# THE TREATMENT OF VOWELS PRECEDING 'R' IN A KEYWORD LEXICON OF ENGLISH

Susan Fitt
*Centre for Speech Technology Research, University of Edinburgh*

## ABSTRACT

Work is progressing on a keyword lexicon aimed at enabling the synthesis of various regional accents of English. This paper focuses on a particular issue, that of vowels before orthographic 'r'. These vowels are discussed with respect to rhotic and non-rhotic accents, in terms of both keyword sets and phonetic realisation. Criteria for the use of keysymbols are discussed, and it is noted that these criteria result in inclusion of post-vocalic |r| in the lexicon, with deletion by rule for non-rhotic accents. It is noted that some keyvowels in our original set have had to be split, while others may prove to be redundant.

## 1. THE KEYWORD LEXICON

As described in a previous paper [1], the keyword lexicon, rather than using conventional phonetic or phonemic symbols or their ASCII equivalents, uses transcriptions based on keywords. This strategy enables one lexicon to represent numerous regional accents. Where an accent has a phonemic distinction, this will be represented in the dictionary; for instance, 'horse' and 'hoarse', although homophones in RP, are distinguished in Scottish English and so must be represented by different keysymbols in the lexicon. These can be described as the NORTH vowel and the FORCE vowel [2].

## 2. DICTIONARY TREATMENT OF 'R'

### 2.1. Terminology

In this paper I shall use the terms 'rhotic' and 'non-rhotic' to refer to accent types.

Although somewhat unsatisfactory, the traditional term 'post-vocalic' is used to describe /r/ in both pre-consonantal and word-final environments:

$$V\ r\ /\_ \quad \left\{ \begin{array}{l} C \\ \# \end{array} \right.$$

In these environments /r/ is consistently pronounced in rhotic accents; in non-rhotic accents /r/ does not exist in pre-consonantal position, and is variable in word-final position depending on both the following word and the regional accent.

'Word-internal pre-vocalic' is used for the environment:

$$r\ /\_\ V$$

In this context /r/ is pronounced whatever the accent type.

### 2.2. Accent-independent Transcription of 'r'

The keyword lexicon covers both rhotic and non-rhotic accents. Initially it was planned to use a special symbol for post-vocalic 'r' as opposed to word-internal pre-vocalic 'r' (c.f. [3], which uses |rr| versus |r|). However, as work has progressed it has proved undesirable to include all regional differences as separate keysymbols in the lexicon, as this results in numerous keysymbols to encode differences which are in many cases predictable. Given that post-lexical rules are in any case necessary for cross-word environmental conditioning, and are useful for within-word allophones such as glottal stops, which have different scopes of application in different accents, we need to decide on the balance between lexical encoding and rule-based derivation (see [1]).

The criteria which have been drawn up are based on the traditional distinction between phonemes, which we need to include in the lexicon, and phones, which are predictable. These criteria are:

> Principle I. For segments,[i] *all phonemes in each accent*, and *only units which have phonemic status in at least one accent*, should be encoded differently in the base lexicon.
>
> Principle II. *If the phonetic realisation of a unit is predictable from the environment* (which includes keysymbols, syllable and morpheme boundaries) then this will be derived by accent-dependent post-lexical processing.

In the case of a conflict, Principle II overrides Principle I as it reduces redundancy in the transcriptions. Following these criteria, we must encode the vowel difference between Scottish 'horse' and 'hoarse' in the base lexica, although many accents of English do not make this distinction; on the other hand, dark and light /l/, which are never contrastive,[ii] will be generated post-lexically.

Principles I and II give us a logical framework for lexical encoding of linguistic information versus derivation by rule. Following the criteria, we must reject the use of separate keysymbols for post-vocalic and word-internal pre-vocalic 'r'. Given the keysymbol transcriptions

| | |
|---|---|
| farm | f * ar r m |
| safari | s @ . f * ar . r iy |

we can predict that the |r| symbol in 'farm' will be realised as [r] (or [ɹ], [ɾ], and so on) in rhotic accents, and as null in non-rhotic accents. On the other hand, the |r| in 'safari' will be realised as an [r] in all accents.

We no longer have the simple mapping of the earlier split-symbol approach, in which |r| was realised as [r] in all accents, and word-internal |rr| was realised as [r] in rhotic accents and null in non-rhotic accents; the balance of description has shifted slightly from the accent-independent lexicon to accent-dependent rules.[iii] However, it should be noted that in the split-symbol approach post-lexical rules are still necessary to predict the realisation of word-final 'r' in non-rhotic accents which use linking [r]. For the word

far f * ar r

we need to know what follows before we can predict the pronunciation of |r| in non-rhotic accents such as RP. (Note that in some accents of English, for instance South African English, linking 'r' is not pronounced, so for these accents this does not apply.) Intrusive 'r', although used to differing degrees by many people, is still regarded as erroneous (though see [4] and related references), and so it is not included in the keyword transcriptions or post-lexical rules; it could, though, be introduced by rule if desired (see [5]).

Although Principles I and II above are based on the traditional phonological distinction between phonemes and allophones, the use of these criteria in a multi-accent lexicon means that the resulting output does not correspond to the traditional division between the two. Keysymbols in the lexicon are not directly equivalent to phonemes, and the output of the post-lexical rules cannot be defined in terms of allophones.

| | RP, phonemic transcription[iv] | RP, phonetic transcription |
|---|---|---|
| **far** | /fɑːr/ | [fɑː] |
| farm | /fɑːm/ | [fɑː] |
| safari | /səˈfɑ.ri/ | [səˈfɑː.ɹi] |

Table 1: Phonemic and phonetic transcriptions of RP (isolated words).

| | Keysymbol transcription | RP, after post-lexical rules |
|---|---|---|
| **far** | \|f * ar r\| | \|f * ar\| |
| **farm** | \|f * ar r m\| | \|f * ar m\| |
| safari | \|s @ . f * ar . r iy\| | \|s @ . f * ar . r iy\| |

Table 2: Keysymbol transcriptions (isolated words).

In the phonemic/phonetic transcriptions shown in Table 1, only 'far' exhibits loss of /r/ in the conversion from phonemes to allophones, whereas in the keysymbol transcriptions in Table 2, both 'far' and 'farm' have |r| in the lexicon but lose this during the application of post-lexical rules.

## 3. VOWELS BEFORE POST-VOCALIC |r|

The vowels I will focus on in this paper are post-vocalic, i.e. they precede |r| plus a consonant, or word-final |r|. Representation of these vowels is complex in an accent-independent lexicon as the loss of |r| in non-rhotic accents is usually accompanied by a change in the quality of the preceding vowel, whereas rhotic accents typically allow the same vowel-set before post-vocalic [r] as before other consonants. For example, Edinburgh English, which is rhotic, has [i] in both 'knees' and 'near', whereas in RP we have [iː] in 'knees' but [ɪə] in 'near'. Vowels before word-internal pre-vocalic |r| do not have the same restrictions; in this position in RP, for instance, we can have [iː], as in 'Leroy', [ʊ] as in 'courier', and so on.

### 3.1. Wells's Keywords

The following keywords from Wells [2] represent those vowel-sets which may occur in non-rhotic accents before a lost |r|: NURSE, NEAR, SQUARE, START, NORTH, FORCE, CURE, and LETTER. These vowel sets are only used before either post-vocalic or intervocalic |r|: 'mar', for example, is |m * ar r|, and 'marring' is |m * ar r . i ng|, while 'ma' belongs with the PALM keyword and is transcribed as |m * aa|.

A further set of vowels may occur either before |r| or elsewhere: PRICE, CHOICE, and MOUTH. In non-rhotic accents these are generally followed by a glide when preceding |r|, for example 'out' |* ow t| is realised in RP as [aʊt], while 'hour' |* ow r| is realised as [aʊə]. Sometimes, though, in both environments these vowels are realised as monophthongs; for more discussion of this and other processes see [2].

The vowels which do not occur before a lost |r| are the short vowels such as [ɪ], which cannot occur in open syllables, and the long close monophthongs [iː] and [uː]. The keyvowels of FACE and GOAT, which in some accents are monophthongs and in others are diphthongs, also do not occur in this position.

### 3.2. Developing the Keysymbols

Focus accents, including RP, General American and some regional accents of Britain and the US have been used as an initial testing-ground for the accent-independent pronunciation lexicon. The original keysymbol set was based on Wells's keywords, as in the above examples, with the addition of a consonant set such as the PEA consonant and the LOCH consonant. It was discovered that Wells's keywords needed expansion to cover the accents involved while adhering to Principles I and II in Section 2.2 above. Wells does note that some accents have failed to undergo certain mergers, or have developed splits in his lexical sets; for a pronunciation dictionary, it is necessary to develop the keysymbol set to take account of this. For example, the PRICE vowel must be split to cover the Scottish phonemic opposition 'tied'/'tide'. This section discusses each of the relevant keywords in turn.

**3.2.1. Nurse.** The [ɜː] vowel does occur in RP in environments which do not precede |r|. The word-set is small and involves borrowed words, such as 'Goethe' or 'chartreuse'. The corresponding American vowel is variable, being [o] in the former and [u] in the latter, so these words are treated as exceptions.

Likewise, General American [ɜː] does not always correspond to RP [ɜː]; some words with word-internal pre-vocalic |r|, such as 'hurry', have [ʌ] in RP, while others such as 'squirrel', have [ɪ]. The 'hurry' set can be treated by rule, since all occurrences of |@@r r| (e.g. 'nurse', 'fur', 'furry') are realised as [ɜː(ɹ)] in both General American and RP, while |uh r| is realised as [ɜːɹ] in General American and [ʌɹ] in RP ('hurry'). Thus, |@@r r| and |uh r| are equivalent in General American but differentiated in RP:

| | |
|---|---|
| Gen. Am. | \|@@r r\|, \|uh r\| → [ɜːɹ] |
| RP | \|@@r r\| → [ɜːɹ] |
| RP | \|uh r\| → [ʌɹ] |

Alternatively, we could use only |uh r| for both 'hurry'/'furry' and recognise a pre-consonantal and morpheme-final conditioning environment in RP:

$$
\begin{array}{lll}
\text{RP} & |\text{uh r}| \rightarrow [\text{ɜːɹ}]\ /\_ & \begin{cases} \text{C} \\ + \end{cases} \\
 & \rightarrow [\text{ʌɹ}]\ /\_ & \text{elsewhere} \\
\text{Gen. Am.} & |\text{uh r}| \rightarrow [\text{ɜːɹ}] &
\end{array}
$$

Principle II above makes the second solution preferential, as the different realisations of |uh r| in RP are predictable from phonetic and morphological environment and so need not be included in the lexicon; on the other hand we would then have a transcription with a short vowel preceding post-vocalic |r|, which complicates the phonotactic specification of the lexicon. At present the first solution is followed.

The 'squirrel'-type examples cannot be treated by rule (c.f. 'Cyril', which has [ɪ] in both accents), but in the current dictionary only 'squirrel', 'stirrup' and their derivatives have the [ɜː]/[ɪ] alternation and so these can be listed as exceptions.

There are, however, more substantial divisions within this keyvowel in Scottish accents. The most common split is between 'word' ([ʌ]) and 'heard' ([ɛ]), although some accents also have [ɪ], for example in 'bird'. The lexicon currently records only the 'word'/'heard' split but future work may include the 'bird' split.

**3.2.2. Near.** A distinction in vowels is possible between 'cereal', 'Cyril' and 'Leroy', necessitating the keyword NEAR. The quality of the NEAR vowel varies across accents, for example Scottish English uses [i] while American English has [ɪ], but as this is a matter of phonetic realisation it does not affect our transcriptions.

In some accents the realisation of this vowel varies by environment. For example, in Leeds the word 'beer' contains a diphthong while 'beery', followed by a vowel, has a monophthong. However, as this is predictable by environment, we can transcribe them both with the same keyvowel. It should be noted that these allophones of |ir| are conditioned by a non-adjacent segment; for diphone synthesis this feature would have to be specified in post-lexical rules. If longer stretches of speech are sampled this will not be necessary.

The lexicon distinguishes between diphthongs such as 'near' |n * ir r| and sequences such as 'skier' |s k * ii @r r|; not all speakers make this distinction, but encoding it in the dictionary allows us to cater for those who do. Such sequences are listed in Table 3 below as combinations of |@r| and |r|, rather than |ir| and |r|.

**3.2.3. Square.** Not all accents distinguish the SQUARE vowel. In Liverpool, for example, SQUARE and NURSE have the same vowel. On the other hand, many New Zealand speakers merge SQUARE and NEAR ([6]). In General American, of course, there is a possible 'Mary'/'marry'/merry' merger. Mergers are no problem in a keyword lexicon as they are many-to-one correspondences. They need not even be specified by rule, as extraction of phones from recorded words which utilise these keysymbols will automatically produce the right result.

**3.2.4. Start.** The START vowel could be treated as an instance of the PALM vowel. Unlike 'hurry'/'furry', this would not violate phonotactic structure in the lexicon, as PALM is a long vowel. Unlike |@@r|, the START vowel does occur before non-morpheme-final intervocalic |r|, for example in 'safari', so we could not derive a distinction by rule, but as there is no pronunciation difference between the START vowel and the PALM vowel in our focus accents it does not seem to be necessary to use different keysymbols. The two keysymbols are currently retained but may be merged if future work does not show a distinction to be necessary.

**3.2.5. North, Force.** Many accents have merged NORTH and FORCE or are in the process of doing so. However, as some distinguish the two and the distinction cannot be produced by rule they must be recorded in the lexicon. Wells [2] lists words which fall into the two groups; these were checked with a speaker from Edinburgh, who was in broad agreement. There are a number of words in the lexicon which are missing from [2], for example 'abort', 'California', and 'corset'. We are in the process of verifying which keyvowels these words use, and this group is noted separately in Table 3 below.

**3.2.6. Cure.** Like the vowel of NEAR, the realisation of this vowel varies by phonetic environment in some accents. As with NEAR, a distinction is made between diphthongs (e.g. 'cure' |k y * ur r|) and sequences (e.g. 'queuer' |k y * uu @r r|, one who queues).

One problem in the transcription of this word set lies in the change of some CURE words to be pronounced with the vowel of FORCE. For example, 'poor' is often [pɔ] rather than [pʊə] in non-rhotic British accents. Unfortunately the change is not systematic, so, for example, a speaker may pronounce 'sure' with [ɔ] and 'tour' with [ʊə]. Some words more commonly have [ɔ] than others do, and some environments, such as following a [j] as in 'pure', are more likely to retain [ʊə], but there are no hard and fast rules. However, as can be seen from Table 3 there are not large numbers of post-vocalic CURE words, so we can note in exceptions lists CURE words likely to be pronounced as [ɔ], bearing in mind that the list cannot be definitive.

**3.2.7. Letter.** This set describes schwa preceding |r|; like the START vowel, it may prove to be redundant. In Table 3 below words with a simple schwa, such as 'letter', and sequences such as 'skier' |s k * ii @r r|, are listed separately; the latter set includes words such as 'familiar', in which the LETTER vowel may follow [ɪ] or [j] according to accent and speaking style.

**3.2.8. Price.** As noted earlier, this vowel has been split into |ai| and |ae| to allow for the Scottish ‘tied’/’tide’ distinction. However, before post-vocalic |r| and in open syllables, only the |ae| variant occurs. Non-rhotic accents tend to have an offglide between |ae| and post-vocalic |r|, for example 'fire' |f * ae r|

becomes [faɪə] in RP, though as noted earlier this may also be pronounced as a monophthong.

As with NEAR and CURE a distinction is made in the lexicon between simple PRICE diphthongs, as in 'ire', and sequences with PRICE + schwa, such as 'priory'.

**3.2.9. Choice.** The vowel of CHOICE is rare before post-vocalic |r|; in our dictionary it only occurs in ‘coir’. (Words such as 'employer' are treated as sequences of |oi| and schwa.)

**3.2.10. Mouth.** This vowel is also relatively uncommon before post-vocalic |r| and the sequence mostly occurs morpheme-finally, in words such as 'sour'.

## 4. FREQUENCY OF OCCURRENCE

Table 3 gives some indication of the frequency of occurrence of the keyvowels discussed, in a dictionary of 110,000 words; these are an approximation, as the lexicon is gradually being refined. It should be noted that the figures include derived words and so some keyvowels, such as |@r|, have a high frequency due to their use in common morphemes such as '-or' and '-er'. Occurrences before intervocalic |r| are included for interest; many of these are morpheme-final.

| Wells’s Keyword | My keyvowel | Examples before post-vocalic \|r\| | Frequency before post-vocalic \|r\| | Frequency before intervocalic \|r\|, with example |
|---|---|---|---|---|
| NURSE | @@r | fir, nurse | 2482 | (furry 40) |
|  | er | deter, heard | 2353 | (deterring 41) |
| NEAR | ir | near, weird | 538 | (era 570) |
| SQUARE | eir | square, cairn | 761 | (area 666) |
| START | ar | car, start | 3420 | (atari 86) |
| NORTH | or | war, north | 1139 | (warring 1) |
| FORCE | our | wore, force | 1022 | (glory 584) |
| NORTH/ FORCE |  | Timor, abort | 1744 | (abhorring 83) |
| CURE | ur | cure, insured | 196 | (curio 568) |
| LETTER | @r | letter, | 15159 | (gorilla 5454) |
| LETTER | @r | skier, linearly | 987 | (priory 124) |
| PRICE | ai | N/A | N/A | N/A |
|  | ae | fire, tired | 326 | (viral 308) |
| CHOICE | oi | coir | 1 | (moira 6) |
| MOUTH | ow | hour, sour | 99 | (maori 20) |
| other |  |  | 0 | (carry, Cyril 2985) |
| total |  |  | 30227 | 11536 |

Table 3: Frequency of vowels before |r|

The dictionary contains 69,204 instance of |r|. 38,560 of these are word-internal pre-vocalic (with 12,376 intervocalic), and 30,333 post-vocalic (7482 word-final and 22,851 pre-consonantal).

A small number of vowels preceding post-vocalic and intervocalic |r| are not shown in Table 3. A handful of these are exceptions such as 'clerk', which has a different vowel in British and American English. A further set are reducible in some accents, for example 'record' (noun) is [ˈɹɛ.kɔ(ɹ)d] in most British accents, but [ˈɹɛ.kəɹd] in some others, including Scottish and General American.

## 5. CONCLUSIONS

Consistently following the criteria for lexical inclusion (Principles I and II) results in a number of instances of rules replacing of lexical encoding of distinctions. The primary case noted in this paper is that of post-vocalic |r| itself, whose realisation can be predicted by rule. Some of the predictability, such as allophonic variation of NEAR, relies on information from non-adjacent segments, and care must be taken to incorporate this information when synthesising the transcriptions.

It is also noted that some of the original keyword sets, such as START, may be redundant, while others, such as NURSE, have had to be split to accommodate various regional accents.

**ACKNOWLEDGMENTS**

This work is funded by the UK Engineering and Physical Sciences Research Council through grant EPSRC GR/L53250.

## NOTES

i. The use of single keysymbols to encode phoneme sequences, or multiple keysymbols to represent a single phoneme, is currently under investigation; this is necessary for some units which consist of a single phoneme in one accent and multiple phonemes in another, such as the /ɪʊ/ diphthong (Welsh) vs. /ju/ (most other accents).

ii. Some accents distinguish between pairs such as 'holy' and 'wholly'/'holey' on the basis of light/dark /l/, with associated allophonic variation of the /oʊ/ vowel (['həʊ.li] vs. ['hoʊɫ.i]). However, the phones should be derivable from the syllable structure and/or the morpheme boundary.

iii. Rules need not be stated separately for each accent; for many features, we can use accent-groups, such as 'rhotic', 'non-rhotic linking' and 'non-rhotic non-linking'.

iv. This is only one possible analysis; some would propose an underlying /r/ in 'farm'. This has some justification phonologically, historically and psychologically. However, pronunciation lexica generally use a more surface-level phonemic approach such as that described in the text.

## REFERENCES

[1] Fitt, Susan, and Isard, Stephen 1998. Representing the environments for phonological processes in an accent-independent lexicon for synthesis of English. *Proceedings: ICSLP 98.*

[2] Wells, John C. 1982. *Accents of English.* Cambridge: Cambridge University Press.

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

[4] Fox, Anthony 1978. To 'r' is human? Intrusive remarks on a recent controversy. *Journal of the International Phonetic Association,* Vol. 8, pp. 72-4.

[5] Brown, Adam 1988. Linking, intrusive and rhotic /r/ in pronunciation models. *Journal of the International Phonetic Association,* Vol. 18, pp. 144-51.

[6] Maclagan, Margaret A., and Gordon, Elizabeth 1996. Out of the AIR and into the EAR: Another view of the New Zealand diphthong merger. *Language Variation and Change,* Vol. 8, No. 1, pp. 125-47.