Proceedings of the CLTW 4 @ LREC2022 , pages 52–59
Marseille, 20-25 June 2022

Development and Evaluation of Speech Recognition for the Welsh Language

Dewi Bryn Jones
Language Technologies Unit
Bangor University, Wales
d.b.jones@bangor.ac.uk

Abstract

This paper reports on ongoing work on developing and evaluating speech recognition models for the Welsh language using data from the Common Voice project and two popular open development kits – HuggingFace wav2vec2 and coqui STT. Activities for ensuring the growth and improvement of the Welsh Common Voice dataset are described. Two applications have been developed – a voice assistant and an online transcription service that allow users and organisations to use the new models in a practical and useful context, but which have also helped source additional test data for better evaluation of recognition accuracy and establishing the optimal selection and configurations of models. Test results suggest that in transcription good accuracy can be achieved for read speech, but further data and research is required for improving recognition results of freely spoken formal and informal speech. Meanwhile a limited domain language model provides excellent accuracy for a voice assistant. All code, data and models produced from this work are freely available.

Keywords: speech recognition, Welsh, Common Voice, wav2vec2, coqui STT

1. Introduction

Automatic speech recognition (ASR) is a technology that's transforming how people interact with computers and consume content. New products and services that cater to speakers of larger languages, that are facilitated by highly accurate automatic speech recognition systems, do not exist for speakers of less-resourced languages. The development of speech recognition with accuracies equivalent to that for larger languages has become ever more critical for any less-resourced languages' digital inclusion (Sayers, et.al., 2021).

This paper reports on ongoing work on developing and evaluating speech recognition for Welsh using primarily crowdsourced data and open-source development kits. It reports on how this work has contributed to ensuring the growth and quality of data crowdsourced from an international project as well as from two useful and practical applications developed by the Language Technologies Unit (LTU). The motivation and operation of the voice assistant application, Macsen, as well as the online transcription service, Trawsgrifiwr Ar-lein, are described in sections 1.1 and 1.2 respectively.

All data and source code for training models as well as for both applications are available from the Welsh National Language Technologies Portal (Prys et al., 2018) to any developer or user who may wish to integrate, customize or run local deployments.

1.1 Trawsgrifiwr Ar-lein – Transcription Service Website

Both the COVID pandemic and new United Kingdom Accessibility Legislation (The National Archives, 2018) created a greater demand for Welsh language speech content to be transcribed. The legislation mandates captions and subtitles for all teaching and student support resources used by universities to deliver blended learning[1] and came into effect during the COVID pandemic when provision of all university teaching moved to remote delivery and/or recorded lectures. Lecturers within Welsh universities and the Coleg Cymraeg Cenedlaethol[2] urgently required an application to help ensure compliance for their digital materials.

Figure 1: the Trawsgrifiwr Ar-lein interface for validating and correcting automatic transcriptions of Welsh language speech.

The LTU developed the Trawsgrifiwr Ar-lein website application[3] that allows users to submit an audio file or a link to a YouTube video of Welsh language speech for automatic transcribing. Users are required to accept terms and conditions each time before submitting content for transcription. These state that the service respects all privacy and copyright and automatically deletes submitted content after 30 days. No copies of their data are made in those 30 days nor is any other use made of it.

Each submission is added to a queue for processing. The audio is first segmented with an aggressive Voice Activation Detection algorithm (webrtcvad).[4] Each segment in turn is transcribed by the speech recognition

[1] Blended learning combines in-person and digital delivery of teaching.
(see https://en.wikipedia.org/wiki/Blended_learning)

[2] The Coleg Cymraeg Cenedlaethol plans and supports Welsh language Higher Education provision.
(see https://www.colegcymraeg.ac.uk/en)

[3] https://trawsgrifiwr.techiaith.cymru/

[4] https://github.com/wiseman/py-webrtcvad

model. In the meantime, users are given a unique URL that can be used to access the interface as seen in Figure 1, to listen, validate and correct the transcriptions in each segment. The interface provides a button to playback the segment's audio, its automatic transcription and a text box for entering corrections. If a segment requires no further corrections, the user clicks on the button which displays a tick and a disk icon to commit the correction and move to the next segment. Both next and previous segments are displayed for context, as well as the audio's waveform in order to correct segmentation. An additional button can also correct segments that are too short by merging the current segment to the next segment to form a larger segment. After all segments have been validated or corrected the interface provides buttons to download the transcription as files in SubRip (srt)[5] or TextGrid format - a file format for annotating speech files with Praat (Boersma et al., 2022).
Section 2.2.2 describes the data sourced with the aid of the transcription application.

1.2 Macsen – Voice Assistant App

Previous work on speech recognition for Welsh had been motivated solely by the development of Macsen, a voice assistant for Welsh speakers that can run on Android or iOS devices (Jones, 2020). Despite a lack of speech data, a functioning and everyday useful Welsh voice assistant was achieved, provided the assistant's speech recognition capability was constrained to recognizing only a closed set of commands and questions that trigger a small collection of the most practical and effective skills, such as for retrieving news, providing weather forecasts and playing music. This work was able to update Macsen's speech recognition model with a larger training dataset and expand its ability to support more new skills while not degrading the performance or the practicality of the app.

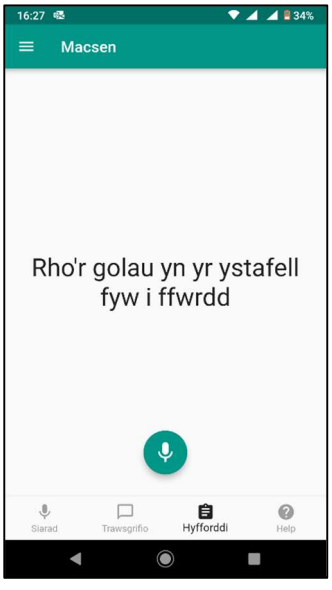

Figure 2 – the Macsen Voice Assistant with the Hyfforddi (Training) bottom navigation bar tab selected and a sentence provided for recording: "Switch off the lights in the living room"

The work also revised the four bottom navigation tabs (see Figure 2) that provide four screens or modes of operation within the app:

- Siarad (*Speak)* – the main screen providing the speech interface to the app's skills
- Trawsgrifio (*Transcribe*) – a new second screen that uses the models trained for transcription. Any speech is converted into text which can then be copied and pasted into any other application on the device, such as for messaging
- Hyfforddi (*Training*) – a screen, as seen in Figure 2, that provides an opportunity for users to record random sentences from the closed set of command and questions
- Help – a screen that lists all sentences from the closed set of commands and questions that the app recognizes. The list is categorized according to skill

Section 2.2.1 describes the data sourced with the aid of the Hyfforddi screen in Macsen voice assistant application.

2. Data

The data used for training Welsh speech recognition models was sourced from popular multilingual open speech and textual datasets. Additional data for evaluating new models was sourced with two applications developed by the LTU.

2.1 Common Voice

The primary speech data resource for training this work's speech recognition acoustic models was the Welsh language subset of Mozilla's Common Voice multilingual speech corpus (Ardila et al., 2020). Following previous attempts in crowdsourcing speech corpora (Prys et al., 2018; Cooper et al., 2019) Welsh has been fortunate to have been supported by the Common Voice project since its first multilingual expansion in June 2018 (Henretty, 2018). Since then, several campaigns to appeal to all speakers of Welsh to voluntarily record and validate recordings have been organized by the community while the LTU has monitored the growth and quality of Welsh Common Voice data for training Welsh language speech recognition.

	Published	Validated (hours)	Other (hours)	Speakers
CV1	Feb 2019	21	1	365
CV2	June 2019	41	6	738
CV3	June 2019	42	6	748
CV4	Dec 2019	59	18	1149
CV5.1	June 2020	83	13	1257
CV6.1	Dec 2020	95	29	1382
CV7	July 2021	110	31	1655
CV8	Jan 2022	116	29	1695

Table 1 – Welsh speech data in Common Voice releases (source: Common Voice website's datasets page).

Table 1 shows how Welsh has progressed through each Common Voice release since June 2018. The total amount of recordings that have been approved by volunteers (validated hours) have increased well with each release.

[5] Further information on the SubRip file format: https://docs.fileformat.com/video/srt/

Whilst the number of recordings that have not yet been validated ('other' hours) have also increased. This may suggest a need to prioritize validation efforts rather than recording new sentences in campaigns for subsequent Common Voice releases. No data from the 'other' split was used in this work since its quality is unknown.
Table 2 shows Mozilla's pre-defined splits of the validated recordings into training, validating and testing sets. Each split contains only one recording per distinct sentence. As noted in Jones (2020), initial versions of Welsh Common Voice consisted of a low number of distinct sentences with a high number of multiple recordings. This consequently created pre-defined splits from Mozilla that were much smaller in size when compared to the overall size of all validated recordings.

	train (minutes)	validation (minutes)	test (minutes)
CV1	34	35	37
CV2	37	37	40
CV3	37	36	41
CV4	66	55	59
CV5.1	311	259	253
CV6.1	557	411	425
CV7	547	439	432
CV8	627	461	469

Table 2 - Welsh data in Mozilla pre-defined splits in all releases.

The only way to remedy such low utilization of contributions into pre-defined splits as set by Mozilla was to ensure that enough distinct sentences are available to the Common Voice website for recording only once by any volunteer. Thus began, after CV3 a concerted effort by members of the LTU to remedy the situation by adding significant amounts of distinct Welsh sentences to Common Voice. Sentences were collected from out-of-copyright materials such as novels and essays as well as from copyrighted texts gifted by individuals[6] and submitted via Mozilla's Common Voice SentenceCollector website. By CV6.1, 14,857 sentences had been added and the size of the Welsh pre-defined training split increased by approximately 1400% from 37 minutes in CV3 to 557 minutes. This provided a larger training set for this work's initial attempts at training models. The next release however, CV7, saw a reduction in the pre-defined training split, with its size decreasing to 547 minutes, indicating there were no new distinct sentence recordings meaning an urgent need for more distinct sentences for improving CV8. Fortunately, the CoVoST 2 corpus (Wang et al., 2020) contains 232,037 unique Welsh sentences created by professional translation of English sentences from version 4 of Common Voice. The CoVoST project conducted sanity checks on all translated sentences by means of language model perplexity and length ratio heuristics with lowest scoring sentences sent for re-translation.

For submission of CoVoST translated sentences back into Welsh Common Voice for recording, this work excluded 130,502 sentences[7] that did not meet the following criteria as set by the Common Voice project or by editors in the LTU:

- sentences should contain less than 15 words
- sentences should not include numbers, acronyms or abbreviations
- all words must be present in a Welsh language lexicon (Prys et al., 2021) or in a list of 20,000 additionally permitted words.

The lexicons facilitated the exclusion of a high number of sentences containing American English proper nouns. In the opinion of the editors in the LTU such sentences were not relevant for speech recognition in a Welsh cultural context. Certain English words, as well as company names and products were judged to be commonly used in Welsh speech and were included in the list of additional permitted words, thus retaining their sentences.
The remaining 101,535 sentences were validated according to Mozilla's recommended method for bulk submissions[8] requiring human editors to proofread a statistically significant random sample of sentences and confirm that a maximum of 5% of sentences were problematic and not appropriate for recording.[9] The sentences were accepted by Mozilla shortly after the CV7 release. Consequently, the size of the pre-defined training split increased by 14.6% a few months later in CV8.

2.2 New Test Sets from LTU Applications

Mozilla Common Voice already provides a test set from its pre-defined splits for researchers to use for measuring their models' recognition accuracy. As shown in Table 2 it is comparable in size to the validation set and by CV8 was 469 minutes in size. This is useful for measuring model accuracy across training sessions and for comparing with models by other researchers. However, since it contains recordings similar in nature to those in the pre-defined training and validation sets, it may not be sufficient for measuring accuracy and suitability in real life application scenarios.
This work collected two test sets from two applications to form a single open resource for testing Welsh speech recognition called the Corpws Profi Adnabod Lleferydd (Speech Recognition Test Corpus) which can be accessed from the LTUs gitlab repositories website.[10]

2.2.1 Voice Assistant Test Set

Within its 'Hyfforddi' (Training) tabbed screen, as shown in Figure 2, the Macsen voice assistant app provides a simple interface that allows users to contribute recordings of sentences randomly selected from the closed set of commands and questions that trigger a response from any of its supported skills. The user touches the microphone button to start and stop recording. Stopping the recording uploads the audio immediately and provides the user with the next sentence for recording. There is no support for

[6] https://github.com/techiaith/brawddegau-adnabod-lleferydd/blob/master/README_en.md
[7] https://github.com/techiaith/brawddegau-adnabod-lleferydd/blob/master/docker/README.md
[8] https://github.com/common-voice/common-voice/blob/main/docs/SENTENCES.md#bulk-submission
[9] https://github.com/common-voice/common-voice/pull/3239
[10] https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/corpws-profi-adnabod-lleferydd

listening and/or re-recording before submitting. The list of possible sentences for recording can be seen within the app under its 'Help' tabbed screen.
Since its release in 2020, approximately 700 recordings have been submitted. Quality control and validation for inclusion into a test set corpus consisted of LTU members listening to each recording and comparing with the original sentence. It was not possible to validate every submission but 300 recordings from 25 users, with a total duration of 17 minutes, were accepted.
The data can be found in the '`data/macsen`' sub-directory of the Corpws Profi Adnabod Lleferydd gitlab repository.[10]

2.2.2 Transcriptions Test Set

As noted in section 1.1, all submitted audio and corrected transcriptions are deleted after 30 days by the Trawsgrifiwr Ar-lein website and are not used for any other purposes in the meantime. The website however does invite users, through a section included in the terms and conditions displayed each time the website is initially opened, to contact the LTU and to discuss providing permission for contributing their audio and corrected transcriptions into the Corpws Profi Adnabod Lleferydd. Another strategy involved commissioning the use of the Trawsgrifiwr Ar-lein website to transcribe recorded sessions from an online conference hosted by the LTU. All speakers had indicated their permission for including transcriptions of their speech into the Corpws Profi Adnabod Lleferydd.
Table 8 in the appendix lists details of 13 YouTube videos that have been transcribed and included into Corpws Profi Adnabod Lleferydd. They include numerous videos from the online conference, but also from teaching resources by various departments at Bangor University, short videos and podcasts for young people by S4C[11] (a Welsh language broadcaster) as well as gaming videos by Menter Iaith Sir Caerffili[12] (a language promotion community group in Caerffili county borough).
Table 9 in the appendix provides information regarding the variations in speech such as gender and accent. All recordings were of native speakers. Accents were generalised as being either 'North' or 'South', although there exist smaller variations of accents for Welsh (Cooper, et al., 2019). This work additionally categorised speech into three types which also took into consideration as to whether transcriptions would be verbatim or non-verbatim, meaning filler words, disfluencies or any small linguistic errors produced during speech were removed or corrected in order to make subtitles as readable as possible.

- Read-Speech – speech by a person reading from a prepared text. Linguistic errors in speech would be minimal with a non-verbatim transcription closer to the actual speech
- Formal-Spoken – speech using a formal register with the assistance of very little or no prepared text. Linguistic errors are more probable, but a non-verbatim transcription would be further from a corrected transcription
- Free-Spoken – speech from speaking freely in an informal register and occasionally some code switching with English. Non-verbatim transcriptions would be furthest from actual speech

Non-verbatim transcriptions may not be as optimal as verbatim transcriptions for evaluating models. A total of 266 segments were found to contain indistinguishable speech, multiple speakers, music, singing or interjections and were therefore excluded from this work's evaluation of models. Table 7 in the appendix lists the tags used to annotate and locate such features in excluded segments.
The transcriptions test set can be found in the '`data/trawsgrifio`' sub-directory of Corpws Profi Adnabod Lleferydd gitlab repositorty.[10]

2.3 Text Corpora

This work also used the following text corpora for training n-gram language models.

2.3.1 Macsen Texts Corpus

The Macsen voice assistant's closed set of questions and commands can serve as a text corpus for training a domain specific language model (Jones, 2020). Sentences can be easily generated from filling slots in template sentences with each possible entry from associated slot value entity files (for example files with lists of topics for the news or names of Welsh language bands). Both template sentences and slot entity values were composed by members of the LTU to facilitate an effective but as natural as possible collection of sentences for users to speak to their Welsh voice assistant. The resulting corpus of 1098 sentences can be downloaded from an API. [13]

2.3.2 OSCAR

The OSCAR corpus (Suárez et al., 2019) of texts crawled from the internet was used to provide a text corpus for training general purpose n-gram Welsh language models. Texts were left deduplicated and unshuffled with no segmentation, special filtering, normalization or tokenization undertaken. This corpus was approximately 23 million words in size.

3. Method

Several acoustic models for Welsh speech recognition have been trained with data from version 8 of Common Voice and open-source speech recognition development kits by coqui STT and HuggingFace. The entry for CV8 in Table 2 provides the duration of each pre-defined split. Common Voice's pre-defined set for testing, as well as the additional test sets as described in section 2.2 were used to measure word and character error rates. Measurements were made of greedy decoding, CTC beam search decoding and decoding with n-gram language model support (Graves et al., 2006).
All training and tests were conducted on a single workstation containing a single NVIDIA Titan 2080 RTX graphics card with 24Gb of RAM.

3.1 coqui STT

Previous work on speech recognition for a voice assistant (Jones, 2020) relied on the then Mozilla DeepSpeech speech recognition kit and its support for transfer learning from an English pre-trained model. In April 2021, Mozilla decided to end all work before its version 1.0 release leaving the start-up coqui AI to continue development.

[11] https://www.s4c.cymru
[12] http://www.mentercaerffili.cymru/
[13] Macsen corpus can be obtained from: https://api.techiaith.org/assistant/get_all_sentences

Previous work had demonstrated that despite a high number of repeated recordings of sentences, risking over fitting to the sentences in Common Voice, training with all 'validated' recordings (116 hours in CV8 as seen in Table 1) and the '`drop_source_layers`' transfer learning hyperparameter value set to 2, was found to be optimal for the Macsen voice assistant app. This work would repeat the same training method to train acoustic models with version 1.2 of the coqui STT kit as well as with more recent and larger datasets from CV8.

All scripts for training and inference, as well as the optimal models produced from this work are available from a LTU GitHub repository.[14]

3.2 wav2vec 2.0

Recent work on wav2vec 2.0 at Facebook AI (Baevski et al., 2020) has made it possible to realise effective speech recognition with smaller quantities of transcribed speech. Representations of speech are initially learnt from large collections of raw speech audio which are then finetuned with transcribed speech data to perform speech recognition. Initial research with English speech recognition demonstrated that just ten minutes of transcribed speech could finetune a model pre-trained with 53,000 hours of raw speech audio and achieve a word error rate of 4.8.

Further work, given the lack of transcribed speech for the majority of the world's 7000 languages, has focused on learning speech representations from multiple languages (Conneau et al., 2020) and has demonstrated that cross-lingual pre-training outperforms monolingual training. The following multilingual models have been pre-trained by Facebook AI and published via the HuggingFace hub[15] for other researchers to finetune for their own languages using their own transcribed speech datasets:

- wav2vec2-large-xlsr-53 (Conneau et al., 2020): pre-trained from 56k hours of raw speech audio in 53 languages.
- wav2vec2-xls-r (Babu et al., 2021): pre-trained from 436k hours of raw speech audio in 128 languages. Models are provided in increasing sizes, from 300 million parameters, to 1 and 2 billion.

Both types of pre-trained models have been exposed to Welsh speech audio from Common Voice. In this work's experiments, all pre-trained models were finetuned for 30 epochs using a concatenation of Common Voice's pre-defined training and validation sets. Identical training hyperparameters values were used for all finetuning training runs with only the name of the pre-trained model varying.

The HuggingFace library support for wav2vec 2.0 speech recognition did not initially support decoding with CTC beam search nor decoding with the support of n-gram language models. Given the urgency for the Trawsgrifiwr Ar-lein transcription application at the time, this work undertook integrating the CTC decoding library from Parlance[16] as well as adding support for training and optimizing n-gram language models.

All scripts for training and inference as well as optimal models are available from a LTU GitHub repository.[17]

3.3 Language Model

Various n-gram language models were created with the KenLM library (Heafield, 2011) using the text corpora described in section 2.3. Optimal values for alpha and beta hyperparameters for CTC with language model decoding were found after 100 trail runs against the CV8 pre-defined test set.

4. Results

Table 3 presents results from evaluating coqui STT and wav2vec2 based models with the CV8 test set. Unfortunately, finetuning a wav2vec2-xls-r-2b pre-trained model, with 2 billion parameters, was not possible due to insufficient GPU hardware. Unsurprisingly however, all wav2vec2 self-supervised based models outperformed the supervised models from coqui STT. A WER as low as 22.4% by a model finetuned from the wav2vec2-xls-r-1b pre-trained model with only greedy decoding is very promising. The addition of a language model trained with the OSCAR corpus with optimized alpha and beta hyperparameters decreased its WER by 39.79% to 13.33 (as highlighted in bold in Table 3). coqui STT's WER, despite having been trained with all of Common Voice's validated recordings, as described in section 3.1, is much higher. However, a larger decrease of 52.01% is achieved with the support of a similar language model. The language model does not decrease each model's CER as significantly - 27.30% decrease for wav2vec2-xls-r-1b and 30.30% for coqui STT.

Model(s)	WER	CER
wav2vec2-large-xlsr-53	24.03	6.74
wav2vec2-large-xlsr-53 + CTC	24.01	6.71
wav2vec2-large-xlsr-53 + CTC + LM	13.79	4.77
wav2vec2-xls-r-300m	25.31	7.01
wav2vec2-xls-r-300m + CTC	25.19	6.98
wav2vec2-xls-r-300m + CTC + LM	14.41	5.03
wav2vec2-xls-r-1b	22.14	6.19
wav2vec2-xls-r-1b + CTC	21.95	6.16
wav2vec2-xls-r-1b + CTC + LM	**13.33**	**4.5**
wav2vec2-xls-r-2b	-	-
coqui STT (AM)	83.33	28.21
coqui STT (AM+LM)	39.99	19.66

Table 3 – Acoustic models test results against CV8 test set. n-gram language model (n=5) trained with the OSCAR corpus.

Table 4 provides recognition results from evaluating coqui STT and wav2vec2 models with the transcription test set from the Corpws Profi Adnabod Lleferydd. Results imply that all models are not as effective and as accurate when applied to a real-world application scenario such as transcribing. As highlighted in bold in Table 4, the best achieved accuracy is a WER of 32.96 by a finetuned wav2vec2-xls-r-1b based model with the support of a language model. Table 6 provides a break-down of results from evaluating the best wav2vec2-xls-r-1b based model with each YouTube video. Accuracy performance varies considerably. Videos of read speech, such as Pl16jPn0Jy4

[14] https://github.com/techiaith/docker-coqui-stt-cy/tree/22.02

[15] https://huggingface.co/models?other=wav2vec2

[16] https://github.com/parlance/ctcdecode

[17] https://github.com/techiaith/docker-wav2vec2-xlsr-ft-cy

and UdWqyWDZ4Y, are transcribed with an accuracy consistent to accuracies reported in Table 3. Other types of speech however are not transcribed as accurately with free spoken speech videos suffering very poor WER scores.

Model(s)	WER	CER
wav2vec2-large-xlsr-53	45.90	16.94
wav2vec2-large-xlsr-53 + CTC	45.66	16.90
wav2vec2-large-xlsr-53 + CTC + LM	34.98	16.47
wav2vec2-xls-r-1b	42.44	15.78
wav2vec2-xls-r-1b + CTC	42.53	15.88
wav2vec2-xls-r-1b + CTC + LM	**32.96**	**15.15**
coqui STT (AM)	92.32	43.26
coqui STT (AM+LM)	71.86	45.68

Table 4 – Model performance on the Transcription test set. n-gram LM with n=5 and trained with the OSCAR text corpus.

Table 5 shows results from using the Corpws Profi Adnabod Lleferydd's Macsen voice assistant test set to evaluate two candidate models for current and future versions of the app. The first candidate was the best performing wav2vec2-xls-r-1b based acoustic model supported by a general-purpose language model. The second candidate was the coqui STT model from previous experiments supported by a domain specific language model trained from the Macsen text corpus as described in section 2.3.1. As highlighted in bold in Table 5, a coqui STT based model with a domain specific language model has considerable better accuracy than the best general purpose wav2vec2-xls-r-1b based models.

Model(s)	WER	CER
wav2vec2-xls-r-1b + CTC + LM	18.06	5.11
coqui STT (AM + domain specific LM)	**4.18**	**2.4**

Table 5 - Model performance on the Macsen Welsh language Voice Assistant test set.

5. Conclusion

This paper has described the development of speech recognition for the Welsh language using speech data from the Mozilla Common Voice project and two popular open-source development kits from HuggingFace and coqui AI. Work on supporting the growth and quality of data in Welsh Common Voice with submissions of thousands of unique and readable sentences is also described. Two new test datasets were constructed from two real world application scenarios – a voice assistant and a transcriber – and used in further evaluation of models.

Results showed that wav2vec2 based models provide impressive accuracy, especially when evaluated with the Common Voice pre-defined test set. This is understandable since models were trained with similar data from other Common Voice pre-defined sets.

Evaluation with a new transcription test set from this work's new Corpws Profi Adnabod Lleferydd suggests that wav2vec2 models may be considered as sufficiently accurate for automatically transcribing read speech. However further research and different types of speech training data is required for improving the accuracy of recognition for free spoken, formal and informal speech. Results suggest that larger models pre-trained from a greater number of hours of raw audio in a greater number of languages can facilitate more accurate acoustic models for Welsh speech recognition after finetuning.

YouTube ID	Decode	WER	CER
Pl16jPn0Jy4	greedy	30.61	9.34
	CTC	30.24	9.19
	CTC+LM	19.91	8.13
4kIby51XL1E	greedy	33.27	10.29
	CTC	33.07	10.26
	CTC+LM	20.99	8.26
0P3VrE-VoOE	greedy	49.87	20.13
	CTC	49.79	20.35
	CTC+LM	40.57	19.8
_UdWqyWDZ4Y	greedy	28.08	8.93
	CTC	27.91	8.92
	CTC+LM	18.41	7.33
TJkVrsNaeY0	greedy	34.48	11.5
	CTC	34.5	11.58
	CTC+LM	27.12	10.62
xSs8TJiD5-Q	greedy	45.4	18.05
	CTC	45.17	17.98
	CTC+LM	37.11	17.69
06Gt5n0BWkw	greedy	49.55	19.35
	CTC	49.34	19.38
	CTC+LM	39.29	18.05
E7qGxNhGP9U	greedy	30.65	10.41
	CTC	30.43	10.43
	CTC+LM	21.3	8.48
BlG0OJ_Kbl4	greedy	54.98	21.0
	CTC	54.68	20.79
	CTC+LM	50.43	25.36
wMMm6rcSpnU	greedy	41.89	14.52
	CTC	40.85	14.44
	CTC+LM	31.08	13.48
C9VnfalWr44	greedy	70.33	26.76
	CTC	69.11	26.9
	CTC+LM	64.74	29.9
yxM1q3AzPJI	greedy	56.35	23.07
	CTC	56.77	23.06
	CTC+LM	44.96	23.25
jdYIrb9L_Tc	greedy	141.2	150.9
	CTC	212.2	189.7
	CTC+LM	102.8	117.4

Table 6 – Test results of a wav2vec2-xls-r-1b based speech recognition model on each video in the transcription test set.

Evaluation of models with the Corpws Profi Adnabod Lleferydd Macsen voice assistant test set suggest coqui STT with a limited domain language model can serve as a very accurate speech recognition component for recognizing sentences for all current skills in the Macsen voice assistant app. Coqui STT's relatively inexpensive computational demands are also attractive since the assistant may be required to run on local and on offline devices. Results have informed on the feasibility of utilizing wav2vec2 models with a general-purpose language model for all current skills. Users would perceive

a significant degradation in recognition of sentences. Further work will aim to improve speech recognition that will allow reliable recognition of a greater number of skills and/or a more open set of commands and questions.

Comparing this work's methods and models with that for other Celtic languages is limited by the fact that only Irish and Breton are supported by the Mozilla Common Voice project. Both coqui STT and HuggingFace wav2vec2 models have been trained and reported for both languages. In Tyers et al. (2021) both Irish and Breton coqui STT models were trained with Common Voice data. By utilizing the same transfer learning mechanism as described in section 3.1, word error rates of approximately 94 were reported for both languages' acoustic models. The addition of an n-gram language model is reported to have improved results to 70.73 for Irish and 68.37 for Breton. Numerous attempts have been made by individuals at finetuning the wav2vec2 pre-trained models listed in section 3.2 for both languages, with word error rates of 42.34 for Irish and 41.71 for Breton for acoustic models reported on the 'Papers With Code' website.[18] Both languages have much smaller total hours of speech than Welsh in Common Voice and would need to ensure both significant amounts of distinct and readable sentences are available as well as to collaborate to appeal to the wider language community for contributions. Other speech data sets may be available and viable for finetuning. Similar approaches to crowdsource data with applications may also be possible using the code from this work.

The best Welsh language coqui STT and wav2vec2 based models from this work have been published to the LTU's GitHub pages[19] as well as to each speech development kit's respective public model repositories.[20] [21] All are licensed with open and permissive licensing in order to provide as many opportunities as possible for discoverability and integration of models by developers of into their software products and services with the least restrictions. Models will be updated for as long as this work continues to improve training of models with both coqui STT and HuggingFace speech recognition development kits. Work will also continue to collect more data through supporting Mozilla Common Voice, the LTUs own applications and from other sources.

6. Acknowledgements

This work was funded by the Welsh Government as part of its implementation of its Welsh Language technology plan (Welsh Government, 2018).

This work has been greatly supported by Rhoslyn Prys who undertook on a voluntary basis several crowdsourcing campaigns, to the Mentrau Iaith, Gwynedd Council, the National Library of Wales who worked with Rhoslyn on some of these campaigns, to the Welsh Government, and to the many participants across Wales and beyond who have contributed their voices to the Welsh Common Voice datasets.

Sentences for Welsh Common Voice were edited and proofread by Professor Delyth Prys and Gruffudd Prys.

The Trawsgrifiwr Ar-lein online transcription website was developed by Stephen Russell and used by Tegwen Bruce-Deans in the construction of the Corpws Profi Adnabod Lleferydd transcription test set.

7. Bibliographical References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., & Weber, G. (2020). Common Voice: A Massively Multilingual Speech Corpus. LREC.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Platen, P. V., Saraf, Y., Pino, J., Baevski. A., Conneau, A., Auli, M. (2021). XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale. ArXiv, abs/2111.09296.

Baevski, A., Zhou, H., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. ArXiv, abs/2006.11477.

Boersma, P., Weenik, D. (2022) Praat: Doing phonetics by computer [Computer Program]. Version 6.2.10. http://www.praat.org/

Conneau, A., Baevski, A., Collobert, A., Mohamed, R., Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. ArXiv, abs/2006.13979

Cooper,S. Jones, D.B. and Prys, D. (2019). Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology. Information, 10(8), p.247. Available at: http://dx.doi.org/10.3390/info1008024

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. p 369-376 ICML 2006 – Proceedings of the 23rd International Conference on Machine Learning.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. WMT@EMNLP.

Henretty, M. (2018). More Common Voices. https://medium.com/mozilla-open-innovation/more-common-voices-24a80c879944 [Accessed March 31, 2022]

Prys,D. Jones, D.B. (2018). Gathering Data for Speech Technology in the Welsh Language: A Case Study. Proceedings of the LREC 2018 Workshop "CCURL 2018 – Sustaining Knowledge Diversity in the Digital Age", p.56. Claudia Soria, Laurent Besacier and Laurette Pretorius (eds.). Available at: http://lrec-conf.org/workshops/lrec2018/W26/pdf/book_of_proceedings.pdf

Prys,D., Jones, D.B. (2018). National Language Technologies Portals for LRLs: A Case Study. In: Vetulani, Z., Mariani, J., Kubis, M. (eds) Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2015. Lecture Notes in Computer Science(), vol 10930. Springer, Cham. https://doi.org/10.1007/978-3-319-93782-3_30

Sayers, D., R. Sousa-Silva, S. Höhn et al. (2021). The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies. Report for EU COST Action CA19102 'Language In The Human-Machine Era'. https://doi.org/10.17011/jyx/reports/20210518/1

[18] https://paperswithcode.com/dataset/common-voice

[19] https://github.com/techiaith/docker-wav2vec2-xlsr-ft-cy/releases

[20] https://huggingface.co/techiaith/wav2vec2-xlsr-ft-cy

[21] https://coqui.ai/models

The National Archive (2018). The Public Sector Bodies (Websites and Mobile Applications) Accessibility Regulations 2018. Available at: https://www.legislation.gov.uk/uksi/2018/852 [Accessed March 31, 2022]

Tyers, F., Meyer, J. (2021). What shall we do with an hour of data ? ArXiv, abs/2105.04674

Welsh Government (2018). Welsh language technology action plan. Available at: https://gov.wales/sites/default/files/publications/2018-12/welsh-language-technology-and-digital-media-action-plan.pdf [Accessed March 31, 2022]

8. Language Resource References

Wang, C., Wu, A., Pino, J. (2020) CoVoST 2: A Massively Multilingual Speech-to-Text Translation Corpus. Available via GitHub: https://github.com/facebookresearch/covost

Prys, D., Jones, D. B., Prys, G., & Watkins, G. L. (2021). Lecsicon Cymraeg Bangor Welsh Lexicon (Version 21.10) [Dataset]. https://github.com/techiaith/lecsicon-cymraeg-bangor

Suárez, P., Sagot, B., Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In CMLC-7 (pp. 9 – 16). Leibniz-Institut für Deutsche Sprache.

9. Appendix

Tag	Meaning
[cerddoriaeth]	Segment contains music
[en_start]	Start of English language speech
[en_finish]	End of English language speech
[canu]	Segment is singing
[siaradwyr lluosog]	Multiple speakers
[ebychiad]	Burst / interjection
[chwerthin]	Laughter
[sŵn y gêm]	Sound of a computer game
[siarad]	Indistinguishable speech

Table 7 - Tags used to annotate the Corpws Profi Adnabod Lleferydd transcriptions test set.

			Duration	
ID	Segments	Voices	Total (min)	Avg (secs)
Pl16jPn0Jy4	79	1	19.22	14.6
4kIby51XL1E	78	1	20.82	16.02
0P3VrE-VoOE	77	1	30.62	23.87
_UdWqyWDZ4Y	182	1	15.95	5.26
TJkVrsNaeY0	144	1	24.42	10.18
xSs8TJiD5-Q	227	1	24.56	6.49
06Gt5n0BWkw	244	1	24.85	6.11
E7qGxNhGP9U	128	7	10.64	4.99
BlG0OJ_Kbl4	8	2	1.74	13.11
wMMm6rcSpnU	70	1	7.42	6.36
C9VnfalWr44	6	12	1.74	17.42
yxM1q3AzPJI	35	2	6.35	10.89
jdYIrb9L_Tc	99	5	4.2	2.55

Table 8 – Corpws Profi Adnabod Lleferydd Transcription Test Set Properties.

ID	Gender	Accent	Type of Speech
Pl16jPn0Jy4	F	S	Read speech
4kIby51XL1E	F	S	Read speech
0P3VrE-VoOE	M	N	Formal Spoken
_UdWqyWDZ4Y	M	N	Read speech
TJkVrsNaeY0	M	N	Formal Spoken
xSs8TJiD5-Q	F	N	Formal Spoken
06Gt5n0BWkw	M	N	Formal Spoken
E7qGxNhGP9U	M+F	N	Formal Spoken
BlG0OJ_Kbl4	M+F	N	Formal Spoken
wMMm6rcSpnU	M	N	Formal spoken
C9VnfalWr44	M+F	N+S	Free spoken
yxM1q3AzPJI	F	S	Free spoken
jdYIrb9L_Tc	M	S	Free spoken

Table 9 - Speech variations in transcription test set - Gender: F=Female, M=Male
Accent: S=South Wales, N=North Wales.