Speech technology and Argentinean Welsh

Elise Bell
The University of Texas at El Paso
eabell2@utep.edu

Abstract

This paper argues for increased efforts to source Welsh language data from the population of Welsh speakers in Argentina. The dialect of Argentinean Welsh is under-resourced even in comparison to other Celtic languages, which are already considered less-resourced languages (LRLs). Argentinean Welsh has been shown to diverge from other dialects of Welsh in the realization of acoustic contrasts such as voice-onset time and vowel duration. These differences potentially obscure phonemic contrasts in the language, creating homophony absent in other dialects. The inclusion of Argentinean Welsh data in training sets for future Welsh speech technology development will increase the applicability of such technology to other speaker communities whose Welsh speech may not align with that currently in use for model training, including second-language and non-fluent speakers.

1 Introduction

The development of speech language technology such as automatic speech recognition (ASR) depends on the availability and accessibility of large-scale language data sets, both spoken and written. The information in these data sets is used to create statistical generalizations that form the basis for speech technologies including speech recognition, text-to-speech systems, and grammatical parsing. Large resources of this type are less available for under-resourced languages, including Welsh and other Celtic languages, making creation of speech technologies for these languages more challenging. As we undertake that challenge, it is vital that we consider the source of the data on which our technology is based. As less-resourced language speech technology becomes more broadly accessible, speakers who deviate from the norms explicitly or implicitly assumed by the technology will begin to come in contact with it. Depending on the variety inherent in the data underlying the system, those marginalized speakers may or may not be able to successfully take advantage of speech technology. The aim of this paper is to highlight the particular areas of speech technology development that may create obstacles or pose problems for users, and to propose the addition of a particular source of acoustic data that lies outside the norm for Welsh language technology. The main speaker group of concern here is speakers of Welsh in Argentina, but the arguments that follow apply to second language (L2) or non-fluent speakers of Welsh as well.

Compared to dialects of Welsh spoken in Wales, Argentinean (or Patagonian) Welsh is extremely under-resourced and under-researched. Documentation efforts amount to a handful of citations (Jones, 1984; Jones, 1998; Sleeper, 2015; Bell, 2017), and to my knowledge, only one speech corpus. Little is known about how the dialect of Welsh spoken in Argentina differs from other dialects of Welsh, although there are several reasons to expect dialectal variation. The effects of bilingualism on speech production are well documented (Flege et al., 1997; Flege et al., 2003; Escudero, 2009), and all adult speakers of Argentinean Welsh are bilingual with Spanish (if not trilingual with English or another language). Dialect differences may also

arise from the effects of second language (L2) acquisition of Welsh. Differences in speech production due to these effects may include the merging of phonemic categories, or the use of different acoustic cues in contrast production. Because these effects are fairly inextricably tied up with the effect of Spanish bilingualism on Argentinean Welsh in general, they will not be treated separately here. This paper presents a brief overview of the state of Welsh language speech technology and resources, followed by a short discussion of the history and modern context of the Welsh language in Argentina. Subsequently, I present evidence that experimentally observed differences between Welsh dialects support the inclusion of Argentinean Welsh data in future speech technology development efforts.

1.1 Welsh speech technology

Speech recognition and speech synthesis technologies rely on (relatively) large amounts of acoustic data, which must be transcribed orthographically (in the case of a grapheme-based speech recognition system) or phonetically (in the case of a phoneme-based system). The collection, analysis, and processing of this data requires resources including people-hours, funding, and often, participation of community members in data crowdsourcing efforts (Prys and Jones, 2018). Currently, available Welsh speech technology is fairly limited (compared to larger-resourced languages like English). Much of what is available has been produced by the Welsh Language Technologies Unit, based at Bangor University.[1] Tools produced by the Language Technologies Unit range from front-end resources accessible to the public (a vocabulary website plugin (Jones et al., 2016), a Welsh language spelling and grammar checker (Prys et al., 2016)) to back-end tools such as a part-of-speech tagger (Prys and Jones, 2015) that are open source and accessible to researchers outside of the unit itself. The unit has also developed speech recognition and synthetic speech technologies that are of particular relevance to this paper. These include the development of Macsen,[2] a Welsh-language personal digital assistant based on data collected by the Languages Technologies Unit using the Paldaruo app[3] and website to crowdsource the collection of Welsh utterances (Prys and Jones, 2018). Utterances were elicited with a set of target words and sentences designed to collect a representative phoneme set. The project is currently available through the Mozilla Common-Voice project[4] where users can contribute and evaluate recordings, and where a portion of the vetted data is available for download.

There are several other text and speech corpora available for the Welsh language. The Language Technologies Unit has created multiple text corpora, including one of social media posts[5] as well as a million-word corpus consisting of various registers of Welsh writing.[6] Researchers at Bangor University's ESRC Centre for Research on Bilingualism in Theory & Practice have also produced two publicly available corpora of Welsh bilingual speech.[7] One of these, the Patagonia corpus, is to my knowledge the only publicly available collection of Argentinean Welsh speech. While such conversational corpora are invaluable for the study of syntactic and morphological phenomena (Carter et al., 2010; Webb-Davies, 2016), the acoustic data they contain is not always of high enough quality, nor is the corpus large enough, to stand alone as the sole source for development of speech technology. A brief history of Welsh in Argentina is presented below, followed by a discussion of the benefits that Argentinean Welsh data may have on future development of Welsh speech technology.

1.2 Argentinean Welsh

The presence of Welsh in Argentina is due to the mid-19th century efforts of a group of Welsh speakers, led by Michael D. Jones, who sought to establish a Welsh colony away from the influence of the English language and British government (Williams, 1975). In 1865, following an agreement with the government of Argentina, a Welsh colony was established in the Patagonia region of of the country. Today, descendants of the original 200 colonists (and of the several thousand who followed in subsequent years) still maintain the Welsh language and culture in Argentina. Modern Welsh speakers are clustered in two areas of

[1] www.bangor.ac.uk/canolfanbedwyr/technolegau_iaith.php.en
[2] http://techiaith.cymru/2016/05/introducing-macsen
[3] https://apps.apple.com/bs/app/paldaruo/id840185808
[4] https://voice.mozilla.org
[5] http://techiaith.cymru/data/corpora/twitter
[6] http://corpws.cymru/ceg/
[7] http://bangortalk.org.uk/

Chubut Province, in Dyffryn Camwy on the Atlantic coast, and Cwm Hyfryd to the west.

Although inter-generational transmission of the language waned during the 20th century, revitalization efforts were spurred in 1965, the centennial of the original colony's establishment. The centennial celebration renewed interest in Welsh culture and language, and by the 1990s several language initiatives were established which still exist today. These include Welsh language medium primary schools, annual Eisteddfodau (traditional poetry and song competitions), and an ongoing teacher exchange program with Wales through the Welsh Language Project.[8]

2 Discussion

Argentinean Welsh is spoken by a population that is separated from Wales by more than a century of sparse contact as well as a language barrier (bilingualism with Spanish, rather than English). These factors have almost certainly contributed to linguistic divergence in many aspects of Argentinean Welsh. The most salient of these aspects for the purpose of this paper is divergence in the language's sound system, in the acoustic realization and phonological representation of speech sounds. Previous research on speech recognition of dialect and accent differences has shown that, given a large enough data set, systems trained on multiple dialects perform better than those trained on a single dialect (Rao and Sak, 2017; Li et al., 2018; Yang et al., 2018). Other work has found that including accent classification when training a multi-accent speech recognition system improved later classification of both accent-classified and accent-unclassified datasets (Jain et al., 2018). Before addressing specific evidence for phonetic and phonological differences between Argentinean Welsh and other dialects of Welsh, the next section discusses the reasoning for including dialectal variation in speech technology models.

The results of previous research indicate that the inclusion of dialectal acoustic variation can provide a more variable and more useful data set for the future development of Welsh language technology. I propose that Argentinean Welsh provides a unique opportunity to broaden the language data base from which Welsh speech technologies are developed. Specific aspects of Argentinean Welsh variation, which may be due to synchronic effects from first language Spanish, the effect of lifelong Spanish bilingualism, or diachronic dialect divergence, are discussed below.

Today, all adult speakers of Welsh are at least bilingual, either with English (in Wales) or with Spanish (in Argentina). This situation complicates what might otherwise be a straightforward dialect comparison between differing varieties of Welsh. Cross-linguistic influence from competing languages Spanish and English is entangled with other linguistic pressures, including effects of first language (L1) on second language (L2) speech, and historical language change as a result of contact. Teasing apart these intertwined factors is far beyond the scope of this paper, and it is sufficient for our purposes to acknowledge that multiple factors exist, and that they likely influence the Welsh language in both regions. Recent work has used experimental methods and corpus analyses to investigate the realization of sound contrasts in Argentinean Welsh that are hypothesized to be susceptible to influence from Spanish contact.

Sleeper (2015) investigated the realization of voice onset time (VOT) in the Welsh voiceless stop series /p t k/. It was hypothesized that while contact with the English system reinforces the retention of the Welsh voiceless aspirated-voiceless unaspirated VOT contrast, contact with Spanish in Argentina may have resulted in a shift to a more Spanish-like voiced-voiceless unaspirated system. Sleeper extracted VOT values from word-initial instances of /p t k/ produced in conversational speech by Welsh bilinguals in Argentina and in Wales, recorded in the Patagonia and Siarad corpora (Deuchar et al., 2014). Results confirmed his hypothesis, with Argentinean Welsh-Spanish bilingual speakers producing shorter Spanish-like VOT in voiceless-stop initial Welsh words, compared to the English-like VOT produced by the Welsh-English bilingual group.

Bell (2018) collected productions of Welsh vowels from Welsh-Spanish bilinguals in Argentina and Welsh-English bilinguals in Wales in order to investigate differences in the acoustic realization of allophonic and phonemic vowel length. Because Spanish does not contrast vowels on the basis of length, nor does duration vary allophonically to the extent that it does in Welsh or English, it was hypothesized that Welsh-Spanish bilinguals were

[8] https://wales.britishcouncil.org/en/programmes/education/welsh-language-project

likely to exhibit differences in their production of long and short Welsh vowels. Results showed that Welsh-Spanish bilinguals produced phonemic vowel length contrasts in much the same way as Welsh-English bilinguals (relying on both vowel duration and spectral quality), but were less similar in production of allophonic duration differences conditioned by following consonant voicing.

Differences in the acoustic realization of the factors mentioned above are likely to prove challenging for an automatic speech recognition system trained only on Welsh produced by fluent speakers in Wales. The differences observed in Argentinean Welsh generally appear to reduce acoustic contrast between Welsh phonemes (the voiceless /p t k/ and voiced /b d g/ stop series, or the vowel length contrast separating minimal pairs such as /mor/ *mor* 'so' and /moːr/ *môr* 'sea'). The collapse of contrasting acoustic cues to these (and potentially other) phonemic differences in Argentinean Welsh is likely to prove challenging for an automatic speech recognition system trained on other dialects of the language.

One solution to this problem, as often seems to be the case in the domain of speech technology, is more data. Natural and lab-produced speech datasets collected from speakers of Argentinean Welsh will serve to diversify the information set from which statistical generalizations about acoustic-phonetic realizations of Welsh are drawn. Knowledge-based approaches to speech recognition that incorporate linguistic generalizations such as phonological rules into the system should also be considered, as they may be well-suited to ASR development from small datasets (Besacier et al., 2006; Gaikwad et al., 2010).

3 Suggestions for future work

While advocating for the inclusion of Argentinean Welsh data in future Welsh speech technology projects is well and good, it must also be acknowledged that there are challenges to doing so. The Welsh-speaking population of Argentina is sparse compared to that of Wales, with speaker numbers in the low thousands, spread throughout the region (Ó Néill, 2005). This problem may be overcome by making use of existing community networks and organizations. Data collection, participant recruitment, and outreach could all potentially be integrated with community events such as the annual Eisteddfod in both the eastern and western Argentinean Welsh communities. Additionally, the Welsh Language Project and Menter Patagonia program involve networks of Welsh language educators in the region who may be interested in integrating speech technology participation into their classrooms of speakers at all levels.

As the resources for Welsh speech technology continue to grow, the opportunity to include data from speakers of Welsh in Argentina increases. Through projects like the speech-collecting Paldaruo app and Mozilla CommonVoice (Prys and Jones, 2018), it is increasingly possible to target and recruit participants who are speakers of Argentinean Welsh (and of course, other minority dialects of the language) through crowdsourcing methods. Encouraging participation in the CommonVoice initiative, which is online and requires little time commitment is a simple first step toward crowdsourcing Argentinean Welsh data.

4 Conclusion

This paper has argued that the dialect of Welsh spoken in Argentina presents a valuable resource for the continuing development of Welsh speech technology. Technologies such as speech recognition benefit from the inclusion of variation (both individual and dialectal) in the data from which they are developed. I have shown that the dialect of Argentinean Welsh is different from other dialects of Welsh in the way speakers acoustically realize underlying phonemic contrasts. This variation, if included in speech technology training data, will serve to develop technologies that will be accessible to more speakers, including those who are less fluent, or who due to language background and L2 cross-linguistic influence are not able to fully take advantage of current Welsh speech technology.

Furthermore, inclusion of the Argentinean Welsh community in the development of Welsh speech technology will strengthen ties between speakers in Wales and Argentina. The goals of Welsh language revitalization programs in both countries include supporting new speakers of the language, and making speech technology accessible and responsive to those new speakers will further progress toward that goal. The Welsh-speaking population of Argentina is a valuable resource, and outreach and integration efforts to and with community members can only stand to benefit future efforts in the development of speech technology for all Welsh speakers.

References

Bell, Elise Adrienne. 2017. Perception of Welsh vowel contrasts by Welsh-Spanish bilinguals in Argentina. In *Linguistic Society of America Annual Meeting*. Linguistic Society of America Annual Meeting.

Bell, Elise Adrienne. 2018. *Perception and Production of Welsh Vowels by Welsh-Spanish Bilinguals*. Ph.D. thesis, The University of Arizona.

Besacier, Laurent, V-B Le, Christian Boitet, and Vincent Berment. 2006. ASR and translation for underresourced languages. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE.

Carter, Diana, Peredur Davies, M. Carmen Parafita Couto, and Margaret Deuchar. 2010. A corpus-based analysis of codeswitching patterns in bilingual communities. In *Proceedings: XXIX Simposio Internacional de la Sociedad Española de Lingüística*, volume 1.

Deuchar, Margaret P., Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto, and D Carter. 2014. Building bilingual corpora. In Thomas, E M and I Mennen, editors, *Advances in the Study of Bilingualism*. Multilingual Matters, Bristol.

Escudero, P. 2009. Linguistic perception of 'similar' L2 sounds. In Boersma, Paul and S. Hamann, editors, *Phonology in perception*, pages 151–190. Mouton de Gruyter, Berlin.

Flege, James E., Ocke-Schwen Bohn, and Sunyoung Jang. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25:437–470.

Flege, James E., Carlo Schirru, and Ian R.A. MacKay. 2003. Interaction between the native and second language phonetic subsystems. *Speech Communication*, 40:467–491.

Gaikwad, Santosh K, Bharti W Gawali, and Pravin Yannawar. 2010. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3):16–24.

Jain, Abhinav, Minali Upreti, and Preethi Jyothi. 2018. Improved accented speech recognition using accent embeddings and multitask learning. In *Proc. INTERSPEECH. ISCA*.

Jones, Dewi Bryn, Gruffudd Prys, and Delyth Prys. 2016. Vocab: a dictionary plugin for web sites. *PARIS Inalco du 4 au 8 juillet 2016*, page 93.

Jones, Robert Owen. 1984. Change and variation in the Welsh of Gaiman, Chubut. In Ball, Martin J. and Glyn E. Jones, editors, *Welsh phonology*, pages 237–261. University of Wales Press, Cardiff.

Jones, Robert Owen. 1998. The Welsh language in Patagonia. In Jenkins, Geraint H., editor, *A social history of the Welsh language: Language and community in the nineteenth century*, pages 287–316. University of Wales Press, Cardiff.

Li, Bo, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4749–4753. IEEE.

Ó Néill, Diarmui. 2005. *Rebuilding the Celtic languages: reversing language shift in the Celtic countries*. Y Lolfa.

Prys, Delyth and Dewi Bryn Jones. 2015. National language technologies portals for lrls: A case study. In *Language and Technology Conference*, pages 420–429. Springer.

Prys, Delyth and Dewi Bryn Jones. 2018. Gathering data for speech technology in the welsh language: A case study. *Sustaining Knowledge Diversity in the Digital Age*, page 56.

Prys, Delyth, Gruffudd Prys, and Dewi Bryn Jones. 2016. Cysill ar-lein: A corpus of written contemporary welsh compiled from an on-line spelling and grammar checker. In *LREC*.

Rao, Kanishka and Haşim Sak. 2017. Multi-accent speech recognition with hierarchical grapheme based models. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4815–4819. IEEE.

Sleeper, Morgan. 2015. Contact effects on voice-onset time (vot) in patagonian welsh. *Journal of the International Phonetic Association*, pages 1–15.

Webb-Davies, Peredur. 2016. Does the old language endure? Age variation and change in contemporary Welsh grammar. In *Welsh Linguistics Seminar*.

Williams, Glyn. 1975. *The desert and the dream: A study of Welsh colonization in Chubut 1865 – 1915*. University of Wales Press, Cardiff.

Yang, Xuesong, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson. 2018. Joint modeling of accents and acoustics for multi-accent speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.