Samrómur Milljón: An ASR Corpus of One Million Verified Read Prompts in Icelandic

Carlos Daniel Hernández Mena, Þorsteinn Daði Gunnarsson, Jón Guðnason
Language and Voice Laboratory, Reykjavík University
Menntavegur 1, 102 Reykjavík, Iceland
{carlosm, thorsteinng, jg}@ru.is

Abstract

The platform samromur.is, or "Samrómur" for short, is a crowdsourcing web application built on Mozilla's Common Voice, designed to accumulate speech data for the advancement of language technologies in Icelandic. Over the years, Samrómur has proven to be remarkably successful in amassing a significant number of high-quality audio clips from thousands of users. However, the challenge of manually verifying the entirety of the collected data has hindered its effective exploitation, especially in the realm of Automatic Speech Recognition (ASR), its original purpose. In this paper, we introduce the "Samrómur Milljón" corpus, an ASR dataset comprising one million audio clips from Samrómur. These clips have been automatically verified using state-of-the-art speech recognition systems such as NeMo, Wav2Vec2, and Whisper. Additionally, we present the ASR results obtained from creating acoustic models based on Samrómur Milljón. These results demonstrate significant promise when compared to other acoustic models trained with a similar volume of Icelandic data from different sources.

Keywords: speech recognition, million prompts, asr corpus, icelandic dataset, automatic verification

1. Introduction

Automatic speech recognition (ASR) systems play a pivotal role in the development of modern communication tools from applications ranging from voice assistants to transcription services. Notwithstanding latest developments in semi-supervised learning (Conneau et al., 2023; Pratap et al., 2023), collecting high-quality parallel text-speech data is crucial in training ASR. This challenge is amplified when it comes to languages with low number of speakers, such as Icelandic. This paper describes the prompt-based parallel text-speech collection effort called Samrómur (Mollberg et al., 2020) and how a large amount of speech data we call Samrómur Milljón was collected and verified making the development of high-quality ASR systems a real possibility for the language.

There are normally three ways of obtaining parallel text-speech data for ASR development each with their advantages and disadvantages. First, speech recordings can simply be transcribed by listening and typing up the recordings. This method does not scale very well as it is hard to crowd-source and is therefore expensive, but it can produce good quality data sets. The second approach is to ask speakers to read predetermined prompts aloud with their recordings are then stored alongside the prompt's text. Lastly, dataset maybe curated from recordings where transcripts exist. These transcripts are often not intended to be used for ASR development so the curation process can get complex. The Samrómur project is based on the prompt-based collection that relies on crowed sourcing to get participation. The process requires a post-processing step where the read prompt is verified against the prompt's text to ensure that the resulting data set is of sufficient quality.

The crowd-sourcing platform Samrómur allows participants to verify prompts manually, hence creating the required high-quality dataset. However, the reading effort was so successful that the verification process left a large amount of prompts unverified. This paper describes how we automatically verified the big proportion of the collected data using a set of state-of-the-art speech recognition technologies such as NeMo (Kuchaiev et al., 2019), Wav2Vec2 (Schneider et al., 2019; Baevski et al., 2020), Whisper (Radford et al., 2023) and Faster-Whisper[1]. Section 2 provides a deeper overview of the Samrómur platform, while section 3 describes Samrómur Milljón in detail. Section 4 explains our methodology to perform the verification, while section 5 presents the acoustic models that we trained using Samrómur Milljón.

2. The Samrómur Platform

This section provides a brief overview of other data collection initiatives, the Samrómur platform, its origin, objectives and achievements in collecting varied voice samples of the Icelandic language.

Some effort has previously been put into collecting ASR data from Icelandic speakers. Hjal (Rögnvaldsson, 2003) was collected in 2002 to build a speaker-independent isolated word recognition

[1] https://github.com/guillaumekln/faster-whisper

LREC-COLING 2024, pages 14305–14312
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

system. In 2011, Reykjavík University, with the support of Google, started collecting (Guðnason et al., 2012) and verifying (Steingrímsson et al., 2017) data for Málrómur, a database of spoken sentences to aid the development of automatic speech recognition for Icelandic. That project resulted in Eyra (Pétursson et al., 2016; Guðnason et al., 2017), a similar data collection platform for distributed speech data collection through a variety of devices.

The Samrómur data collection started in 2019 as part of the Language Technology Programme for Icelandic 2019-2023 (Nikulásdóttir et al., 2020) and remains active until today. Samrómur was built on top of Mozilla's Open Voice platform (Ardila et al., 2020) to collect open speech data from the community in Iceland. The platform offers two different crowd sourcing tasks. Firstly, to donate your voice into the database by reading into your microphone a handful of sentences at a time, and secondly, to verify already submitted voice samples. Speakers use their own devices, with their own microphone, whether it be a desktop computer, laptop or a mobile phone, resulting in a large variety of input quality. The prompts are drawn from the Icelandic Gigaword Corpus (Steingrímsson et al., 2018), a large corpora of Icelandic text. Suitable sentences were filtered out of the large corpora to use in the collection.

Throughout the collection, more prominence was put on collecting additional voice samples rather than verifying the ones that had been gathered. To gather recordings, a lot of effort was put in advertising the initiative. Alongside sponsored posts on social media and press releases to increase awareness of the project, the president of Iceland as well as other politicians supported the project and encourage the citizens of Iceland to participate. All this helped to gather a great number of recordings from a varied group of people.

To further increase the size of the collection, competitions for elementary school students were held. Both students, teachers and parents could participate and the students that donated the most samples got an award and a meeting with the president of Iceland. This kind of competition was held three times, in 2020, 2021 and 2022. Each time the competition greatly increased the number of recordings donated for the period, as well as awareness of the project. An additional competition was held in 2021 where workplaces competed to donate as many voice samples as they could.

The Samrómur collection platform has produced multiple parallel ASR corpora in Icelandic which have been published with open source licenses. These include "Samrómur Icelandic Speech 1.0" (Mollberg et al., 2020) and "Samrómur Children" (Mena et al., 2022). The platform is still up running and used to collect data if needed, albeit not in the same volume as before. For example, Samrómur recently been used to collect speech from second language learners of Icelandic for Computer Aideded Pronunciation Training (Richter et al., 2022; Richter and Guðnason, 2023). To date, the initiative has collected 4,156 hours of voice samples in Icelandic from 27,928 people of varied ages, genders and nationalities. 495 hours or about 12% of the samples have been manually verified.

3. The Samrómur Milljón Corpus

Samrómur Milljón Hernández Mena and Guðnason (2023) is a monolingual ASR corpus in Icelandic with a size of 967 hours, comprised of 1,002,157 speech recordings of 16,604 unique speakers with ages between 4 to 79 years old. It is the result of the automatic verification of a larger corpus known as "Samrómur Unverified 22.07" Hedström et al. (2022) which contains 2,159,314 (2,233 hours) unverified speech recordings.

3.1. Corpus Characteristics

Samrómur Milljón has the same structure as the other datasets collected using the Samrómur platform, see for example "Samrómur Icelandic Speech 1.0" (Mollberg et al., 2020) or "Samrómur Children" (Mena et al., 2022). It contains speech files with a size between one to ten seconds long in FLAC format, encoded at a sampling rate of 16 kHz with a depth equal to 16 bit of linear PCM in one single channel. The reference transcriptions and other relevant information associated to each individual recording are contained in a metadata file in a format known as "Tab Separated Values", or TSV for short.

Samrómur Milljón differs from previously published datasets in its composition, as it can be inferred in Table 1, which presents some basic statistics of the corpus. This shows that Samrómur Milljón has a much larger age range than previous datasets, the gender balance is different with female contributors predominant and just a minimum number of speakers are labeled as unknown (UNK), whose gender, age range or both are not known. This uncertainty comes about due to contributors not filling out the optional fields in their submission forms. Either way, the reference transcriptions that are the most valuable asset of the corpus are guaranteed by our verification methodology that will be explained in a later section.

As the main classification criteria in Samrómur Milljón apart from gender, is the age range of the speakers, data is classified in three main groups: strictly under 18 years old (yrs<18), from 18 to 49 (18>=yrs<=49) and strictly over 49 years old

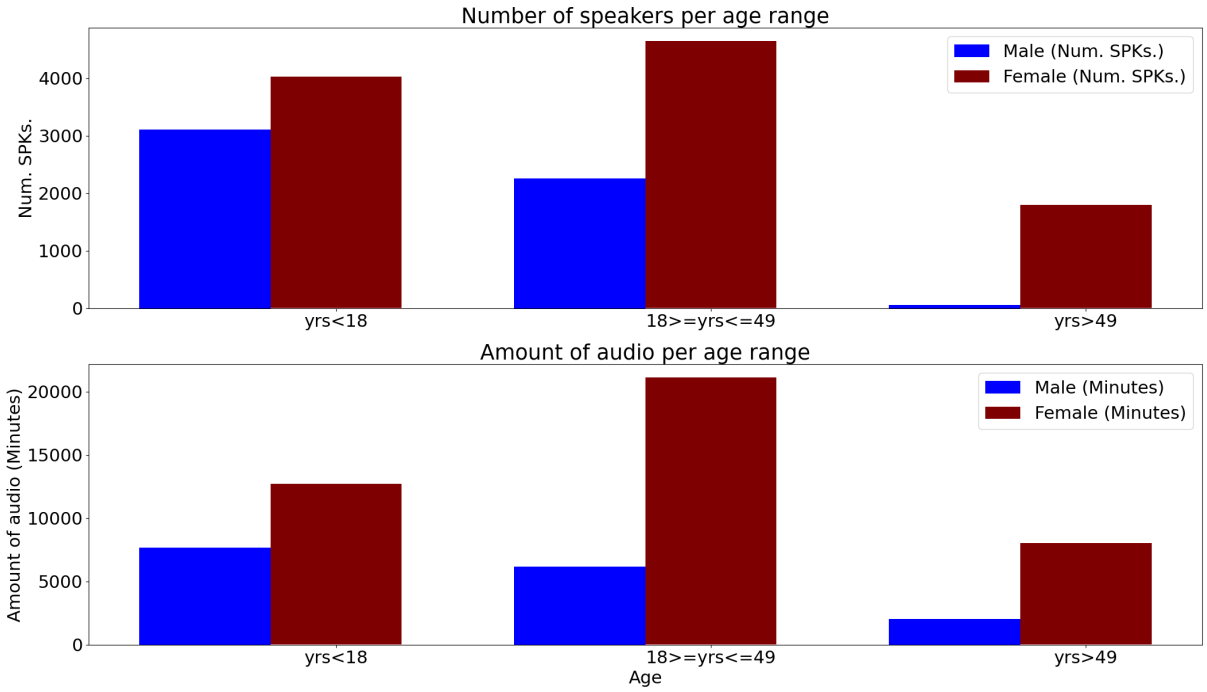

Figure 1: Bottom) Amount of Audio per age range shown in minutes. Top) Number of speakers per age range.

Gender	Female	Male	UNK
Duration	697h22m	264h28m	5h16m
Utterances	714,564	282,499	5,094
Speakers	10,447	5,948	209

Table 1: Gender distribution of Samrómur Milljón. Recordings with unknown age, unknown gender or both are stored under the category of "UNK".

(yrs>49). It has to be emphasized that in the case of participants under the age of 18, it was required one of their guardians to fill a consent form, in accordance with the provisions of the European Commission through the General Data Protection Regulation (GDPR) (European Commission, 2018). Figure 1 shows the data distribution of the corpus in terms of number of speakers (top) and in terms of amount of audio (bottom). What can be learned from this figure is that the relationship between amount of audio and the number of speakers is consistent, which means that in general, most of the speakers contribute with some average amount of audio, and there are no speakers with an excessive contribution of recordings. Another relevant aspect of Figure 1 is that the contribution coming from male speakers over 49 years old is poor when comparing with female speakers in same age range. Future collection campaigns should focus on the tendencies revealed by Figure 1, in order to collect more of the data that is needed to get a balance through all the speech groups.

4. Verification

The automatic verification of the transcriptions belonging to Samrómur Milljón was performed using state-of-the-art speech recognition systems such as NeMo, Wav2Vec2, Whisper and Faster-Whisper. Our first idea was to transcribe the 2 million recordings of "Samrómur Unverified 22.07" using these systems and keep just the ones that perfectly matched the corresponding reference transcription with 2 systems or more. However, due to time and hardware limitations, we had to change this strategy in order to maximize the number of verifications and at the same time, not to spend an excessive amount of time in doing so.

Regarding to the acoustic models utilized to perform the verification, it has to be pointed out that all of them are publicly available at `Huggingface.co` and they were the best ASR models for Icelandic that we could find at the moment of doing the verification. It is also relevant to mention that all of these models were trained with the following datasets:

- Althingi Corpus (Helgadóttir et al., 2017). Available at Helgadóttir et al. (2021).
- Malrómur Corpus (Steingrímsson et al., 2017; Guðnason et al., 2012). Available at Steingrímsson et al. (2017).

- Samrómur Corpus (Mollberg et al., 2020). Available at Mollberg et al. (2020).
- Samrómur Children Corpus (Mena et al., 2022). Available at Hernández Mena et al. (2021).

The total duration of the sum of these four datasets is 875 hours. The verification process was performed using one acoustic model at a time running in a GPU Tesla A100.

4.1. Verification with NeMo

The NeMo model Hernandez Mena (2022a) chosen for the verification task was created using the version 1.10.0 of the NeMo Toolkit[2]. According to the developers, the model was trained just with the four datasets listed in section 4.
The process of transcribing the 2 million recordings with this model took around 7 hours. We decided not to use a language model because we wanted to be as strict as possible; otherwise, it could be the case that some mispronunciations found in the recordings were masked by the language model. At the end, we got a total of 348,295 perfect matches which is too low compared to the 2 million input recordings. Nevertheless, it has to be pointed out that the NeMo model is one of the highest in terms of speed but it is the poorest in terms of precision.

4.2. Verification with Wav2Vec2

The Wav2vec2 model Hernandez Mena (2022b) chosen was created by fine-tuning the checkpoint wav2vec2-large-xlsr-53[3] with the Icelandic datasets listed in section 4 plus 126 hours of unpublished material coming from L2 speakers; in other words, a total of 1,000 hours of Icelandic was used to create this model. It is important to mention that although the datasets: Samrómur, Samrómur Children and the data from the L2 speakers come from the Samrómur platform, the fact is that they don't share speakers with Samrómur Milljón, so they can be considered as different data. In addition, the contribution of those datasets to the 1,000 hours is less than half (around 366 hours). The major contributor is the Althingi corpus with 514 hours of Parliament speeches.
This time, the process of transcribing 2 million files took 7 days. We did not use a language model for the same reasons explained for the NeMo model in section 4.1. The number of matches was 1,002,218 which is a better result than the one obtained with NeMo. A subsequent analysis revealed that the NeMo matches is a subset of the Wav2Vec2 matches. This may mean that there are one million of utterances that are hard to transcribe due to several reasons as the Wav2Vec model has a decent performance, even compared to the Whisper one, which is the best of all in terms of precision.

4.3. Verification with Whisper and Faster-Whisper

The Whisper model Hernandez Mena (2023b) was created by fine-tuning the checkpoint whisper-large[4] with exactly the same data used for the Wav2Vec2 model of section 4.2.
As mentioned before, the original plan was to use this model to transcribe the 2 million recordings, but soon we noticed that this was going to be unrealistic due to the amount of time required for Whisper and also because, unfortunately, we worked for some time with a corrupted copy of the model. According to our calculations, the transcription of the 2 million recordings would require around 36 days with our available set up. So, we initiated the process starting from data previously validated by Wav2Vec2. After two and a half days, we noticed that the model was producing a considerable amount of trash and hallucinations; so we decided to stop the process in order to investigate what was occurring. During this process, we transcribed a total of 138,992 recordings obtaining just 18,839 perfect matches. We contacted the developers and other users of that same model and they did not report issues. At that point we decided to find an alternative solution as Whisper was proving to be very slow for our needs at that time; that is when we discovered Faster-Whisper.
As the Whisper model described was publicly available at HuggingFace with a CC-BY-4 license, we simply cloned it and converted it into a Faster-Whisper model with 2 lines of code[5]. We used this new model to transcribe just the data validated by Wav2Vec2 obtaining a total of 863,220 perfect matches. This high number of matches reveals that our copy of the model was indeed corrupted as it seems not plausible that a wrong model (the original one) is capable of producing a correct one (the Faster-Whisper one).
At the end, the system that determined the final number of recordings contained in Samrómur Milljón was Wav2Vec2.

4.4. Verification Results

Inspecting the metadata file included in the Samrómur Milljón, which comes in TSV format, one

[2] https://github.com/NVIDIA/NeMo/releases/tag/v1.10.0

[3] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[4] https://huggingface.co/openai/whisper-large

[5] That model is also available at HuggingFace: Gunnarsson and Hernandez Mena (2023)

will find a column called "verified_with" that informs about which ASR systems matched the reference transcription of one particular recording. The abbreviations found in this column are as follows: N is for NeMo, V is for Wav2Vec2, W is for Whisper and F is for Faster-Whisper. So, the value V+N+W associated to a particular recording means that such recording found perfect matches with Wav2Vec2, NeMo and Whisper. Table 2 shows the number of perfect matches performed by the distinct ASR systems over the whole Samrómur Milljón.

Sys.	Matches	Percent.
V+N+F	325,713	32.50%
V+N+W	4,449	0.44%
V+F	537,453	53.62%
V+N	18,072	1.80%
V+W	14,390	1.43%
V	102,080	10.18%

Table 2: Perfect matches performed by the distinct ASR systems over the whole Samrómur Milljón.

As it can be seen in Table 2, the majority of the speech recordings belonging to Samrómur Milljón (53.62%) were verified by 3 different ASR systems, while more than 80% were verified by Wav2Vec2 along with Faster-Whisper. It is also worth to mention that 100% of the speech recordings were verified by Wav2Vec2 at least. It corresponds to the users the selection of recordings that better fits their particular needs.

5. Acoustic Models Trained with Samrómur Milljón

Two derivatives produced with Samrómur Milljón are a Whisper acoustic model and a Wav2Vec2 acoustic model. As it will be seen in the next sections, these models are comparable in performance to the models used to make the verification. This fact can be taken as an indirect proof that the data in Samrómur Milljón is correct, otherwise, the models produced with it would suffer an inferior performance.

5.1. Whisper Model Trained with Samrómur Milljón

The Whisper model trained with Samrómur Milljón Hernandez Mena (2023c) is the result of fine-tuning the checkpoint "openai/whisper-large" during 62,640 steps. Table 3 shows the Word Error Rate (WER) results obtained with this model over known datasets while comparing with the WER results of the model used to perform the verification.

Dataset	V. M.	S. M.
Samrómur (Test)	8.47%	7.76%
Samrómur (Dev)	7.29%	7.03%
Samrómur Children (Test)	7.74%	7.04%
Samrómur Children (Dev)	4.59%	4.42%
Malrómur (Test)	5.11%	11.51%
Malrómur (Dev)	5.28%	11.00%
Althingi (Test)	8.25%	16.18%
Althingi (Dev)	7.99%	16.00%

Table 3: WER comparison between the Whisper model trained with Samrómur Milljón (S. M.) versus the Whisper model used for the verification (V.M.).

In addition to this, a Faster-Whisper model Hernandez Mena (2023d) was created using this Whisper model as a base.

5.2. Wav2Vec2 Model Trained with Samrómur Milljón

The Wav2Vec2 model trained with Samrómur Milljón Hernandez Mena (2023a) is the result of fine-tuning the checkpoint "facebook/wav2vec2-large-xlsr-53" during 30 epochs. Table 4 shows the WER results obtained with this model over known datasets while comparing with the WER results of the model used to perform the verification.

Dataset	V. M.	S. M.
Samrómur (Test)	9.84%	7.69%
Samrómur (Dev)	8.73%	6.78%
Samrómur Children (Test)	9.39%	6.46%
Samrómur Children (Dev)	6.05%	4.23%
Malrómur (Test)	5.64%	6.63%
Malrómur (Dev)	6.15%	5.83%
Althingi (Test)	11.43%	17.90%
Althingi (Dev)	11.09%	17.93%

Table 4: WER comparison between the Wav2Vec2 model trained with Samrómur Milljón (S. M.) versus the Wav2Vec2 model used for the verification (V.M.).

6. Discussion

The methodology applied to perform the automatic verification of the recordings that constitute the Samrómur Milljón only detects perfect matches between the reference transcriptions and one or more of the ASR systems utilized, but it does not inform about the correctness of other relevant information included in the dataset such as gender, range of age or speaker id. From the perspective of the ASR field and from most of the ASR systems that can be trained by users, it is sufficient to count with pairs of recording/transcription to create accurate acoustic models, even for

real world scenarios. However, there are some examples of systems that can use more information to improve their performance. For instance, Kaldi (Povey et al., 2011) applies Speaker Adaptation Training (SAT) which requires knowledge about the gender of the speakers in the training corpus. It would be worth to perform further verifications to the data of Samrómur Milljón in order to expand its effectiveness to other applications.

7. Conclusions and Further Work

In this paper we have presented the Corpus Samrómur Milljón, a monolingual ASR corpus in Icelandic comprised of 1,002,157 speech recordings (967 hours) that is the result of the automatic verification of the more than 2 million recordings contained in the dataset "Samrómur Unverified 22.07" using state-of-the-art ASR systems such as NeMo, Wav2Vec2, Whisper and Faster-Whisper. In addition to this, we have trained and released acoustic models in Wav2Vec2, Whisper and Faster-Whisper that are independent of the models used to perform the verification.

As a further work, we have identified the necessity of deploying more data collection campaigns in order to balance the bars shown in Figure 1. It will also worth to perform automatic verifications on other information included in Samrómur Milljón such as gender, range of age and speaker id, in order to expand the effectiveness of the corpus to other fields. It will also worth to apply our same verification methodology to the rest of the data in "Samrómur Unverified 22.07" that remains unverified but using bigger acoustic models trained with the combination of both the data used to train the models that we used to perform the verification plus the data coming from Samrómur Milljón (around 2,000 hours in total).

However, despite the pending tasks, we strongly think that the work presented in this paper is a relevant contribution to the speech technologies in Icelandic, a language that will not be under represented any more.

8. Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur, is funded by the Icelandic Ministry of Education, Science and Culture.

9. Bibliographical References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

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.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

European Commission. 2018. General data protection regulation (gdpr) – official legal text. web page: https://gdpr-info.eu/.

Jón Guðnason, Oddur Kjartansson, Jökull Jóhannsson, Elín Carstensdóttir, Hannes Högni Vilhjálmsson, Hrafn Loftsson, Sigrún Helgadóttir, Kristín M Jóhannsdóttir, and Eiríkur Rögnvaldsson. 2012. Almannaromur: An open icelandic speech corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

Jón Guðnason, Matthías Pétursson, Róbert Kjaran, Simon Klüpfel, and Anna Björk Nikulásdóttir. 2017. Building asr corpora using eyra. In *INTERSPEECH*, pages 2173–2177.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2017. Building an asr corpus using althingi's parliamentary speeches. In *Interspeech*, pages 2163–2167.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Carlos Daniel Hernandez Mena, David Erik Mollberg, Michal Borskỳ, and Jón Guðnason. 2022. Samrómur children: An icelandic speech corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 995–1002.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Thorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jón Guðnason. 2020. Samrómur: Crowd-sourcing data collection for icelandic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3463–3467.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for icelandic 2019-2023. *arXiv preprint arXiv:2003.09244*.

Matthias Petursson, Simon Klüpfel, and Jon Gudnason. 2016. Eyra-speech data acquisition system for many languages. *Procedia Computer Science*, 81:53–60.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

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.

Caitlin Richter and Jón Guðnason. 2023. Relative dynamic time warping comparison for pronunciation errors. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Catlin Richter, Branislav Bédi, Ragnar Pálsson, and Jón Guðnason. 2022. Computer-assisted pronunciation training in icelandic (captini): developing a method for quantifying mispronunciation in l2 speech. *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, page 334.

Eiríkur Rögnvaldsson. 2003. The icelandic speech recognition project hjal. *Nordisk Sprogteknologi. Årbog*, pages 239–242.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Steinþór Steingrímsson, Jón Guðnason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2017. Málrómur: A manually verified corpus of recorded icelandic speech. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 237–240.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large icelandic text corpus. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

10. Language Resource References

Thorsteinn Dadi Gunnarsson and Carlos Daniel Hernandez Mena. 2023. Acoustic model in icelandic: whisper-large-icelandic-30k-steps-1000h-ct2. HuggingFace.

Staffan Hedström, Judy Y. Fong, Ragnheiður Þórhallsdóttir, David Erik Mollberg, Smári Freyr Guðmundsson, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Eydís Huld Magnúsdóttir, and Jon Gudnason. 2022. Samromur unverified 22.07. CLARIN-IS.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2021. Althingi asr, audio and transcriptions. HuggingFace.

Carlos Daniel Hernandez Mena. 2022a. Acoustic model in icelandic: stt_is_quartznet15x5_ft_ep56_875h. HuggingFace.

Carlos Daniel Hernandez Mena. 2022b. Acoustic model in icelandic: wav2vec2-large-xlsr-53-icelandic-ep10-1000h. HuggingFace.

Carlos Daniel Hernandez Mena. 2023a. Acoustic model in icelandic: wav2vec2-large-xlsr-53-icelandic-ep30-967h. HuggingFace.

Carlos Daniel Hernandez Mena. 2023b. Acoustic model in icelandic: whisper-large-icelandic-30k-steps-1000h. HuggingFace.

Carlos Daniel Hernandez Mena. 2023c. Acoustic model in icelandic: whisper-large-icelandic-62640-steps-967h. HuggingFace.

Carlos Daniel Hernandez Mena. 2023d. Acoustic model in icelandic: whisper-large-icelandic-62640-steps-967h-ct2. HuggingFace.

Carlos Daniel Hernández Mena, Michal Borsky, David Erik Mollberg, Smári Freyr Guðmundsson, Staffan Hedström, Ragnar Pálsson, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Jóhanna Vigdís Guðmundsdóttir, Eydís Huld Magnúsdóttir, Ragnheiður Þórhallsdóttir, and Jón Guðnason. 2021. Samrómur children asr, audio and transcriptions. HuggingFace.

Carlos Daniel Hernández Mena and Jón Guðnason. 2023. Samrómur milljón, audio and transcriptions. HuggingFace.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jon Gudnason. 2020. Samrómur asr, audio and transcriptions. HuggingFace.

Steinþór Steingrímsson, Jón Guðnason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2017. Málrómur asr, audio and transcriptions. HuggingFace.