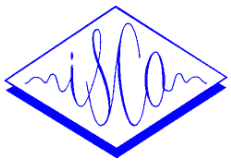

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

DIPHONE SYNTHESIS FOR THE WELSH LANGUAGE

Briony Williams

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

ABSTRACT

The Welsh language is comparatively little researched, and this work represents the first attempt to develop speech technology for Welsh. A list of pseudo-Welsh nonsense words was generated. Certain linguistic features of Welsh, such as the relationship between stress location and phonological vowel length, made this task more complicated than for English. A native speaker was recorded reading this list. Over ten percent of the speech was segmented by hand. The segmentation was carried out at the "demi-phoneme" level, from the beginning of a phoneme to its midpoint, in order to train a segmenter to find plausible diphone boundaries automatically. The segmentations were used to train a set of Hidden Markov Models, which automatically segmented the rest of the recordings. The segmentations were corrected by hand, and pitchmarking was carried out. An index of diphone locations was produced, together with a diphone dictionary. The resulting synthesised speech can be used either for Welsh, or for English spoken with a Welsh accent.

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 descibed 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.

Welsh	*RP*	*Welsh*	*Description*	*RP*
p	p	x	voiceless uvular fricative	(x)
t	t	lh	voiceless lateral fricative	-
k	k	rh	voiceless alveolar trill	-
b	b	mh	voiceless labial nasal	-
d	d	nh	voiceless alveolar nasal	-
g	g	ngh	voiceless velar nasal	-
f	f	th	voiceless dental fricative	th
h	h	dh	voiced dental fricative	dh
s	s	ch	v'less palato-alv affricate	ch
z	z	jh	v'd palato-alv affricate	jh
v	v	j	voiced palatal glide	j
l	l	w	voiced labial-velar glide	w
r	r	sh	v'less palato-alv fricative	sh
m	m	(zh)	voiced palato-alv fricative	zh
n	n	ng	voiced velar nasal	ng

Table 1: Welsh consonant phonemes, with RP equivalents.

Welsh	*RP*	*Welsh*	*RP*
i	i	@i	ai
e	e	ai	-
a	a	oi	oi
o	o	ui	-
u	u	iu	j uu
@	@	eu	-
ii	ii	au	-
ee	ei	@u	au
aa	aa	-	oa
oo	ou	-	@@
uu	uu		

Table 2: (South) Welsh vowels, with RP equivalents.

2 DESIGNING THE SPEECH 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 aim was to lend as much naturalness as possible to the text, in order to obtain from the speaker a truly Welsh pronunciation not influenced by knowledge of English. 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

Voiceless nasals (/mh, nh, ngh/) appear only at the start of words, often (for the nasals) as a result of the lexically-cued nasal mutation of the consonant. Voiceless stops may appear both initially and 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. 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. In general, a monosyllable is preferable as a diphone source, due to its greater acoustic salience compared to a penult, but it was not possible to use one in all cases.

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). Sets of phonemes were defined, such as the set of consonants that can initiate a syllable-initial cluster (/m, n, ng, mh, nh, ngh, x, th, b, d, g, v, dh, f, p, t, k, s/), or the set of vowels that can follow the grapheme 'i' when this corresponds to a palatal glide (/e, a, o, ee, aa, oo, ai, au, @i/). The first draft was edited by hand, to 'double up' phoneme pairs where possible, so that some items formed the source of more than one diphone. This meant that some items could then be deleted, thus saving on resources. The final text 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).

3 DERIVING THE DIPHONES

A male native speaker of South Welsh was recorded uttering the items described above. He sat in a soundproofed recording booth wearing a headset microphone and cued by a computer screen just beyond the booth (to eliminate paper-rustling). He wore a laryngograph electrode in a collar round the neck. Communication between subject and supervisor was by means of microphones and headphones. The outputs from microphone and laryngograph were multiplexed, digitised and stored directly to disk. Each item was recorded as a two-second speech file in recording sessions stretching over three days.

3.1 Preliminary processing

To begin with, the speech and laryngograph signals were separated and end-point detection (manually checked) was carried out on them. 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, using specially-chosen utterances to ensure that at least ten examples of each phoneme were included. The segmentation was carried out at the 'demi-phoneme' level. This meant that, for example, there were separate units for the closure phase and release/aspiration phase of stops, and separate units for the two halves of diphthongs. An automatic HMM-based segmenter was trained over this material, training units corresponding to 'demi-phonemes' (unlike the otherwise identical autosegmenter used in [7], which used phoneme units only).

3.2 Diphone extraction

The automatic segmenter was then 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 draft of the diphone boundaries. The result of automatic segmentation was edited by hand for the portion of interest in each speech file. A boundary between two 'demi-phonemes' corresponded to a diphone boundary and was given special attention, as follows. In cases where the waveform was periodic at that point (mainly vowels, glides and nasals), the boundary was located by hand at the point of minimum amplitude (ie. the most negative value). This was in order to minimise audible clicks on subsequent concatenation duraing synthesis of two units from different origins. To the same end, the boundary between the closure and release phases of stops was located a little before the release burst in all cases, so that the amplitude, around zero, would have no potential to cause clicks on concatenation. For the same reason, the boundary during fricatives was located as near to the zero-amplitude point as possible. Hand-editing to this degree of detail meant that deriving diphone boundaries took a great deal of time and effort. However, synthesis using these units displayed a gratifying absence of clicks and extraneous noises that have the potential to detract markedly from the quality of synthesiser output.

3.3 Pitchmarking

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. The tool incorporates a simple peak-picking algorithm. The pitchmark files are used together with the speech files by the PSOLA algorithm during synthesis. There was no laryngograph waveform for the duration of voiceless consonants, and so a software tool was used to fill these gaps and obtain an unbroken pitchmark file for each speech file.

4 USE OF THE DIPHONES

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', a text file where each line contains, in addition, the relevant speech file name. This will be different for each new speaker.

4.1 Producing the diphone dictionary

The diphone dictionary file was then produced. 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. Special-purpose software was written to derive these numbers from the output of the segmentation process.

Existing software was then used to compile the speech files and dictionary file into the 'grouped waveform' format, in order to conserve disk storage space. The relevant portions of the 2824 speech files were concatenated into one large file of nearly 16 megabytes (a saving of over 45 megabytes), while the new form of the dictionary file contained information relating to this single compiled speech file. This form of the data is important when the target architecture is a PC, since a PC's storage space is limited compared to a larger machine.

4.2 Synthesising utterances

Preliminary checks of the output involved synthesising utterances from manually-composed input files. Each row of these input files contained up to three items: a phoneme, the phoneme duration, and (if relevant) an F0 specification giving both the percentage of the duration of the segment at which this F0 was located, and an F0 value. At this stage of development, the phonemes were provided manually.

The durations were provided manually, but were derived from a very simple algorithm that would be straightforward to automate, as follows. The hand-segmented part of the database was run through a statistics program, and for each phoneme the mean and standard deviation of the duration was derived. These duration values represented segment durations found in isolated words and short phrases. In order to acquire duration values more suited to continuous speech, each mean duration was multiplied by 0.75, and the resultant values were used in the input files. Each clause-final segment was then multiplied by two to allow for final lengthening. As regards the segmental durations, initial impressions of the output were favourable, both for the Welsh and the English output. The overall tempo is measured rather than fast, but the speech is wholly intelligible. 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 [8].

The F0 information was likewise added manually but according to a simple algorithm that would be easy to automate. Since the original recording subject had a fairly high speaking voice, it was decided to replicate this feature in the F0 values chosen. An F0 target was allocated to the stressed syllable of every content word and to the immediately following unstressed syllable. Since the default location for stress in Welsh polysyllables is the penult, there was always a following unstressed syllable in the sentences used at this stage. Each F0 target was located at a point 25% into the duration of the relevant vowel. Silence segments, located at the start and end of the input, and between clauses, also had F0 targets allocated (sometimes two such targets, one at the start and one at the end of the silence segment).

The F0 values were determined as follows. The clause-initial silence segment was allocated a target of 120 Hz. The first non-silence target of the utterance (ie. on a stressed syllable) was allocated 170 Hz, and the immediately following syllable was given 190 Hz, ie. 20 Hz higher. This replicates the very characteristic Welsh intonation of a slight rise beginning on the accented syllable, with each new accent starting on a slightly lower F0 than the previous accented syllable (see [9]). The second accented syllable was given the target of 150 Hz (20 Hz below the previous accented syllable), and the syllable after it was 20 Hz higher at 170 Hz. The third accented syllable, if any, was given 130 Hz and the syllable following it 150 Hz, continuing the pattern. If there was a fourth accented syllable, this was given 120 Hz (ie. only 10 Hz lower) and the following syllable was 130 Hz (ie. only 10 Hz higher). The silence segment was allocated 90 Hz if at the end of an utterance, 120 Hz otherwise. A second clause in the same utterance was allocated slightly lower values (10 Hz lower in every case), in an attempt to replicate paragraph-level intonational downdrift. Initial impressions of the resultant intonation were favourable. The synthesised intonation is distinctly Welsh without being exaggeratedly "sing-song". A similar procedure was followed for the synthesis of English, and produced synthetic intonation that was eminently intelligible and (subjectively speaking) aesthetically pleasing. It appears that very little extra work will be required to produce the final form of the F0 algorithm.

5 REMAINDER OF TTS SYSTEM

The work reported here is intended to form part of a text-to-speech synthesis system for Welsh. Existing software for English TTS is being adapted for Welsh.

5.1 Letter-to-sound rules

Letter-to-sound rules for Welsh have been written [10] and are being incorporated into the software. 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.

5.2 Lexicon

A small lexicon of function words has been developed. 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 lexicon is not absolutely required. However, it may be useful in order to save processing time during synthesis.

6 FUTURE WORK

6.1 Prosodic processing

The remaining tasks are to develop the duration and F0 algorithms along the lines indicated in section 4.2 above, and integrate them into the existing English software. This should be comparatively straightforward, since the results obtained with the simple algorithms indicate that good synthesised Welsh speech can be obtained with the minimum of durational and F0 modelling.

6.2 Anglo-Welsh synthesis

It would also be desirable to increase the coverage of the TTS system to include full English capabilities. This would require the rewriting of the pronunciations in an existing large English lexicon, which would also involve the de-reduction of many vowels that are reduced to schwa in RP English but which receive their full vowel quality in Welsh English. The task would also require an extensive recasting of existing letter-to-sound rules for English, in a way which would reflect the known correspondences between English and Welsh phonemes. This adaptation would be a major task.

6.3 Other tasks

It would be useful to add English proper names to the lexicon for Welsh synthesis, as these are the only English words that are found extensively in Welsh speech. In addition, 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.

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.

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), vol. **1**, pp. 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*, vol. **2**, pp. 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, pp. 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), vol. **2**, pp. 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] C.H. Thomas (1967) "Welsh intonation -- a preliminary study", *Studia Celtica*, vol. **2**, pp. 8-28.

[10] B. Williams (1994) "Welsh letter-to-sound rules: rewrite rules and two-level rules compared". *Computer Speech and Language*, 1994 (in press).