Developing Live Welsh Speech Recognition Models for a Commercial Product - a case study

Preben Vangberg[1], *Leena Sarah Farhat*[1], *Dewi Bryn Jones*[1], *Sean Kinahan*[2]

[1]Language Technologies Unit, Bangor University
[2]Haia Communications Ltd.

prv21fgt@bangor.ac.uk, lnf20dhp@bangor.ac.uk, d.b.jones@bangor.ac.uk

1. Introduction

This paper reports on the work undertaken to develop Welsh speech recognition within a wider pilot project for adding a Welsh speech-to-translated English text capability into Haia Communication Ltd's (Haia)[1] innovative hybrid events platform.

A hybrid event system is a platform that allows for the integration of physical and virtual elements in an event. It combines the advantages of both physical and virtual event experiences, allowing participants to interact and engage from a physical venue as well as remotely via digital methods. In comparison to platforms such as Microsoft Teams, Zoom, and other video conferencing options, hybrid event systems provide more complete capabilities including multi-camera setups, live streaming capabilities, virtual attendee interaction tools, interactive Q&A sessions, virtual exhibitor booths, networking capabilities, and analytics to gauge attendee engagement and satisfaction. Hybrid events conducted by Welsh public sector organisations require facilitating bilingual Welsh and English language interactions, which at present are provided through human interpretation.

Consequently, a key requirement by the Haia platform for facilitating bilingual interactions via technology is support for accurate and efficient online or live automatic speech recognition (ASR).

2. Methodology

This project implemented Welsh language Automatic Speech Recognition (ASR) with Coqui STT. Coqui STT is an open-source deep-learning toolkit for training and deploying speech-to-text models. Implemented as an unidirectional recurrent neural network (RNN) for streaming [1], it can transcribe in near real-time, producing quick text output that can be displayed or used for live closed captioning. This made it much well suited for Haia's requirements.

Recently proposed multilingual speech-to-translated text [2, 3, 4] have reported BLEU scores for Welsh speech to translated English text as required by the wider pilot project. However, all are transformer-based [5] so were disqualified from our investigation as they require entire clips of sound for processing and thus cannot provide live and near real-time inference. These models are large in size, computationally expensive and slower.

Due to Coqui STT's support for transfer learning, models for languages lacking thousands of hours of transcribed speech data can be bootstrapped from models for larger better-resourced languages. The accuracy of Coqui STT acoustic models can be improved with n-gram language models trained from text corpora.

Welsh language Coqui STT models had previously been trained however, further training data was used to reduce WER.

2.1. Training Data

The speech training data sources are listed in Table 1.

Data source	Hours of data
Common Voice	121
Banc Trawsgrifiadau	20
YouTube dataset	23

Table 1: *List of data sources used for the ASR models with the number of hours rounded to the nearest hour.*

Common Voice [6] is a crowd-sourced dataset of recordings of unique and random sentences read by volunteers. Participation by the Welsh-speaking community has provided a useful and open resource for basic Welsh speech recognition research. Banc Trawsgrifiadau Bangor (*Bangor Transcriptions Bank*) is a recent resource that consists of verbatim transcriptions of spontaneous natural speech and is distributed under a CC-0 license. Despite Coqui STT's support for transfer learning [7, 8], large amounts of data are still required for training an accurate model for streamed or online speech recognition. Unfortunately, no such large speech dataset resource exists for Welsh. This work attempted to resolve the gap by building a dataset from YouTube.

2.1.1. Additional Data from YouTube

Previous experiments suggested that synthetic training data built from automatic transcriptions of YouTube videos could be utilised as training data. [9] This project expanded on this work by using a more accurate and bilingual wav2vec2 transformer-based model [10] to automatically transcribe videos and streams from selected channels containing a high number of Welsh and English speech. The bilingual wav2vec2 model used is a pre-trained wav2vec2-large-xlsr-53 [11] model, fine-tuned with a 50/50 balance of English and Welsh data from version 11 of Common Voice, achieving a WER of 17.07 for both languages on the Common voice test set - 7.13 for Welsh and 27.54 for English. Transcription results were kept with the accompanying confidence scores provided by the speech model's CTC beam search decoder and were additionally tagged with the language detected by Google's Compact Language Detector v3[2]. The detected language's probability was also kept with each transcription. Initial attempts to train Coqui STT models on

[1]https://haia.live

[2]https://github.com/google/cld3

the entire collection of automatically transcribed speech from YouTube (920 hours of Welsh speech) produced models with worse WERs, underlining its poor overall quality. To improve data quality, all transcripts were discarded except for those with a speech model confidence score of more than 0.5 and a correct language detection probability of 100%. The resulting dataset contained approximately 23 hours of Welsh language speech data.

2.2. Training the models

All the Coqui STT models were trained on a single NVIDIA RTX A6000 GPU card, using a learning rate of 0.001, a dropout rate of 0.25, and a batch size of 64. Additionally, the first 2 source layers of the base model were dropped. The Coqui STT 1.4.0 English model was used as a base for all the other models.

A KenLM scorer [12] was trained using the cy deduplicated instance of Oscar corpus [13, 14] with $n = 5$.

3. Results

Three Welsh Coqui STT models were trained; One using only the data from Common Voice 13 (cv13), one adding the data from YouTube (yt), and a further one using data from Common Voice, YouTube, and Banc Trawsgrifiadau (bt). All models were tested with the Common Voice 13 test set. The test results for these models with a scorer can be seen in Table 2.

Model	**WER**	**CER**
cv13	24.89%	10.48%
cv13-yt	*23.67%	*9.74%
cv13-yt-bt	23.22%	9.67%

Table 2: *Word- and character- error rates for the 3 models with a scorer. Results with a star are statistically significantly better ($p < 0.05$) than the result above.*

Table 2 shows that the addition of the 23 hours from YouTube improved the performance of the model in a statistically significant way. However, adding Banc Trawsgrifiadau as well did not further statistically significantly improve the results. Yet, when testing the models on the Banc Trawsgrifiadau test set, the addition of the Banc Trawsgrifiadau training data did improve the word error rate from 56.26% (cv13-yt) to 49.63%. This indicates that the addition of Banc Trawsgrifiadau, while it did not improve the overall results, did improve "in domain" results and indicates that the difference in the domain might have played a role.

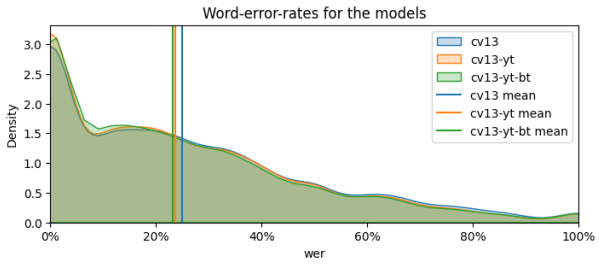

Figure 1: *A kernel density plot showing the distribution of word-error rates within the test results*

When looking at the distribution of WERs within the results in Figure 1, we can see that all three models barely differ. Yet both cv13-yt and cv13-yt-bt perform marginally better than cv13 across the board, resulting in lower overall mean WERs for the models.

	cv13	cv13-yt	cv13-yt-bt
cv13		3.3***	4.5***
cv13-yt	-3.3***		1.2
cv13-yt-bt	-4.5***	-1.2	

Table 3: *The difference in means between the word error rates of the 3 models. * means $p < 0.05$, ** means $p < 0.01$, and *** means $p < 0.001$.*

	cv13	cv13-yt	cv13-yt-bt
cv13		4.1***	4.4***
cv13-yt	-4.1***		0.44
cv13-yt-bt	-4.4***	-0.44	

Table 4: *The difference in means between the character error rates of the 3 models. * means $p < 0.05$, ** means $p < 0.01$, and *** means $p < 0.001$.*

The difference in mean between cv13 and cv13-yt and cv13-yt-bt is statistically significant ($p < 0.001$) as highlighted in Table 3 and Table 4, however, the difference between cv13-yt and cv13-yt-bt is not.

These results are significant improvements upon the state-of-the-art for Welsh Coqui STT models. Previous Coqui STT models have had CERs of 28.21% (without scorer) and 19.66% (with scorer) [15], with the best Welsh Coqui results available have a CER of 16.01%(without scorer) [7]. The best results for this project have a CER of 13.25% (without scorer) and 9.67% (with scorer).

4. Conclusion

The models presented in this work improved upon existing state-of-the-art Welsh Coqui STT models. They fulfil the requirement for a fast and as accurate as possible online or streaming speech recognition model for Haia's hybrid events platform.

Further work is required to achieve significantly lower WERs as well as mitigate errors in the wider speech to translated text pipeline by creating more robust machine translation models and/or denoising autoencoders[16][17].

The open source toolkits, datasets, and methodology developed in this work demonstrate a strong feasibility for Haia to collect and generate its own speech data to improve its speech recognition models in support of their wider business strategy of catering to the bilingual and multilingual hybrid events market.

5. Acknowledgements

This work was funded by the Innovate UK AKT2I Pilot Scheme (Project ID Number 245).

We would like to thank Tom Burke and Andy Esser at Haia Ltd. for their cooperation, as well as Chris Woods, Stefano Ghazzali, Matthew Russell and Professor Delyth Prys at Bangor University for their support.

The primary authors of this work are supported by the UKRI Centre for Doctoral Training in Artificial Intelligence, Machine Learning and Advanced Computing (AIMLAC), funded by grant EP/S023992/1.

6. References

[1] Coqui. Coqui - streaming RNNs in TensorFlow. [Online]. Available: https://coqui.ai/blog/stt/speech-recognition-deepspeech

[2] C. Wang, A. Wu, and J. Pino, "CoVoST 2 and massively multilingual speech-to-text translation." [Online]. Available: http://arxiv.org/abs/2007.10310

[3] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. XLS-r: Self-supervised cross-lingual speech representation learning at scale. [Online]. Available: https://arxiv.org/abs/2111.09296v3

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision." [Online]. Available: http://arxiv.org/abs/2212.04356

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing." Association for Computational Linguistics, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[6] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively- multilingual speech corpus." in *Proceedings of the 12th Conference on Language Resources and Evalua- tion*, pp. 4211–4215.

[7] P. Vangberg, "Transfer learning for speech-to-text: Investigating the impact of the base language on the performance of models."

[8] F. M. Tyers and J. Meyer, "What shall we do with an hour of data? speech recognition for the un- and under-served languages of common voice." [Online]. Available: http://arxiv.org/abs/2105.04674

[9] Leena Sarah Farhat, "Applying a teacher-student learning approach to improve welsh STT frameworks."

[10] D. B. Jones. A welsh and english bilingual speech recognition wav2vec2 based model. [Online]. Available: https://huggingface.co/techiaith/wav2vec2-xlsr-ft-en-cy

[11] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition." [Online]. Available: http://arxiv.org/abs/2006.13979

[12] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, pp. 187–197.

[13] P. J. O. Su'arez, B. Sagot, and L. Romary, "Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures," ser. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. L"ungen, and C. Iliadi, Eds. Leibniz-Institut f"ur Deutsche Sprache, pp. 9 – 16. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215

[14] P. J. Ortiz Su'arez, L. Romary, and B. Sagot, "A monolingual approach to contextualized word embeddings for mid-resource languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1703–1714. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.156

[15] D. Jones, "Development and evaluation of speech recognition for the welsh language," in *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pp. 52–59.

[16] Y. Cheng, Z. Tu, F. Meng, J. Zhai, and Y. Liu, "Towards robust neural machine translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1756–1766. [Online]. Available: http://aclweb.org/anthology/P18-1163

[17] H. Khayrallah and P. Koehn, "On the impact of various types of noise on neural machine translation," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics, pp. 74–83. [Online]. Available: http://aclweb.org/anthology/W18-2709