SYNTHESIS OF REGIONAL ENGLISH USING A KEYWORD LEXICON

Susan Fitt and Stephen Isard
Centre for Speech Technology Research,
University of Edinburgh, 80 South Bridge, Edinburgh, UK
sue@cstr.ed.ac.uk, stepheni.cstr.ed.ac.uk
http://www.cstr.ed.ac.uk/projects/unisyn.html

ABSTRACT

We discuss the use of an accent-independent keyword lexicon to synthesise speakers with different regional accents. The paper describes the system architecture and the transcription system used in the lexicon, and then focuses on the construction of word-lists for recording speakers. We illustrate by mentioning some of the features of Scottish and Irish English, which we are currently synthesising, and describe how these are captured by keyword synthesis.

Keywords: lexicon, accents, regional pronunciation, synthesis

1. INTRODUCTION

Different accents of English can have different pronunciations for the same word, for example 'bother' is /ˈbɒ.ðə/ in RP but /ˈbɑ.ðəɹ/ in General American. Speech synthesisers that store their lexicons in the form of phonetic transcriptions need separate lexicons for different accents. Rather than using phonetic symbols, our lexicon contains pronunciations transcribed in terms of keywords based on [1]. Abstracting away from the phonetics in this way means that a single lexicon can represent numerous different accents. For example, the vowel in 'bother' is represented by a single symbol occurring in the class of words that are pronounced with /ɒ/ in RP and /ɑ/ in General American. By contrast, 'horse' and 'hoarse', although homophones in RP, are distinguished in Scottish, Irish and some other accents of English and so their vowels must be represented by different keysymbols in the lexicon.

Our method does require the use of some accent-dependent post-lexical rules, as described below (see also [2]) but they constitute a small set for any given accent, and are far less laborious to compile than a new phonetically transcribed lexicon; furthermore, many of these rules apply to several accents.

For purposes of concatenative synthesis, we get a speaker of a given accent to pronounce a word set covering the diphones (or some other concatenative unit) corresponding to the set of keysymbols in the lexicon, and at synthesis time we retrieve the sounds that that speaker produced for the keysymbols.

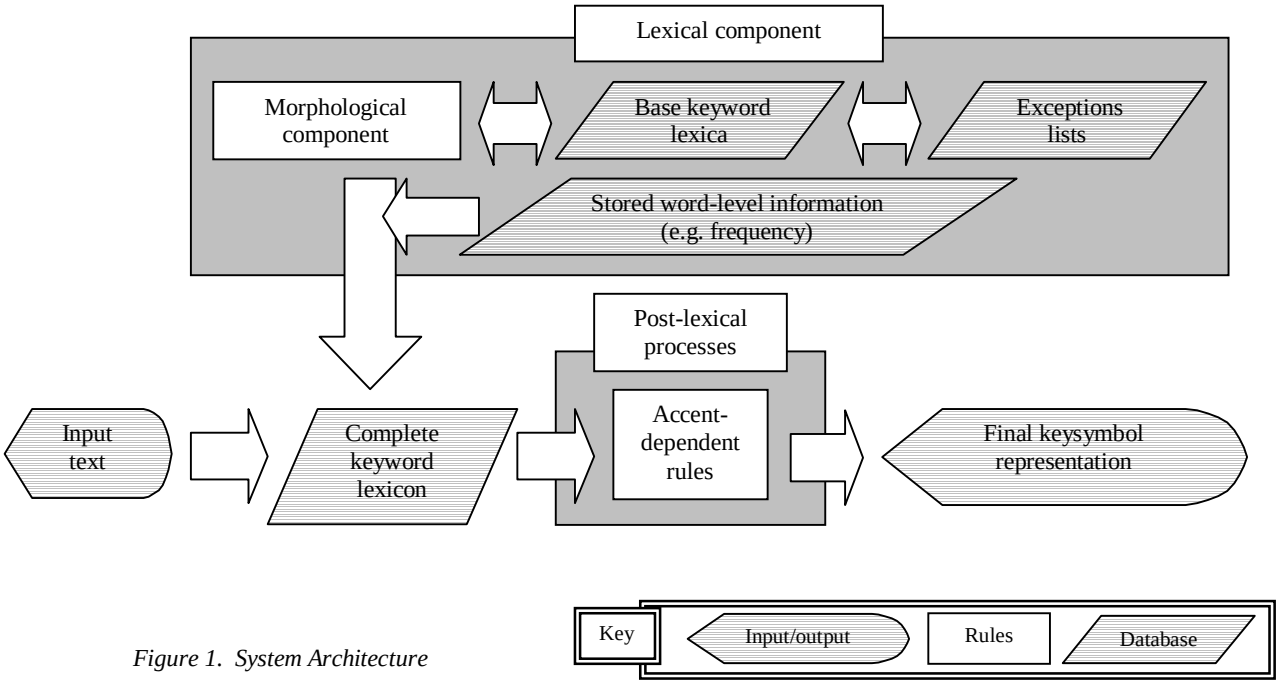

Figure 1. System Architecture

Vowels			
Symbol	**Example**	**Transcription**	**Well's Keyword**
i	tin	t * i n	kit
e	ten	t * e n	dress
a	tan	t * a n	trap
ao	Gandhi	g * ao n . d iy	-
ah	task	t * ah s k	bath
o	top	t * o p	lot
u	took	t * u k	foot
uh	touch	t * uh ch	strut
au	toss	t * au s	cloth
ii	tea	t * ii	fleece
ei	tape	t * ei p	face
aa	ta	t * aa	palm
oo	tall	t * oo l	thought
ou	toe	t * ou	goat
uu	two	t * uu	goose
iu	chew	ch * iu	-
ai	tight	t * ai t	price
ae	tie	t * ae	-
oi	toy	t * oi	choice
ow	town	t * ow n	mouth
i@	idea	ae . d * i@	-
@@r	turn	t * @@r r n	nurse
er	term	t * er r m	-
ir	dear	d * ir r	near
ar	tar	t * ar r	start
eir	dare	d * eir r	square
or	torch	t * or r ch	north
our	torn	t * our r n	force
oour	historic	h i . s t * oour . r i k	-
ur	tour	t * ur r	cure
oou	yogurt	y * oou . g @r r t	-
@r	rotor	r * ou . t @r r	letter
iy	pity	p * i . t iy	happy
@	rota	r * ou . t @	comma

Consonants		
Symbol	**Example**	**Transcription**
p	pea	p * ii
t	tea	t * ii
k	key	k * ii
b	bee	b * ii
d	Dee	d * ii
g	geese	g * ii s
m	me	m * ii
n	knee	n * ii
ng	sing	s * i ng
f	fee	f * ii
th	thief	th * ii f
s	sea	s * ii
sh	she	sh * ii
x	loch	l * o x
v	veal	v * ii l
dh	thee	dh * ii
z	zeal	z * ii l
zh	gite	zh * ii t
ch	cheese	ch * ii z
jh	gee	jh * ii
r	reed	r * ii d
y	yeast	y * ii s t
w	we	w * ii
hw	wheel	hw * ii l
l	lea	l * ii
ll	Llewelyn	ll @ . w * e . l i n
h	he	h * ii

Table 1: Basic keysymbols used in transcriptions

We are now testing the system on various accents by synthesising real speakers from different regions. This enables first-hand checking of the validity of the output pronunciations; due to a lack of comprehensive and modern dictionaries of regional pronunciation, it is not possible to do such checking using the literature.

2. OVERVIEW OF LEXICON

There are a number of levels in the architecture of the system, as illustrated in *Figure 1*. The lexical component produces off-line a keyword lexicon which consists of a word-list with keyword transcriptions and other information such as frequency. During synthesis, input text goes to the keyword lexicon for lookup, and a keyword transcription is output. This keyword lexicon is accent-independent, apart from isolated exceptions such as 'imp<u>o</u>rtant', which is pronounced in Scotland with |our| and in some US accents with |or|. The keysymbol output, whether words or sentences, then goes through the post-lexical processes, which are accent-dependent. These post-lexical processes are similar to traditional allophonic rules, and deal with such features as |t|/|d| tapping in US English, or r-linking in RP English.

2.1 Transcription System

The number of keysymbols and their use is not fixed, as ongoing investigation of more accents suggests new divisions or realignments of symbols and what they represent. At present, however, the set consists of the keysymbols shown in *Table 1*. These are augmented by symbols not used in the master lexicon, but introduced by post-lexical rules, for example taps (see 4.2). There are also, of course, stress, syllable and morphological markers; morpheme boundaries do not necessarily coincide with syllable boundaries.

The basic symbols are augmented by:

- square brackets: used to denote a segment which is deletable in certain accents, e.g.
  sentence s * e n . t [@] n s
- numbers: sub division of a primary group, e.g.
  blue b l * iu3

- capitals: used to denote a segment which is reducible in certain accents, e.g.
  fragile f r * a . jh AI l

These notations can be combined with any basic keysymbols or suprasegmental markers, and may also be combined with each other, for example [@] denotes a schwa usually omitted in US Englishes, while [@1] denotes a schwa usually omitted in British Englishes.

3. WORD-LISTS

Word-lists for recording diphones were constructed from the keyword lexicon. Unlike CSTR's usual diphone word-lists, which consist of nonsense words, these word-lists were constructed using real words from the dictionary. When using naive speakers to make recordings, it is obviously preferable to use real words, as the speaker will not have to learn the transcription system. This is especially advantageous for keysymbol transcriptions, as these will contain distinctions which are not used by the speaker and so can be confusing.

3.1 Preparation of Lexicon

It is necessary to run the lexicon through the post-lexical rules before extracting diphones, in order to ensure that all relevant pairs are extracted. This results in different word-lists for recording different accents.

For example, in the base lexicon 'greed' and 'agreed' both contain the keysymbol |ii|, but after the Scottish post-lexical processing rules the morpheme-boundary in 'agreed' will trigger use of a long vowel:

greed g r * ii d → g r * ii d
agreed @ . g r ii $ d → @ . g r * ii: d

Instead of running the post-lexical rules we could use the initial keyword transcriptions but include all morpheme-boundary permutations in the diphone list, but as most accents do not use morpheme boundaries distinctively this would result in a high degree of redundancy and long lists for speakers to read.

It should also be noted that some symbols are deleted by the post-lexical rules, for example word-internal pre-consonantal |r| in non-rhotic accents (see [3]) e.g. 'card' |k * ar r d| → |k * ar d|, giving us the combination |ar d| which does not occur in the original lexicon.

Bracketed or capitalised symbols are mostly re-written by the post-lexical rules, for example, |OU1| represents either |ou| or |@| depending on the accent, and so do not need to be included as symbols in the diphone list. Words containing these in the original transcriptions are best avoided in the word-lists, as some of the words using these symbols vary by style as well as accent, for example 'obey' |OU1 . b * ei|, and it is obviously desirable to select words which have a high likelihood of being pronounced as intended. On the other hand, lower-case numbered symbols represent subdivisions of a primary keysymbol group and so for some accents may represent phonemes not listed elsewhere in the lexicon; these are therefore included in the word-lists.

3.2 Word Selection

As noted above, some types of words were avoided in the word-list. Homographs were also excluded from consideration and words which are treated as exceptions in the lexicon (for example 'important') were avoided. It was decided that the most important criterion for word-list construction was word-frequency, rather than attempting to match syllable-patterns or to get the maximum number of diphones per word. Word-frequencies were extracted from several on-line texts, and these were used to order the lexicon. A script then selected the first example of each diphone from the ordered lexicon.

3.3 Diphone Extraction

Initially all symbol-pairs were extracted, including all boundary positions. Some diphone pairs were then discarded from the list as they can be represented by other diphone pairs, for example the syllable boundary is important for 'hatrack' as opposed to 'Patrick', but not for 'fanzine' versus 'fans'. In cases such as the latter the more frequent word was selected, so |v . z| in 'evzone' was discarded in favour of |v z| in 'gives'. There were also a number of potential diphone pairs which were not found in the dictionary. The script produced a list of these, which were then divided into cross-syllable and within-syllable pairs.

Cross-syllable pairs which could not be represented by existing within-syllable pairs were checked against symbols found at word-boundaries to see if word-pairs could be constructed. This was especially common with vowel-vowel pairs such as |oi|-|e|, which is not found word-internally but occurs across word-boundaries, for example 'toy elephant'. Some keysymbols are not found at word-boundaries, for example the symbol |e| does not occur word-finally, so |e|-|oi| should never be needed.

Within-syllable pairs will obviously only occur within words, discounting fast speech processes. We might assume that if they do not occur in the lexicon we do not need them, but it is important to check whether they might occur when the lexicon is expanded to contain new words. Even if we are reasonably sure that the lexicon has good coverage, names can go from almost unheard-of to well-known in a very short time, for example 'Kosovo'. It is preferable to find words containing possible pairs in case they are needed in the future, but of course there can be problems in finding such words, and problems in obtaining a suitable pronunciation if the speaker is unfamiliar with them.

Use of keysymbols rather than phonemes does result in speakers recording pairs which are not distinct for them, which leaves some redundancy in the word-list; however, reducing the diphone set would involve writing rules to remove the redundant symbols in each accent, and as such rules are not needed for other purposes this approach would increase rather than decrease the work involved in making recordings.

4. RESULTS

Scottish and Irish English are, of course, not single systems. Each region contains a number of geographical accents as well as social stratification. However, there are some features common to all or most of their accents, and some of these will be discussed here. It should be noted that the current work is focussing on segmental differences between accents; there are of course differences in intonation and segment duration as well, but these are beyond the scope of the current work.

4.1 Scottish English

Some features of the transcription system were originally motivated by pronunciation features of Scottish English, though some of these are also found elsewhere. For example, the |ae|-|ai| distinction ('tied' versus 'tide') is a well-known feature of Scottish English. |or|-|our| ('horse' versus 'hoarse') is also a notable phonemic distinction in Scottish English. This appears in other accents of English too, although in many cases, for example certain US dialects, it is recessive; however, it shows no evidence of dying out in Scottish English.

Division of Wells's NURSE keyword into |@@r| and |er| was motivated by Scottish English, as it is necessary to distinguish such pairs as 'Hurd' and 'heard' or 'cur' and 'Kerr'; in RP these are all pronounced /ɜ/, but in Scotland they are usually /ʌ/ and /ɛ/ respectively. Some Scottish accents also have a third division, /ɪ/, in words such as 'bird'; this has not at present been included in our transcriptions. It is also worth noting that middle-class Edinburgh speakers may have a slightly different system, using /ɜ/ in many of these words.

There are also features of Scottish English which involve mergers of phonemic distinctions usual in other accents of English. For example, Scottish English does not distinguish 'pull' and 'pool', or 'cot' and 'caught'. This means that if we transcribe 'pull' as |p * u l|, and 'pool' as |p * uu l|, a speaker recording these keywords will realise both the |u| keysymbol and the |uu| keysymbol with the same phoneme. All words transcribed with either |u| or |uu| in the lexicon will then be synthesised using this phoneme, generally pronounced [ʉ].

Postlexical rules for Scottish English include the vowel-lengthening rule mentioned above, and t-glottaling, the extent of which varies by accent and social class. Unlike RP the dark/light |l| contrast is not needed as all |l|'s are generally dark, and unlike many British accents h-dropping ('hat' → |* a t|) does not occur.

4.2 Irish English

Like Scottish English, Irish English is rhotic. It also retains some of the vowel distinctions often lost before orthographic 'r' in other accents, such as 'horse'/'hoarse' and 'Hurd'/'heard'. Our (Southern) Irish speaker, however, who does not have a particularly broad accent, distinguishes |or|-|our| but not |@@r|-|er|.

Lack of /θ/ is a noticeable feature of Irish English; in fact there is generally a distinction in place of articulation between |th| words and |t| words, though this is often so small as to be virtually inaudible, which is the case for our speaker. Use of the |th|-|t| keysymbols, however, results in separate diphones being recorded for these pairs, so any difference will be retained.

A number of keysymbols have a different phonetic realisation in Irish English from RP or other Englishes; for example, |a|-|aa| ('Pam'/'palm') are contrasted in RP by quality and length as /æ/-/ɑː/, while in Irish English the distinction is mainly in length (/a/-/aː/). Vowels such as |ei| and |ou| are generally monophthongs rather than diphthongs as in RP. Such features are of course captured in recording the diphones.

For post-lexical rules, like Scottish English, there is no distinction between dark and light |l|, but for Irish English |l| is light. In common with US English, Irish English uses tapped |t|/|d| in certain phonetic environments. However, our speaker only uses taps at word-ends preceding a vowel, for instance 'What a waste of time;' in such cases word-boundary diphone pairs need to be included. The need for such diphones in a given accent is signalled by the specification of cross-word environments in the post-lexical rules, along with the introduction of keysymbols not used word-internally. The speaker also generally uses [t] word-finally before a consonant, rather than the glottal stop which is common in other accents, for example 'But what do they know?', but this diphone will be included in within-word pairs.

5. CONCLUSION

We have described the structure of a keyword lexicon and its use in the construction of a diphone word list. To test and debug the lexicon, we are currently synthesising Scottish and Irish speech from diphones collected in this way, using the same lexicon for synthesis of both accents. The design of the lexicon should make it applicable to many other accents of English as well.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

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

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

[3] Fitt, Susan (1999). The treatment of vowels preceding 'r' in a keyword lexicon of English. *Proceedings: ICPhS 99.*