Developing automatic speech recognition for Scottish Gaelic

Proceedings of the CLTW 4 @ LREC2022 , pages 110–120
Marseille, 20-25 June 2022

Developing Automatic Speech Recognition for Scottish Gaelic

Lucy Evans, William Lamb, Mark Sinclair, Beatrice Alex
University of Edinburgh, University of Edinburgh, Quorate Technology Ltd., University of Edinburgh
lucy.evans9@hotmail.com, w.lamb@ed.ac.uk, mark.s.sinclair@gmail.com, b.alex@ed.ac.uk

Abstract

This paper discusses our efforts to develop a full automatic speech recognition (ASR) system for Scottish Gaelic, starting from a point of limited resource. Building ASR technology is important for documenting and revitalising endangered languages; it enables existing resources to be enhanced with automatic subtitles and transcriptions, improves accessibility for users, and, in turn, encourages continued use of the language. In this paper, we explain the many difficulties faced when collecting minority language data for speech recognition. A novel cross-lingual approach to the alignment of training data is used to overcome one such difficulty, and in this way we demonstrate how majority language resources can bootstrap the development of lower-resourced language technology. We use the Kaldi speech recognition toolkit to develop several Gaelic ASR systems, and report a final WER of 26.30%. This is a 9.50% improvement on our original model.

Keywords: Scottish Gaelic, Automatic Speech Recognition, Low-Resource ASR, Alignment

1. Introduction

For a minority language with 57,100 speakers at the last census (National Records of Scotland, 2015), Scottish Gaelic has a surprising level of language technology provision. Over the past ten years, researchers have developed: a part-of-speech tagger (Lamb and Danso, 2014), a lemmatiser and word-embedding model (Lamb and Sinclair, 2016), a derivation of a categorical grammar (Batchelor, 2016; Batchelor, 2019), a syntactic parser (Boizou and Lamb, 2020), a Gaelic to Irish machine translation system (Murchú, 2019),[1] a wordnet (Bella et al., 2020) and a text-to-speech system.[2] Data sparsity is a major challenge for most minority languages attempting to gain entry to more advanced NLP tools and methodologies. In some ways, Gaelic is in a fortunate situation in this regard: the fieldwork efforts of the School of Scottish Studies (University of Edinburgh), along with a century's worth of Gaelic broadcasting by the BBC (Lamb, 1999, 143), have produced sizeable corpora of natural language data. At the same time, most are in the form of raw audio and paper-based text (typed and handwritten).[3] In order to move towards more involved NLP tasks and applications, we must first solve the issues of automatically and accurately recognising text and audio. The current paper focuses on the latter problem: automatic speech recognition (ASR).

ASR is already integrated into the lives of many majority language speakers. English speakers, for example, can take advantage of voice assistants like Alexa, Siri and Google Home, which recognize verbal commands and perform tasks in response. ASR is also used, of course, to enhance existing audio-visual resources by generating automatic transcriptions and subtitles. ASR methods are key to improving accessibility for certain users: many with dyslexia find it easier to dictate to a computer than to write, and those with physical challenges may find voice methods more accessible than touchscreens or keyboards. At a sociolinguistic level, building ASR systems for minority languages allows for their inclusion in new, technologically-mediated speech domains and encourages existing speakers to continue using them. Ultimately, this work has a key role in language revitalisation.

In this paper, we discuss efforts to develop a full ASR system for Scottish Gaelic, from a starting point of limited resource. We present a novel cross-lingual approach to creating acoustic model training examples, and describe several Gaelic NLP resources that were developed as secondary outcomes of the project.

2. The Low Resource Problem

The problem of low-resource ASR is widespread, as demonstrated by the small number of languages supported by speech assistant technologies. For example, Siri[4] and Google Home[5] each support only 12 languages out of the over 7,000 languages in the world. Their linguistic limitations are due, in part, to the strict requirements on the datasets and resources needed to build an ASR system. Of course, majority languages have much larger commercial potential, as well.

[1] Google added Scottish Gaelic to its Translate system in 2016.

[2] Developed by the University of Edinburgh spin-out, Cereproc: `https://www.cereproc.com`.

[3] A notable exception is Corpas na Gàidhlig – the 30M word corpus of historical and contemporary text based at the University of Glasgow (O Maolalaigh, 2016)

[4] Siri: Arabic, Cantonese, Dutch, English, Finnish, French, German, Hebrew, Italian, Malay, Mandarin Chinese, Spanish

[5] Google Home: Danish, Dutch, English, French, German, Hindi, Italian, Japanese, Korean, Norwegian, Spanish, Swedish

2.1. The Ideal Dataset

Constructing a conventional[6] ASR system requires 3 components: an acoustic model (AM), a language model (LM) and a pronunciation lexicon. The AM is trained on transcribed speech data, and learns to discriminate between the acoustic features of a target language's phonemes. The LM is trained on text data only, and learns typical sequences of words. Finally, a pronunciation lexicon is a list of words accompanied by phonetic transcriptions. Effectively, the lexicon is an intermediate between the two models at inference time. The AM uses the lexicon to map phonemes it recognizes to the words they form a part of, and the LM then estimates which combination of those words is most likely to have been spoken.

It is important that the transcriptions used to train the AM are verbatim, i.e., only containing the words that were spoken. This is because, in training, every frame of speech in the recording must be mapped to a component phone of a word in the transcription. The audio frame is then used as an example of how that phoneme is pronounced. If non-verbatim words are present in the transcript, some speech frames will be used as examples of phonemes that were never spoken. This leads to inaccuracy when recognising those phonemes. As a requirement for creating AM training examples, every transcribed utterance must also be time-aligned, i.e., assigned a start and end time within its corresponding recording. This would be laborious to perform manually, so it is usually done with an automatic aligner.

The text, both in terms of the transcriptions and the LM training data, has further requirements. Firstly, non-linguistic data, such as HTML tags or page numbers, must be removed. This is because they do not form part of a written sentence in the target language. Additionally, it must be possible to retrieve any word in the text from the lexicon. This enables the AM to map that word to its component phonemes to learn, and later recognize, the acoustic features of those phonemes. It follows that the pronunciation lexicon should contain at least one entry[7] for each distinct word in the training data. To avoid duplication of pronunciations in the lexicon, punctuation, capitalisation and digits in the text must be normalised. For example, if the tokens '9' and *naoi* ('nine') both occurred in a text, it would lead to ambiguity in the system; they would be mapped to the same pronunciation.

2.2. Low-Resource ASR

Modern approaches to ASR use deep neural network (DNN) models, which generally require hundreds of hours of transcribed audio and millions of tokens of text as training data. For this reason, data sparsity is a common hindrance in ASR modelling, especially with minority languages. Therefore, data augmentation techniques (Tüske et al., 2014; Renduchintala et al., 2018; Yılmaz et al., 2018) have become popular in low-resource ASR. These techniques strive to increase the quality and quantity of speech data by synthetically modifying existing data with noise, speed perturbation and other forms of variability. Other experimental methods, such as combining training data from multiple languages, are discussed further in section 3.

The collection and transcription of speech data is a significant challenge for most languages. As noted, most gathered text data requires cleaning and normalisation. For many majority languages, a wealth of NLP resources are available to facilitate this. English, for example, benefits from **num2words** (Dupras, 2022), a tool for verbalising digits in text, and **NLTK** (Bird et al., 2009), a natural language toolkit with modules for text cleaning and normalisation. Unfortunately, these kinds of tools rarely exist for minority and lower-resourced languages. Consequently, it takes more effort to acquire and prepare appropriate training data in 'low-resource ASR' contexts.

Typically, the pronunciation lexicon is even more difficult to obtain than the ASR training data. This is because the lexicon must be manually constructed by a language expert. Considering the number of tokens in a single language, this is an extensive and time-consuming task. As a result, comprehensive pronunciation lexicons do not exist for most minority languages.

3. Background

Popular approaches to tackling speech data sparsity in ASR involve using data from greater-resourced languages to bootstrap the low-resource system. One such approach applies the idea of multi-task learning (Caruana, 1997). This is where a single model simultaneously learns to perform multiple related tasks. For example, an AM learns to discriminate between phonemes from multiple different languages. Huang et al. (2013), for example, used a shared-hidden-layer multilingual DNN, in which the hidden layers of the model are trained on data from multiple languages. In this case, only the top, classifying layer is language-specific. Klejch et al. (2021) trained a similar multilingual acoustic model with language-specific output layers, and then fine-tuned the full model on monolingual data from each of its target, low-resource languages. This type of approach enables the feature extraction layers of the model to benefit from learning *global* discriminative speech features, while the output layer specialises in the target language.

Fully multilingual acoustic models have also been explored. Grézl et al. (2014) trained an acoustic model on multiple non-target languages, with the output layer

[6] Some modern ASR construction techniques, such as end-to-end and CTC, do not require a lexicon, or even a language model. They do, however, require quantities of data that far exceed the resources available for most minority languages.

[7] Multiple entries are used to recognise alternative pronunciations.

corresponding to all phonemes present in all of the training languages. The model was then adapted to its target low-resource language, reducing the number of outputs and shifting the model's weights towards the acoustic space of that language. Even before adaptation, the multilingual system was shown to outperform a monolingual target language system. This is a consistent finding in ASR research (Huang et al., 2013; Liu et al., 2018), and is likely due to an improvement in the model's ability to generalize to unseen speech data (Chen and Mak, 2015). From these results, we can conclude that non-target language materials are key to facilitating low-resource speech recognition.

Despite the aforementioned advances, previous work on Gaelic ASR is limited. Rasipuram et al. (2013) tackled the absence of a well-developed Gaelic pronunciation lexicon by exploring the use of grapheme-based ASR. The approach uses the Kullback-Leibler Hidden Markov Model (KL-HMM), in which graphemes are used instead of phonemes as the sub-word unit of the acoustic model. This exploited the fairly regular relationship between Gaelic graphemes and phonemes, and was shown as an effective approach to the problem. However, in years since, a substantial phoneme-based pronunciation lexicon has, in fact, been developed. Am Faclair Beag (Bauer and MacDhonnchaidh, 2022)[8] contains over 35,000 Gaelic words with IPA-style pronunciations, and is regularly maintained and updated. The existence of a large Gaelic lexicon enables a more traditional ASR approach to be undertaken, since numerous standard word-to-phoneme mappings have become available. In the sections that follow, we describe the development of such a system and demonstrate how non-target language resources can help prepare speech training data.

4. Resources

4.1. Collection of Resources

To train our AM, we collected transcribed speech data from the following sources:

- Clilstore,[9] an open-source repository of teaching videos,
- transcriptions made by Tobar an Dualchais (TaD),[10] from recordings of traditional narrative held by the School of Scottish Studies Archives (University of Edinburgh: UoE),
- output transcriptions from the Scottish Gaelic Automatic Handwriting Recognition Project, which utilised manuscripts of Gaelic traditional narrative at the School of Scottish Studies Archives (UoE),
- recordings of multi-speaker Zoom calls,
- audio books,
- and finally roughly 1000 short videos from LearnGaelic,[11] a language teaching resource created by MG Alba, the Gaelic media service.

Most of the data collected was from non-scripted interviews, with the exception of the pre-defined prompts, and as such can be classed as spontaneous speech. However, a sizeable proportion was also collected from oral narrative or lectures and so is less spontaneous. Written text data for training the language model (LM) was collected from all of the above transcriptions, as well as from: 1) An Crúbadán (Scannell, 2007), a web-scraped corpus of Gaelic text; and 2) short summaries of all of the Gaelic audio available on TaD. Finally, we used the aforementioned Gaelic pronunciation lexicon, Am Faclair Beag (Bauer and MacDhonnchaidh, 2022), as the starting point for the ASR system's lexicon.

4.2. Suitability of Resources

A substantial amount of Gaelic training data was collected, but it was by no means purpose-built for ASR. The text data included digits, page numbers, HTML tags, and notes, as well as punctuation and capitalisation. The transcriptions contained speaker labels and other non-verbatim text, and, most significantly, were not time-aligned to their audio recordings. To our knowledge, neither a text normalisation tool, nor an automatic aligner, existed for Gaelic. The data preparation stage, therefore, constituted a large proportion of the project time, and is described in the following sections.

In addition to the training data requiring cleaning, the lexicon was in need of modification. While the original lexicon included pronunciations for 35,000 words, this was for base-forms only; many morphological permutations were not present. The training data, however, contained over 150,000 distinct tokens. As we required an entry in the lexicon for every distinct token in the training data, we needed to augment the lexicon to accommodate out-of-vocabulary (OOV) tokens.

4.3. Solving the Suitability Problem

We removed capitalisation, punctuation, page numbers, speaker labels and other junk strings from texts using regular expressions implemented in Python. Our aim was to extirpate all non-verbatim or non-linguistic text, but any that did not match the specified patterns remained in place.

To tackle the presence of digits in the text, we developed a Gaelic digit verbaliser. One complexity of this task is that Gaelic uses both the decimal and vigesimal numbering systems. For many digits then, more than one verbalisation is possible. The token '80', for example, may verbalise to both *ceithir fichead* ('four twenties' - vigesimal system) and *ochdad* ('eighty' - deci-

[8] https://www.faclair.com
[9] https://clilstore.eu/clilstore/
[10] https://www.tobarandualchais.co.uk
[11] https://learngaelic.scot

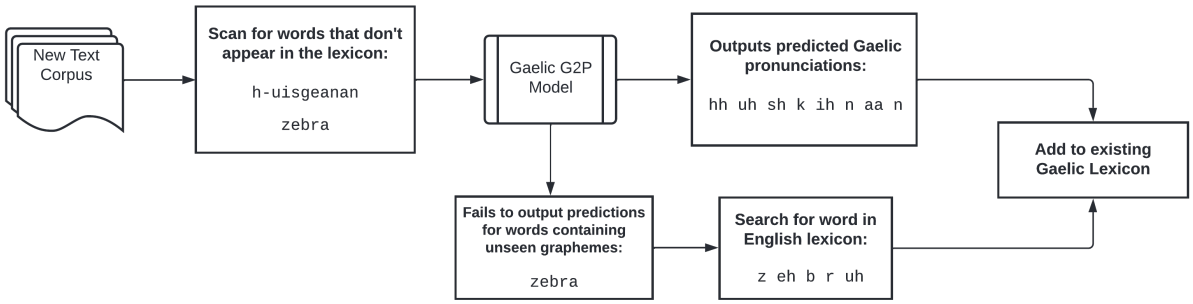

Figure 1: Diagram to show the grapheme-to-phoneme (G2P) process for adding words to the pronunciation lexicon.

mal system). It was important, therefore, that the verbaliser was compatible with both systems. The problem with verbalising transcribed digits with multiple verbalisation options is that, without listening to the audio, it is impossible to be certain which was actually spoken. Listening and manually transcribing each occurrence of a digit would be too time-consuming, so we required an automated solution. The numbering systems correlate with particular contexts, users and periods of times.[12] After examining each text type, and taking its age and context into account, we estimated its distribution of decimal to vigesimal verbalisations. Digits in the corpus were then verbalised at the estimated distribution.

For augmenting the number of pronunciation lexicon entries, a Grapheme-to-Phoneme (G2P) model was trained. This is a statistical model that learns the relationship between graphemes and phonemes in a given language. It is trained on pairs of words and pronunciations, and can be used to predict pronunciations for OOV words. We used the **Sequitur G2P** Python toolkit (Bisani and Ney, 2008) to train a G2P model on 90% of the original Gaelic lexicon entries. The model achieved a promising string error rate of 3.82% when tested on the remaining 10% of the words in the lexicon. We extracted the full list of words in the training data that did not appear in the lexicon (around 115,000 words), and used the G2P model to predict a pronunciation for each. With some words, the model failed to output a predicted pronunciation. This was often because the word contained graphemes, such as 'z', that are not in the Gaelic alphabet, and were hence unseen to the model during training. We deduced that most of these words were English. We looked them up in an English lexicon, provided by Quorate Technology Ltd., and their English pronunciations were added to the Gaelic ASR lexicon. The resulting lexicon was, therefore, bilingual. This does increase the risk of the ASR system substituting a Gaelic word for an English word in a transcription, however, this was not a noticeable consequence in our experiments. Figure 1 details the full lexicon augmentation process.

The final stage of data preparation was to align each transcription to its corresponding audio. Given the lack of a Gaelic automatic aligner model, this was our most challenging task.

5. Solving the Alignment Problem

5.1. What is Alignment?

Alignment is the process of assigning each word in the transcript a start and end time in its corresponding recording. An automatic aligner does this by mapping words in the transcription, via their component phonemes, to audio frames in the recording. Similar to an AM from speech recognition, the aligner learns the typical low-level acoustic features of each phoneme in the target language. At inference time, each word in the transcript is looked up in the pronunciation lexicon to generate a sequence of phonemes that are known to occur in the recording. The aligner then uses its learned acoustic knowledge to map each frame of speech to a phoneme in that sequence. This way, every word in the transcription is assigned a start and end time via its component phonemes.

5.2. Seed Model for Alignment

As an aligner is trained to recognize language-specific phonemes, it follows that a language-specific aligner is usually required. No Gaelic aligner model existed, and training our own would have required time-aligned training data. Manual alignment was a possible solution, but it would have been too laborious and expensive for the project. To mitigate this circular dependency, we experimented with a non-target language model to seed the alignment process.

Considering that the aligner is provided with a known sequence of words, which can be converted to phoneme sequences via the lexicon, its only task is to predict at which precise times those sequences occur. This is in contrast with speech recognition, where the model must also predict *which* phonemes, and consequent words, are spoken. As the aligner is not required to do

[12]For example, writers in more technical domains, and younger speakers at large, are more likely to use the decimal system, while older speakers tend to use the vigesimal one.

this, it follows that cross-linguistic phonological variation (e.g. differences in phonemes versus allophones), may not be too problematic for the task. Take, for example, an aligner that has been trained on a language which does not distinguish between /k/ and its aspirated equivalent, /kʰ/. If that aligner is used for a language which *does* distinguish between the two, it will at some point be faced with a recording in which /kʰ/ occurs. In this case, the aligner would be able to pick up on the more global features of the /k/ phoneme to make a confident estimate at when its aspirated variant is pronounced. We hypothesised that using a non-target language aligner model would be a viable solution to the task of aligning the Gaelic data. This approach to the task is further described in the next sections.

5.3. Lexicon Phoneset Mapping

We used an English alignment model, provided by Quorate Technology Ltd., to seed the alignment process. The aligner uses a set of 29 English phonemes. The problem with this phoneset is that Gaelic has more phonemes than English. For example, where Gaelic distinguishes between /kʲ/, /kʰ/ and /kʲʰ/, English simply classes these as allophones of the phoneme /k/. The issue arises when the lexicon is used to map the words in the transcription to their known sequence of phonemes in the recording. Because the Gaelic pronunciation lexicon uses the additional Gaelic phonemes, these will be present in the resulting sequence of phonemes to be aligned. Upon encountering /kʲ/, /kʰ/ and /kʲʰ/ in that sequence, the aligner would fail, as these phonemes are not present in its phoneset. For this reason, it is important to match the phoneset used in the pronunciation lexicon to the phoneset that the aligner is able to recognize. We therefore created a mapping between the Gaelic and English phonesets to account for the additional phonemes in Gaelic.

The English aligner uses a computer-friendly English phoneset that is based on ARPABET (Klautau, 2001). Am Faclair Beag, on the other hand, uses a Gaelic adaptation of IPA. Both phonesets can be directly mapped back to Standard IPA (Brown, 2012), making it possible to convert between the two. The Gaelic IPA phonemes were first restored back to their Standard IPA equivalents, which can be found in the 'About' section of the lexicon's website (Bauer and MacDhonnchaidh, 2022). Then, a new mapping was created from the Standard IPA Gaelic phonemes to the subset of those phonemes available to our English aligner model. For phonemes that were shared between the two languages, this was trivial. For each of the Gaelic-exclusive phonemes, however, we decided on an English 'closest equivalent' phoneme. Taking the above example, the closest English phoneme for each of the 3 distinct Gaelic phonemes, /kʲ/, /kʰ/ and /kʲʰ/ was /k/. Each of these Gaelic phonemes was mapped, accordingly, to a single English phoneme. The full phoneset mapping is shown in Table 1. Once the phoneset mapping had been constructed, every Gaelic phoneme in the pronunciation lexicon was converted into its English equivalent. This meant that the phoneset used by the lexicon matched the phoneset used by the aligner. The *pseudo*-Gaelic phoneset, therefore, allowed us to use an English AM towards Gaelic alignment.

GD	IPA	EN	GD	IPA	EN
b	p	p	dʲ	tʲ	tʃ
p	pʰ	p	tʲ	tʲʰ	tʃ
ḃ	ḃ	g	ð	ð	ð
ɣ	ɣ	g	rʲ	ɾʲ	ð
ç	ç	k	r	ɾ	ɾ
g	k	k	R	rˠ	ɹ
gʲ	kʲ	k	a	a	ɑ
k	kʰ	k	aː	aː	ɑ
kʲ	kʲʰ	k	ɛ	ɛ	ɛ
x	x	k	e	e	eɪ
t	t̪ʰ	t	eː	eː	eɪ
d	t̪	t	i	i	i
l	l	l	iː	iː	i
Lʲ	ʎ	l + j	ɪ	ɪ	ɪ
L	l̪ˠ	ɬ	j	j	j
m	m	m	o	o	oʊ
n	n	n	oː	oː	oʊ
ŋ	ŋ	ŋ	ɔ	ɔ	ɔ
ŋʲ	ŋʲ	ŋ	ɔː	ɔː	ɔ
Nʲ	ɲ	n + j	u	u	u
N	n̪ˠ	n̩	uː	uː	u
v	v	v	ɯ	ɯ	ʉ
f	f	f	ɯː	ɯː	ʉ
s	s̪	s	ɤ	ɤ	ʊ
h	h	h	ʊː	ʊː	ʊ
ʃ	ʃ	ʃ	ə	ə	ə

Table 1: Phoneset Mapping. GD = Gaelic Adaptation of IPA, IPA = Standard IPA, EN = English IPA, i.e. Standard IPA phonemes present in the English phoneset

The phoneset mapping was carried out with the assistance of Gaelic language experts, but their expertise was not necessarily a requirement for the task. This is because IPA is a set of phonemes described by their various qualities, such as place and manner of articulation. This information enables those who may not be familiar with certain phonemes, for example, because they are not speakers of a language that uses them, to understand which other phonemes they are related to. In addition, IPA is a fairly global phoneset, making the task possible for a large number of languages.

5.4. Training Data Alignment

Once the lexicon phoneset had been adapted to the aligner's one, the alignment could begin. As the aligned data would be used to train the acoustic model, it was important that the data were aligned accurately. Aligners have two outputs: word-level timings

and a word confidence score for each aligned word. Word confidence scores (see Kemp and Schaaf (1997); Gillick et al. (1997)) measure the probability that a certain word is actually spoken at its given start and end times. The scores can be used to evaluate the accuracy of the alignment – the higher the score, the more likely it is to be accurate. We, therefore, used average word confidence scores to filter the aligned utterances for our final training set.

While aiming for high alignment quality, it is also important to keep in mind that the DNN models used for speech recognition require a large *quantity* of training data. Data that aligns well tends to be less noisy, so including only the best-aligned data would prevent the model from adapting to noisy audio conditions. When filtering the data, therefore, it was necessary to find a balance between quality of alignment and quantity of retained data. We judged that any utterances with an average word confidence score of $< 70\%$ should be discarded. Initially, only a subset of the full training corpus (the Clilstore dataset) was aligned with the English model. The frequency of average word confidence scores for utterances in this initial dataset can be seen in Figure 2.

Given the selection criteria, the initial yield of data was substantial: from 27 hours of data, 21.2 hours, or 78.5%, were retained. This is an indicator of the overall quality of the alignment, which is promising, given the novel cross-lingual approach used. We trained a Gaelic AM using this initial aligned dataset, and, because an AM can also be used as an automatic aligner, we were then able to *re-align* the data using a Gaelic-specific model. We did this twice: first using the Gaelic model trained on the s5b dataset (see Table 2), and again using the model trained on the s5c dataset. As model performance improved from training on a larger dataset, so did the quality of the alignment. This resulted in a higher yield of aligned audio being collected with every re-alignment, as shown in Table 2.

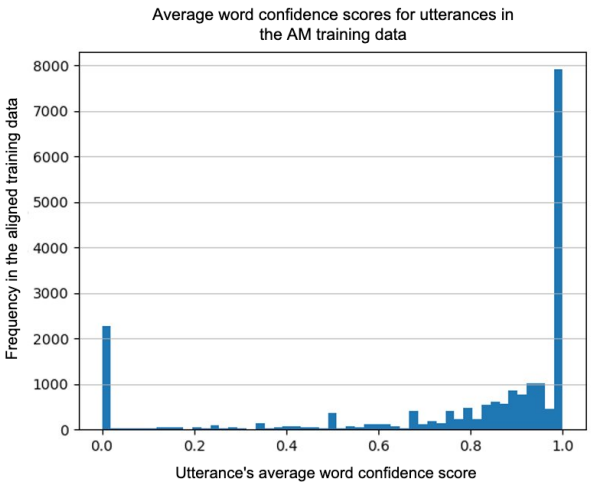

Figure 2: Histogram showing the frequency of average word confidence scores for aligned utterances in the AM training data. These statistics are used to filter well-aligned examples.

6. ASR Model Building

6.1. System Overview

We constructed a number of Gaelic ASR systems using the **Kaldi** speech recognition toolkit (Povey et al., 2011). Kaldi is an open-source toolkit that includes scripts, or 'recipes', that can be used to build and evaluate full ASR systems. Our ASR systems were constructed in an iterative manner. As explained, an initial speech dataset was first aligned with the English aligner. The resulting data was used to train our first Gaelic AM, which could then, itself, be used for alignment. After this point, every new speech corpus obtained was aligned with our latest and most accurate model. The yield from this filtered alignment was added to the AM training data, and used to retrain the AM. The full alignment and training cycle is shown in Figure 3. Additionally, the entire process of data preparation and Gaelic ASR system development is visualised in Figure 4 in Appendix A.

6.2. Acoustic Models

We used the Kaldi AMI recipe (Carletta, 2006) as a starting point for our AM architecture. The recipe, based on Swietojanski et al. (2013), constructs a 15-layer time-delay neural network (see Peddinti et al. (2015)), which increases the number of input context frames at every layer. The initial input to the model is one audio frame *t0* with six surrounding context frames, corresponding to *t-3* and *t+3*. The frames are input as high-dimensional MFCCs (80-dimensions) with 100-dimensional i-vectors. Training ran for 15 epochs. This setup was used for the s5, s5b and s5c models. After s5c, the full set of AM training data was finalised, and so we began experimenting with the model's architecture. This is further detailed in the results section.

6.3. Language Models

We trained various 4-gram language models using the KenLM language modelling toolkit (Heafield et al., 2013). Each model was trained on 90% of the full available text dataset, and evaluated for its perplexity score on the remaining 10%. Two models were used in our final experiments, their only difference being number of tokens of training data, shown in Table 3.

7. Evaluation

For the ASR evaluation dataset, we aimed to extract a set of utterances with a range of speakers, dialects, topics and acoustic environments. This is because our goal was to build a system that performed well on varied Gaelic speech. We extracted utterances from a larger

	Hours			
Dataset	**s5**	**s5b**	**s5c**	**s5d**
Clilstore	21.2	21.2	22.7 **(+1.5)**	23.5 **(+0.8)**
TaD		17.9	22.2 **(+4.3)**	29.6 **(+2.9)**
TaD dump 2			4.5	
Handwriting			13.7	17.6 **(+3.9)**
Zoom Calls			0.2	0.5 **(+0.3)**
Audiobooks				0.9
MG Alba				31.4
Total	21.2	39.1	63.3 **(+5.8)**	103.5 **(+7.9)**

Table 2: Yield of aligned data from each re-alignment. Model training occurred for each new dataset, and re-alignment occurred for dataset s5c and s5d. Bold is the additional hours of data gained from re-alignment.

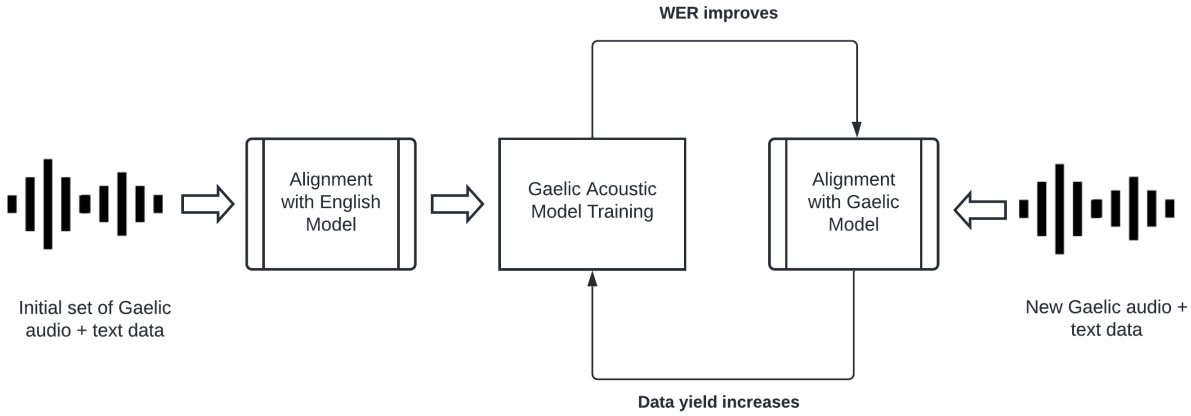

Figure 3: The alignment and training process carried out to iteratively train new models

Model	**Tokens in training data**	**Perplexity**
H	7,378,450	90.1
I	8,593,567	81.3

Table 3: Language Model Perplexity Results

number of short sessions to ensure that the final set had wide variability. We extracted an hour of speech data from the initial dataset that had been aligned with the English model. Of course, this was non-overlapping with the training data. Because the evaluation set is used to assess the performance of the final ASR system, we aimed for a dataset with greater alignment quality than our training data. To facilitate this, the word confidence score filtering threshold was increased from the original 70% to 95%. Once the aligned utterances had been extracted, a Gaelic expert manually corrected the automatic alignments to ensure 100% alignment accuracy. The final evaluation set amounted to 56 minutes of speech data with high quality reference transcripts. We used this evaluation set to generate a word error rate (WER) of each new ASR system, measuring its performance. WER is the standard evaluation metric for ASR, and measures how much the transcription output by the ASR system differs from a reference transcription (Jurafsky and Martin, 2021). WER can be considered similar to $1 - accuracy$.

8. Results

As shown in Table 4, our first Gaelic ASR system achieved a WER of 35.8%. Considering this model was trained on only 21.2 hours of speech data that had been aligned with an English model, this result was promising. As noted previously, the model architecture and training conditions were maintained for models s5, s5b and s5c. In ASR research, increases in training data tend to correlate with improved performance. We report the same: our system's WER improved by 7.6% by simply increasing our training set quantity from 21.2 to 63.3 hours (see model s5c, Table 4).

After training the s5c model, we received new speech data from MG Alba. This increased our AM training set to over 100 hours. It also increased our LM training data by over 1 million tokens. As this would be our final training set, we retrained the LM and began experimenting with the AM architecture. We first decided to

reduce the number of training epochs from 15 to 4 as the training logs suggested many of the later epochs were redundant. Having too many training epochs also risks over-fitting to the training data. Combined with the new LM, we expected a fairly substantial WER reduction (WERR) from the s5c model to the s5d model. However, the WER only decreased by 0.8%. This led us to believe that the size and capacity of the model itself may have been a cause of over-fitting. The model was, therefore, retrained using 11, as opposed to 15 layers, again for 4 epochs. The dimensionality of the MFCCs was also reduced from 80 to 40, as we suspected that the extra input resolution likely did not add much value. This model, shown as s5d-small in Table 4, attained a more substantial WERR from the s5c model: 1.9%. The resulting WERR from our initial to final ASR systems is 9.5%, which is a significant relative improvement of 26.54%.

Model	AM data (hrs)	LM	WER(%)
s5	21.2	H	35.8
s5b	39.1	H	31.0
s5c	63.3	H	28.2
s5d	103.5	I	27.4
s5d-small	103.5	I	26.3

Table 4: ASR Results

9. Discussion

The performance improvements achieved for the Gaelic ASR system are very promising. WER is still high when compared to majority language ASR systems, however, and would not be classed as suitable for production-level ASR. That said, fully automatic transcription tasks have a much more demanding WER threshold than other related tasks. For example, the WER that we achieved is within the threshold required for machine-assisted transcription. Thus, the system could be used, for example, to align a transcription to a video and create subtitles. This would give much added-value to existing Gaelic language resources, and some of the project collaborators have already used the system to do just that. See, for example, the Island Voices videos on Youtube (Wells, 2012), which have been augmented with Gaelic subtitles using the Gaelic aligner model.

In addition to improving the quality of existing resources, the creation of new time-aligned Gaelic transcriptions also creates the opportunity for a feedback loop. This is where the Gaelic system is used to assist in transcribing and aligning new data that can be added to the training dataset. Thus, as the quantity of training data is increased, the performance of the ASR system improves. As shown in our re-alignment process, improvements in the ASR performance also increase the yield of data that can be extracted for training.

Regarding future work, we suggest that a multilingual approach, similar to those described in Section 3, is implemented for the AM. In particular, it could be beneficial to exploit the resources available for Irish. With 1,761,420 speakers in the 2016 census (Central Statistics Office, 2020), Irish is better resourced than Gaelic. It also benefits from dedicated Irish speech and language technology research centres at Trinity College Dublin (Trinity College Dublin, 2019) and Dublin City University. Not only would the incorporation of Irish increase the quantity of data available for training, it would also enable the use of a number of useful language tools that have been built for Irish. Finally, given that the language is closely related to Gaelic, we believe the addition of Irish to the training data would be beneficial: the similarity between the languages would facilitate the recognition of Gaelic phonemes specifically, whilst their differences would improve generalisability to unseen data.

10. Acknowledgements

We gratefully acknowledge funding from the Soillse Research Fund and DDI/SFC's BEACON Build Back Better Open Call: COVID-19 Response and Accelerating Economic and Social Recovery in Edinburgh & South East Scotland. We are also indebted to the following groups and individuals, who provided valuable language data and other support: Am Faclair Beag, Ceòlas Uibhist Ltd, European Ethnological Research Centre, Grace Note Publications, Guthan nan Eilean / Island Voices, LearnGaelic (MG Alba), National Folklore Collection (University College Dublin), Ruairidh MacIlleathain, Sabhal Mòr Ostaig, The National Library of Scotland, The School of Scottish Studies Archives, Tobar an Dualchais / Kist o Riches, University of the Highlands and Islands and Prof Wilson McLeod.

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). [13]

11. Bibliographical References

Batchelor, C. (2016). Automatic derivation of categorial grammar from a part-of-speech-tagged corpus in scottish gaelic. *PARIS Inalco du 4 au 8 juillet 2016*, page 1.

Batchelor, C. (2019). Universal dependencies for scottish gaelic: syntax. In *Proceedings of the Celtic Language Technology Workshop*, pages 7–15.

Bauer, M. and MacDhonnchaidh, U. (2022). Am faclair beag. `https://www.faclair.com/index.aspx?Language=en`. [Online; accessed 19-February-2022].

Bella, G., McNeill, F., Gorman, R., O Donnaile, C., MacDonald, K., Chandrashekar, Y., Freihat, A. A., and Giunchiglia, F. (2020). A major Wordnet for

[13] `http://www.ecdf.ed.ac.uk/`

a minority language: Scottish Gaelic. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2812–2818, Marseille, France, May. European Language Resources Association.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Boizou, L. and Lamb, W. (2020). An online linguistic analyser for scottish gaelic. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020*, volume 328, pages 119–122. IOS Press.

Brown, A. (2012). International phonetic alphabet.

Carletta, J. (2006). Announcing the ami meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Central Statistics Office. (2020). Irish language and the gaeltacht. https://www.cso.ie/en/releasesandpublications/ep/p-cp10esil/p10esil/ilg.

Chen, D. and Mak, B. K.-W. (2015). Multitask learning of deep neural networks for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1172–1183. DOI: 10.1109/TASLP.2015.2422573.

Dupras, V. (2022). num2words. https://github.com/savoirfairelinux/num2words. [Online; accessed 14-April-2022].

Gillick, L., Ito, Y., and Young, J. (1997). A probabilistic approach to confidence estimation and evaluation. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 879–882. IEEE.

Grézl, F., Karafiát, M., and Veselý, K. (2014). Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658. DOI: 10.1109/ICASSP.2014.6855089.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

Huang, J.-T., Li, J., Yu, D., Deng, L., and Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308. DOI: 10.1109/ICASSP.2013.6639081.

Jurafsky, D. and Martin, J. H. (2021). Speech and language processing (3rd edition draft). https://web.stanford.edu/~jurafsky/slp3/. [Online; accessed 13-April-2022].

Kemp, T. and Schaaf, T. (1997). Estimating confidence using word lattices. In *Fifth European Conference on Speech Communication and Technology*. Citeseer.

Klautau, A. (2001). Arpabet and the timit alphabet. https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf. [Online; accessed 14-April-2022].

Klejch, O., Wallington, E., and Bell, P. (2021). The CSTR System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Proc. Interspeech 2021*, pages 2881–2885. DOI: 10.21437/Interspeech.2021-1035.

Lamb, W. and Danso, S. (2014). Developing an automatic part-of-speech tagger for Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 1–5.

Lamb, W. and Sinclair, M. (2016). Developing word embedding models for Scottish Gaelic. In *Actes de la conférence conjointe JEP-TALN-RECITAL*, volume 6, pages 31–41.

Lamb, W. (1999). A diachronic account of Gaelic news-speak: The development and expansion of a register. *Scottish Gaelic Studies*, 19:141–171.

Liu, D., Wan, X., Xu, J., and Zhang, P. (2018). Multilingual speech recognition training and adaptation with language-specific gate units. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 86–90. DOI: 10.1109/ISCSLP.2018.8706584.

Murchú, E. P. Ó. (2019). Using intergaelic to pretranslate and subsequently post-edit a sci-fi novel from Scottish Gaelic to Irish. In *Proceedings of the Qualities of Literary Machine Translation*, pages 20–25.

National Records of Scotland. (2015). Scotland's census 2011: Gaelic report (part 1).

O Maolalaigh, R. (2016). DASG: Digital Archive of Scottish Gaelic/Dachaigh airson Stòras na Gàidhlig. *Scottish Gaelic Studies*, 30:242–262.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*, pages 3214–3218. DOI: 10.21437/Interspeech.2015-647.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Sig-

nal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Rasipuram, R., Bell, P., and Magimai.-Doss, M. (2013). Grapheme and multilingual posterior features for under-resourced speech recognition: A study on Scottish Gaelic. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7334–7338. DOI: 10.1109/ICASSP.2013.6639087.

Renduchintala, A., Ding, S., Wiesner, M., and Watanabe, S. (2018). Multi-modal data augmentation for end-to-end asr. *arXiv preprint arXiv:1803.10299*.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental*, 5:1.

Swietojanski, P., Ghoshal, A., and Renals, S. (2013). Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 285–290. IEEE.

Trinity College Dublin. (2019). Irish speech and language technology research centre. https://www.tcd.ie/slscs/itut/. [Online; Accessed 13-April-2022].

Tüske, Z., Golik, P., Nolden, D., Schlüter, R., and Ney, H. (2014). Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer.

Wells, G. (2012). Series 1 (Gaelic) Island Voices playlist. https://www.youtube.com/playlist?list=PL2770777DF19FEFAF. [Online; accessed 13-Apil-2022].

Yılmaz, E., Heuvel, H. v. d., and van Leeuwen, D. A. (2018). Acoustic and textual data augmentation for improved asr of code-switching speech. *arXiv preprint arXiv:1807.10945*.

Appendix A: Gaelic ASR System Development Process

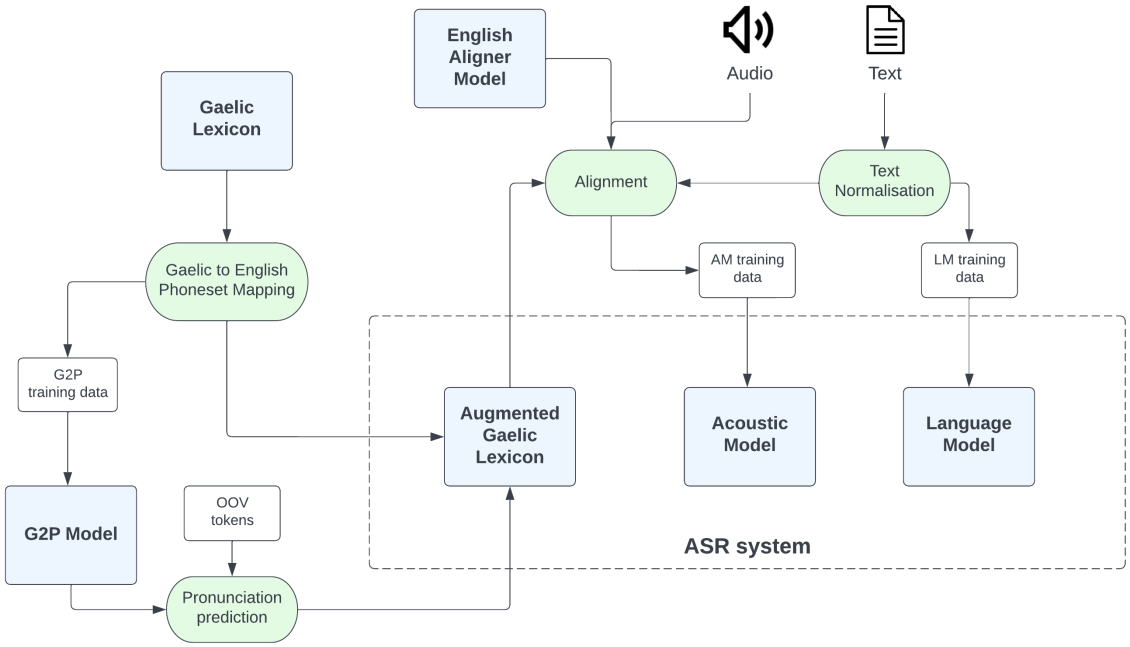

Figure 4: Diagram to show the full Gaelic ASR system development process. The lexicon is adapted (to use a different phoneset) and augmented (using G2P pronunciation prediction). Additionally, audio-to-text alignment creates acoustic model training examples, and text normalisation creates language model training examples. The full ASR system is composed of the augmented lexicon, the acoustic model, and the language model. OOV = Out of Vocabulary.