

HOW I BUILT ASR FOR ENDANGERED LANGUAGES WITH A SPOKEN DICTIONARY*

Christopher Bartley, Anton Ragni

School of Computer Science, University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK
 {csjbartley1, a.ragni}@sheffield.ac.uk

ABSTRACT

Nearly half of the world’s languages are endangered. Speech technologies such as Automatic Speech Recognition (ASR) are central to revival efforts, yet most languages remain unsupported because standard pipelines expect utterance-level supervised data. Speech data often exist for endangered languages but rarely match these formats. Manx Gaelic (~2,200 speakers), for example, has had transcribed speech since 1948, yet remains unsupported by modern systems. In this paper, we explore how little data, and in what form, is needed to build ASR for critically endangered languages. We show that a short-form pronunciation resource is a viable alternative, and that 40 minutes of such data produces usable ASR for Manx (<50% WER). We replicate our approach, applying it to Cornish (~600 speakers), another critically endangered language. Results show that the barrier to entry, in quantity and form, is far lower than previously thought, giving hope to endangered language communities that cannot afford to meet the requirements arbitrarily imposed upon them.

Index Terms— low-resource, critically-endangered, automatic speech recognition

1. INTRODUCTION

More than 1,400 languages already have fewer than 1,000 speakers. At one loss every two weeks, these would disappear in roughly 54 years. Reversing this trend depends on community effort to strengthen transmission, expand learning opportunities, and raise visibility. Automatic Speech Recognition (ASR) has been shown to be central to these efforts [1], making oral archives more accessible to learners, educators, and researchers. Yet many endangered languages remain unsupported because they lack what is believed to be the necessary resource; an *utterance-level* corpus of segmented speech–text pairs (5–15 s clips with per-segment transcripts). Nevertheless, developing speech technology for these languages remains one of the most pressing areas of research.

By contrast, endangered language communities tend to create transcribed speech in two different formats. The first we refer to as *short-form*; isolated word or brief-phrase recordings, typically up to 5 seconds long. For example, Forvo [2] is a crowdsourced pronunciation resource spanning 430+ languages, substantially more than modern multilingual speech datasets [3–5]. The second is continuous recordings that run for minutes or even hours, such as radio broadcasts, interviews, and folklore (henceforth *long-form*). In both cases, transcription text may be verbatim, but is often interspersed with translations, explanations, and other meta language. These “in-the-wild” data exist because they meet the most immediate needs of

*Additional long-form speech-transcription is segmented and used to refine our models.

the community. One could argue that the challenge for speech technology is to adapt to such resources, rather than asking communities to build datasets whose immediate value to them is uncertain.

Languages with small speaker populations, such as Manx and Cornish, are often termed *low-resource*, as though the challenge were simply a lack of data. Consequently, speech technologies for *low-resource* languages tend to rely on crowdsourced utterance-level datasets [5, 6] and untold amounts of compute poured into massive pretrained models [7]. However, these requirements are not always feasible for *endangered* language communities, particularly for the smallest speaker populations whose languages are continually unrepresented in modern multilingual speech datasets and benchmarks, such as FLEURS [4], and CommonVoice [5].

Prior work in this area has explored extremely low-resource scenarios, such as the IARPA BABEL program, which used as little as three hours of manually aligned and segmented telephone speech [8, 9] to build ASR. Although unsupervised learning and zero- or few-shot adaptation have pushed requirements lower [10–12], success often depends on similarity to languages seen during pre-training. [13]. In contrast, NLP research often alleviates supervision requirements in low-speaker settings by leveraging rudimentary resources common to many languages. For instance, bilingual dictionaries have proven highly effective in machine translation, both as the primary source of supervision [14] and as components integrated directly into model architectures [15]. Similarly, MTOB (Machine Translation from One Book) is a benchmark that asks large language models to translate Kalamang (~200 speakers) by reading a single grammar book [16]. To the best of our knowledge, no one has explored using such rudimentary resources for ASR.

In this work, we ask *What are the minimum requirements for developing ASR for critically endangered languages?* and *Can short and long-form resources substitute for an utterance-level speech corpus?* We address these questions with a series of experiments and make the following contributions:

- i) We present the first ASR systems for two critically endangered languages, Manx and Cornish.
- ii) We show how 40 minutes of short-form speech can produce ASR (<50% WER) for endangered languages.
- iii) With just 8 minutes of short-form speech, we train an initial ASR that can automatically segment long-form recordings for further ASR refinement.
- iv) We achieve usable transcription technology (<25% WER) with 40 additional minutes of segments.
- v) We create the first utterance-level speech corpora for Manx (18 hours) and Cornish (39 hours).

The rest of this paper is organised as follows. Section 2 discusses the background of this paper, section 3 presents a guide to data preparation, section 4 model development and results, and section 5 concludes our findings.

2. BACKGROUND

2.1. End-to-end (E2E) Models

Fine-tuning end-to-end (E2E) models has emerged as the dominant paradigm for extending speech technologies to low-resource languages. Meta’s MMS claims ‘coverage’ of 1,000+ languages [7]. They define ‘coverage’ as the share of languages whose in-domain test sets achieve <5% CER under a single model trained on New Testament readings across 1,107 languages (MMS-lab). Broad-domain ASR is shown only for a smaller subset of 102 languages. MMS is just one example of approaches that rely on extensive pre-training followed by fine-tuning. However, these methods come with broader downsides. Firstly, fine-tuning requires utterance-level speech data, which most endangered languages do not have. Secondly, popular fine-tuning toolkits like HuggingFace Transformers [17], limit control over model tuning to basic hyperparameters (e.g., learning rate, batch size), thus ignoring the benefits one could gain from finer-grained controls, such as incorporating specialised lexicons and external language models. Whilst it is possible to extend the standard recipes with these capabilities, doing so is difficult for non-experts. Lastly, fine-tuning E2E models is commonly believed to lead to state-of-the-art results universally, but this assumption does not hold for many low-resource languages [18].

2.2. Hidden Markov Models (HMMs)

In contrast, Hidden Markov Model (HMM) systems could be far more data and compute-efficient [19]. For ASR tasks they are typically paired with a Gaussian mixture model (GMM) or a deep neural network (DNN). Their modularity requires finer-grained control over components such as acoustic modelling, pronunciation lexicons and language models, the latter of which enable greater integration of linguistic knowledge via web-based text data [20] (news, Wikipedia, digitised books), which are often more abundant than speech data for written languages. However, these systems are often regarded as difficult to use, with the dominant implementation in the Kaldi toolkit [21] demanding significant expertise and engineering effort. They too require utterance-level speech data, and while their modularity remedies some of the shortcomings of E2E modeling, their degree of success is often language-specific [22], depending heavily on the availability of linguistic resources like lexicons and text corpora.

2.3. Community Efforts for Speech Technology

Both paradigms inadvertently leave behind endangered languages; utterance-level speech–text corpora are rare to non-existent, and small speaker communities have limited capacity to produce them. However, hope resides in the fervour and sustained commitment that these communities have for their respective languages. Manx, for instance, has made a remarkable recovery from near extinction, growing its speaker base from just a few hundred in the 1960s to around 2,200 today. In the process, they have made public over 300 hours of educational material, podcasts, and recorded interviews. Similarly, Cornish revival efforts have rekindled an active speaker base of around 600, and prospective learners benefit from a variety of audible educational resources and media. Many remaining endangered languages likely have similar resources, and few—if any—have utterance-level recordings. A truly universal speech technology would make the most of these efforts and convert them into tools that serve the immediate needs of each community.

Table 1. Summary of Manx (gv) and Cornish (kw) data.

Set	Lang.	Source	Domain	Sp. Style	#Speakers	Dur. (mins)
<i>Short-form data</i>						
A	gv	Learnmanx	Education	Careful	>10	102.13
B	gv	Forvo	Education	Careful	>10	15.95
C	kw	Forvo	Education	Careful	9	8.5
<i>Long-form data</i>						
D	gv	Learnmanx	Education	Careful	>10	183.68
E	gv	Clilstore	Religion	Read	2	288.97
F	gv	YouTube	Interview	Conv.	>10	261.81
G	gv	Learnmanx	Literary	Read	1	237.30
H	kw	Skeulantavas	Literary	Read	1	2581.84

3. DATA PREPARATION

3.1. Short- and Long-form Speech

To follow our approach, collect short-form speech and text files where the text corresponds exactly to the word or phrase spoken. Audio will typically be ≤ 2 seconds in length, and while longer clips are acceptable, the likelihood of a verbatim match tends to decline with duration. Any non-verbatim transcripts should be filtered out but kept aside for later steps. For sourcing these data, community “spoken dictionaries” and pronunciation sites are useful. Forvo [2] has these for 430+ languages, and whilst valuable, this remains but a fraction of the world’s $\sim 7,000$ languages. They do, however, support crowdsourced additions to their collections. Failing this, we estimate it would take a few hours to collect a few hundred verbatim word or phrase clips with a prompt list and a microphone. For total duration and number of files, we show that as little as 8.5 minutes (433 files) can be effective. Our spoken dictionary collections for Manx and Cornish are summarised by the top block of Table 1.

The second stage of our approach involves creating new speech segments (utterance-level) from long-form speech recordings (≥ 30 seconds). Verbatim transcriptions of professional quality are preferred but not strictly required. So long as some portion of the text matches parts of the audio, new segments can be created. Gather as many recordings as possible that have accompanying text, whether in captions, descriptions, subtitles, or external documents. Choose sources that match your intended use. If the goal is to create usable transcription technology (<25% WER) for read-speech (audio-book style) recordings, then prioritise that data as domain-matched training usually gives the biggest gains. For a more general-purpose ASR, prioritise a mix of domains and speaking styles. In proceeding experiments, we investigate the amount of long-form data needed to adapt models to selected domains. The data we collect and use to build Manx and Cornish ASR come from publicly available web sources^{1 2} and are summarised in the bottom block of Table 1.

3.2. Text Data

For written languages, web-based text data are often more plentiful than speech. For instance, bible.com has translations of the Bible in 2,300+ languages. Leveraging text data like these can substantially improve ASR by direct decoding or rescoring with external language models (LMs). Here we outline a simple approach to leveraging 4.8 million Manx words derived from web-based sources

¹Manx speech sources: learnmanx.com, clilstore.eu, [@ManxNationalHeritage](http://youtube.com/@learnmanx)

²Cornish speech sources: forvo.com/languages/kw/, skeulantavas.com/audio, [@Wikitongues](http://youtube.com/@Wikitongues)

Table 2. Summary of test sets.

#	Lang.	Domain	Sp. Style	#Speakers	#Utts.	Dur. (mins)
T1	gv	Education	Careful	>10	341	16.95
T2	gv	Religion	Read	2	278	16.67
T3	gv	Interview	Conv.	>10	279	16.03
T4	kw	Podcast	Read	1	131	14.60
T5	kw	Interview	Conv.	1	37	5.18

³. After collection, we derive a 72k list of Manx words by extracting and filtering a frequency-based wordlist from our text data and supplement rare words using a Manx dictionary [23]. The resulting out-of-vocabulary (OOV) rate was 1.66% across all Manx test sets. In the experiments that follow, we assess per domain how much gain a general LM trained on a mix of domains can provide when applied to ASR models (4.2). For unwritten languages or languages where little to no text data is available, this stage is impractical. To simulate this case, for Cornish we use only the paired transcription text (no web text), and for Manx we report runs where the external LM is disabled.

3.3. Test Sets

In order to evaluate HMM and E2E modeling techniques, it was necessary to produce test sets, summarised by Table 2. For Manx, we selected recordings from three distinct domains and speaking styles: enunciated *careful* speech from educational recordings intended for learners, traditional read-speech of religious texts, and spontaneous conversational speech from interview recordings. Recordings were randomly selected until their total length met or exceeded 15 minutes. We then cut each recording into timestamped audio intervals and matched each to its transcript line via Label Studio [24]). A similar approach was taken for Cornish but for two test sets; a read-speech test from the same distribution as our training data and an out-of-domain spontaneous speech test (5 minutes) from a Youtube interview.

3.4. Preprocessing

We converted all audio to a common format of 16-bit PCM, single-channel WAV, with a sampling rate of 16 kHz. We normalised all text, including transcription and unstructured text, by converting to uppercase and standardising whitespace. All punctuation was removed, except for intra-word hyphens and apostrophes (e.g. *mother-in-law*) to preserve lexical meaning. We replaced diacritic markers with their canonical variants (e.g., $\mathcal{C} \rightarrow C$; $\mathcal{E} \rightarrow E$; $\mathcal{N} \rightarrow N$). Numerals from 0-30 were expanded to their word forms using language-specific number lists from omniglot.com, which covers 2,246 languages. In some cases, a single transcription document covered multiple recordings. While not strictly necessary, we manually split these to enforce a one-to-one mapping between each long-form recording and its transcription to improve the chances of successful segmentation.

To use our long-form audio, we must first convert them into utterance-length segments. To do this, we use the Kaldi toolkit [21] and follow the forced alignment pipeline presented in LibriSpeech [25] with two changes. Firstly, we adopt a Unicode graphemic lexicon [26] as, like most other languages, Manx and Cornish do not

³Manx text sources: manxradio.com, learnmanx.com, corpus.gaelg.im, gv.wikipedia.org, culturevannin.im

Table 3. %WERs for Manx source comparison of training sets with different formats, speaker diversity, and quantity.

Quantity	Sets	Source	Format	#Speakers	T1	T2	T3
16 mins	A	Learnmanx	short-form	>10	53.88	79.26	85.96
	B	Forvo	short-form	>10	97.44	96.19	99.30
	D, E, F	multi	utterance-level	>10	53.15	62.18	79.79
	G	Learnmanx	utterance-level	1	75.92	85.39	97.56
102 mins	A	Learnmanx	short-form	>10	42.73	73.80	81.11
	D, E, F	multi	utterance-level	>10	34.94	37.89	59.31
	G	Learnmanx	utterance-level	1	79.37	87.48	97.79

have a phonemic pronunciation dictionary. Secondly, instead of using a pre-built acoustic model, we train a monophone⁴ GMM–HMM on our short-form speech (Manx 102 mins, Cornish 8 mins) which is then used to bootstrap Viterbi forced alignment [27] of the long-form audio. We had an alignment success rate between 70-80% for both systems. Substantially lower than this may point to missing/incomplete transcripts which require revision. Using our alignments, we train a triphone model. The Smith–Waterman alignment and segmentation steps remain the same as in LibriSpeech [25].

4. EXPERIMENTS

We begin by establishing baselines for short-form and utterance-level Manx data to investigate source diversity, form, and quantity. Building on this, we ask how little supervision is needed to achieve usable transcription technology (<25% WER), with and without external LMs. Finally, we extend our approach to Cornish to assess generality to other endangered languages when text is scarce.

4.1. Source Comparison

Setup – Using the LibriSpeech recipes from the Kaldi toolkit [21], we train a single GMM–HMM baseline with speaker-adaptive transforms⁵. During decoding, we use a word-based four-gram language model trained on 4.8M Manx words, using the wordlist described in 3.2. These remain constant and evaluation is carried out on the same three test sets (T1–T3). We then vary data format (short- vs. utterance-level), speaker diversity (>10 vs. 1), and quantity (16 vs. 102 mins). At 16 minutes, we compare random samples of single- and multi-speaker utterance-level data with short-form sets from *LearnManx* and *Forvo*, the latter included because its clips are shorter on average and its speaker and lexical diversity are lower (despite >10 speakers, 1 speaker accounts for most of this data). At 102 minutes, we further compare single- and multi-speaker utterance-level data with the whole *LearnManx* short-form set.

Results – Our 102-minute, multi-speaker utterance-level setting achieved the best results across all three test set, followed closely for T1 by the *LearnManx* spoken dictionary. We suspect this may be due to some speaker overlap between these sets. Source diversity in speakers, styles and domains is the the most determining factor; the single-speaker utterance-level systems were outperformed at both quantities by the *LearnManx* spoken dictionary systems, and the much less diverse *Forvo* system performed markedly worse than all others at 16-minutes. Quantity had the least impact, with the single-speaker system performing worse with more data.

⁴Training a robust triphone (context-dependent) system at this scale is challenging mainly due to data sparsity.

⁵LDA+MLLT, then SAT/fMLLR. Speaker adaptive features are estimated per utterance in our case because speaker identities are unavailable

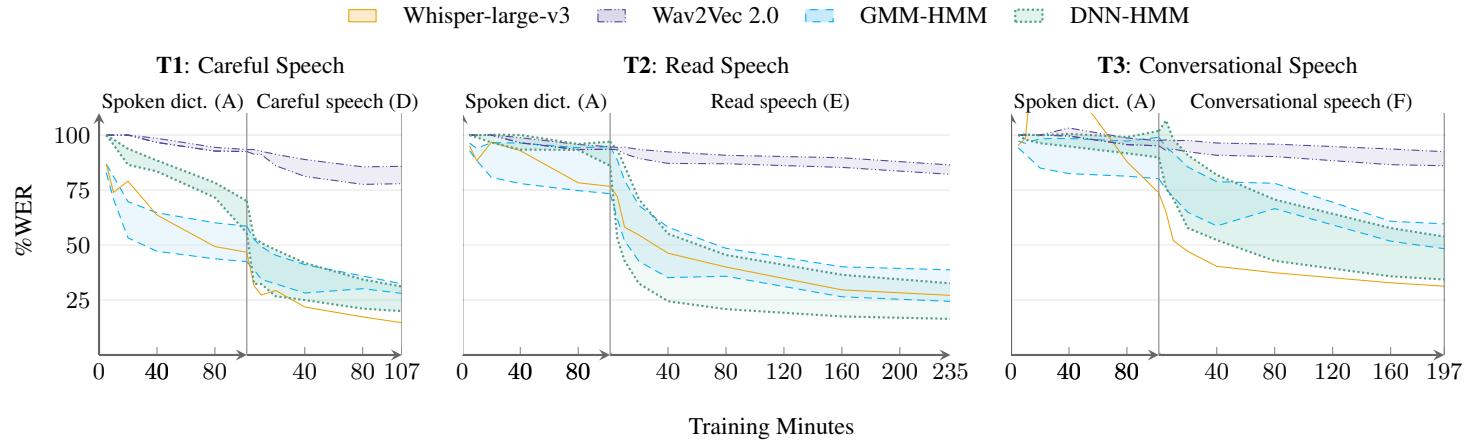


Fig. 1. %WER performance of models trained on progressively larger amounts of Manx spoken dictionary data and new utterances from specific domains, assessed across careful, read, and conversational speech test sets. Shaded areas denote the gain from external LM integration (top vs. bottom line).

4.2. Minimum Supervision Requirements

Setup – Using 5, 10, 20, 40, 80, 102-minute subsets of the short-form *LearnManx* set, we train and assess models across our careful, read, and conversational speech test sets. Once the short-form data are exhausted, we branch each model into three domain-specific runs. For each branch, we add multi-speaker utterance-level data from a single domain (careful, read, or conversational) in progressively larger subsets, assessing each against its in-domain test set. We evaluate two HMM systems—a speaker adaptive GMM-HMM and a TDNN-F model trained with LF-MMI from scratch (no GMM bootstrap)—and two E2E models (Wav2Vec 2.0 and Whisper-large-v3). We fine-tune our pre-trained models according to the Speech-Brain [28] LibriSpeech recipes (LoRA). To investigate how well each model leverages text data, we incorporate the LM from the previous experiment. For the HMMs and wav2vec 2.0, the LM is integrated directly during decoding, whereas for Whisper it is used for n-best list rescoring.

Results – From the spoken-dictionary data alone (Figure 1), only the GMM-HMM (with LM) and Whisper generalise meaningfully to careful speech (<50% WER). The GMM-HMM was the most impressive here, achieving <50% WER on careful speech after just 40 minutes of short-form data. Adding domain-matched utterances drives clear specialisation. Whisper finishes best on careful (14.66% WER) and conversational (31.30% WER) speech, but is overtaken on read speech by the DNN-HMM with external LM integration (16.41% WER). The shaded areas for each show that our LM contributes most to the HMMs, typically ~20 absolute WER points (largest on read/conversational), whereas it resulted in 2 point average gain for wav2vec 2.0 in direct decoding mode and none for Whisper in n-best list rescoring mode.

4.3. Cornish

Setup – Guided by these findings, we apply a similar approach to Cornish using 8.5 minutes of short-form data from Forvo. Despite the issues with this source highlighted in 4.1, we are able to use it to perform the same alignment process described in 3.4. As a result, we create a 39 hour speech corpus from the long-form Cornish resources

in Table 1. These involve only one speaker, which we expect to be detrimental to generalisation, however no other long-form resources are available at this time. To simulate a scenario where text resources are scarce, we neglect the use of an external LM and instead fine-tune whisper-large-v3 on this corpus. Evaluation is performed on a 14-minute in-domain read speech set (T4) and a short out-of-domain interview-style recording (T5).

Results – Our Whisper system achieves an in-domain test set performance of 7.72% WER (1.65% CER) which passes Meta’s MMS threshold of “coverage” [7]. However, it also achieves 72.55% WER (34.17% CER) on the out-of-domain test, indicating severe domain and speaker mismatch. We suspect that this discrepancy comes mainly from the fact that the 39 hour corpus used for training all comes from a single speaker. However, research from the IARPA BABEL program shows that ASR systems with out-of-domain WERs of ~70% are still useful for tasks such as Keyword Spotting [29].

5. CONCLUSION

ASR support for endangered languages remains sparse, principally due to a lack of utterance-level supervision. Addressing this technological gap, we have shown that usable ASR can be attained for endangered languages that have no utterance-level supervised data by leveraging rudimentary resources such as a spoken dictionary. We show that just 40 minutes of spoken dictionary data from multiple speakers can produce a viable ASR baseline (<50% WER). Our results highlight the flexibility of HMMs in integrating short- and long-form supervision and text information to deliver usable ASR (both HMM and E2E based) from unconventional resources. On the other hand, Whisper generalises markedly better to conversational speech, although the same may not be true for other E2E models such as wav2vec 2.0. Ultimately, this paper has shown that developing speech technology for low-resource and endangered languages does not require utterance-level corpora as a starting point. Collecting new and clean resources is an important task, but future work would do well to make sense of the data that already exist and continue to be created. Doing so may mean difference to the most urgent cases of language endangerment.

6. REFERENCES

- [1] Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, “Speech corpus of ainu folklore and end-to-end speech recognition for ainu language,” *CoRR*, vol. abs/2002.06675, 2020.
- [2] “Forvo: The pronunciation dictionary,” <https://forvo.com/>, Accessed: 2025-09-04.
- [3] Karen Livescu Lee and Shinji Watanabe, “The ml-superb 2.0 challenge: Towards inclusive asr benchmarking for all language varieties,” .
- [4] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna, “Fleurs: Few-shot learning evaluation of universal representations of speech,” 2022.
- [5] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, “Common voice: A massively-multilingual speech corpus,” *CoRR*, vol. abs/1912.06670, 2019.
- [6] Yerbolat Khassanov, Saida Mussakhojayaeva, Almas Mirzakhmetov, Alen Adiyev, Mukhamet Nurpeissov, and Huseyin Atakan Varol, “A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline,” in *Proceedings of the 16th Conference of the European Chapter of the ACL: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, Eds., Online, Apr. 2021, ACL.
- [7] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, “Scaling speech technology to 1,000+ languages,” 2023.
- [8] Mark JF Gales, Kate M Knill, and Anton Ragni, “Low-resource speech recognition and keyword-spotting,” in *International Conference on Speech and Computer*, 2017.
- [9] Samuel Thomas, Kartik Audhkhasi, Jia Cui, Brian Kingsbury, and Bhuvana Ramabhadran, “Multilingual data selection for low resource speech recognition,” Tech. Rep., 2016.
- [10] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, “Unsupervised speech recognition,” 2022.
- [11] Alexander H. Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski, “Towards end-to-end unsupervised speech recognition,” 2022.
- [12] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei, “Speak foreign languages with your own voice: Cross-lingual neural codec language modeling,” 2023.
- [13] Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass, “Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages,” 2023.
- [14] Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang, “Bilingual dictionary based neural machine translation without using parallel sentences,” in *Proceedings of the 58th Annual Meeting of the ACL*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, Eds., Online, July 2020, ACL.
- [15] Xing Jie Zhong and David Chiang, “Look it up: Bilingual dictionaries improve neural machine translation,” 2022.
- [16] Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi, “A benchmark for learning to translate a new language from one grammar book,” 2024.
- [17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew, “Huggingface’s transformers: State-of-the-art natural language processing,” *CoRR*, vol. abs/1910.03771, 2019.
- [18] Ondřej Klejch, William Lamb, and Peter Bell, “A practitioner’s guide to building asr models for low-resource languages: A case study on scottish gaelic,” 2025.
- [19] Keyu An, Hongyu Xiang, and Zhijian Ou, “Cat: A ctc-crf based asr toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency,” 2020.
- [20] Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu, “A comparison of techniques for language model integration in encoder-decoder speech recognition,” 2018.
- [21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Veselý, “The kaldi speech recognition toolkit,” in *ASRU*, 2011.
- [22] Mark John Francis Gales, Kate Knill, Anton Ragni, and Shakti Prasad Rath, “Speech recognition and keyword spotting for low-resource languages: Babel project research at cued,” in *Workshop on Spoken Language Technologies for Under-resourced Languages*, 2014.
- [23] Phil Kelly, “Fockleyreen: Manx–english dictionary,” 2014, Mirror of Phil Kelly’s Manx vocabulary (Fockleyreen); ~130,000 entries; mirror created 2 December 2014.
- [24] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov, “Label Studio: Data labeling software,” 2020–2025, Open source software; accessed 15 Sep 2025.
- [25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *ICASSP*, 2015.
- [26] M.J.F. Gales, K.M. Knill, and A. Ragni, “Unicode-based graphemic systems for limited resource languages,” in *ICASSP*, 2015.
- [27] L.R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, 1989.
- [28] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.
- [29] Jan Trmal, Guoguo Chen, Dan Povey, Sanjeev Khudanpur, Pegah Ghahremani, Xiaohui Zhang, Vimal Manohar, Chunxi Liu, Aren Jansen, Dietrich Klakow, David Yarowsky, and Florian Metze, “A keyword search system using open source software,” in *SLT*, 2014.