Scottish long vowels, the allophones do not automatically fall out from producing all possible segment combinations.

3.2. Style and Speaking Rate

There remains the question of how much detail would be included in accent-specific rules. If we wish to include pronunciations for different styles and speaking rates, there are still more options. Many of these overlap with accent-specific processes, for example glottalisation of /t/ is more common in casual speaking styles than in formal speech, and elision of segments, such as schwa in 'secretary', is more common in fast than in slow speech. Such variants are evidently useful for speech recognition, and also for some applications in speech synthesis which require particularly slow or fast speech.

It should be noted that while there is a considerable amount of research on the effects of speaking rate or styles on certain features in various accents (for example, Reid [4] and Romaine [5] for Edinburgh English), there is no comprehensive study of all the pronunciation variants occurring in different styles or speaking rates for any one accent, let alone the many different accents covered by a keyword dictionary. This practical difficulty suggests that for the moment, only well-attested phenomena such as glottalisation or flapping should be included in a rule set; it also suggests that as much morphological and other relevant information as possible should be included in the dictionary to facilitate rule-development at a later` stage.

Certain common alternations cause difficulty if we are to provide naturalistic pronunciations. One of the most prevalent is optional vowel reduction (as opposed to the use of reduced vowels as phonemes, discussed above). Examples of this in RP are:

Word	**Phone string**
autocrat	[ˈɔ.tə.kɹat] or [ˈɔ.tou.kɹat]
ovation	[əʊˈvɛɪ.ʃn] or [əˈvɛɪ.ʃn]

It is obviously desirable to record the most prevalent pronunciation in the base lexicon, but for some words the reduced vowel is more common, while for others the unreduced one is more widespread. While we can transcribe a full vowel in keyword symbols, and then allow phonological processes to reduce unstressed vowel phonemes to a schwa in fast speech, it is more complex to write rules for especially careful pronunciation which would transform reduced to full vowels. If we are creating naturalistic lexica, we would not wish to record only full forms.

3.3. Cross-word Phenomena

While the sub-lexica approach may be appropriate in certain circumstances, there are also cross-word phenomena which must be handled at a different stage of processing, necessitating the inclusion of phonological rules in either the synthesiser or the recogniser. One of these is the use of word-final /r/ in non-rhotic accents. Williams and Isard [2] use a specific symbol, 'rr', for rhotic /r/ in words such as 'card' or 'car'. For a non-rhotic accent such as RP, the 'rr' would be automatically be converted to a null phone in a word such as 'card', where it is followed by a consonant. However, for a final 'rr' in a word such as 'car', we need to know whether this is followed by a vowel, a consonant or a pause before it can be converted to either a null phone or [ɹ]. This information obviously cannot be contained in accent-specific sub-lexica, which deal only with single words. It is thus apparent that some phonological rules must be contained in the synthesiser or recogniser.

4. LEXICAL EXCEPTIONS

There remain some words which have to be treated as exceptions if we are to produce accurate pronunciations. 'Tomato', for example, can only be dealt with realistically by treating it as an exception in either British English or American (RP /təˈmɑ.təʊ/ versus General American /təˈmeɪ.toʊ/). The /ɑ/-/eɪ/ pairing in these accents only occurs

in this word, so it would not be profitable to set up a keyword vowel for this one case. If all such exceptions were encoded with key-vowels, this would vastly increase the complexity of the lexicon. If all keywords had to be recorded each time a new accent was synthesised, this procedure would also be made more time-consuming by the addition of more key-vowels.

5. STRESS AND SYLLABIFICATION

As well as the segments themselves, stress and syllabification vary across accents and must be considered in the production of an accent-independent lexicon.

5.1. Stress Variation

Some stress variation across accents is random, such as 'ballet' in British English (/ˈba.leɪ/) and American English (/baˈleɪ/). However, there are some cases for which a number of words follow the same pattern, for example 'mutate', 'frustrate' and so on, with primary stress on the first syllable in American English and on the second in British English. It may be worthwhile to extend the notion of keyword vowels and have keyword stress, rather than deal with these words by rule or list them as exceptions. Another example is secondary stress in the two accents, in words such as 'secondary'. This word forms a particularly complex example, as it typically has four syllables in American English and three in British English:

Accent	**Phoneme string**
American English:	/ˈsɛ.kənˌdɛ.ɹi/
British English:	/ˈsɛ.kən.dɹi/

While a simple solution would be to transcribe the four syllables for British English, giving /ˈsɛ.kən.də.ɹi/, this is a rather stilted pronunciation and so is not ideal if we are aiming for naturalistic speech synthesis.

5.2. Syllabification

Syllabification and morphological boundaries are important for the keyword dictionary as they condition some of the phonological processes which apply to the base pronunciations. However, it appears that syllabification can vary across accents. For example, Wells [2, Vol. 3 p. 537] claims that the /s/ in the Canadian pronunciation of 'bicycle' must belong with the first syllable, but for Southern American English it belongs with the second syllable; this can be inferred from the allophones which occur in this word. For Canadian English the syllable must be closed for raising to occur, while in Southern American a diphthong occurs in a closed syllable and a monophthong in an open syllable. This means that phonetic transcriptions in the two accents must be as follows:

Accent	**Phone string**
Canadian English:	[ˈbəɪs.ɪ.kl]
Southern American English:	[ˈba.sɪ.kl]

It is not clear at present how widespread such cases are; if they occur in a minority of words they can be given a syllabification which suits most accents, and treated as exceptions elsewhere, but if they are more prevalent this issue may need to be re-examined.

6. CONCLUSIONS

There are a number of theoretical and practical issues to be resolved in keyword synthesis. It is proposed that phonemic variation within accents be encoded in the lexicon by use of keyword symbols, while allophonic differences be derived by rule. Morphological information needs to be included in the lexicon as this forms the environment for some allophones. It has been noted, though, that there are sometimes difficulties in determining phonemic or allophonic status and that sometimes the solution should be chosen on practical rather than theoretical grounds.

If we wish to include some variation according to style or speaking rate this makes the lexicon more complex, and at present our knowledge of the linguistic processes occurring in different accents is somewhat limited. Furthermore, even in keyword synthesis exception lists cannot be avoided. Despite these reservations, it is hoped that much regional variation can be covered by keyword lexica.

7. ACKNOWLEDGEMENTS

We are pleased to acknowledge the support of the UK Engineering and Physical Sciences Research Council through grant EPSRC GR/L53250.

8. REFERENCES

1. Fitt, Susan. "The generation of regional pronunciations of English for speech synthesis." *Proceedings: Eurospeech 97*, Vol. 5, pp. 2447-50. Patras. 1997.

2. Williams, Briony J., and Isard, Stephen. "A keyvowel approach to the synthesis of regional accents of English." *Proceedings: Eurospeech 97*, Vol. 5, pp.2435-8. Patras. 1997.

3. Wells, John C. (1982). *Accents of English.* Cambridge: Cambridge University Press.

4. Reid, Euan (1978). "Social and stylistic variation in the speech of children: some evidence from Edinburgh." In: Peter Trudgill (ed.), *Sociolinguistic patterns in British English*, pp. 158-71. London: Edward Arnold.

5. Romaine, Suzanne (1978). "Postvocalic /r/ in Scottish English: sound change in progress?" In: Peter Trudgill (ed.), *Sociolinguistic patterns in British English*, pp. 144-57. London: Edward Arnold.