Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 995–1002
Marseille, 20-25 June 2022

Samrómur Children: An Icelandic Speech Corpus

Carlos Mena, David Erik Mollberg, Michal Borský, Jón Guðnason
Language and Voice Lab, Reykjavík University
Menntavegur 1, 102 Reykjavík, Iceland
{carlosm, de14 ,michalb, jg}@ru.is

Abstract

Samrómur Children is an Icelandic speech corpus intended for the field of automatic speech recognition. It contains 131 hours of read speech from Icelandic children aged between 4 to 17 years. The test portion was meticulously selected to cover a wide range of ages as possible as we aimed to have exactly the same amount of data per age range. The speech was collected with the crowd-sourcing platform `samromur.is`, which is inspired on the "Mozilla's Common Voice Project". The corpus was developed within the framework of the "Language Technology Programme for Icelandic 2019 − 2023"; the goal of the project is to make Icelandic available in language-technology applications. Samrómur Children is the first corpus in Icelandic with children's voices for public use under a Creative Commons license. Additionally, we present baseline experiments and results using Kaldi.

Keywords: children's speech corpus, children's speech recognition, icelandic children's speech, icelandic corpus.

1. Introduction

1.1. Project Background

The creation of "Samrómur Children" is part of the "Language Technology Programme for Icelandic 2019 − 2023" (Nikulásdóttir et al., 2020) which aims to make Icelandic available in language-technology applications, as well as in all areas of communication, with all the deliverables of the programme published under open licenses to encourage the use of them in commercial products such as pronunciation apps to help children to improve their read-out-loud skills or to learn Icelandic. Samrómur Children is a subset of the data gathered by crowd-sourcing using the web application "Samrómur" (Mollberg et al., 2020), an ongoing project of speech data collection, based on the "Mozilla's Common Voice Project" for open-source voice collection (Ardila et al., 2019). The goal of Samrómur is to build a large-scale speech corpus for Automatic Speech Recognition (ASR) for Icelandic as it is contemplated in the programme. It was therefore necessary to set up a separate platform from the Mozilla's one, in order to have more control over the data gathering and the distribution of the datasets when it comes to consent forms and meta-data collected.

Samrómur is the largest prompted speech collection effort for Icelandic so far. However, previous efforts in this respect have been deployed before. One example of this, is the "Icelandic Speech Recognition Project Hjal" (Rögnvaldsson, 2003) with the main goal of collecting sufficient material to train a speaker-independent isolated word recognition system. Another example is the "Almannarómur Project" (Guðnason et al., 2012), started as a collaboration between the Reykjavik University and the Icelandic Centre for Language Technology in cooperation with Google, that provided technical support. One of the main products delivered by the Almannarómur Project is the Málrómur corpus (Steingrímsson et al., 2017; Guðnason et al., 2017), which consists in 136 hours of manually evaluated speech utterances with correct transcriptions. The "Althingi Parliamentary Speech Corpus" (Helgadóttir et al., 2017) consists of 542 hours of parliamentary speech which has been automatically aligned. The "Samrómur 21.05"[1] is a corpus derived from the Samrómur web application, consisting in $100,000$ utterances (around 114 hours) and containing adult's speech only (18+ years). Samrómur Children is the first corpus dedicated exclusively to children's speech in Icelandic, and even though it is a dataset derived from the Samrómur, it will be treated separately, as children's speech represents a major challenge compared to adult's speech, and it also requires special considerations as we will see in the following section.

1.2. The Special Case of Children's Speech

It is possible to find in the literature numerous studies regarding children's speech and what are the differences when compared to adults (Giuliani and Gerosa, 2003; Wilpon and Jacobsen, 1996). This problem has been addressed from a variety of perspectives such linguistics (Palethorpe et al., 1996; Fringi et al., 2015; Li and Qian, 2019), acoustics (Gerosa et al., 2006; Gerosa et al., 2007), signal processing (Lee et al., 1997; Ghai and Sinha, 2009; Ghai, 2011) or even from so technical approaches, such as the one trying to determine the relationship between bandwidth and recognition accuracy in children's speech (Russell et al., 2007; Li and Russell, 2001).

There are many distinct points of view, but it seems to be a consensus that children's speech is challenging from the perspective of the speech recognition (Li and Russell, 2001; Li and Russell, 2002) because of the changes in the phonatory apparatus of the children through time (Lee et al., 1999; Mugitani and Hiroya, 2012; Hämäläinen et al., 2014).

Already in the seventies, Kent pointed out that the accuracy of motor control improves with age until adult-like performance is achieved at about 11 or 12 years (Kent, 1976), which is concordance with Potamianos, who noticed that the speech recognition performance is a function of the speaker's age (Potamianos et al., 1997). Studies of children's speech starting from the early ages of 3 and 4 years

[1] `https://www.openslr.org/112/`

old (Hämäläinen et al., 2014; Elenius and Blomberg, 2005) support that same hypothesis.

1.3. Children's Speech Recognition

Nowadays, the field of ASR is growing at a speed never seen before. We have passed from creating simple isolated word recognizers (Furui, 1986) to develop real time speech-to-speech translation systems (Bangalore et al., 2012) in just few decades. Nonetheless, not all the spectrum of human speech is benefited by such advances at the same rate. Normally, the breakthroughs occur in the range of adult's speech and then leak to for example, children's speech, through certain techniques such like transfer learning (Tong et al., 2017; Matassoni et al., 2018), data augmentation (Fainberg et al., 2016; Sheng et al., 2019; Chen et al., 2020) or model adaptation (Serizel and Giuliani, 2014; Qian et al., 2016).

In addition, different paradigms of ASR have been used to directly address the problem of children's speech. Such is the case of Hidden Markov Models (Das et al., 1998; Potamianos and Narayanan, 2003), Support Vector Machines (Safavi et al., 2018; Retico et al., 2016), Neural Networks (Giuliani and BabaAli, 2015; Wu et al., 2019) and (in recent times) End-to-End Systems (Shivakumar and Narayanan, 2021; Hu et al., 2020). Nevertheless, none of this would be possible without an appropriate amount of children's speech data.

1.4. Children's Speech Corpora

In recent years, the popularization of the so-called "Challenges", aiming to improve the speech recognition accuracy given specific circumstances and with the use of children's speech (Yu et al., 2021; Ng et al., 2020; Lo et al., 2020; Chen et al., 2020), has been beneficial for the ASR field, especially because it reflects an increase in the availability of children's data.

According to Claus (Claus et al., 2013), there are a number of children's speech corpora in many different languages available for ASR, but none of them in Icelandic. He also points out that the majority of the resources come from children between 6 to 18 years old. Thus the Samrómur Children is the first corpus of this type which is in Icelandic. The children in the corpus are aged between 4 to 17 years, making it comparable to other datasets, and therefore representing a valuable contribution for the ASR field and the language technologies.

2. The Samrómur Children Corpus

As mentioned in the abstract, Samrómur Children was collected using the website `samromur.is`. For this reason, Samrómur Children is in the format of the Samrómur corpus (Mollberg et al., 2020) which is convenient and straightforward from a computational perspective.

The Samrómur data collection has been ongoing since late 2019. Crowd-sourcing data from children poses as a special challenge as parental consent is needed. Therefore, a considerable amount of work went into making the consent form for participants in order to be in accordance with the provisions of the European Commission, which stipulates that all individuals under the age of 18 are required to have consent from a guardian to participate for the General Data Protection Regulation (GDPR) compliance (European Commission, 2018).

The Samrómur collection platform was described in a previous publication (Mollberg et al., 2020), the prompts that participants read are sourced from various text corpora; 1) the Icelandic Web of Science (`visindavefur.is`), 2) the Icelandic portion of Wikipedia, 3) the MIM corpus (Helgadóttir et al., 2012), 4) the Icelandic Gigaword corpus (Steingrímsson et al., 2017), and books donated from the authors. The code for the text preprocessing prompts is available on GitHub[2]. The prompts were designed to fit the intent reader on the platform based on their age. Special care was taken to make prompts short and easy to read for children and prompts that could include profanity or inappropriate language for minors were removed. The design choices are shown in Table 1.

Age group	Sentence length	Word max length
10 and under	2-8	8
11-15	6-10	17
16 and older	5-15	35

Table 1: Rules for how the prompts were divided for different age groups. At age 16 and up, the user is expected to be a fully proficient reader.

Samrómur Children is pending publication on `ldc.upenn.edu` but already released at `clarin.is` with a Creative Commons Licence (CC-BY 4.0).

2.1. Corpus Format

The corpus comprises audio snippets, a metadata file provided in a "tab-separated values" (tsv) format and a README file with relevant information about the dataset.

2.1.1. Directory Organization

The metadata file is located at the root directory of the corpus next to the README file. The audio files are spread in their corresponding directories of train, development and test. Within the previous folders, one can find a number of directories with a numeric ID, one per every speaker in the corpus. Inside the speaker folders, one can find the audio files associated to the speaker.

2.1.2. Audio Format

The distributed audio files are encoded at a 16 kHz sampling rate with 16 bit of linear PCM, and 1 channel. Every audio file contains one sentence uttered by the voice of one single speaker. The minimum length of a sentence is 3 words and the maximum is 14. In terms of time-length, the average is around 3.4 seconds approximately. The audio files of the corpus are named according to the following convention: $< speaker\_ID > - < utterance\_ID > . < file\_extension >$.

2.1.3. Metadata

The metadata file contains 19 columns with relevant information about the speakers such as gender, age, dialect and

[2] `https://github.com/cadia-lvl/samromur-tools/tree/master/ScriptMaker`

native language as well as the prompts in their original text form (as the participants saw them) and in their normalized form (in lower case with no punctuation marks).

2.2. Corpus Design

As discussed in Section 1.2, it is known that children's speech is particularly hard to recognise due to its high variability, which is due to the developmental changes in children's anatomy and speech production skills (Hämäläinen et al., 2014). Therefore, the children's age has to be taken into account when creating the train, development and test portions. Nonetheless, Samrómur Children is an unbalanced corpus in terms of gender and age of the speakers. Figure 1 plots two bar graphs. The one at the bottom shows the amount of training data broken down by age and gender of the speakers. As it can be seen, there are some ages with more training data than others and, in general, the amount of data provided by female speakers is greater.

This intrinsic unbalance impose conditions in the type of the experiments than can be performed with the corpus. For example, an equal number of female and male speakers through certain ranges of age is impossible. For this reason, we took special care in creating at least a test portion that maximizes the amount of data that can be even along the majority of ranges of age. So, the test portion of the Samrómur Children covers the ages between 6 to 16 years in both female and male speakers. Every of these ranges of age in both genders have an exact duration of 5 minutes each. Notice that Samrómur children includes ages from 4 to 17 years, but the ages 4, 5 and 17 contain so little data that it was not possible to include them in the test or development sets.

The development portion of the corpus contains only speakers with an unknown gender information. The reason is to leave more quality data for the test portion, which is more important when reporting experiments. Both test and development sets have a total duration of 1 hour and 50 minutes each.

In order to perform fairer experiments, speakers in the train and test sets are not shared. Nevertheless, there is only one speaker shared between the train and development sets due to the relatively big amount of audio files associated to it (1000 speech files). It can be identified with the speaker $ID = 010363$. However, no audio files are shared between these two sets.

2.3. Corpus Statistics

The Samrómur Children is comprised of $137,597$ utterances from $3,175$ speakers, having a total duration of 131 hours of speech data. Table 2 shows the same information but broken down into gender of the speakers. However, there is a few number of speakers with unknown gender information.

2.3.1. The Train Portion

The training portion of the corpus has a total of $134,394$ utterances from $2,517$ speakers. The total duration of this portion is 127 hours and 25 minutes. Table 3 shows the statistics of the train portion broken down into gender of the speakers.

Gender	Female	Male	Unknown
Duration	$73h38m$	$52h26m$	$05h02m$
Num. Utterances	$78,993$	$53,927$	$4,677$
Num. Speakers	$1,667$	$1,412$	96

Table 2: Statistics of the whole corpus

Gender	Female	Male	Unknown
Duration	$72h43m$	$51h30m$	$03h11m$
Num. Utterances	$78,148$	$53,058$	$3,188$
Num. Speakers	$1,357$	$1,097$	63

Table 3: Statistics of the train portion

2.3.2. The Test Portion

The test portion of the corpus has a total of 845 utterances from 625 speakers. The total duration of this portion is 1 hour and 50 minutes. Table 4 shows the statistics of the test portion broken down into gender of the speakers.

Gender	Female	Male	Unknown
Duration	$00h55m$	$00h55m$	$0h0m$
Num. Utterances	845	869	0
Num. Speakers	310	315	0

Table 4: Statistics of the test portion

2.3.3. The Development Portion

The development portion of the corpus has a total of $1,489$ utterances from 34 speakers. The total duration of this portion is 1 hour and 50 minutes. Table 5 shows the statistics of the development portion broken down into gender of the speakers.

Gender	Female	Male	Unknown
Duration	$00h00m$	$00h00m$	$01h50m$
Num. Utterances	0	0	$1,489$
Num. Speakers	0	0	34

Table 5: Statistics of the development portion

Notice that in this portion, despite the speaker with $ID = 010363$ is shared with the training set, the audio files belonging to this speaker that are spread in both portions are different. So, the speaker is shared but the files associated to it are divided between the two splits.

3. Experiments

The following section is intended to describe the construction of an ASR system using Kaldi (Povey et al., 2011) in a typical architecture consisting of language model, pronunciation dictionary and acoustic model, in order to provide some baseline results and, at the same time, demonstrating that the Samrómur Children is suitable for training off-the-shelf ASR engines.

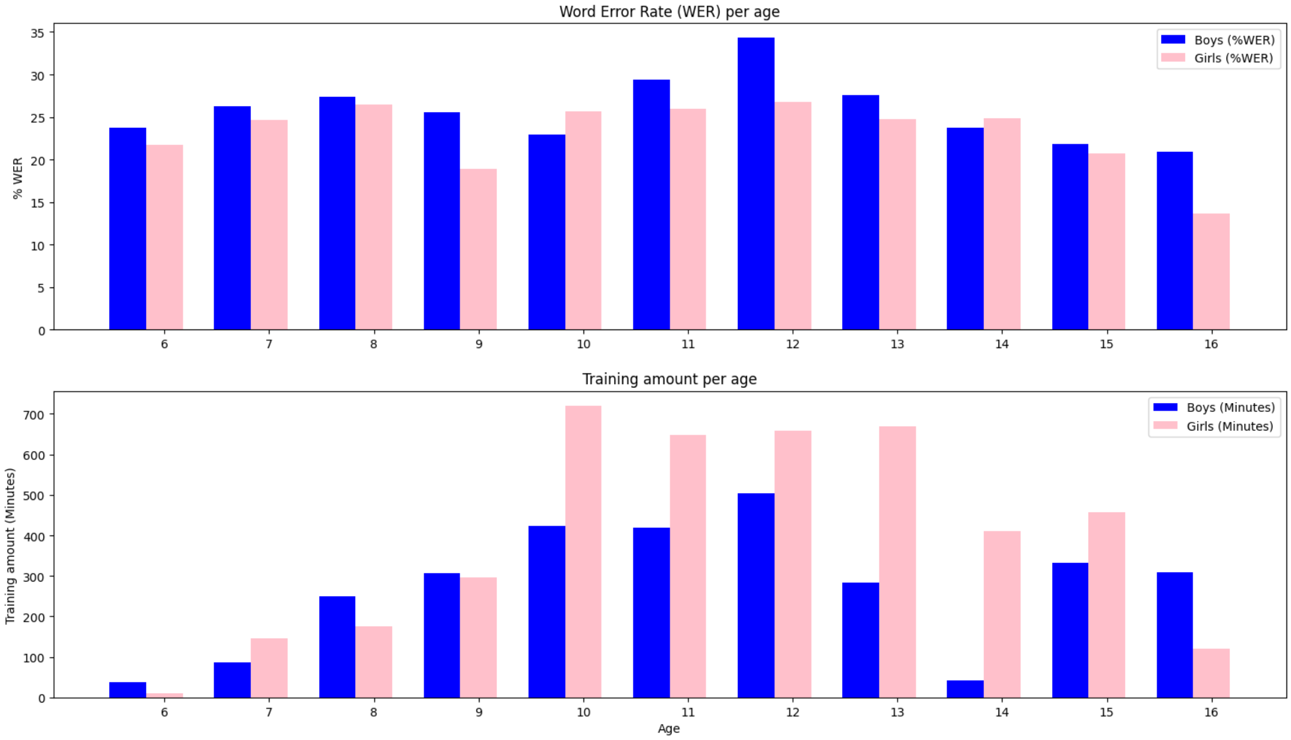

Figure 1: Bottom) Training amount per age and gender shown in minutes. Top) WER per age and gender.

When the data in Samrómur Children is divided through the different ranges of age the amount of data for an specific range of age is reduced dramatically. For this reason we decided to perform this baseline experiments using the whole data. In the future, when we gather more children's data it will be possible to perform more detailed experiments taking into account more narrow ranges of age.
The code to generate both the language model and the pronouncing dictionary is available in GitHub[3].

3.1. Kaldi Setup

We followed the Kaldi recipe[4] designed for the TED-LIUM corpus (Rousseau et al., 2012). As a first stage, this recipe generates an HMM triphone model with LDA/MLLT (Saon et al., 2000) and SAT (Tomashenko, 2017) training adaptations. Next is to augment the training data using speed perturbation (Ko et al., 2015) in two different ratios with respect to the original speed (0.9 and 1.1), then calculating iVectors (Peddinti et al., 2015) for the whole corpus (including the augmented data) and finally, implementing a TDNN-LSTM network (Graves et al., 2013). The result of this process produced the best word error rate (WER) of 21.11% in the development set and 24.47% in the test set (see Table 6).

3.2. Language Model

The language model was created using the Icelandic Gigaword corpus (Steingrímsson et al., 2018) as well as the training prompts. The Gigaword corpus contains text from newspaper articles, parliamentary speeches, adjudications, books, transcribed radio/television news and more. The normalization process of the sentences utilized to generate the language model included to allow only characters belonging to the Icelandic alphabet, expanding numbers and abbreviations, and removing punctuation marks (Nikulásdóttir et al., 2018b). The resulting text has a length of more than 44 million lines of text (5.3 GB approximately), and it was used to create a 3-gram (for decoding) and a 4-gram (for re-scoring) language models with the SRILM toolkit (Stolcke, 2002). Notice that there is no need to create an special language model suitable for children as they read prompts from the same sources as all the versions of Samrómur; the only difference is that children read shorter prompts in terms of words as shown in Table 1.

3.3. Pronunciation Dictionary

The pronouncing dictionary or "lexicon" was created with the words extracted from the text for the language model (section 3.2) and with the help of the Sequitur-G2P (Bisani and Ney, 2008; Nikulásdóttir et al., 2018a), which is a trainable grapheme-to-phoneme tool. The resulting pronouncing dictionary contains more than 960,000 entries.

3.4. Results

Table 6 shows the overall WER obtained with our Kaldi setup in the test and development portions of the corpus applying both the HMM model and the LSTM network.
Table 7 shows the best WER results obtained with the LSTM network in the test portion of the corpus but bro-

[3] `https://github.com/cadia-lvl/samromur-asr/tree/master/s5_base`

[4] `https://github.com/kaldi-asr/kaldi/tree/master/egs/tedlium`

Experiment	(%WER) Dev	(%WER) Test
Kaldi HMMs	32.76	43.71
Kaldi LSTM	21.11	24.47

Table 6: WER results obtained with Children Corpus

ken down in gender and age range through SCLITE (NIST, 2018). These results correspond to the upper bar graph shown in Figure 1.

Age (years)	(%WER) Female	(%WER) Male
6	21.7	23.7
7	24.7	26.3
8	26.5	27.4
9	18.9	25.6
10	25.7	22.9
11	26.0	29.4
12	26.8	34.3
13	24.8	27.6
14	24.9	23.7
15	20.7	21.8
16	13.7	20.9

Table 7: WER results per age using a LSTM network on the test portion

4. Discussion

Results presented in Table 6 do not correspond to the amount of training data utilized to generate them ($127h25m$) when comparing to the results in Table 8, which shows the WER we obtained with the adult's version of Samrómur with a similar Kaldi setup.

Portion	Duration	Best %WER
Dev	$15h16m$	11.48
Test	$15h51m$	12.98

Table 8: WER obtained with adult's speech of Samrómur

Results in Table 8 were generated with $114h34m$ of training data and, as it can be seen, the development and test portions are considerably larger than the ones in Samrómur Children. So, why do we have worse results with Samrómur Children than with Samrómur adults? The answer could rely on the speech variability in children discussed in section 1.2, and the uneven distribution of training data through the different ranges of age.

Figure 1 shows two different bar graphs aligned by age (horizontal axis). As it can be seen, the bars at the bottom come in a variety of heights, reflecting how unbalanced the corpus is in this respect. Nevertheless, the bars of WER in the upper graph are distributed more evenly. This particular behavior of the WER could be explained by a sort of transfer learning from an age range to another. As a example of this, see the ages of 6 and 7 years and how poor they are in terms of training data. Now notice how they have a corresponding WER which is not as different as the WER of the rest of age ranges.

However, these observations open the door to the idea of doing further experiments intended to explain in depth the phenomenon described here.

5. Conclusion and Further Work

In this paper, we have performed an historical review of previous ASR developments for Icelandic and we have seen the context in which Samrómur Children was created. We have discussed the issues that come when working with children's speech and we have presented a literature review defending the exposed ideas.

Later on, the Samrómur Children corpus was presented in detail. We explained how meticulously the test portion of the corpus was created to cover a wide range of age ranges as possible with the exact same amount of data, in order to perform fairer experiments with the corpus.

We have shown graphs and tables pointing out the intrinsic unbalance of the corpus and how despite that, the WER distributes in an even fashion across the ranges of age. According to our point of view, this particular behavior of the WER is not described in the literature as clear as it is shown in Figure 1.

Regarding to the ASR experiments performed with our Kaldi setup, the best WER we were able to achieved in the test set is 24.47% and 21.11% in the development set. Unfortunately, this WER is higher when comparing with a similar amount of training data from adult's speech.

Further experiments are needed to clarify the questions addressed by this paper and to establish a lower WER. Luckily, Samrómur is an active project; it is just a matter of time for it to provide with new data releases, and not only of children's, but also with adult's and elder's speech. We will also be able to release data that compensate any imbalance in age or gender that is present in the current version of Samrómur Children.

Finally at the experiments section, it has been demonstrated that the Samrómur Children is suitable for ASR engines and it is a valuable resource that contributes to the advances of the speech technology not only in Iceland but in the rest of the world.

6. 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.

The verification for the dataset was funded by the the Icelandic Directorate of Labour's Student Summer Job Program in 2020 and 2021. Special thanks to the summer students for all the hard work.

7. References

Ardila, R., Branson, M., Davis, K. Henretty, M., Kohler, M. Meyer, J., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Bangalore, S., Sridhar, V. K. R., Kolan, P., Golipour, L., and Jimenez, A. (2012). Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 437–445.

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

Chen, G., Na, X., Wang, Y., Yan, Z., Zhang, J., Ma, S., and Wang, Y. (2020). Data augmentation for children's speech recognition–the" ethiopian" system for the slt 2021 children speech recognition challenge. *arXiv preprint arXiv:2011.04547*.

Claus, F., Rosales, H. G., Petrick, R., Hain, H.-U., and Hoffmann, R. (2013). A survey about databases of children's speech. In *INTERSPEECH*, pages 2410–2414.

Das, S., Nix, D., and Picheny, M. (1998). Improvements in children's speech recognition performance. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 433–436. IEEE.

Elenius, D. and Blomberg, M. (2005). Adaptation and normalization experiments in speech recognition for 4 to 8 year old children. In *Ninth European Conference on Speech Communication and Technology*.

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

Fainberg, J., Bell, P., Lincoln, M., and Renals, S. (2016). Improving children's speech recognition through out-of-domain data augmentation. In *Interspeech*, pages 1598–1602.

Fringi, E., Lehman, J. F., and Russell, M. (2015). Evidence of phonological processes in automatic recognition of children's speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59.

Gerosa, M., Giuliani, D., and Narayanan, S. (2006). Acoustic analysis and automatic recognition of spontaneous children's speech. In *Ninth International Conference on Spoken Language Processing*.

Gerosa, M., Giuliani, D., and Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49(10-11):847–860.

Ghai, S. and Sinha, R. (2009). Exploring the role of spectral smoothing in context of children's speech recognition. In *Tenth Annual Conference of the International Speech Communication Association*.

Ghai, S. (2011). *Addressing pitch mismatch for children's automatic speech recognition*. Ph.D. thesis, Indian Institute of Technology Guwahati.

Giuliani, D. and BabaAli, B. (2015). Large vocabulary children's speech recognition with dnn-hmm and sgmm acoustic modeling. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Giuliani, D. and Gerosa, M. (2003). Investigating recognition of children's speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–137. IEEE.

Graves, A., Jaitly, N., and Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.

Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsson, H. H., Loftsson, H., Helgadóttir, S., Jóhannsdóttir, K. M., and Rögnvaldsson, E. (2012). Almannarómur: An open icelandic speech corpus. In *Proceedings of the Third International Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2012, Cape Town, South Africa*.

Guðnason, J., Pétursson, M., Kjaran, R., Klüpfel, S., and Nikulásdóttir, A. B. (2017). Building asr corpora using eyra. In *INTERSPEECH*, pages 2173–2177.

Hämäläinen, A., Candeias, S., Cho, H., Meinedo, H., Abad, A., Pellegrini, T., Tjalve, M., Trancoso, I., and Dias, M. S. (2014). Correlating asr errors with developmental changes in speech production: A study of 3-10-year-old european portuguese children's speech. In *Workshop on Child Computer Interaction-WOCCI 2014*, pages pp–1.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The tagged icelandic corpus (mím). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, pages 67–72.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. (2017). Building an asr corpus using althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

Hu, K., Bruguier, A. J., Sainath, T. N., Prabhavalkar, R. P., and Pundak, G. (2020). Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models, November 5. US Patent App. 16/861,190.

Kent, R. D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of speech and hearing Research*, 19(3):421–447.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Lee, S., Potamianos, A., and Narayanan, S. (1997). Analysis of children's speech: Duration, pitch and formants. In *Fifth European Conference on Speech Communication and Technology*.

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Li, C. and Qian, Y. (2019). Prosody usage optimization for children speech recognition with zero resource children speech. In *Interspeech*, pages 3446–3450.

Li, Q. and Russell, M. J. (2001). Why is automatic recognition of children's speech difficult? In *Seventh European Conference on Speech Communication and Technology*.

Li, Q. and Russell, M. J. (2002). An analysis of the causes of increased error rates in children's speech recognition. In *Seventh International Conference on Spoken Language Processing*.

Lo, T.-H., Chao, F.-A., Weng, S.-Y., and Chen, B. (2020). The ntnu system at the interspeech 2020 non-native children's speech asr challenge. In *INTERSPEECH*, pages 250–254.

Matassoni, M., Gretter, R., Falavigna, D., and Giuliani, D. (2018). Non-native children speech recognition through transfer learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6229–6233. IEEE.

Mollberg, D. E., Jónsson, Ó. H., Thorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Guðnason, J. (2020). Samrómur: Crowd-sourcing data collection for icelandic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3463–3467.

Mugitani, R. and Hiroya, S. (2012). Development of vocal tract and acoustic features in children. *Acoustical Science and Technology*, 33(4):215–220.

Ng, S.-I., Liu, W., Peng, Z., Feng, S., Huang, H.-P., Scharenborg, O., and Lee, T. (2020). The cuhk-tudelft system for the slt 2021 children speech recognition challenge. *arXiv preprint arXiv:2011.06239*.

Nikulásdóttir, A. B., Guðnason, J., and Rögnvaldsson, E. (2018a). An icelandic pronunciation dictionary for tts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.

Nikulásdóttir, A. B., Helgadóttir, I. R., Pétursson, M., and Guðnason, J. (2018b). Open asr for icelandic: Resources and a baseline system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language technology programme for icelandic 2019-2023. *arXiv preprint arXiv:2003.09244*.

NIST. (2018). Sctk, the nist scoring toolkit. available at https://github.com/usnistgov/sctk, November 11.

Palethorpe, S., Wales, R., Clark, J. E., and Senserrick, T. (1996). Vowel classification in children. *The Journal of the Acoustical Society of America*, 100(6):3843–3851.

Peddinti, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Potamianos, A. and Narayanan, S. (2003). Robust recognition of children's speech. *IEEE Transactions on speech and audio processing*, 11(6):603–616.

Potamianos, A., Narayanan, S., and Lee, S. (1997). Automatic speech recognition for children. In *Fifth European Conference on Speech Communication and Technology*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Qian, Y., Wang, X., Evanini, K., and Suendermann-Oeft, D. (2016). Improving dnn-based automatic recognition of non-native children speech with adult speech. In *WOCCI*, pages 40–44.

Retico, A., Giuliano, A., Tancredi, R., Cosenza, A., Apicella, F., Narzisi, A., Biagi, L., Tosetti, M., Muratori, F., and Calderoni, S. (2016). The effect of gender on the neuroanatomy of children with autism spectrum disorders: a support vector machine case-control study. *Molecular autism*, 7(1):1–20.

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

Rousseau, A., Deléglise, P., and Esteve, Y. (2012). Tedlium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Russell, M., D'Arcy, S., and Qun, L. (2007). The effects of bandwidth reduction on human and computer recognition of children's speech. *IEEE Signal Processing Letters*, 14(12):1044–1046.

Safavi, S., Russell, M., and Jančovič, P. (2018). Automatic speaker, age-group and gender identification from children's speech. *Computer Speech & Language*, 50:141–156.

Saon, G., Padmanabhan, M., Gopinath, R., and Chen, S. (2000). Maximum likelihood discriminant feature spaces. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 2, pages II1129–II1132. IEEE.

Serizel, R. and Giuliani, D. (2014). Deep neural network adaptation for children's and adults' speech recognition. In *Italian Computational Linguistics Conference (CLiC-it)*.

Sheng, P., Yang, Z., and Qian, Y. (2019). Gans for children: A generative data augmentation strategy for children speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 129–135. IEEE.

Shivakumar, P. G. and Narayanan, S. (2021). End-to-end neural systems for automatic children speech recognition: An empirical study. *arXiv preprint arXiv:2102.09918*.

Steingrímsson, S., Guðnason, J., Helgadóttir, S., and Rögnvaldsson, E. (2017). Málrómur: A manually verified corpus of recorded icelandic speech. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 237–240.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Gudnason, J. (2018). Risamálheild:

A Very Large Icelandic Text Corpus. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Tomashenko, N. (2017). *Speaker adaptation of deep neural network acoustic models using Gaussian mixture model framework in automatic speech recognition systems*. Ph.D. thesis, Le Mans.

Tong, R., Wang, L., and Ma, B. (2017). Transfer learning for children's speech recognition. In *2017 International Conference on Asian Language Processing (IALP)*, pages 36–39. IEEE.

Wilpon, J. G. and Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, volume 1, pages 349–352. IEEE.

Wu, F., García-Perera, L. P., Povey, D., and Khudanpur, S. (2019). Advances in automatic speech recognition for child speech using factored time delay neural network. In *INTERSPEECH*, pages 1–5.

Yu, F., Yao, Z., Wang, X., An, K., Xie, L., Ou, Z., Liu, B., Li, X., and Miao, G. (2021). The slt 2021 children speech recognition challenge: Open datasets, rules and baselines. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1117–1123. IEEE.