A WELSH SPEECH DATABASE: PRELIMINARY RESULTS

Briony Williams
CSTR, University of Edinburgh
80 South Bridge,
Edinburgh EH1 1HN, Scotland, UK.
briony@cstr.ed.ac.uk
http://www.cstr.ed.ac.uk/~briony/

ABSTRACT

A speech database for Welsh was recorded in a studio from read text by a few speakers. The purpose is to investigate the acoustic characteristics of Welsh speech sounds and prosody. It can also serve as a resource for future work in speech synthesis and recognition.

The speech is labelled by hand at the acoustic phonetic level, and labelled semi-automatically at the phoneme, syllable, and word levels. Statistical analysis of the database resolves some long-standing questions in the phonetics of Welsh stress, and yields more data on Welsh speech sounds. The overall procedure could be useful for work on other minority languages where very little basic acoustic phonetic research has yet been done.

Keywords: database, phonetics, Welsh

1. INTRODUCTION

Welsh is a minority language spoken in Wales. In the 1991 census, about 500,000 people claimed to speak Welsh (19% of the population). There has been less research on Welsh than on the major languages, and few examples of speech or language technology.

A prerequisite for development in speech technology is a labelled speech database. This is also needed in basic phonetic research into the characteristics of the sounds and prosody of a language. Most corpora collected for Welsh are not machine-readable, and not segmented. A recent example is [1], where speech from five aged near-monoglot speakers of Welsh was transcribed from tape orthographically. Another recent example of a speech database for Welsh, SpeechDat Cymru [2], is segmented both phonemically and orthographically. However, it is based on telephone speech, and so is less suitable for use in deriving basic phonetic information, due to the low quality of the signal over the telephone.

The database described here, although it is almost conteporaneous with SpeechDat Cymru, concentrates on high-quality recorded speech using lengthy read texts, for the purposes of basic phonetic research. It could also serve as a seed database for the initial training of HMMs in speech recognition work, as well as a source of speech units for concatenative speech synthesis.

2. SCRIPT AND RECORDING

The recording script was made up from texts taken from various Welsh periodicals, as follows:

- *Y Cymro* (a newspaper with a very formal style): one long text.
- *Golwg* (a magazine on arts/culture/current affairs, with an informal style): 15 texts.
- *Y Wawr* (a traditional women's magazine, with a medium formal style): 6 texts.

The speakers were selected through personal contacts, with the aim of covering different accent areas of Wales. So far, three speakers have been recorded. Recordings took place in a recording studio at 16 kHz sampling rate, 16-bit quantisation, in ESPS sound file format.

3. ANNOTATION

3.1 Acoustic phonetic level

The speech files were labelled by hand at the acoustic phonetic level, using the ESPS "xwaves" software. The characteristics of elements at this level were as follows:

1) The closure, burst, and aspiration phases of stops were separately labelled.
2) The closure, burst and approximation phases of alveolar trills were separately labelled.
3) The "voiceless nasals" usually had separate phases: voiced nasal and voiceless glottal fricative.
4) Partially-aspirated sections of vowels were labelled as "voiced [h]" at this level. These can occur at the edge of a vowel when next to a voiceless consonant.
5) Phonologically geminate continuants were *not* split, given the absence of acoustic grounds for doing so.
6) The two elements of diphthongs were separately labelled in those cases where there were clear and fast-moving formant transitions.
7) Stretches of dysfluent speech, or of silence, were marked as such. They were subsequently ignored.
8) Most monophthongs (those without a diacritic in the orthography) were not marked for phonological length, even though this affects not only duration but also formant frequencies. Phonological length was considered to be conditioned by higher-level linguistic context in nearly all cases (see [3]).

3.2 Higher linguistic levels

Most of the higher linguistic levels were then labelled semi-automatically, using rule-based tree-building and manual post-editing. The database thus becomes an example of the kind of multi-tiered database advocated in [4]. The software used to accomplish this was an extension of the EMU speech database labelling software [5]. EMU is available at no charge from [6].

3.2.1 Phonemic level

The phonemic level was built up using phonetic-to-phonemic rules. The mapping could be many-to-many. In each rule, the initial string specified the input, and the second string the output, as in the following:

- {{<s=stop> q h} {<s>}} A seqence of "stop closure" (defined elsewhere as one of [p,t,k,b,d,g]), "release burst" ([q]) and "aspiration" is mapped to the appropriate stop phoneme.
- {{t sh} {ch}} The sequence of [t] and [sh] is mapped to the affricate phoneme /ch/.
- {{a} {cr}} The sequence of [a] plus the creaked segment [cr] is mapped to the vowel phoneme /a/.

After all rules had applied to the utterance, a certain amount of post-editing was needed in most cases. For example, an extra segment was added at the phonemic level in the case of geminate consonants: this is one case where a single phonetic segment could have two phonemic "parents". However, despite the need for post-editing, the overall procedure was considerably quicker than fully manual labelling at the phonemic level.

3.2.2 LexPhon level

The LexPhon level was largely a convenient fiction to ensure that each word took the same phonemic form every time, for use by the syllable-building rules. The LexPhon rules mostly copied the phoneme segments. Some post-editing was required, especially in cases of assimilation (where the LexPhon level shows the form *before* assimilation). Other cases concerned dialectal variation. For example, the plural suffix "*-au*" for nouns is usually pronounced as /e/ in South Welsh accents and /a/ in North Welsh accents in an informal style, while in a formal style all accents would pronounce it as /ai/ (North) or /ai/ (South). At the LexPhon level, therefore, the representation of this morpheme was coerced to /ai/, whatever the form that was actually produced.

3.2.3 Syllable level

The next rules built syllable-level elements from the LexPhon elements. This set of rules was specific to each text, being a form of syllable dictionary containing only those syllables that occur in the text. This was in order to restrict the number of syllables in the dictionary so as to improve the accuracy of the syllabic parsing.

The text was initially passed through a set of existing letter-to-sound rules for Welsh (see [7]). The output strings were passed through syllabification rules, to obtain a list of unique syllables for that text. Syllables beginning in a given phoneme were ordered with the longest string first, so that syllable-final consonants were not missed in the parsing process. The syllable-level labels were a simple concatenation of LexPhon units separated by hyphens, as follows:

- {{a r n} {a-r-n}}
- {{g w ai th} {g-w-ai-th}}

Post-editing at the syllable level was the most extensive of all the levels. This was because consonants would often need to be reassigned to the preceding syllable, due to an intervening word boundary (which, of course, the rules had no information about at this stage).

Other labels were associated with the Syllable level:

- **PitchAcc**: For manual labelling of the pitch accent.
- **SyllStat**: For labelling of the syllable's position in the word. In polysyllabic words, this is done by hand, labelling the ultima (final syllable), penult (penultimate), antepenult, and preantepenult. All remaining syllables are monosyllabic words, and a small C program adds "mono" labels afterwards.

3.2.4 Word level

The next set of rules mapped syllables to words. These rules were likewise based on a text-specific dictionary containing rules such as the following:

- {{d-i-r w-e-dh} {dirwedd}}
- {{dh-e m-o k-r-a t-ai-dh} {ddemocrataidd}}

Little or no post-editing was required at this level.

3.2.5 Stress level

The labels at this level applied to word units, and were assigned automatically by rules based on a text-specific dictionary. Labels indicated both the word stress pattern and the content/function word type. Examples are:

- PenC: Content word stressed on the penult.
- MFun Monosyllabic unstressable function word.
- MCep Monosyllabic content word with epenthetic vowel. Hand-edited to MCep1 or MCep2, according to whether or not the epenthetic vowel was present, forming a second syllable.

Little or no post-editing was required at this level.

3.2.6 Intermediate, Intonational and Utterance levels

The prosodic levels (for intermediate and intonational phrases) require manual labelling. The "Utterance" level is a formality, gathering all Intonational level units into one, and is performed automatically.

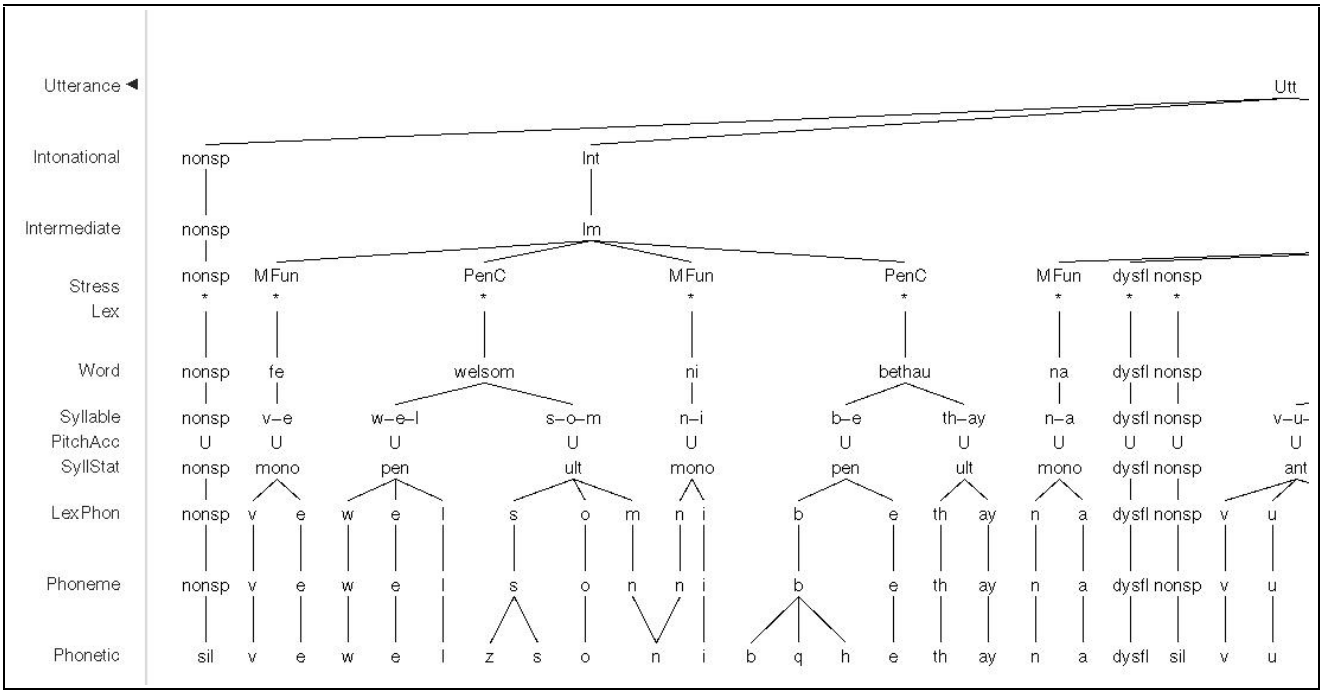

Figure 1: Hierarchical labelling of a speech file at several linguistic levels

3.2.7 Example of levels

Figure 1 shows a partially-annotated speech file (it still lacks PitchAcc labels). The following features are seen:

- At the phonetic level, a voiced [z] and a following voiceless [s] are subsumed into a single voiceless /s/ at the phonemic level above.
- At the phonetic level, a sequence of [b]-closure, release burst ([q] here) and aspiration ([h]) is subsumed into a /b/ phoneme at the phonemic level.
- On the other hand, a single [n] at the phonetic level is given two /n/ parents at the phonemic level, as this is a geminate nasal at a word boundary.
- The first of these two /n/ phonemes is marked as the /m/ unit at the LexPhon level. This corresponds to a slightly different verbal inflection, and is the form of this word used in the text-specific dictionary.

4. QUERYING AND ANALYSIS

4.1 Database querying

The EMU software is equipped with a database query language. Queries were formulated on questions of Welsh phonetics. For example, an auditory study ([8]) had stated that a voiceless stop after a stressed penult vowel was longer than a similar consonant after a vowel in any other syllable. This claim was later supported by acoustic measurements taken of all consonant types in the relevant context [9]. The relevant query in the EMU database query language was as follows:

[[Phoneme=vowel] -> [[[#Phoneme=cons -> Phoneme=vowel] ^ SyllStat=ult] ^ Stress=PenC|PenF]]

This query outputs a list of all phonemic consonants that are followed by a vowel, in the case where both segments are dominated by an ultima syllable which is in turn dominated by a word with either PenC or PenF stress (i.e. it is stressed on the penult) and where the consonant of interest is preceded by a vowel.

Similar queries were formulated for consonants after other syllable types for purposes of comparison. In this preliminary exercise, the queries were run over a very small amount of data from one speaker only (a female speaker with a North Welsh accent).

The output takes the form of a list in which each row contains the segment label, segment start and end times, and utterance label. This list was used as the input to the Splus statistics software, in order to derive descriptive statistics on the duration of the given consonants.

4.2 Statistical analysis

4.2.1 Post-penult consonant duration

In the example above, running t-tests over consonants in various contexts showed that the post-stressed-penult consonant did indeed show greater duration, the mean being 82 ms. This was significantly longer ($p < 0.01$) than all other cases, which showed no significant differences among themselves. The other contexts were: post-antepenult (72 ms), post-preantepenult (63 ms), word-inital (69 ms) and -final (66 ms), monosyllable-initial (63 ms) and -final (64 ms).

4.2.2 Vowel duration in stressed penults

A more interesting query concerned the duration of the vowel in a stressed penult. Welsh phonology is unusual in that the intrinsically shortest vowel, schwa (/ə/), may appear in a stressed syllable in the case of polysyllabic words. Since polysyllables are regularly stressed on the penult, the curious outcome is that stress often lies on a syllable which is shorter than the following, supposedly unstressed, final syllable.

A query was formulated that would factor out any artefacts of vowel distribution. Phonologically long vowels occur only in stressed syllables, whereas short vowels may occur in both stressed and unstressed syllables. Therefore only short vowels were considered. In addition, schwa was excluded from the analysis, as this is the only short vowel with a restricted distribution (non-final syllables only). The remaining vowels were: /a, e, i, o, u, ɨ/ (where /ɨ/ is the high central unrounded vowel). Lists of these segments were obtained for several syllabic contexts, and t-tests were run over the segment durations. The results are in Table 1 below.

Vowel context	Mean duration
Stressed monosyllable	124 ms
Unstressed ultima	92 ms
Unstressed monosyllable; Stressed penult; Unstressed antepenult or preantepenult	80 ms 77 ms 75 ms

Table 1: Duration of non-schwa short vowels

Phonologically short non-schwa vowels were longest in stressed monosyllables. This duration was significantly longer ($p<0.01$) than that of vowels in unstressed ultimas, which in turn was longer ($p<0.01$) than that of all other such vowels. Thus, stress does not appear to be a factor in duration for these vowels, and the only relevant factor appears to be word-final lengthening.

4.2.3 Other searches

Duration is not the only acoustic parameter that may be analysed. It is also possible to investigate the fundamental frequency, formant frequencies and bandwidths, and energy of the signal. This is done by first processing the speech files to obtain the necessary parameter tracks (e.g. using ESPS). Database querying is then carried out as before, and an EMU command ("get_track") is used to obtain the parameter values for the specific segments indicated. These values can then be subjected to statistical analysis in the same way as indicated for duration values. In addition, given the full labelling, not only phonemes may be investigated, but also higher-level units such as syllables, words, or intonational phrases.

5. CONCLUSIONS

A method has been presented for more rapid processing of speech databases by using a semi-automatic rule-based system for transcribing at higher linguistic levels. The annotated database can then be subjected to searches according to various lingusitic criteria, and the results of those searches can be analysed statistically. Some preliminary examples have been given of how this methodology may be used in the phonetic description of a language. Future work will concentrate on labelling more data from further speakers, and running other phonetic queries and analyses over the labelled data.

6. REFERENCES

[1] Ahmad, K. and A. Davies (1993), The Creation of a Corpus of Spoken Welsh. University of Surrey, Dept. of Mathematical and Computing Sciences, Computing Sciences Report.

[2] SpeechDat Cymru: http://galilee.swan.ac.uk/home/SDCymru/welcome.html

[3] Awbery, G.M. (1984) Phonotactic Constraints in Welsh. In: M.J. Ball and G.E. Jones, (eds.) *Welsh Phonology*. Cardiff: University of Wales Press.

[4] Barry, W.J. & A.J. Fourcin (1992) Levels of labelling. *Computer Speech and Language*, **6**, pp. 1-14.

[5] Cassidy, S. and J. Harrington (1996), EMU: an Enhanced Hierarchical Speech Data Management System. Speech Science and Technology Conference 1996 (SST96), Adelaide, Australia.

[6] EMU software: http://www.shlrc.mq.edu.au/emu

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

[8] Jones, R. (1967) A structural phonological analysis and comparison of three Welsh dialects. MA thesis, University College of North Wales, Bangor.

[9] Williams, B. (1985) Pitch and duration in Welsh stress perception: the implications for intonation. *Journal of Phonetics*, **13**, pp. 381-406.