Levels of Annotation for a Welsh Speech Database for Phonetic Research

Briony Williams
CSTR, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland, UK
[briony@cstr.ed.ac.uk]

Abstract

A Welsh speech database intended for use in phonetic research requires careful annotation at several linguistic levels. The initial stage is that of labelling at the acoustic phonetic level, where the closure, burst and aspiration phases of a stop consonant are all separately labelled. The next is the phonemic stage, which can be derived from the former in most cases. Next is the syllabic stage, where each syllable is labelled in terms of its word status and in terms of lexical stress. The final stage is the lexical stage, where each word is labelled according to its word class. A statistical package can then be run over this data to yield information on the acoustic characteristics of Welsh speech sounds, and also about the nature of lexical stress in Welsh. In addition, it is hoped to derive rules for intonation patterns for use in an existing Welsh text-to-speech synthesiser.

1 Background

The Welsh language is spoken in many parts of the principality of Wales, within the United Kingdom. It belongs to the Celtic language family, which in turn forms part of the Indo-European group of languages. It is related to Breton and Cornish, and more distantly to Irish and Scottish Gaelic.

Welsh as a separate language dates from about the late 6th century AD. It was originally spoken throughout Wales and much of northern Britain, but has declined in use, especially over the first part of the present century. However, there are indications of a recent growth in the number of Welsh speakers. The 1991 Census showed that 508,000 people (18.7% of the population) claimed to be able to speak Welsh. More importantly for the future of the language, the proportion of Welsh speakers among young people is growing, and its use is increasing in the mass media (especially television).

1.1 Motivation for an annotated speech database

Compared to the major languages of Europe, little acoustic phonetic research has been carried out on Welsh. The SpeechDat project is making recordings of a limited subset of Welsh words over telephone lines (Jones et al., 1998), but SpeechDat databases are intended for use in developing voice-driven teleservice products, rather than for basic phonetic research (Tropf, 1997). The work described in this paper concerns a speech database with a much wider vocabulary, recorded in a recording studio, and labelled by hand, which is to be used for purposes of phonetic research. It is hoped that the database will also yield intonation rules for use in an existing Welsh text-to-speech synthesiser (Williams, 1994).

1.2 Characteristics of the recorded speech

The speech to be recorded (from a small number of speakers) will have the following characteristics:

- **Monologue** rather than dialogue. This yields non-overlapping clear speech where individual segments can be clearly located and accurately measured.
- **Continuous speech** rather than isolated words. This will yield information on the acoustic characteristics of the intonation patterns of running speech.
- **Recording studio** environment rather than telephone line or office environment. This will yield a clear acoustic signal that provides the maximum amount of acoustic phonetic information.
- **Read** speech rather than spontaneous. Reading aloud written texts will minimise the number of hesitation pauses and slips, ensuring that the intonation is as close as possible to its intended form.
- **Manual** segmentation rather than automatic. This yields accurate segment boundaries that can be relied on in the subsequent statistical analysis of duration. They can also be used to train an automatic segmenter for future work, if desired.
- The six **speakers** to be recorded will be divided between the two sexes, and also between the six major accent areas of Wales. They will be neither children nor very old, in order to avoid problems connected with vocal defects or extremely high F0.

The above considerations are based on the discussion of spoken corpus design in ch. 3 of Gibbon et al. (1997).

1.3 Choice of linguistic style

The choice of text style is not as straightforward for Welsh as for English. Welsh has two distinct varieties:

- The literary language, used in formal writing and in formal public speech, but not in daily conversation. It is considered somewhat old-fashioned.
- The spoken language, which differs markedly from the former in grammar, morphology and vocabulary. It has some comparatively minor regional variants, and can also vary a little in its level of formality.

It was decided to use the spoken variety, with a minimum of regionalisms and of the contractions used in very informal registers. This style is used by magazines such as “Golwg”, a popular current affairs magazine, and the recording texts have been taken largely from issues of “Golwg”. This style probably resembles the variety that would be used for human interaction with computers. The corpus is intended to yield insights for use in speech technology: hence this style was chosen.

Twenty texts have been selected, making a total of about 25,000 words. Most are from “Golwg”, while others are taken from “Y Wawr” (“The Dawn”), the magazine of the Welsh-speaking Women’s Institute “Merched y Wawr”. These are likewise in a popular style.

2 Annotation levels

The levels of labelling are based on some of those described in Barry and Fourcin (1992). They are exemplified in Figure 1, which shows a recording of the sentence "Mae'n cael cwpanaid o goffi cyn mynd i'r gwely" ("He has a cup of coffee before going to bed"), recorded by a male speaker from South Wales. The figure shows the waveform, the orthographic form, and various levels of segmentation and labelling.

2.1 Acoustic phonetic

The acoustic phonetic level of labelling concerns events describable in terms of such phonetic descriptors as: Release burst; Fricative noise; Nasal. These labels make no claims about the linguistic function or distinctiveness of the segments, but refer to events identifiable in acoustic terms alone. For example, separate phases of a stop (closure, release burst, and aspiration) are labelled separately. In Fig. 1, these phases are labelled using a symbol showing the phone type (p,t,k,b,d,g) followed by the phase label (c=closure, b=burst, h=aspiration).

2.2 Phonemic

The phonemic level shows the individual phonemes, with acoustic events grouped into a unit corresponding to a linguistically meaningful segment. For example, the separate acoustic phases of a stop are here grouped into one segment, as in Fig. 1. Also, an assimilated nasal is shown as the nasal in its lexical form. In Fig. 1, this is seen in the case of the first /n/, shown on the acoustic phonetic level as /ng/ as it is assimilated before /k/.

2.3 Syllabic

On this level, phonemes are grouped into syllables. In Fig. 1, the syllables are labelled as in Table 1 below.

Symbol	Meaning
m	Unstressed monosyllabic word
M	Stressed monosyllabic word
a	Unstressed antepenultimate syllable
P	Stressed penult syllable
u	Unstressed ultima syllable

Table 1: Symbols used for labelling at the syllabic level

2.4 Lexical

At the lexical level, syllables are grouped into units the size of orthographic words, labelled according to part of speech. The labels used in Fig. 1 are as in Table 2 below.

Symbol	Meaning
BOD	verb "to be" ("*bod*")
-p	(infixed) particle
VN	verb-noun (uninflected verb, also noun)
N	noun
P	preposition
A	adjective

Table 2: Symbols used for labelling at the lexical level

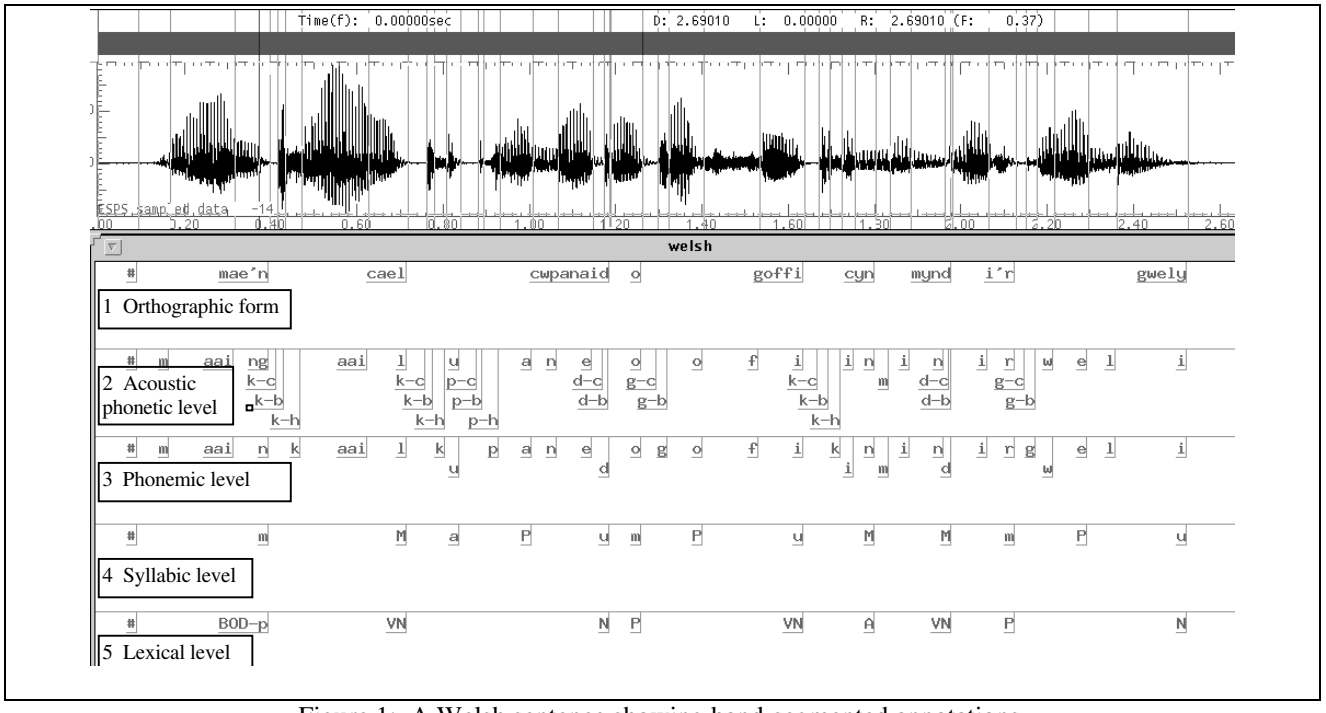

Figure 1: A Welsh sentence showing hand-segmented annotations.

2.5 Intonation

The intonation of each utterance will also be labelled, using the widespread "ToBI" system of intonation labelling (Pitrelli *et al*., 1994)[1]. This system, originally developed for American English intonation, has proved easy to adapt for use in describing the intonation of other varieties of English. It will be necessary to adapt it more radically for describing the intonation of Welsh, which differs greatly from that of most varieties of English (see Thomas (1967) for an auditory analysis using a different kind of framework). A first attempt at adapting the system in this way has been made by Evans (1997), using spontaneous speech elicited during the "Map Task".

3 Questions to be investigated

The annotated corpus will be analysed statistically, using such acoustic parameters as duration, fundamental frequency, and formant frequncies. It is hoped that the analysis will shed light on certain long-standing questions concerning Welsh, as follows.

3.1 Acoustic characteristics of speech sounds

A certain amount of basic research has been carried out already on the acoustic characteristics of Welsh speech sounds (see Ball (1984) for an intial survey). However, a great deal remains to be done, and this type of speech corpus (clear speech and a high-quality recording) is the most appropriate kind for this type of investigation. Basic data will be gathered on such topics as vowel durations in given linguistic contexts, and segment formant frequencies and bandwidths.

3.2 Welsh lexical stress

The acoustic characteristics of lexical stress in Welsh have long been an interesting research question. Williams (1985) found that word-level stress in Welsh was dependent not so much on increased duration or F0 in the stressed syllable (as in English), but rather on such indirect cues as the longer duration of the post-stress consonant, the functioning of the stressed syllable in the rhythmic unit, and the interaction of lexical stress placement with the intonation pattern.

In that case, only stressed penultimate syllables were considered, and stressed monosyllables were disregarded. Since stressed monosyllables in Welsh do indeed show greater duration and intensity than unstressed syllables (as do irregularly-stressed ultimas) it is necessary to include these syllable types also in a new study of Welsh lexical stress, in order to determine common factors that hold across all syllable types. In addition, the phonotactic constraints of Welsh appear to be that the shortest vowel, schwa, is never found in word-final syllables of polysyllabic words. This skewed distribution pattern may have conditioned the previous finding (Williams, 1985) that stressed syllables (nearly always the penult) actually showed *shorter* mean duration than unstressed syllables. In order to determine whether this finding was merely an artefact of vowel distribution patterns, it is necessary first to factor out the effects of phonological constraints on vowel distribution across different syllable types.

This will be done by making measurements directly of, for example, the mean duration of a phonologically short /e/ in an unstressed antepenult, unstressed ultima, and stressed penult. The results will resolve the question of whether the anomalous patterning of vowel duration and stress is merely an artefact of vowel distribution, or is a real reflection of the effects of word-level stress.

3.3 Intonation patterns

As indicated, the intonation patterns in the corpus will be transcribed manually, using the ToBI system. It will then be possible to study such questions as the following:

- The relation between the location of the highest F0 peak and the corresponding accented syllable.
- The relation between simple (H*, L*) and complex accents (L+H*, etc), and the effects on the direction of F0 change within the accented syllable.
- The frequency of the various accent types, and any co-occurrence restrictions there might be.

4 Potential speech technology applications

Although the annotated corpus is primarily intended as a resource for basic phonetic and phonological research, it will also form a useful resource in any future projects developing speech technology applications for Welsh.

4.1 Text-to-speech synthesis

A diphone-based text-to-speech synthesiser for Welsh already exists (Williams, 1994). However, the duration rules used are minimal in the extreme, and the intonation rules do not differ from the English system on which it was based. Therefore, there is a need to extract from the corpus rules for prosody for use in the synthesiser. It is hoped to derive the following kinds of information:

- Rules for the duration of segments in given linguistic contexts. This will be in terms of statistical "z-scores" related to the mean duration of each phoneme type, as in Campbell & Isard (1992).
- F0 frequencies for the intonation maximum and minimum frequency, for each sex.
- Characteristic accent types to use for declarative sentences, yes/no questions, and wh-questions.
- The default relationship between a H* accent and the corresponding peak in F0.

4.2 Speech recognition

The corpus will also be of use for the training and testing of a set of Hidden Markov Models. These models can then be used in the initial automatic segmentation of a future corpus of spontaneous or telephone-based speech, for use in developing a speech recogniser. This, however, would be a separate research project in its own right, requiring as it does a completely different kind of database to reflect the kind of speech signal that a speech recogniser is likely to encounter.

[1] Information on ToBI is available at the following URL: http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html

5 Acknowledgements

This work is supported by research grant no. GR/K84356 to Mr. S. Isard, University of Edinburgh, from the UK's Engineering and Physical Sciences Research Council.

6 References

Ball, M.J. (1984) Phonetics for Phonology. In: Ball, M.J. & Jones, G.E. (eds.) *Welsh Phonology*. Cardiff: University of Wales Press.

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

Campbell, W.N. & Isard, S.D. (1992) Segment durations in a syllable frame. *Journal of Phonetics*, 19: 37-47

Evans, C.L. (1997) "C_ToBI: Towards a system for the prosodic transcription of Welsh". Unpublished MSc dissertation, Dept. of Linguistics, Univ. of Edinburgh.

Gibbon, D., Moore, R. & Winski, R. (1997) *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Jones, R.J., Mason, J.S., Jones, R.O., Helliker, L. & Pawlewski, M. (1998) SpeechDat Cymru: A Large-Scale Welsh Telephony Database. Proceedings of the Workshop on "Language Resources for European Minority Languages", May 27th 1998, Granada, Spain.

Pitrelli, J.F., Beckman, M.E., & Hirschberg, J. (1994) Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, Japan. Vol. 1: 123-126.

Thomas, C.H. (1967) Welsh intonation — a preliminary study. *Studia Celtica*, 2: 8-28.

Tropf, H. (1997) SpeechDat: European Speech Databases for Creation of Voice-Driven Teleservices. *ELRA Newsletter*, March 1997, pp. 9-10.

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

Williams, B. (1994) Diphone Synthesis for the Welsh Language. *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, Japan. Vol. 2: 739-742.