Proceedings of the Institute of Acoustics

THE DESIGN OF A SPEECH DATABASE FOR WELSH DIPHONE EXTRACTION

Briony Williams

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

1 INTRODUCTION

The work reported in this paper forms part of a project aiming to develop a text-to-speech synthesis system for Welsh. Spoken Welsh takes the form of many local accents and dialects, which fall into two main groups: North Welsh and South Welsh. The work reported below is based on a general South Welsh accent: adapting the output for a North Welsh accent would be straightforward.

1.1 Diphone Synthesis

The two main types of output algorithm for a speech synthesiser are the rule-based formant synthesiser (as in [1]) and the diphone synthesiser (as in [2]). The formant synthesiser stores the parameters of individual phones, concatenating and smoothing between them at run-time. The diphone synthesiser, on the other hand, stores units (diphones) which run from halfway through one phone to halfway through the next, thus storing the transition information which is vital to phone identification, and smoothing at the acoustically less variable mid-points of phones. This method requires less time and less detailed acoustic phonetic knowledge of the language in question than does the formant synthesis method. As relatively little is known about the acoustic realisation of Welsh speech sounds, compared with the situation for other European languages, the diphone method seems more appropriate in this case.

1.2 Diphone Extraction

The procedure for obtaining the raw material of diphones is outlined in [2] and [3], and consists essentially of recording a native speaker pronouncing a series of nonsense words that conform to the phonological system of the language. The diphone segments are then excised from the recordings and stored, together with the location of the phone boundary within the diphone. Strictly speaking, the units obtained are 'allodiphones', as different variants of certain phonemes may be stored where there is a phonetic difference in realisation due to context. For example, in English, diphones for the voiceless stops followed by a vowel may be obtained both for contexts where the stop is syllable-initial (and therefore probably aspirated) and contexts where it is syllable-final (unaspirated and possibly glottalised), as well as contexts where it follows /s/ in a syllable-initial cluster (unaspirated but unglottalised). Considerations such as this necessitate a detailed knowledge of the language's phonological system on the part of the designer of the recording materials.

Originally, the location of diphone boundaries, and of the phone boundary within a diphone, was carried out by hand. This was a demanding and time-consuming task, open to human error. However, a recent diphone database for English was automatically segmented by training a Hidden Markov Model speech recogniser to the speaker's voice, inputting the relevant phones contained in the database, and running the recogniser as a segmenter. The segmenter located the phone boundaries, while a spectral mismatch minimisation algorithm located the optimal diphone boundaries [3]. Given this procedure, it would be possible to produce diphone sets for both North Welsh and South Welsh (and even for more specific Welsh accents) relatively quickly, after the initial hurdle of the database design has been overcome.

1.3 Procedure for English

The procedure for designing the recording database for the extraction of English diphones is outlined in [2]. Phonemic forms of nonsense words were generated by a program, and were represented by an alphabetic phonemic transcription with which the speaker had to familiarise himself, eg. `uhMAHtuh'. In [4], the phonemic forms of nonsence words were displayed to the speaker in Machine Readable Phonemic Alphabet (MRPA) form: this approach presupposes a substantial amount of linguistic knowledge on the part of the speaker. In [2], nonsense words were chosen rather than real words, as algorithmically-generated nonsense words held less risk of missing a diphone out of the recording materials. Isolated words were chosen, as more then one word at a time was unnecessary for English, and would waste recording time. One syllable in each word was stressed, and the diphone was located in this syllable (except in the case of diphones with schwa), in order to obtain maximum acoustic salience for the unit in question. Syllables and segments not involved in the diphone were realised by appropriate combinations of /l/ and /a/ or /@/ (schwa): these filler segments were chosen as being neutral with respect to coarticulations such as lip rounding. Since the position of the stressed syllable in an English word is not fixed, the stressed (and capitalised) syllable could be located word-initially (for diphones with initial silence), word-finally (for diphones with final silence), or word-internally (for most diphones): such variation in position did not occasion any redesign of the pattern of the nonsense word. The only distributional constraints on consonants in English are as follows:

(1) /h/, /w/, /y/ (palatal glide) cannot be syllable-final.
(2) /ng/ (velar nasal) cannot be syllable-initial.
(3) /ng/ cannot be preceded by a phonologically tense vowel (except for the case of *boing*, /b oi ng/).

The first two constraints were easily incorporated into the program generating the recording database, while the third would have been among the few manual emendations made to the program's output. It was not necessary to allow for any further constraints, unlike the situation for Welsh (see 3.1 below).

2 GENERATING THE DATABASE

A similar strategy was followed in designing the recording materials for Welsh diphones. A program was written to generate the nonsense words *ex nihilo*, with a certain amount of manual editing of the program's output. This section will discuss the generation program, while section 3 will examine the Welsh-specific problems facing the designer of the database, and section 4 will outline the manual post-editing phase. Section 5 will sketch out some general principles for diphone database design.

2.1 Generation Program

2.1.1 Overall strategy: The decision was taken to make the output as Welsh-like as possible. To this end, certain 'real' function words were employed, notably those conditioning certain consonantal mutations at the start of a following word (see 3.2.2 below). Since adult Welsh speakers are also able to speak English, it was felt to be important to discourage any influence of English that might creep in if the materials were seen by the speaker as being in some way artificial or un-Welsh. For example, a word ending in a voiceless stop, or beginning with a /v/ phoneme (other than when preceded by a function word causing the Soft Mutation) is a rare occurrence in Welsh, and so might be treated as an English loan-word by the speaker. This would run the risk of eliciting a non-authentic pronunciation. Therefore every effort was made to ensure that such words did not occur, and that consonants were found in contexts in which they would be frequent (eg. voiced stops in Soft Mutation context). This policy decision

had the effect of greatly complicating the writing of the program: however, the output subjectively appeared less artificial than the example nonsense words produced for English in [2].

2.1.2 PASCAL program: A PASCAL program was written by the author to generate nonsense words conforming to the phonological system of Welsh. PASCAL is a high-level programming language of a highly structured type that lends itself to the task of producing blocks of nonsense words patterned on the same template. In most cases, each line of output consisted of more than one 'word', as the nonsense word of interest was often preceded by a 'real' function word conditioning one of the three types of word-initial consonant mutation that occur in Welsh (soft mutation (lenition), nasal mutation, and spirant mutation). The program was provided with certain constants, some of which follow:

(4) Consonants that can occur syllable-finally (/p, t, k, b, d, g, f, th, x, v, dh, s, ch, jh, lh, l, r, m, n, ng/
(5) All consonants (the above plus /h, z, sh, rh (voiceless /r/), w, j (palatal glide), mh, nh, ngh (the three voiceless nasals)/
(6) Consonants that can follow /s/ syllable-initially (/p, t, k/).
(7) Consonants that can initiate a cluster syllable-initially (/m, n, ng, mh, nh, ngh, x, th, b, d, g, v, dh, f, p, t, k, s/): some of these do so only as mutated forms of radical (root) consonants.
(8) Short vowels (/i, e, a, o , u, @/).
(9) Vowels that can follow the grapheme 'i' when this corresponds to a palatal glide (/e, a, o, ee, aa, oo, ai, au, @i/).
(10) Consonants that, if not followed by another consonant, may follow a phonologically long vowel if this is in a stressed syllable (/s, lh, b, d, g, v, dh, f, th, x, m, n, ng, l, r/).
(11) Consonants characteristically preceded by a function word causing soft mutation (/b, d, v, g, dh/)
(12) Consonants characteristically preceded by a function word causing nasal mutation (/m, n, ng, mh, nh, ngh/).
(13) Consonants characteristically preceded by a function word causing spirant mutation (/x, th/).
(14) Consonants characteristically not preceded by any mutation context, ie. radical consonants (/k, f, p, s, t/).
(15) Consonants not entering into any clusters (/z, sh, ch, jh, th, rh, j (palatal glide), h/).

These subsets of phonemes were required for various blocks of nonsense words, in which the program would process each of the above strings one at a time, in order to ensure that all possible permutations had been covered. For the sake of ease of programming, all phonemes were represented by a single character, either upper-case or lower-case. An auxiliary program was then written, to convert the output of the program to the Machine-Readable Phonemic Alphabet for Welsh (MRPAW) that had been devised, in which each phoneme is represented by one or more lower-case graphemes, and is separated by a space from other phonemes. The output of the auxiliary program was more suitable for human inspection.

2.1.3 Orthography versus phonemic form: It was decided that the prompts file available to the recording subject would be in the form of orthography rather than a phonemic transcription or an attempted phonetic respelling. This is possible because Welsh orthography conforms to Welsh pronunciation far more closely than English spelling conforms to English pronunciation. The advantage of this approach is that it would not be necessary for the speaker to have any specialist linguistic knowledge (thus greatly increasing the pool of potential recording subjects), and also that there would be less risk of a slip of the tongue due to task difficulty. Therefore, a second auxiliary program was written, converting the output of the generation program into Welsh orthography.

2.2 Global decisions

As with the design of the English database referred to above, certain global decisions were made which had a far-reaching effect on the pattern of the output. The decisions taken differed from those used for English, as Welsh imposes its own particular constraints.

2.2.1 Default vowel and consonant: The default vowel used for 'filler' purposes was /a/ or /aa/, a low front unrounded vowel, phonologically long or short according to the nature of the following consonant(s). As was the case for English, this was chosen as being neutral with respect to lip rounding. The default consonants for 'filler' purposes were /t/ (following short vowels) and /d/ (following long vowels or schwa, or word-initially and word-finally). Since /d/ is more frequent in Welsh than /t/, it was allowed to occur in more contexts. An example nonsense phrase with filler vowels and consonants is *dydo gad*, /d @* d o g a* d/, containing the 'o-g' diphone.

2.2.2 Number of words: As stated in 2.1.2 above, the number of words per output line was greater than one (two or three words were output). In some cases, this was because of the presence of a mutation-causing function word, which was necessary in order to add credibility to the word of interest (eg. *yn dam-cad*, /@ n d a* m k a* d/, with soft-mutation-causing '*yn*'). The other type of case is discussed in section 3.2.

3 WELSH-SPECIFIC FEATURES

Certain features of Welsh phonology necessitated certain decisions in the design of the database, given that the aim was to make the nonsense words as naturalistically Welsh as possible (see 2.1.1). These features are discussed below.

3.1 Consonant Distribution Restrictions

3.1.1 Word-initial consonants: Voiceless nasals and voiceless /rh/ in Welsh appear only at the start of words or morphemes, often (for the nasals) as a result of the nasal mutation (eg. after *fy*, /v @/, 'my'). Voiceless stops may appear both word-initially and word-finally, but are far more common initially. Therefore both these classes were treated as exclusively word-initial, together with /h, w, j (palatal glide), z, sh/. In addition, the labialised consonants /lw, nw, rw/ have a very restricted distribution, appearing in certain of the words where initial /g/, /ng/ or null precedes and a vowel follows. These were therefore classed together and given special treatment in their own block of output lines.

3.1.2 Consonant clusters: The number of possible syllable-initial consonant clusters in Welsh is large. The number may even be larger than in English, as initial /t l -, d l -, n l -, k n -, g n -, v l -, v r -, m l -, m r -/, among others, are all possible. Some of these forms are the result of the mutation of a radical (root) form, and do not appear themselves as a radical form (eg. no words begin with /d l -/ in the radical form). The acoustic difference between syllable-initial and syllable-boundary forms of clusters is only significant in the case of initial /s/ and initial stops [8]. Therefore only these phonemes have separate diphones for syllable-initial and syllable-boundary versions of consonant clusters, as in the following examples, where underscore indicates syllable affiliation:

(16)	*y tlad yma*	/@ t l aa* d @ m a/	t_-_l diphone
(17)	*y dat-lad*	/@ d a* t l aa* d/	t-l diphone
(18)	*y stlad yma*	/@ s t l aa* d @ m a/	s_-_t and _t_-_l diphones

In the diphone representations above, the underscores reflect the fact that a left or right context is being more closely specified than in the case without underscores. The nonsense phrase in (16) contains the diphone with a possibly aspirated, unglottalised /t/, and a possibly partially devoiced /l/. The phrase in (17) contains the diphone with an unaspirated, possibly glottalised /t/, and fully voiced /l/. That in (18) contains the diphone with an unaspirated, unglottalised /t/ and fully voiced /l/. In the case of most cluster-initial consonants (ie. not /s/ or stops), the pattern shown in (17) (ie. the syllable-boundary form) is the only one used in the database. No separate diphones are needed for clusters of three consonants, as these will be composed of a sequence of two diphones.

3.1.3 Non-Welsh consonants: Certain consonants (/ch (voiceless palato-alveolar affricate), jh (voiced palato-alveolar affricate), and z/) occur almost exclusively in loan-words from English. However, a full range of diphones for these consonants also had to be included, as it was not possible to predict which vowel/consonant or consonant/vowel diphones involving these consonants did not actually occur in any existing loan-words. Therefore there is the capacity to handle new loan-words containing the full range of consonant/vowel combinations.

3.2 Vowel Distribution Restrictions

3.2.1 Schwa distribution: The vowel schwa does not occur in the ultima of a polysyllable nor in a stressed monosyllable. Therefore, while other vowels may be located in a monosyllabic word, a schwa must be located in a non-final syllable of a polysyllable. A monosyllable with final schwa is a stressless function word, and because of the interaction between these and initial consonantal mutations in the following word, it was not possible to allow for the full range of consonants in words following such a 'real' function word. Therefore it was decided to locate schwa in the penult of a polysyllable, as in *bydau di*, /b @* d e d i/, for the b-@ diphone. This contrasts with the form *bewd di*, /b eu* d d i/, for the b-eu diphone, where a monosyllable is used. In *bydau di*, the final -*au* is intended to be perceived as the plural morpheme -*au*, thus increasing the naturalness of the material for the speaker.

3.2.2 Interaction of vowel length, consonants and syllable position: In stressed syllables, monophthongs are long before a single consonant of a certain type (eg. /dh/) or no consonant [5], and are short before a single other consonant or more than one consonant. In unstressed syllables, monophthongs are always phonologically short. Long and short versions of most monophthongs differ in quality as well as duration [8], therefore it is necessary to ensure that separate diphones are available for long and short monophthongs. In most polysyllables, the penult is stressed. If the desired diphone is a combination of a short monophthong and one of the monophthong-lengthening consonants (eg. the o-dh diphone), then the monophthong cannot be located in a penult nor in a monosyllable (as, being stressed, it ought then to be long before that consonant) Since the final syllable of a polysyllable is normally unstressed, this is the place to locate a vowel when the diphone in question demands a short vowel whatever the nature of the following consonant, eg. *dydo ddad*, /d @* d o dh a* d/ for the o-dh diphone. If only the single nonsense word *(C)odd* had been used, the vowel would necessarily have been pronounced long before a voiced dental fricative [5]. Although this vowel could have been located in the unstressed antepenult of a single word (eg. *doddada*, /d o dh aa* d a/), the ultima was preferred. This is because the ultima is frequently more acoustically salient (in terms of duration , intensity, and higher frequency) even than the stressed penult [7], and so more and more accurate acoustic information would be available if the ultima were used.

It would also be possible to locate the monophthong in the penult if a further consonant followed the lengthening consonant (as all monophthongs are short before a cluster). This was in fact done in a few cases, namely those consonants (voiced and voiceless stops) where separate syllable-final and syllable-

boundary forms were required for the V-C diphones. Thus, for example, the nonsense phrase *dy dogsi*, /d @ d o* g s i/ yielded the syllable-final diphone o-g, and the nonsense phrase *dydo gad*, /d @* d o g a* d/ yielded the syllable-boundary diphone o-$g (where '$' refers to the syllable boundary).

4 PRODUCING THE SCHEME FILE

4.1 Adding Diphone Representations

Apart from the orthographic prompt file, to be used to prompt the speaker when recording, it was also necessary to produce a 'scheme' file. This is a version of the database in phonemic form, where each nonsense phrase is followed by the diphone(s) to be extracted from it. This forms a vital record of the origin of each diphone for later checking during development. Therefore the program was amended very slightly to produce a representation of the relevant diphone(s) on each output line, following the nonsense phrase. One of the two auxiliary programs then converted the program's character-coded output to phonemic form in the Machine-Readable Phonemic Alphabet for Welsh outlined in 2.1.2 above.

4.2 Manual Editing

The program's output was then edited manually. This was in order to cut down the number of nonsense phrases for recording, by reassigning some diphones to other phrases and deleting the original phrase. For example, most diphones involving the segments /a/, /d/ and /l/ were reassigned to other phrases where the segment occurred as a 'filler': the more specific phrase could then be deleted. This enabled 63 phrases to be deleted, leaving 2487 nonsense phrases. This represents a not inconsiderable saving in recording time and segmentation effort.

4.3 Use of the Scheme File

This scheme file may now be used for any (South) Welsh speaker. For each speaker, it will be necessary to run software (written by P. Taylor at the CSTR) to produce a 'link' file, where each line contains an individual speech file name, the phonemic representation of the nonsense phrase recorded in that file, and a representation of the diphone(s) to be extracted from that portion of speech. In the algorithmic production of the link file, the latter information is derived automatically from the scheme file [4]. Thus the production of the scheme file is by far the biggest hurdle in the development of a diphone set for a new language or accent, as it requires considerable knowledge of the phonological system of the language. In the case of North Welsh, it would be possible to add the extra diphones manually to produce a new scheme file for a North Welsh accent, since the vowels specific to North Welsh accents are additional to those found in South Welsh accents. Thus adaptation should not require re-running the full process.

5 INSIGHTS GAINED ON DATABASE DESIGN

Nonsense word speech databases for diphone extraction have to date been generated by means of a program tailored to the requirements of the particular language. This necessitates a certain amount of knowledge on the part of the programmer, concerning the phonological system and phonotactic constraints of the language. Naturally, it would be easier if the process of database design could be automated, at least to some extent. It might be that languages differ too much for this procedure ever to be fully automated, but at least in the case of non-tone languages there are certain general requirements that can be distinguished, and these are outlined below. Any algorithm for database design that might be produced should have the capacity to accept values of these parameters as input.

5.1 Consonant Distribution

To begin with, it is necessary to know which consonants are syllable-initial only, and which (if any) are syllable-final only. In this respect, consonants which belong overwhelmingly to one category are to be treated as belonging exclusively to that category, in order to preserve naturalness in the output (eg. voiceless stops in Welsh seldom end a syllable).

5.2 Positional Allophones of Consonants

It is also necessary to know the syllabic contexts in which positional allophones of phonemes are to be expected, and also which phonemes are principally affected (eg. for Welsh, voiceless stops in syllable-initial, syllable-final, and post-/s/ contexts). It is not necessary to know the precise acoustic details of the differing realisations, as it would be if formant synthesis were being attempted: merely the contexts are needed. It is also not necessary to know the acoustic details of intrinsic allophones (ie. coarticulation) such as the n-p diphone (probably realised phonetically as [m p]) since these are automatically produced by the speaker without explicit specification.

5.3 Vowel Distribution

It is necessary to know the distribution of vowels with respect to stressed and unstressed syllables (eg. in English, schwa cannot appear in stressed syllables), and also with respect to syllable position in the word (eg. in Welsh, schwa may appear in stressed penults, but cannot appear in stressed monosyllables or in final syllables of polysyllables). This parameter, unlike the previous two, is likely to have a far-reaching effect on the appearance of the nonsense material, as (among other things) it can determine whether one or more nonsense words are used per item.

5.4 Vowel Length

An important factor is the possible interaction between the phonological length of monophthongs and the type and number of any following consonants. In English, there is no such interaction,except possibly in the case of /ng/ (which must be syllable-final and cannot be preceded by a phonologically long vowel). In Welsh, there is extensive interaction, and this has implications for the location of vowel phonemes with respect to word boundaries, when before certain consonants.

5.5 Fixed/Free Stress

The question of whether word stress is fixed or free is another important consideration. In English, word-level stress is free, and can in principle fall on any syllable of a polysyllable. In Welsh, it is fixed, and falls on the penultimate syllable of polysyllables (irregular cases will be ignored for the purposes of database design, in order to enhance naturalness). This fact has implications for vowel diphones with initial or final silence (eg. #-a or a-#, utterance-initial and utterance-final diphones respectively), since it is preferable to locate these vowels in an acoustically salient syllable (in order to obtain as much acoustic vowel information as possible). Therefore, in the case of #-a, the relevant vowel must fall in a penult or stressed monosyllable (rather than an antepenult) eg. *at yma*, /a* t @* m a/, while in the case of a-#, the relevant vowel must fall in an ultima or stressed monosyllable. Since, in Welsh, the ultima is of considerable acoustic salience, it is sufficient to use this (phonologically) unstressed ultima, so a suitable nonsense phrase for a-# is *yn doda*, /@ n d oo* d a/. In fact this (ultima) form is the only possible one in the case of /a/ (which is phonologically short), since orthographic 'a' at the end of a stressed monosyllable would be pronounced as phonologically long /aa*/.

5.6 Accent Variation

It would also be desirable to have a means of mapping between different accents of the same language, so that databases may easily be designed for related accents. Therefore information on additional vowels and consonants (or segments to be deleted) for other accents should also be available.

6 ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of a BP Research Fellowship, awarded by the Royal Society of Edinburgh and funded by British Petroleum. Thanks are also due to Paul Taylor and Ian Nairn of the CSTR, for insight into the workings of an English diphone synthesiser.

7 REFERENCES

[1] J ALLEN, M S HUNNICUTT & D KLATT (1987) *From text to speech: The MITalk system.* Cambridge: Cambridge University Press.
[2] S D ISARD & D A MILLER (1986) `Diphone synthesis techniques'. *IEE Conference Publication no. 258:* 77-82.
[3] P A TAYLOR & S D ISARD (1991) `Automatic diphone segmentation', *Proceedings of the Second European Conference on Speech Communication and Technology (Eurospeech 91)*, September 24-26 1991, Genova, Italy, vol. 2: 709-711.
[4] P TAYLOR (1992) `The Osprey Speech Synthesis System'. Unpublished MS.
[5] G M AWBERY (1984) `Phonotactic Constraints in Welsh', In [6].
[6] M J BALL & G E JONES (1984) *Welsh Phonology.* Cardiff: University of Wales Press.
[7] B J WILLIAMS (1989) `Stress in Modern Welsh'. Distributed by Indiana University Linguistics Club.
[8] G E JONES (1984) `The Distinctive Vowels and Consonants in Welsh'. In [6].