Proceedings of the Institute of Acoustics

DIPHONE SYNTHESIS FOR WELSH

Briony Williams

Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland, UK.

1 INTRODUCTION

The Welsh language is one of the lesser-used and lesser-researched languages of Europe. This work represents the first known attempt at developing a speech synthesiser for Welsh. Because comparatively little is known about the acoustic characteristics of Welsh speech sounds, it was decided to use diphone concatenation rather than rule-based parametric synthesis. The software of an existing text-to-speech synthesis system for English (described in [1]) was adapted for use with Welsh. This software uses the PSOLA synthesis technique, as described in [2], [3]. The software can run on a SUN workstation or on a PC with an LSI DSP board.

The number of Welsh phonemes included was 51, including 3 used only in English loanwords (/z/, affricates /ch/ and /jh/) and 3 used in restricted contexts (labialised /lw, nw, rw/). In total, there were 32 consonants and 19 vowels. Also, it was decided that the synthesiser should be able to handle English as well, due to the number of English names that appear in Welsh speech. So three phonemes were added to cover English sounds that had no equivalent in Welsh (equivalences derived from [4]). These phonemes were: /zh/ (voiced palato-alveolar fricative), /oa/ (as in RP "paw"), and /@@/ (as in RP "purr"). Tables 1 and 2 show the (South) Welsh phonemes, with their equivalents in RP (often phonetically very different), and the extra English phonemes.

2 PRODUCING THE DIPHONE DATABASE

The text of a speech database was designed in the form of pseudo-Welsh nonsense words and short phrases, in order to ensure that all possible phoneme pairs were included. Welsh spelling is a very reliable guide to pronunciation, so the text was in normal orthography. Thus it was not necessary to use a phonetically-trained subject, unlike the case for English described in [5]. There were two or three words in each item, often a mixture of pseudo-Welsh words and "real" function words. The consonant used as a "filler" was /d/, as this is a common consonant in Welsh and has no lip rounding that might affect surrounding segments. The vowels used as "fillers" were short /a/ and long /aa/, as these had no lip rounding and were neutral as regards front or back tongue position.

2.1 Relevant linguistic features of Welsh

Word-initial consonants of lexical words in Welsh may change according to syntactic and lexical context. For example, *cath*, /k *aa th/, "cat", a feminine noun, may appear as *y gath*, /@ g *aa th/, "the cat" (soft mutation); *ei chath*, /i x *aa th/, "her cat" (spirant mutation); or *fy nghath*, /v @

ngh *aa th/, "my cat" (nasal mutation). Certain phonemes appear word-initially only as the result of a mutation, such as the voiceless nasals (/mh, nh, ngh/), which occur only morpheme-initially. Voiceless stops may appear both word-initially and word-finally, but are far more common initially than in other contexts. Therefore both these classes were treated as word-initial only, together with the consonants /rh (voiceless alveolar trill), h, w, j (palatal glide), z, sh/.

Welsh monophthongs may be either long or short, differing in duration and vowel quality [6]. Vowels in unstressed syllables are short. In stressed syllables, monophthongs are long if followed by one of /b, d, g, v, dh, f, th, x, m, n, ng, l, r/ (unless marked explicitly as long by a circumflex in the orthography). This meant that combinations of "long monophthong plus a consonant not in this list" had to be derived using an orthographic circumflex on the vowel grapheme. The pair "short monophthong plus a consonant from this list" had to be located in an unstressed syllable to ensure shortness: this was done by placing it in an unstressed final syllable, as in *dydo ddad*, /d @* d o dh aa* d/, for the /o-dh/ diphone (where asterisk indicates lexical stress in the phonemic transcription). In Welsh, the unstressed final syllable of a polysyllable has great acoustic salience and long duration (see [8]), and so is a suitable point from which to derive a diphone.

Schwa in polysyllables appears only in non-final syllables. In monosyllables, it appears only in a few unstressed function words. Since such words most often interact with the initial consonant of the following word (according to the mutation system described above), it was not possible to allow for the full range of consonants in words following such a "real" function word. Thus it was decided to locate schwa in the (stressed) penult of a polysyllable, as in *bydau di*, /b @* d e d i/, for the /b-@/ diphone (in Welsh, schwa may be stressed). This contrasts with the pseudo-word in *bewd di*, /b *eu d d i/ for the /b-eu/ diphone, where a monosyllable is used.

2.2 Pseudo-word generation program

A PASCAL program was written to generate a first draft of pseudo-Welsh words and short phrases that contained all possible two-phoneme combinations (ie. all possible Welsh diphones). The final text, after some hand-editing, contained 2824 items, covering 2973 diphones. These included a few specifically English diphones, added to ensure that any English word could also be synthesised, ie. diphones involving the phonemes /zh, @@, oa/ (see section 1 above).

2.3 Recording and processing the database

A male native speaker of South Welsh was recorded uttering the items described above, wearing a laryngograph collar in order to obtain the glottal waveform. Recording was carried out at a sampling rate of 10 kHz, using 16-bit quantisation. End-point detection (manually checked) was carried out on the recordings. This reduced the 2824 speech files to a total of just over 61 megabytes in all. A little over ten percent of this database was then segmented by hand. An automatic HMM-based segmenter was trained over this material, and was run over the remainder of the database. The result of automatic segmentation was edited by hand for the portion of interest in each speech file. Full details of processing are given in [9]. Both hand-segmenting and automatic segmenting were carried out at the "demi-phoneme" level, that is, half a phoneme. Software was written to convert the output label files into files giving diphone boundaries. This method enabled the diphone boundaries to be estimated by the automatic segmenter so that less manual post-editing was required.

A software tool was run over the end-point detected laryngograph output files to derive pitchmark files, these being text files giving the location of the peak of each pitch period in ms from the start

of the file. The tool incorporates a simple peak-picking algorithm. The pitchmark files are used together with the speech files by the PSOLA algorithm during synthesis.

2.4 Producing the diphone dictionary and related files

A "scheme file" was then produced. This is a text file where each line is a phonemic form of the utterance together with a phonemic form of the diphone(s) contained in it. The scheme file can be used for several speakers, given the same accent and same recording text. From this was derived the "link file", which is a text file where each line contains, in addition, the relevant speech file name. This file is different for each new speaker.

Software was then written to derive the diphone dictionary file. This is a text file in which each line corresponds to a diphone. Each line contains a diphone representation (eg. /p-aa/), a speech file name, and three numbers giving the location in that speech file of the diphone start, diphone mid point (corresponding to the phoneme boundary), and diphone end point. This file is used to locate the required diphones during synthesis.

3 LETTER-TO-SOUND RULES

The work reported here forms part of a text-to-speech synthesis system for Welsh. Existing software for English TTS has been adapted for Welsh, in particular the letter-to-sound rules and the dictionary, as follows.

Letter-to-sound rules for Welsh have been written [11]. In contrast to English, the orthography of Welsh is a very reliable guide to its pronunciation, especially for the consonants. There are problems in the case of the grapheme "i", which can represent either a vowel or a palatal glide, and also in the case of the grapheme "w", which can represent either a vowel, a labial-velar glide, or even a labialisation marker for certain alveolar consonants. These graphemes are the source of most of the complexity in the rules.

Four sets of rules were written, corresponding to four passes through the input word. This modular approach was chosen partly because it allowed for easy adaptation of the rules to different accents of Welsh. The output of each rule module (or each pass through the input word) formed the input to the next one. The software used was the publicly-available "phon" software written by Greg Lee of Hawaii University and implemented by him for English. This software converts the linguistic rules, written in the form of context-sensitive rewrite rules, to a header file for a "C" program. The software is then compiled as a "C" program and hence runs quickly. Some modifications were made to the software to adapt it to Welsh, in particular the addition of a class of consonant graphemes before which stressed monophthongs are phonologically long.

3.1 Epenthesis and diaeresis rule module

The first set of rules adds epenthetic vowels and any diaereses. Epenthetic vowels are vowels which are pronounced in some dialects of Welsh but not shown in the spelling. They take on the quality of the vowel of the preceding syllable (or the second part of a diphthong in the preceding syllable), and occur morpheme-finally. For example, *llyfr*, "book", is pronounced /lh *i v i r/ in some dialects, while *aml*, "often", would be pronounced /*a m a l/. Given the input form cefn (for "back") the first set of rules would output cefen, using the rule in (1) below. In this rule, the input grapheme is enclosed in square brackets flanked by left and right contexts, and the output

is given to the right of the equals sign. The hash sign indicates a word boundary, while the "R" represents the class of "l, n, r" graphemes.

(1) e [f] R # = fe

Diaereses are also added by the first set of rules between two vowel graphemes that cannot be part of the same diphthong and which cannot be a combination of a glide and a vowel. For example, *lleoedd*, "places", would be input as lleoedd and output as lle"oedd by the first set of rules. These rules are required in order to supply the correct contexts to later rules that locate word stress and add syllable boundary markers. The first set of rules contains a total of 170 rules (including rules for punctuation characters).

3.2 Word stress rule module

The second set of rules locates the position of word-level stress, and also disambiguates between the vocalic and consonantal realisations of the "i" and "w" graphemes. Stressed vowels are output in capitalised form by these rules, while consonantal "i" and "w" are output as "J" and "M" respectively. These rules, in effect, perform a limited parse of the input word into syllables, in order to locate the penultimate syllable (which is the default stressed syllable in Welsh polysyllabic words). For example, input cefen would be output as cEfen by these rules, and input gwaca+u (stress shown orthographically by an accent mark on the vowel, here represented by a plus sign) would be output as gMacAU. The second set of rules contains a total of 731 rules (including punctuation rules).

3.3 Grapheme-to-phoneme rule module

The third set of rules performs the conversion from graphemes to phonemes. This is relatively straightforward, although the rules for vowel symbols have the task of determining phonological vowel length based on stress level and the type and number of any following consonants. Thus input cEfen would be output as /k *ee v e n/, and input gMacAU would become /g w a k *ai/. This third set of rules contains a total of 356 rules (including punctuation rules).

3.4 Syllabification rule module

The final rule module inserts syllable boundaries (shown as a dot). Thus input /k *ee v e n/ and /g w a k *ai/ become output /k *ee . v e n/ and /g w a . k *ai/ respectively. This set of rules contains a total of 18 rules.

4 PRONUNCIATION LEXICON

A small lexicon of function words was developed to begin with. It also includes words with irregular pronunciation and stressing (eg. polysyllables with stress on the final syllable that is not orthographically marked, such as *mwynhau*, /m ui n h *ai/, "to enjoy"). Unlike English, Welsh does not have variable lexical stress placement, and so does not have the potential for stress-related minimal pairs such as "DIgest - diGEST". This means that there is not the same need in Welsh for a full lexicon and syntactic parsing in order to derive a word's syntactic class and thereby its pronunciation. Also, in Welsh the orthography is a reliable guide to pronunciation. This means that, for the vast majority of Welsh words, a large pronunciation lexicon is not absolutely required. However, it is useful in order to save processing time during synthesis.

4.1 Production of seed wordlist

After the small function word lexicon had been provided with pronunciations by hand, a much larger lexicon of lexical words was produced. This was the first machine-readable pronunciation dictionary of Welsh ever produced, as even hard-copy Welsh dictionaries do not contain pronunciation, and so was necessarily limited in scope. The raw material was a collection of postings to the WELSH-L electronic discussion list. The postings were either in Welsh or contained some Welsh among the English text, giving a total of 63,394 words of text which, however, included extraneous material such as mail headers. This was edited by hand to remove all English text and mail headers, etc.

The next stage was to work through the text removing all intial consonant mutations by hand, so that only the "radical" forms of words remained. This was in order to feed the later mutation rules, so that all possible forms of each word would be captured. Using standard Unix text-processing tools, a list of all unique words was then produced. This was edited by hand to remove numerals, remaining English words, spelling errors and other irregularities. The result was filtered against the existing function word lexicon to remove words already in that lexicon. The number of remaining words in this seed wordlist was 3636 Welsh words.

4.2 Expansion of wordlist

English placenames were filtered out from the wordlist by hand, as these generally do not participate in the mutation system. This left 3326 words. Rules were then written (using the "phon" software referred to above) to perform all possible mutations on initial consonants. These rules were then run over the list of 3326 words, giving an output of 8975 words (including the input 3326 words). The four-stage letter-to-sound rules were then run over this expanded wordlist, and the default word class of "nil" (ie. lexical word) was assigned to each entry. The resulting output was merged with the function word lexicon to give a master lexicon of 10,552 entries (including English placenames and some irregularly-pronounced Welsh words).

4.3 Supplementary wordlist

The initial raw text corpus had been made up of postings to WELSH-L that contained Welsh text from its inception in November 1992 up to the end of July 1993. A further text corpus was amassed, containing Welsh postings from August 1993 to August 1994. This was reduced to a list of unique words using Unix text-processing tools. This list was filtered against a list of over 25,000 English words to remove the bulk of the English words contained in it. It was also filtered against the master Welsh wordlist to remove all Welsh words already captured. The remainder was split by hand into extra English words (873 words) and extra Welsh words (2864 words). The latter list was edited by hand to undo all mutations and arrive at the radical forms of words. Since the bulk of the raw text consisted of words already present in the lexicon, it was not considered to be worth the effort of undoing every mutation by hand while still in running text form. The result was processed by the mutation rules referred to earlier, such that the original 2864 words yielded 7687 words (including the original 2864 words). The letter-to-sound rules were run over this wordlist, and the result added to the existing lexicon, making a new lexicon of 18,212 entries. This was compiled into a form usable by the text-to-speech system.

5 OTHER PROCESSING

Other adaptations were made which are not as far advanced as the modules described above, and these areas are the subjet of ongoing work.

5.1 Prosodic rules

Initial duration values for phonemes were derived by taking a value for each phoneme that was three-quarters of the mean duration of that phoneme in the small corpus of isolated words that had been hand-segmented (see section 2.3 above). Rules were written to adjust phone and syllable durations in various contexts. For example, a phonologically long monophthong in a penultimate syllable will be shortened, while a consonant after a stressed vowel will be lengthened (the latter feature having been observed in auditory studies of Welsh and also in [8]). Rules were also written to lengthen syllables that were final in the utterance, the clause, and the (polysyllabic) word. No stress-related durational adjustment was required, as this does not appear to be a feature of Welsh in the same way that it is in English [8].

A start was made on adapting the software for English intonation. Currently, only the three pitch levels have been reset, these being the "floor" (90 Hz), "reference" (140 Hz) and "ceiling" (230 Hz) pitch levels. The values were chosen to reflect the fact that the original speaker had a fairly high-pitched speaking voice, and also the fact that Welsh intonation tends to utilise a wider frequency range than does English. This area will require further research.

5.2 Anglo-Welsh synthesis

It was thought desireable to enable the synthesiser to handle input English text as well, to be spoken in English with a Welsh accent. The existing English dictionary of over 25,000 entries was adapted for Welsh by editing the phonemic strings to show their Welsh equivalents. In addition, the English letter-to-sound rules were edited to reflect a Welsh accent of English. Existing English final-r processing was enabled, as Welsh English is generally non-rhotic [4]. The Welsh prosodic rules were retained. The resulting synthesised speech is clear and intelligible, with a noticeable Welsh accent.

5.3 Other tasks

The Welsh TTS system could be extended to cover a North Welsh accent, by recording a North Welsh speaker. This would entail adding new items to the recording text, as North Welsh represents a net increase of seven vowels over South Welsh. The lexicon and letter-to-sound rules would have to be slightly modified accordingly, but this would not be a large task. In addition, a female speaker could be recorded, in order to provide a choice of speaker sex.

REFERENCES

[1] P.A. Taylor, I.A. Nairn, A.M. Sutherland & M.A. Jack (1991) "A real time speech synthesis system", *Proceedings of the Second European Conference on Speech Communication and Technology (Eurospeech 91)*, 1: 341-344.

[2] C. Hamon, E. Moulines & F. Charpentier (1989) "A diphone synthesis system based on time-Domain modifications of speech", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing.*

[3] F. Charpentier & E. Moulines (1989) "Pitch-synchronous waveform processing techniques for text-to speech synthesis using diphones", *Proceedings of the European Conference on Speech Communication and Technology*, 2: 13-19.
[4] J.C. Wells (1982) *Accents of English*, vol. 2: The British Isles. Cambridge: Cambridge University Press.
[5] S.D. Isard & D.A. Miller (1986) "Diphone synthesis techniques", *IEE Conference Publication* no. 258: 77-82.
[6] M.J. Ball & G.E. Jones (eds.) (1984) *Welsh Phonology*. Cardiff: University of Wales Press.
[7] P.A. Taylor & S.D. Isard (1991) "Automatic diphone segmentation", *Proceedings of the Second European Conference on Speech Communication and Technology (Eurospeech 91)*, 2: 709-711.
[8] B. Williams (1985) "An acoustic study of some features of Welsh prosody", in C. Johns-Lewis (ed.), *Intonation in Discourse*. London: Croom Helm.
[9] B. Williams (1994) "Diphone synthesis for the Welsh language", *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, September 19-22 1994.
[10] C.H. Thomas (1967) "Welsh intonation — a preliminary study", *Studia Celtica*, 2: 8-28.
[11] B. Williams (1994) "Welsh letter-to-sound rules: rewrite rules and two-level rules compared". *Computer Speech and Language*, 8: 261-277.

ACKNOWLEDGEMENTS

This work was carried out while the author was in receipt of a three-year Research Fellowship from the Royal Society of Edinburgh, funded by BP.

DIPHONE SYNTHESIS FOR WELSH

Table 1: Welsh consonant phonemes ("W"), with RP equivalents where appropriate

W	Description	RP	W	Description	RP
p	voiceless labial stop	p	zh	voiced palato-alveolar fricative	zh
t	voiceless alveolar stop	t	ch	voiceless palato-alveolar affricate	ch
k	voiceless velar stop	k	jh	voiced palato-alveolar affricate	jh
b	voiced labial stop	b	l	voiced alveolar lateral	l
d	voiced alveolar stop	d	r	voiced alveolar trill	r
g	voiced velar stop	g	m	voiced labial nasal	m
f	voiceless labio-dental fricative	f	n	voiced alveolar nasal	n
th	voiceless dental fricative	th	ng	voiced velar nasal	ng
h	voiceless glottal fricative	h	lh	voiceless lateral fricative	-
s	voiceless alveolar fricative	s	rh	voiceless alveolar trill	-
x	voiceless uvular fricative	(x)	mh	voiceless labial nasal	-
v	voiced labio-dental fricative	v	nh	voiceless alveolar nasal	-
dh	voiced dental fricative	dh	ngh	voiceless velar nasal	-
z	voiced alveolar fricative	z	j	voiced palatal glide	j
sh	voiceless palato-alveolar fricative	sh	w	voiced labial-velar glide	w
lw	labialised voiced alveolar lateral	-	nw	labialised voiced alveolar nasal	-
rw	labialised voiced alveolar trill	-			

Table 2: South Welsh vowels ("W"), with RP equivalents where appropriate

W	Description	RP	W	Description	RP
i	centralised close front unrounded (short)	i	@i	diphthong starting at /@/	ai
e	half-open front unrounded (short)	e	ai	diphthong starting at /a/	-
a	open central unrounded (short)	a	oi	diphthong starting at /o/	oi
o	half-open back rounded (short)	o	ui	/u/ is vocalic element	-
u	centralised close back rounded (short)	u	iu	/i/ is vocalic element	j uu
@	-mid central unrounded (ie. schwa)	@	eu	/e/ is vocalic element	-
ii	close front unrounded (long)	ii	au	diphthong starting at /a/	-
ee	half-close front unrounded (long)	ei	@u	diphthong starting at schwa	au
aa	open central unrounded (long)	aa	-	half-open back rounded	oa
oo	half-close back rounded (long)	ou	-	mid central unrounded (long)	@@
uu	close back rounded (long)	uu			