TEXT-TO-SPEECH SYNTHESIS FOR WELSH AND WELSH ENGLISH

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

email: briony@cstr.ed.ac.uk

ABSTRACT

This work represents the first known attempt to develop a text-to-speech synthesiser for Welsh. A list of pseudo-Welsh nonsense words was generated, allowing for certain difficulties particular to Welsh. Diphones were derived semi-manually from the recorded nonsense words. The first known phonemic lexicon for Welsh was derived from an electronic corpus. Letter-to-sound rules for Welsh were written, differentiating between the vocalic and consonantal realisations of two graphemes, and assigning lexical stress. An existing English text-to-speech synthesis system was adapted for Welsh. Some simple duration and F0 rules were written that gave pleasing results with the minimum of rules. The resulting system can be used for Welsh or for English spoken with a recognisable Welsh accent.

1 THE WELSH LANGUAGE

Welsh is one of Europe's lesser-used languages, and dates from the late sixth century AD. It belongs to the Celtic family of Indo-European languages, and is spoken in parts of Wales. This work represents the first known attempt to develop text-to-speech synthesis (TTS) for Welsh. Comparatively little is known about the acoustic characteristics of Welsh speech sounds, and so diphone concatenation was used rather than rule-based synthesis. An existing TTS system for English [1] was adapted for use with Welsh. The software uses the PSOLA synthesis technique when concatenating diphones, as described in [2], [3].

The phonemes included three that are used only in English loanwords (/ʒ/, /t͡ʃ/, /d͡ʒ/), and also three used in restricted contexts (/lʷ, nʷ, rʷ/), together with three extra segments that were purely English. The total was 51 segments, split between 32 consonants and 19 vowels. The three extra segments were added to cover English sounds that had no equivalent in Welsh (equivalences derived from [4]). These three segments were: /ʒ/, /ɔː/, and /ɜː/. Tables 1 and 2 show the South Welsh phonemes, with their equivalents in RP and the extra English segments.

The accent modelled was a South Welsh one, though it would be possible to develop a similar system for a North Welsh accent (which has a larger number of vowels). The general strategy chosen was to adapt an existing diphone-based TTS system for RP English.

W	*RP*		*W*	*RP*		*W*	*RP*
p	p		t	t		k	k
b	b		d	d		g	g
f	f		θ	θ		h	h
χ	-		s	s		ʃ	ʃ
t͡ʃ	t͡ʃ		v	v		ð	ð
z	z		ʒ	ʒ		d͡ʒ	d͡ʒ
l	l		r	r		ɬ	-
r̥	-		j	j		w	w
m	m		n	n		ŋ	ŋ
m̥	-		n̥	-		ŋ̊	-
lʷ	-		nʷ	-		rʷ	-

Table 1: Welsh consonant phonemes, with RP equivalents: "*W*" = Welsh.

W	*RP*		*W*	*RP*		*W*	*RP*
ɪ	ɪ		ɛ	ɛ		a	æ
ɔ	ɒ		ʊ	ʊ		ə	ə
i	i		e	eɪ		aː	ɑ
u	u		əɪ	aɪ		aɪ	-
oɪ	oɪ		uɪ	-		iʊ	ju
eʊ	-		aʊ	-		əʊ	ɑʊ
o	oʊ		-	ɔː		-	ɜː

Table 2: (South) Welsh vowels, with RP equivalents: "*W*" = Welsh.

2 TEXT-TO-PHONEME STAGE

2.1 Letter-to-sound rules

Letter-to-sound (LTS) rules for Welsh were written [5]. In contrast to English, Welsh orthography is a reliable guide to pronunciation. Also, word stress is fixed on the penult in the majority of cases. However, there are problems in the case of "w" and "i", which cause most of the complexity in the rules. This is because "i" can be either a vowel or a palatal glide, while "w" can be a vowel, a labial-velar glide, or a labialisation marker.

The LTS rules were written in three blocks, corresponding to three passes through each word. This structure was advisable in order to cover different accents. The second ruleset was valid for all accents, but the others would require modification according to accent. The initial accent modelled was a South Welsh one.

2.1.1 Epenthetic vowels and diaeresis

The first rule set (170 rules) took the orthographic form, and output a modified form with epenthetic vowels inserted. These are vowels pronounced in some accents, but not shown in the spelling: eg. *cefn*, "back", /kévɛn/ (shown as **'keven'** at this stage). A symbol was also inserted between two vowels that could not form a diphthong, in order to assist the later syllabification rules: eg. *heol*, "road", /héɔl/, was shown as **'he"ol'**. An example rule is: **e[f]R# = fe**. Here, the target segment is "f", with a preceding "e" and a following "l", "n" or "r" (symbolised by "**R**") and a word boundary (#). The output is "fe", as in **'cefen'** from input **'cefn'**.

2.1.2 Disambiguation and stress placement

The second ruleset was by far the most complicated, containing 731 rules. Searching left-to-right, the stressed vowel was located and output in capitalised form. Also, "w" and "i" had varying output forms, as shown in Table 3 below.

Condition:	Input "w"	Input "i"
Consonant	**M**	**J**
Stressed vowel	**W**	**I**
Unstressed vowel	**w**	**i**

Table 3: Output forms of "w" and "i".

2.1.3 Grapheme-to-phoneme conversion

The third ruleset comprised the grapheme-to-phoneme rules proper (356 rules), including rules for determining the phonological length of stressed monophthongs. The combination "**sJ**" was output as /ʃ/, and the labialised consonants /lʷ, nʷ, rʷ/ were output in certain contexts.

2.2 Lexicon

There was no existing machine-readable pronunciation lexicon for Welsh. In a Welsh TTS system, a lexicon is not absolutely required for most words. This is because the lack of stress-related minimal pairs means there is never a need to know a word's class in order to find the correct pronunciation. However, a lexicon is included in order to speed up the processing by avoiding the LTS rules for the bulk of the words. Full details are given in [7].

2.2.1 Small lexicon

A small lexicon of 331 function words was developed. It included 466 words with irregular pronunciation, such as those with stress on the final syllable that was not orthographically marked: eg. *mwynhau*, /muɪnhái/, "to enjoy". The small lexicon was produced mainly from a Welsh grammar-book [6]. Rules were run over the exception words to generate all possible mutated forms (where the initial consonant undergoes one of three types of change). The LTS rules were applied, and the pronunciations edited by hand, for each of the 1262 words.

2.2.2 Main lexicon

A machine-readable corpus of Welsh text was gathered from the WELSH-L discussion list, which featured a colloquial style of Welsh that is in everyday use. The corpus was hand-edited to remove English words, mail headers, etc., and to convert all mutated initial consonants to unmutated form. A list of 3626 unique words not in the small lexicon was produced. Mutation rules were applied to this wordlist to produce all mutated forms, giving an expansion factor of about 2.5. The LTS rules were run over this list, and a default wordclass (lexical word) was assigned to each entry. With the small lexicon, there were now 10552 words. Further text from the same source was processed to derive more words. The final lexicon contained 18212 words (including mutated forms). This was compiled into a binary form for use in the TTS system.

3 PHONEME TO SPEECH STAGE

3.1 Designing the recording script

The text of a recording script for diphone extraction was designed in the form of pseudo-Welsh nonsense words. These were in normal orthography, so it was not necessary to use a phonetically-trained subject.

Certain linguistic features of Welsh made the design of this list of pseudo-words more difficult than it would be for English. For example, Welsh monophthongs may be either long or short, differing in duration and vowel quality ([8]). Vowels in unstressed syllables are short, while in stressed syllables monophthongs are long if followed by one of /b, d, g, v, ð, f, θ, χ, m, n, ŋ, l, r/. This meant that combinations of "a short monophthong plus a consonant from this list" had to be located in an unstressed syllable to ensure shortness: This was done by placing the vowel in an unstressed final syllable, (as in *dydo ddad*, /də́dɔ ðá:d/, for the ɔ-ð diphone). Further examples appear in [7].

A program was written to generate a list of pseudo-Welsh words containing all Welsh diphones. Sets were defined, such as the set of consonants initiating a syllable-initial cluster (/m, n, ŋ, m̥, n̥, ŋ̊, χ, θ, b, d, g, v, ð, f, p, t, k, s/). The final text contained 2824 items, covering 2973 diphones, including a few specifically English diphones.

3.2 Recording and segmenting

A male native speaker of South Welsh was recorded. He sat in a soundproofed booth wearing a headset microphone and a laryngograph electrode. The outputs from microphone and laryngograph were stored directly to disk. The sampling rate was 10000 samples per second.

Ten percent of this database was segmented by hand at the "demi-phoneme" level. Thus, for example, there were separate units for the closure and aspiration phase of stops. An automatic HMM segmenter was trained over this material, with units corresponding to "demi-phonemes" (unlike the otherwise identical autosegmenter used in [9], which used phoneme units only).

The automatic segmenter was run over the remainder of the database. The advantage of hand-segmenting and training HMM's at the "demi-phoneme" level was that the automatic segmenter yielded an initial estimate of the diphone boundaries. The result was edited by hand using a "glitch-minimisation" procedure. Hand-editing diphone boundaries took a great deal of time and effort, but the resulting synthetic speech displayed a gratifying absence of clicks and extraneous noises.

A software tool was run over the laryngograph files to derive pitchmark files, these being text files giving the location of the peak of each pitch period. The pitchmark files are used together with the speech files by the PSOLA algorithm during synthesis. There was no laryngograph waveform during voiceless consonants, and so a software tool was used to fill these gaps.

3.3 Diphone dictionary

A "scheme file" was then produced. This is a text file where each line gives the phonemic form of the utterance together with 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", a similar file where each line also contains the relevant speech file name.

The diphone dictionary was produced. This is a text file where each line corresponds to a diphone, and contains the following:

a) A diphone representation (eg. p-a:, or "p-aa" in orthographic form);
b) A speech file name:
c) Three numbers giving the location in that speech file of the diphone start-point, diphone mid-point and diphone end-point.

3.4 Duration and F0 modelling

The initial segment durations were 0.75 of the mean duration for each segment in the hand-segmented isolated words, in order to derive durations more suited to running speech. Rules were written such that each clause-final segment was multiplied by two to allow for final lengthening, while word-final and post-stress segments were also lengthened. The fact that practically no duration modelling is required is probably due to the fact that Welsh, unlike English, does not exhibit stress-related vowel reduction or vowel lengthening. Indeed, it has been found that vowel duration in Welsh has very little relation to stress [10].

The F0 algorithm was minimally different from that used for English. Only the three pitch levels were reset, these being the "floor" (set to 90 Hz), the "reference" level (140 Hz) and the "ceiling" (230 Hz). These values reflect the fact that the recorded subject had a fairly high-pitched voice, and that Welsh tends to use a wider frequency range than does English. The resulting F0 trace does not sound particularly Welsh, but neither does it sound English. To those who speak no Welsh, the resulting impression is adequate to suggest a Welsh accent.

4 WELSH ENGLISH

An existing English lexicon of 25000 words was edited to convert the phonemes to their Welsh equivalents (as shown in Tables 1 and 2 above). The LTS rules were similarly edited to reflect a Welsh accent of English. The existing English (RP) processing of word-final "r" was re-enabled, as Welsh English, like RP, is non-rhotic [4]. The Welsh duration and F0 rules were retained. The resulting synthesised speech was clear and intelligible, with a most noticeable Welsh accent.

5 FUTURE WORK

Tasks still remaining include the conversion of the TTS system to a North Welsh accent. North Welsh contains eight vowels not found in South Welsh: /ɨ, ɨː, ɨʊ, aɨ, aːɨ, oɨ, uɨ, əɨ/). Converting to North Welsh would require the following steps:

a) Add extra nonsense words to the recording script to cover the extra vowels.
b) Record a North Welsh speaker.
c) Perform semi-manual segmentation of the recorded files to derive the diphones.
d) Add the extra material to the scheme file and link file, and compile the diphone dictionary.
e) Edit the LTS rules to include the new vowels as output where appropriate.
f) Run the new LTS rules over the lexicon to produce a new lexicon for North Welsh.

These tasks, although tedious and time-consuming, are not particularly difficult, given that a South Welsh TTS system has already been produced.

Another possibility for future work would be to develop an English TTS system with a Scottish accent. This would require less work than a North Welsh system, since much of the existing English TTS system could be re-used.

6 REFERENCES

[1] P.A. Taylor, I.A. Nairn, A.M. Sutherland & M.A. Jack (1991) "A real time speech synthesis system", EUROSPEECH, 1991, pp. 341-344.

[2] C. Hamon, E. Moulines & F. Charpentier (1989) "A diphone synthesis system based on time-Domain modifications of speech", ICASSP 1989.

[3] F. Charpentier & E. Moulines (1989) "Pitch-synchronous waveform processing techniques for text-to speech synthesis using diphones", EUROSPEECH, 1989, pp. 13-19.

[4] J.C. Wells (1982) *Accents of English, vol. 2*. Cambridge: CUP.

[5] B. Williams (1994) "Welsh letter-to-sound rules: rewrite rules and two-level rules compared". *Computer Speech and Language*, vol. 8, pp. 261-277.

[6] Uned Iaith Genedlaethol Cymru (1976) *Gramadeg Cymraeg Cyfoes* (Contemporary Welsh Grammar). Y Bontfaen, Morgannwg: D. Brown a'i Feibion Cyf.

[7] B. Williams (1994) "Diphone synthesis for Welsh". *Proceedings of the Institute of Acoustics*, vol. 16, pp. 359-366.

[8] M.J. Ball & G.E. Jones (eds.) (1984) *Welsh Phonology*. Cardiff: University of Wales Press.

[9] P.A. Taylor & S.D. Isard (1991) "Automatic diphone segmentation", EUROSPEECH, 1991, pp. 709-711.

[10] B. Williams (1985) "An acoustic study of some features of Welsh prosody", in C. Johns-Lewis (ed.), *Intonation in Discourse*. London: Croom Helm.

7 ACKNOWLEDGEMENTS

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