

Fotheidil: an Automatic Transcription System for the Irish Language

Liam Lonergan¹, Ibon Saratxaga², John Sloan¹, Oscar Maharog¹,
Mengjie Qian³, Neasa Ní Chiaráin¹, Christer Gobl¹, Ailbhe Ní Chasaide¹

¹Phonetics and Speech Laboratory, Trinity College Dublin,

²HiTZ Basque Center for Language Technology, AhoLab,
University of the Basque Country UPV/EHU,

³Department of Engineering, University of Cambridge

¹{llonerga, sloanjo,mahargbo, neasa.nichiarain, cegobl, anichsid}@tcd.ie, ²ibon.saratxaga@ehu.eus, ³mq227@cam.ac.uk

Abstract

This paper sets out the first web-based transcription system for the Irish language - Fotheidil, a system that utilises speech-related AI technologies as part of the AB AIR initiative. The system includes both off-the-shelf pre-trained voice activity detection and speaker diarisation models and models trained specifically for Irish automatic speech recognition and capitalisation and punctuation restoration. Semi-supervised learning is explored to improve the acoustic model of a modular TDNN-HMM ASR system, yielding substantial improvements for out-of-domain test sets and dialects that are underrepresented in the supervised training set. A novel approach to capitalisation and punctuation restoration involving sequence-to-sequence models is compared with the conventional approach using a classification model. Experimental results show here also substantial improvements in performance. The system will be made freely available for public use, and represents an important resource to researchers and others who transcribe Irish language materials. Human-corrected transcriptions will be collected and included in the training dataset as the system is used, which should lead to incremental improvements to the ASR model in a cyclical, community-driven fashion.

1 Introduction

Artificial intelligence (AI) has become a pervasive part of today’s world. While AI undoubtedly brings many benefits, these benefits are felt primarily by speaker communities of the world’s major languages. Speakers of minority languages have not been adequately serviced with technology that works for them and is appropriate for their needs.

Automatic speech recognition (ASR), the process of automatically transcribing speech into text, is a prime example of this disparity. While modern systems for English or Chinese approximate, or even improve upon, the performance of hu-

man transcription, for most languages Speech-to-Text does not exist. One of the largest barriers to developing ASR systems for minority languages is a lack of large, transcribed speech corpora. Recently, approaches leveraging large unlabelled speech corpora, such as semi-supervised learning (Zhang et al., 2020; Radford et al., 2023) and self-supervised learning (Baevski et al., 2020) have achieved state-of-the-art performance for common ASR benchmarks, and were beneficial for low-resource languages (DeHaven and Billa, 2022).

Speech-to-Text integrated technologies like automatic closed captioning on platforms such as YouTube and TikTok have been widely adopted by users of major languages. However, lesser-resourced languages are not included in such services. In light of this gap, we present Fotheidil¹ - a freely available web-based transcription system for the Irish language that utilises various speech-related AI components to transcribe long audio and video files. The structure of the paper is as follows: Section 2 outlines relevant background information; Section 3 details the system Interface; and section 4 describes the transcription pipeline, and the experiments carried out to improve Irish ASR performance using semi-supervised learning (SSL), as well as experiments carried out to train a Capitalisation and Punctuation Restoration (C&PR) model, which improves the legibility of the ASR outputs for the end-user; and finally, Section 5 contains the discussion and conclusions.

2 Background

2.1 Irish language

Irish, a Goidelic or Gaelic language, is a member of the Celtic branch of the Indo-European language family. Today, the Gaelic languages are spoken in small communities scattered mostly along the western seaboard of Ireland and the western islands

¹<https://fotheidil.abair.ie>

of Scotland. The almost extinct Manx, which is currently being revived, is also a Goidelic language and is spoken on the Isle of Man.

The Irish language is highly inflected and has a complex phonological system. The language is diverse in its dialects and accents, with three regional dialects of Ulster (Ul), Connaught (Co) and Munster (Mu) and further sub-dialects, as well as the accents of non-native speakers i.e., learners and new speakers (Nn). The dialects vary significantly in terms of pronunciation, vocabulary and grammar and the phonology and syntactic structure of non-native speakers can often approximate that of English. Speaker variety is used here to describe the dialect or accent of a speaker.

2.2 ABAIR

The ABAIR initiative has been developing speech technology and applications to close the technology gap for the Irish language. Synthetic voices for the 3 major dialects of the languages of Ulster (Ul), Connacht (Co) and Munster (Mu) have been developed, with plans to expand this to further sub-dialects. Additionally, speech recognition systems for Irish have been developed with a sociolinguistic focus, by ensuring that we have adequate coverage of the different varieties of the language where possible and by evaluating our systems for their performance on speakers of different varieties.

2.3 Automatic Speech Recognition

ASR, the task of converting speech into text, has seen significant progress in recent years, due to advances in deep learning, access to hardware such as graphical processing units (GPU) and the increasing use of very large speech corpora. There are two conventional approaches to ASR - the traditional, modular approach, where the speech-to-text task is broken into the distinct components of acoustic modeling, pronunciation lexicon, and language modeling. The sub-modules are modeled independently and then combined in a decoding graph as a weighted finite-state transducer. In contrast, End-to-End (E2E) systems handle the entire speech-to-text task within a single model, directly learning the mapping from audio to text without the need for separate modules, offering a more streamlined but data-intensive approach. While E2E systems have surpassed modular systems in most performance benchmarks for ASR, the need to use large training corpora makes them less suitable for low-resource languages (Loneragan et al., 2024).

2.3.1 Semi-supervised learning

One of the most significant bottlenecks to the development of speech recognition for an under-resourced language is the availability of transcribed audio material to train an ASR system in a supervised manner. However, untranscribed speech is more readily available, due to the increasing proliferation of audio and video materials on the internet. Semi-supervised learning (SSL) is a paradigm that seeks to incorporate large unlabelled datasets in the learning framework to reduce the reliance on a large amount of labelled data. Among various SSL techniques applied to ASR, Noisy Student Teacher training (NST) has gained significant attention, achieving state-of-the-art performances across various datasets (Zhang et al., 2020; Park et al., 2020). Moreover, it has improved performance in code-switching ASR (Xi et al., 2024) and in low-resource ASR (Li and Vu, 2024).

Noisy Student Teacher training: in NST training, a teacher model is trained with the available labelled data. This model is used to generate pseudo-labels for the unlabelled dataset, which is combined with the labelled data to create a new training set for the student model. Noise is introduced to the new training set, forcing the student model to learn to reproduce the teacher model's outputs under noisy conditions, steering the model to learn more robust features that may better match the variability of real use.

2.4 Capitalisation and Punctuation Restoration

The output of the ASR system consists of raw text, using just lowercase characters, without punctuation symbols, acronyms or digits. This format is not very suitable in terms of readability and thus an additional processing is needed to restore proper capitalization and punctuation. Typically, capitalisation and punctuation restoration (C&PR) systems are word-level classifiers which implement either two separate classifiers for capitalisation and for punctuation, or a joint one. A review on these systems can be found in Păiș and Tufiș (2022).

In the case of Irish language, while punctuation rules are analogous to other western languages, capitalisation has its particularities, due to initial mutation, which is indicated orthographically by attaching different particles at the beginning of a word. These particles are one or two letters and are always written in lowercase, while the word keeps

its original capitalisation: *i nGaeilge* (in Irish), *ón bhFrainc* (from France).

These specific cases are not covered by word level capitalisation models for other languages, which usually have just two classes indicating if the initial letter of the word should be lower or upper-cased. There are also character level capitalisation systems, but they perform worse and they are not so common.

Besides capitalisation and punctuation, the readability of the text is improved if numbers are written in digits instead of their textual form and acronyms are written in their condensed form instead of as they have been uttered by the speaker. Additionally, the use of specific symbols like percentage, currency, ordinal markers is also desirable. These fully formatted texts are often referred to as rich transcriptions.

3 Interface

User experience (UX) has been a key concern for the development of the Fothaidil interface. Users of the platform are likely to experience significant wait times while the files they uploaded undergo processing and recognition. These wait times have been shown to exhibit a negative logarithmic relationship with user satisfaction (Egger et al., 2012; Reichl et al., 2010). Mitigating the negative effects of long loading times in UX is typically tackled in two ways: speeding up processing; and reducing the user’s frustration or perception of wait times. Increasing the processing speed is largely dependent on available hardware but can be aided by efficient infrastructure design. Reducing users’ perception of waiting times can be achieved on the front end through effective use of loading visualisations (Kim et al., 2017). A description of both is provided below, with reference to Figure 1.

3.1 Backend

Three main back-end functionalities are hosted on separate Virtual Machines (VMs) to avoid competition for CPU resources. Media processing is carried out on one VM, voice activity detection, speaker diarisation and ASR on another, with database storage on a third. When a user uploads a media file, it is first directed to the Media Server where the audio is stripped/converted to a wav file with a 16000 Hz sampling rate and video compression takes place if necessary. The converted wav file is then sent to the Recognition VM

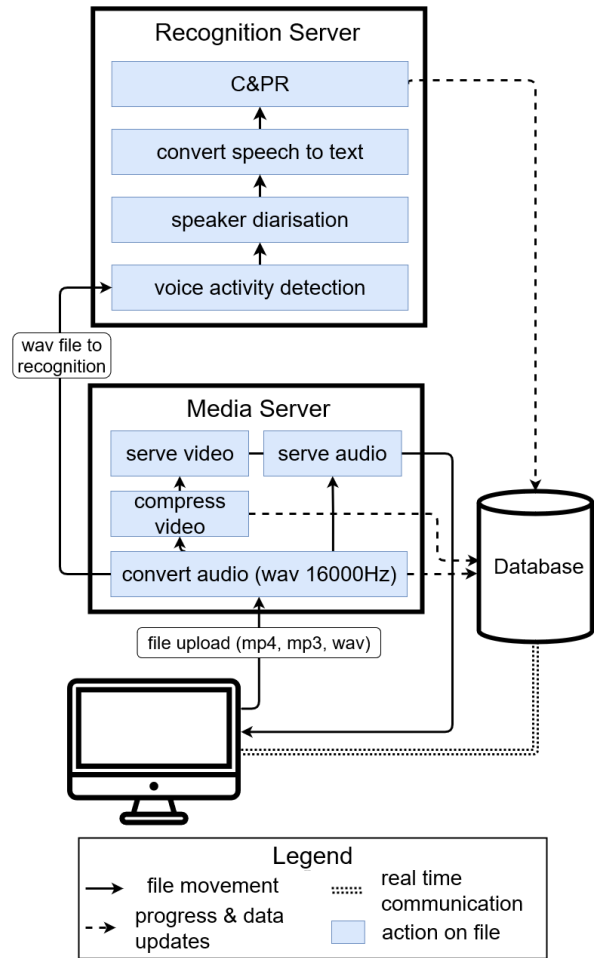


Figure 1: Infrastructure Diagram

where speaker diarisation and recognition models are hosted. Updates on the progress for each of the processes with potentially long wait times are continuously stored on the database.

Real Time Communication (RTC) between the front end and the database enables the user to view progress for each of these back-end processing steps. The main dashboard for interacting with, and editing, the processed data is shown in Figure 2. Users are able to edit the text, times and speaker as well as download the output in pdf, docx or srt format.

4 Transcription Pipeline and Experiments

The transcription pipeline is a multi-step process, which brings together different systems to transcribe long audio files into text. The choice of the models used in some of these steps is made by weighing up their performance and efficiency.

The process is as follows:

- i. Upload audio or video file.
- ii. Extract or convert audio to 16kHz mono wav.

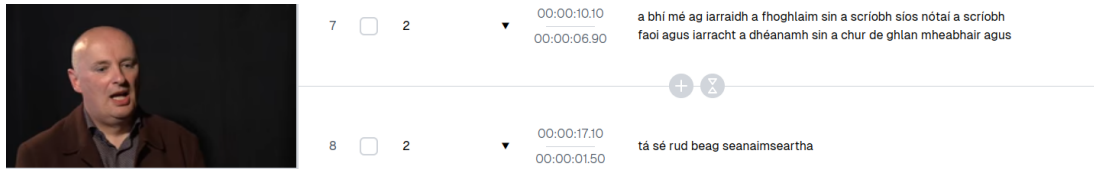


Figure 2: Main User Interface

- iii. Voice activity detection to create segments.
- iv. Speaker diarisation to assign speaker labels within speech segments.
- v. Continuous segments of same speaker are joined.
- vi. Segments are decoded with ASR system.
- vii. ASR output is enhanced using a C&PR model.

While there are better performing alternatives that make use of GPUs, our web-based service is limited to CPU usage only. The voice activity detection and speaker diarisation systems are off-the-shelf, pretrained models and are detailed briefly in Sections 4.1 and 4.2, while the ASR and Punctuation and Capitalisation models, which have been trained specifically for Irish, and these experiments are described in Sections 4.3 and 4.4.

4.1 Voice Activity Detection

Voice activity detection is the process of finding speech segments within an audio file. The voice activity detection (VAD) module used is Silero-VAD (Silero, 2024) - a robust, lightweight, pre-trained model. While PyAnnote is the conventional off-the-shelf choice for both VAD and speaker diarisation, it requires a GPU to be used efficiently. Silero-VAD offers a CPU only alternative with competitive VAD performance.

4.2 Speaker Diarisation

The goal of speaker diarisation is to assign a speaker label to each speech segment. A pretrained Kaldi-based speaker diarisation x-vector (Snyder et al., 2018) model² that is trained using augmented VoxCeleb1 and VoxCeleb2 datasets is used as part of this pipeline. The model has a reported EER performance of 3.7% on the Speakers in the Wild speaker identification test set.

4.3 Automatic Speech Recognition

Modular ASR approaches are often more suited to low-resource domains, as they do not require the

²<https://kaldi-asr.org/models/m8>

same amount of data as E2E approaches. Additionally, modular systems are optimised to run efficiently on the CPU. Therefore, Kaldi-based DNN-HMM ASR models are used in our system and in the following experiments.

A baseline supervised model M_0 is trained using the supervised training set of 398h as described in Section 4.3.1 and detailed in Table 6 in the Appendix. To explore the usefulness of SSL for Irish ASR, a version of the approach outlined in Manohar et al. (2018), modified to include the noising element of NST, is tested in these experiments. The teacher model M_0 is used to decode the unlabelled set of 3230h in an undetermined fashion, preserving the full decoding lattices. These lattices are rescored using a large n-gram language model (LM) and the best path through the rescored lattices is found. The best paths are taken as pseudo-labels and are combined with the supervised training set to create a semi-supervised training set. The student model M_1 is trained with the semi-supervised training set using the noising technique Spectral Augment (Park et al., 2015).

4.3.1 Data

The supervised acoustic training set comprises various datasets, as described in Table 6 in the Appendix with a breakdown of duration by speaker variety. Recordings used for ABAIR synthetic voices of the three dialects are used, totaling 41.4h (Syn). MíleGlór (MG) is an initiative for recording Irish speakers in the field and online using dialect-specific prompts, and a portion of 17.3h of this set is used. Additionally, two spontaneous speech (SS) corpora are combined, the large *Corpas na Cainte Beo* and the smaller *Comhrá*, totaling 259.7h. Audiobooks consisting of both professional and home recordings make up 36.6h. *Caint Chonamara*³ (CCh), is a collection of conversations that was recorded in the Conamara area in 1964, representing rich dialectal speech of the Co dialect. *Báiliúchán Béaloidis Árann* (BBhÁ) is a folklore

³<https://www.sksk.de/index.php/de/veroeffentlichungen-2/materialien/33-caint-chonamara>

collection of conversational speech from the Aran Islands⁴. Datasets SS, AB, CCh and BBhÁ were aligned using the alignment protocol set out in Lonergan et al. (2024).

The unsupervised acoustic data consists of broadcast recordings from four Irish language radio shows featured on Raidió na Gaeltachta: Barrscéalta, which mainly features speakers of Ul Irish; Adhmhaidin, which primarily contains speakers of the Co dialect; An Saol Ó Dheas which largely features Mu speakers; and Nuacht a hAon, which has a mix of dialects. These recordings are downloadable in MP3 format from Raidió na Gaeltachta’s podcast page. Silero-VAD, as described in Section 4.1, is used to find speech chunks for decoding and resulted in 3230h. A breakdown in duration by radio show is provided in Table 8 in the Appendix.

Five test sets are used to evaluate the system and details for these sets are given in Table 9 in the Appendix. The first two are portions of MG and SS taken from the training set corpora, ensuring no data leakage, and can be considered as in-domain tests. These sets are 10.2h and 28.2h in length respectively and their speaker variety breakdown in duration is detailed in Table 7 in the Appendix. Two additional out-of-domain test sets are the Irish test portions of CommonVoice (CV) (Ardila et al., 2020) and Fleurs-R (FL) (Ma et al., 2024) datasets. The quality of these datasets is markedly poor. Both datasets feature predominately non-native (Nn) speakers, and the texts for FL seem to be machine-translated English texts, which contain many foreign proper nouns. However, as they are publicly available and out-of-domain, we have included them here. They are 0.6h and 2.2h long respectively. Finally, 10 minutes each from the four radio shows (0.7h) from which the unsupervised dataset is created, were hand-labelled by the authors and are used here for evaluation (HL). While these do not appear in the unsupervised set, there is overlap in terms of speakers and likely content.

The text corpus of 36.6 million words used for LM training is comprised of normalised versions of the New Corpus of Ireland (Kilgarriff et al., 2006) (c. 30m words), the Bible (c. 0.1m words), Irish language Wikipedia texts (2.9m) and the supervised training set texts (4.6m).

⁴<https://bba.duchas.ie/en/about/bba>

4.3.2 Experiment

The acoustic model (AM) in the baseline ASR system M_0 is a Time-Delay Neural Network (TDNN) (Peddinti et al., 2025), trained using the 398h supervised train set (see Table 6) for 4 epochs. The initial alignment is produced by a triphone GMM-HMM trained with standard MFCC features, applying linear discriminative analysis (LDA), maximum likelihood linear transformation (MLLT), feature space maximum likelihood linear regression (fMLLR) and speaker adaptive training (SAT). The features for training the TDNN model are 40-dimensional high-resolution MFCCs stacked with 100-dimensional online extracted i-vectors. Two widely used on-the-fly data augmentation techniques for ASR – speed perturbation (Ko et al., 2015) with factors of 0.9, 1 and 1.1, and spectral augmentation (Park et al., 2015) were applied to augment the AM training data. The TDNN model consists of 6 TDNN layers with a hidden dimension size of 768. A pronunciation dictionary based on the Global rules, as described in Qian et al. (2022) and Lonergan et al. (2023a), which capture cross-dialect variation in the pronunciation of phonemes and morphemes, is used, along with a 4-gram LM, trained using the text corpus described in the last paragraph of Section 4.3.1.

As described in Section 4.3, pseudo-alignments for the unsupervised data are acquired by decoding the data using M_0 , rescoring the undetermined decoding lattices and finding the best path for each utterance. Rescoring is done using a 5-gram LM trained with the same texts described in Section 4.3.1. The resulting unsupervised and supervised alignments are combined with equal weighting. These semi-supervised alignments are then used to train M_1 with SpecAug for 6 epochs, with the same AM structure, lexicon and LM as M_0 .

Recurrent neural network LMs (RNNLM) are beneficial in rescoring n-best lists generated by an ASR system (Xu et al., 2018). An RNNLM is trained on the text corpora listed in Section 4.3.1 and is used in these experiments. Where results including RNNLM are reported, they are labelled as (+LM).

4.3.3 Results

The WERs for in-domain test sets MG and SS have a relative improvement of 9% and 2% respectively with M_1 . For the out-of-domain test sets CV and FL, the performance improves by 14% and 7% relatively. For HL, which more closely matches the

	MG	SS	CV	FL	HL
M ₀	14.1	27.3	27.5	51.9	22.1
M ₁	12.8	26.7	23.7	48.5	16.1
+LM	10.9	24.0	19.6	44.5	14.1

Table 1: ASR performance breakdown of models M₀ and M₁ of test sets and RNNLM rescoring (+LM).

	Overall	Ul	Co	Mu	Nn
M ₀	14.1	18.5	14.0	10.8	13.2
M ₁	12.8	15.3	13.1	10.2	12.6
+LM	10.9	12.7	11.6	8.8	10.4

Table 2: ASR performance breakdown by speaker variety of MíleGlór test set.

unsupervised data, there is a more dramatic relative performance improvement of 27%. RNNLM rescoring improves performance across the board and is complementary with the improved, semi-supervised acoustic model. Table 2 provides a breakdown of the performance on the MG set by speaker variety. The starkest improvement brought by the SSL approach to MG is the relative WER reduction of 17% for Ul speakers.

From these results, it is clear that SSL most significantly impacts performance on out-of-domain datasets, or domains more similar to the unsupervised training set (i.e. HL). Another noteworthy result is the boost in performance of Ul speakers, which is the least represented of the three dialects in the supervised training set (see Table 6). Previous studies on Irish dialect bias in ASR have shown that the Mu and Co dialects reinforce each other in terms of performance, whereas Ul, being a more distant dialect, is an outlier (Lonergan et al., 2023b). The improvement can be explained by Ul being well represented in the unsupervised set, indicating that such dialect bias can be alleviated using SSL. The improvements could be increased further by repeating this experiment multiple times, using the student of a previous experiment as the teacher for the next, or by increasing the size of the unlabelled dataset.

4.4 Capitalisation and Punctuation Restoration

In this work, we propose a novel approach to tackle the punctuation and capitalisation task, namely, a sequence-to-sequence (S2S) approach which will target all the rich transcription features in a unified way. The input for such a model is uncapitalised and unpunctuated i.e., as close as possible to the

actual output of the ASR system. The output is the same text in its rich transcription format with correct capitalisation and punctuation, while also including digits and acronyms. The conventional approach is to use a classifier, which for each word in the input text, predicts whether the word should be followed by punctuation or should be capitalised, however as the input and output texts do not have a one-to-one word correspondence, a S2S architecture is more appropriate.

To that aim, our proposed model is a transformer based machine translation model, implementing the original model by Vaswani et al. (2017), which is based on attention mechanisms. We used the MarianNMT implementation (Junczys-Dowmunt et al., 2018) of this architecture.

For comparison purposes we have also tested a baseline system using a classification model (CLAS), Nvidia’s Nemo Punctuation and Capitalisation Model⁵. This model features two token-level classifiers on top of a pre-trained BERT LM. For each word in the input text, the model predicts a punctuation mark that should follow the word, if any, and predicts also if the word should be capitalized or not. The output text is then regenerated applying the classification results to each input word. The classes of the original capitalisation model are expanded to include two additional classes for 2nd and 3rd letter capitalisation. The punctuation classifier has been trained with seven classes: commas, periods, question marks, exclamation marks, semicolons, colons and none.

4.4.1 Data

A text corpus of 5 million Irish sentences has been used to train the model. This corpus consists of the Irish section of the Paracrawl corpus (PC_ga) (Esplà et al., 2019), and the already mentioned New Corpus of Ireland (NCE), the Bible (BI) and the spontaneous speech corpus texts from the supervised training set, excluding the sentences used for testing (SS). The details are shown in table 11.

Four additional datasets have been used for evaluation: The Irish Language part of the FLoRes evaluation dataset (FO) (Team et al., 2022), commonly used for machine translation evaluation for low resourced languages and the MiléGlór (MG), Fleurs-R (FL) and CommonVoice (CV) evaluation datasets, employed also for the evaluation of

⁵https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/nlp/punctuation_and_capitalization.html

the ASR. The details of these databases are summarised in Table 10 in the Appendix.

The original text corpora were cleaned to create the training and evaluation datasets, removing non-standard characters, brackets, curly brackets and parenthesis, and standardising the use of spaces, quotes and so on. This clean text is the ground truth that will be used as target dataset in the case of our machine translation model. It will be referred as rich transcription (RT) dataset.

In order to obtain an input text as similar as possible to plain text output of an ASR system, the ground truth target was processed by the normalisation module of the Abair (Murphy et al., 2023) speech synthesis system. The normaliser converts every digit, acronym and some symbols into pronounceable texts, keeping the punctuation and capitalisation of the text. We will refer to this dataset as normalised rich transcript (NR). This dataset is used as ground truth to train the classifier system. The normalised text is then stripped out from any non-alphabetic character and lower cased, obtaining the input dataset (IN) for both of the systems.

4.4.2 Experiment Set-up

The proposed S2S model has a transformer architecture, with 8 heads, 6 encoding and 6 decoding layers, transformer dropout of 0.1 and tied embeddings. The training was done using label smoothing of 0.1, learning rate of $3 \cdot 10^{-4}$, warm-up stage and early stopping using cross-entropy, perplexity, BLEU detok, and CE-mean-words as validation metrics and a beam size of 6.

The baseline classifier system used the standard architecture of the NeMo model. It was trained using Google’s pretrained BERT-base-uncased⁶.

As the approach is a sequence to sequence task, to evaluate the systems, our main metric is a modified word error rate that uses the rich transcript as ground truth (instead the usual uncapitalised, unpunctuated text). We will denote it as WER_{pc} to distinguish it from the usual WER in ASR. We are also using character error rate (CER) to mean calculated with the rich transcript as target. Along with these metrics, we also use the BLEU score (Post, 2018), a common machine-translation metric which measures the similarity between generated and reference translations using n-grams.

⁶<https://huggingface.co/google-bert/bert-base-uncased>

	S2S			CLAS		
	Capt	Punct	WER_{pc}	Capt	Punct	WER_{pc}
FO	0.98	0.96	7.87	0.97	0.96	7.93
MG	0.98	0.95	8.36	0.97	0.95	9.69
CV	0.97	0.89	15.17	0.96	0.89	16.68
FL	0.97	0.96	8.27	0.97	0.96	8.53
ALL	0.98	0.95	8.40	0.97	0.95	9.38

Table 3: Capitalisation and punctuation accuracy and WER_{pc} using normalised rich transcripts (NR) as target.

4.4.3 Experiments and Results

We have performed two experiments to evaluate the proposed model using two metrics. Firstly, we compare the S2S approach with the baseline classifier approach. Due to the more limited capabilities of the classifier, and to allow a fair comparison of the performance of both models, we have used the normalised rich transcripts (NR) datasets as targets and the lower-cased, punctuation removed versions as input (IN). In this setup the input and output text are exactly the same with the only difference of punctuation marks and capitalisation.

The main metric used for comparison here is accuracy. This gives a general idea of the performance of the systems although the classes are severely unbalanced. The results in Table 3 show the accuracy and resulting WER_{pc} of both classifiers: capitalisation and punctuation. Both systems perform well with the proposed S2S system showing slightly higher accuracy and better WER_{pc} .

The second experiment setup reflects the actual use case of the restoration system: plain text at the input (the IN dataset) and full rich transcription (RT) at the output. Table 4 presents 3 groups of results: No C&P correspond to the comparison between the input (IN) and target outputs (RT) without any C&PR system and gives an idea of the disparity of both datasets, defining the maximum error level (or minimum BLEU) that will be corrected by the restoration systems. S2S and CLAS groups correspond to the results of both systems.

The results show that both systems are effective in C&PR, obtaining important reductions in the WER_{pc} and CER metrics. Our proposed S2S system reduces the WER_{pc} and CER by more than 50% and improves the BLEU more than 20 points for all datasets. S2S clearly outperforms the baseline classifier in this experiment, because it not only restores the casing and punctuation more effectively, but also changes digits or acronyms to a textual output that is closer to the target rich tran-

	No C&PR			S2S			CLAS		
	WER _{pc}	CER	BLEU	WER _{pc}	CER	BLEU	WER _{pc}	CER	BLEU
FO	22.2	7.7	64.7	7.9	1.9	85.1	13.62	5.31	80.1
MG	18.5	5.0	66.1	8.34	1.69	88.4	10.18	2.25	84.7
CV	25.6	5.7	61.4	18.29	4.94	82.9	16.78	3.19	79.9
FL	21.7	7.2	64.1	8.52	1.99	84.3	13.33	4.83	80.3
ALL	19.7	5.8	65.5	8.50	1.83	87.4	11.29	3.14	83.6

Table 4: C&PR performance with rich transcription (RT) as target.

	ASR	No C&PR			S2S		
	WER	WER _{pc}	CER	BLEU	WER _{pc}	CER	BLEU
MG	10.9	26.11	9.65	55.2	18.78	6.76	72.3
CV	19.6	40.43	16.59	45.5	34.75	15.63	63.0
FL	44.5	54.58	30.35	24.9	50.70	28.65	31.9

Table 5: Performance of the S2S C&PR system on ASR generated text with rich transcriptions (RT) as target.

scription. The error rates for S2S are below 9% for all the databases, except CV. This database contains a large number of very short, fragmentary sentences with inconsistent casing and punctuation, which may be interpreted as titles.

Finally Table 5 shows the effect of applying the S2S system to the actual output of the ASR. The input text is the plain text generated by the M_1 (+LM) ASR model. The WER of these texts compared to the uncapitalised and unpunctuated references as shown previously in Table 1 is shown in the column ASR for readability. The results in column No C&PR show the WER_{pc} of the same texts when they are compared to the rich transcription. The columns under S2S show the results when the restoration system is applied. The final WERs are always lower than the accumulated WERs of the ASR and the S2S, suggesting that the degradation in the input text generated by the ASR does not impact in the performance of the S2S system.

5 Conclusion and Future Work

This paper constitutes an important step towards democratising speech-related AI technologies for the Irish language and its speakers. The ASR experiments have demonstrated that SSL learning is an attractive solution to improve performance for out-of-domain datasets and underrepresented dialects in the supervised training set. As stated, this improvement can be increased by iteratively repeating this process or increasing the size of the unlabelled dataset. Future work will explore SSL further.

The S2S model offers an elegant solution to the C&PR problem, improving significantly over the

baseline due to its ability to effectively deal with the lack of a one-to-one relationship between the outputs of an ASR system and the rich transcriptions. The S2S model could additionally be trained using the output texts of the ASR systems so that it will be able to correct some of the ASR errors. Furthermore, it can be trained to convert specific keywords, such as punctuation symbol names, currency or units, which would be very useful for dictation applications.

Acknowledgments

This work is part of the AB AIR initiative, which is supported by the Department of Tourism, Culture, Arts, the Gaeltacht, Sport and Media, with funding from the National Lottery, as part of the 20-year Strategy for Irish. The publication was partly funded by the Irish Research Council under grant number 214047. Part of this work was supported by the José Castillejo Mobility Programme of the Spanish Ministry of Science, Innovation, and Universities (CAS23/00294). The C&PR module is partially based on findings from the HiTZketan project of the UPV/EHU (COLAB22/13). We would like to especially acknowledge Dr. Gorka Labaka for his collaboration in the initial implementation of this system.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the*

- Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.
- Mitchell DeHaven and Jayadev Billa. 2022. Improving low-resource speech recognition with pre-trained speech models: Continued pretraining vs. semi-supervised training. *arXiv preprint arXiv:2207.00659*.
- Sebastian Egger, Tobias Hossfeld, Raimund Schatz, and Markus Fiedler. 2012. Waiting times in quality of experience for web based services. In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 86–96. IEEE.
- Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the new corpus for ireland. *Language Resources and Evaluation*, 40:127–152.
- Woojoo Kim, Shuping Xiong, and Zhuoqian Liang. 2017. Effect of loading symbol of online video on perception of waiting time. *International Journal of Human-Computer Interaction*, 33(12):1001–1009.
- Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.
- Chia-Yu Li and Ngoc Thang Vu. 2024. Improving noisy student training for low-resource languages in end-to-end ASR using CycleGAN and inter-domain losses. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 133–142.
- Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023a. Cross-dialect lexicon optimisation for an endangered language ASR system: the case of Irish. In *Proc. Interspeech 2023*, pages 4865–4869.
- Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2024. Low-resource speech recognition and dialect identification of Irish in a multi-task framework. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 67–73.
- Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023b. Towards dialect-inclusive recognition in a low-resource language: are balanced corpora the answer? In *Proc. Interspeech 2023*, pages 5082–5086.
- Min Ma, Yuma Koizumi, Shigeki Karita, Heiga Zen, Jason Riesa, Haruko Ishikawa, and Michiel Bacchiani. 2024. Fleurs-r: A restored multilingual speech corpus for generation tasks. In *Proc. Interspeech 2024*, pages 1835–1839.
- Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur. 2018. Semi-supervised training of acoustic models using lattice-free MMI. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4844–4848. IEEE.
- Andy Murphy, Liam Lonergan, Mengjie Qian, Harald Berthelsen, Christoph Wendler, Neasa Ní Chiaráin, Ailbhe Ní Chasaide, and Christer Gobl. 2023. Abair & Éist: a demonstration of speech technologies for Irish. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 121–123.
- Daniel Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc Le. 2020. Improved noisy student training for automatic speech recognition. In *Proc. Interspeech 2020*, pages 2817–2821.
- Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2015. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech 2015*, pages 2613–2617.
- Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2025. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proc. Interspeech 2024*, pages 3214–3218.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Vasile Păiș and Dan Tufiș. 2022. Capitalization and punctuation restoration: a survey. *Artif. Intell. Rev.*, 55(3):1681–1722.
- Mengjie Qian, Harald Berthelsen, Liam Lonergan, Andy Murphy, Claire O’Neill, Claire Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2022. Automatic speech recognition for Irish: testing lexicons and language models. In *Proceedings of the 33rd Irish Signals and Systems Conference (ISSC)*, pages 1–6.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Peter Reichl, Sebastian Egger, Raimund Schatz, and Alessandro D’Alconzo. 2010. The logarithmic nature of QoE and the role of the Weber-Fechner law in QoE assessment. In *2010 IEEE International Conference on Communications*, pages 1–5. IEEE.

Team Silero. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. <https://github.com/snakers4/silero-vad>.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2005.09629*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu Xi, Wen Ding, Kai Yu, and Junjie Lai. 2024. Semi-supervised learning for code-switching asr with large language model filter. *arXiv preprint arXiv:2407.04219*.

Hainan Xu, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur. 2018. A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5929–5933. IEEE.

Yu Zhang, James Qin, Daniel Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.

A Appendix: Additional Tables

	Full	UI	Co	Mu	Nn
Syn	41.4	22.2	8.8	10.4	-
MG	17.3	3.1	5.6	3.4	5.2
SS	259.7	55.5	104.9	95.4	3.9
AB	36.6	10.3	2.5	14.6	9.2
CCh	25	-	25.0	-	-
BBhÁ	17.9	-	17.9	-	-
Total	397.9	91.1	164.7	123.8	18.3

Table 6: Duration breakdown in hours of ASR training set by speaker variety.

	Total Dur (h)	UI	Co	Mu	Nn
MG	10.2	2.5	3.4	3.0	1.3
SS	28.2	7.5	10.4	10.3	-

Table 7: Duration breakdown by speaker variety of MíleGlór and Spontaneous Speech test sets.

	Dialect	Dur (h)
Adhmhaidin	Co	779.0
Barrscéalta	UI	1002.4
Saol Ó Dheas	Mu	993.9
Nuacht	mix	400.0
Total	-	3230.0

Table 8: Duration in hours and dialect information of unsupervised set by radio show

	MG	SS	CV	FL	HL
#Utts	8,423	19,266	516	548	198
Dur (h)	10.2	28.2	0.6	2.2	0.7

Table 9: Number of utterances and total duration of test sets.

	#lines	#words	#chars
FO	1012	25772	163254
CV	515	3423	19617
FL	548	13634	86282
MG	8423	98463	559675

Table 10: Features of the databases used for evaluation

	#lines	#words	#chars
PC_ga	3.2	63.1	417.2
NCE	1.8	30.4	181.5
BI	0.3	0.8	4.5
SS	1.4	3.6	20.6

Table 11: Features of the databases used for training the Cap&Punct systems (numbers in millions)