Open ASR for Icelandic: Resources and a Baseline System

Anna Björk Nikulasdóttir, Inga Rún Helgadóttir, Matthías Pétursson, Jón Guðnason
Reykjavik University
Reykjavik Iceland
{annabn, ingarun, matthias, jg}@ru.is

Abstract

Developing language resources is an important task when creating a speech recognition system for a less-resourced language. In this paper we describe available language resources and their preparation for use in a large vocabulary speech recognition (LVSR) system for Icelandic. The content of a speech corpus is analysed and training and test sets compiled, a pronunciation dictionary is extended, and text normalization for language modeling performed. An ASR system based on neural networks is implemented using these resources and tested using different acoustic training sets. Experimental results show a clear increase in word-error-rate (WER) when using smaller training sets, indicating that extension of the speech corpus for training would improve the system. When testing on data with known vocabulary only, the WER is 7.99%, but on an open vocabulary test set the WER is 15.72%. Furthermore, impact of the content of the acoustic training corpus is examined. The current results indicate that an ASR system could profit from carefully selected phonotactical data, however, further experiments are needed to verify this impression. The language resources are available on http://malfong.is and the source code of the project can be found on https://github.com/cadia-lvl/ice-asr/tree/master/ice-kaldi.

Keywords: language resources, automatic speech recognition, Icelandic

1. Introduction

With advances in automatic speech recognition (ASR) technology and access to open source ASR tools like Kaldi[1] (Povey et al., 2011), the threshold for the development of ASR systems has been lowered substantially. The core ASR algorithms, including state-of-the-art neural network architectures, are already implemented, leaving the development of language resources as the main task.

The development of sufficient language resources, however, poses a challenge for less-resourced languages. Two specialized resources are needed for automatic speech recognition: a speech corpus with matching units of speech and text, and a pronunciation dictionary. Additionally, a large text corpus is needed for the development of a language model.

In this paper we describe the development of language resources for an Icelandic large vocabulary speech recognition system (LVSR), and the setup of a baseline system in Kaldi.

The first speech recognizer for Icelandic was *Hjal*, developed in 2003 as an isolated word recognizer (Rögnvaldsson, 2003). The *Hjal* project was successful as a pilot project, but has not been maintained. Currently, three ASR systems are being developed for Icelandic using Kaldi: an ASR system for parliamentary speeches (Helgadóttir et al., 2017), a system for automating transcriptions of radiology dictations[2], and the LVSR project described in this paper. Google already supports Icelandic in its speech recognition applications[3].

We inspect available language resources for Icelandic and prepare them for use in a speech recognition system. After defining the characteristics of the resources and testing an initial setup of an LSVR system, we make suggestions for further development of the resources. In particular we address three questions: a) is the available speech corpus sufficiently large for the development of a state-of-the-art LVSR system? b) is the pronunciation dictionary large enough for an LVSR system? and c) what is the impact of the content of the speech corpus on the results of the system?

[1] http://kaldi-asr.org/

[2] http://laeknaromur.is/en.html

[3] see e.g. https://cloud.google.com/speech/

2. Data

To develop and train an ASR system three kinds of language resources are needed: a speech corpus to train an acoustic model, a text corpus for language modeling, and a pronunciation dictionary. For the current project available resources for Icelandic were examined.

2.1. Málrómur Speech Corpus

During the years 2010-2011 the Reykjavik University collected Icelandic speech samples in cooperation with Google (Guðnason et al., 2012; Steingrímsson et al., 2017). Over 550 speakers of both genders and various age recorded a collection of sentences and isolated words using smartphones. The version of Málrómur used in this project has been cleaned of recordings marked as 'faulty' by the speakers themselves (normally empty recordings), and contains about 122,000 utterances or approx. 150 hours of speech. The recorded utterances represent 33,135 unique prompts read by the speakers. This list of prompts is composed as follows: the largest part are short sentences from news articles, then carefully selected sentences regarding diphone and triphone distribution (developed during the *Hjal* project (Rögnvaldsson, 2003)), proper names, Icelandic streets and locations, international cities and countries, numbers, miscellaneous entries, and a list of URLs. We call the sentence list from the *Hjal* project *phonotactical sentence list* since its purpose is to thoroughly represent the phonotactics of Icelandic. As shown in Fig. 1, the news sentences are 66% of all entries in the list of prompts, location names and proper names are about 20% and 10% respectively, and

other categories have much fewer entries. In the Málrómur speech corpus the distribution of categories changes. Most notably the few entries from the phonotactical list and the list of URLs are recorded multiple times, leading to their increasing partition of the speech corpus.

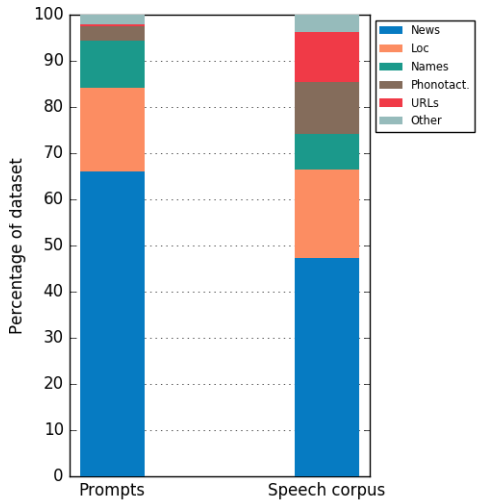

Figure 1: Distribution of prompt categories in the prompts list (~33k entries) and the Málrómur speech corpus (~122k entries). The three smaller categories, numbers, international cities and countries, and miscellaneous entries are subsumed under *Other*

The amount of repetitions is not evenly distributed among the prompts as shown in Table 1: 23% of the utterances have only one representation in the corpus and 74% are recorded 2 to 10 times, the remaining 3% are recorded up to over 100 times. When looking at the absolute numbers of recordings, 7,450 (6%) are unique, 83,752 (69%) recordings belong to the group of 2 to 10 times repeated utterances and almost 10,000 or 8% represent the 86 utterances repeated more than 100 times in the corpus.

Repetition	No. of prompts	No. of recordings
1x	7,450	7,450
2x	8,751	17,502
3-4x	11,090	37,428
5-10x	4,671	28,822
11-20x	937	12,866
21-50x	82	2,321
51-100x	68	6,001
>100x	86	9,495
TOTAL	33,135	121,822

Table 1: The repetition of prompts in the Málrómur corpus. Most prompts are recorded 1-4 times.

The prompts from news articles and the phonotactical sentences are multi word sentences but most of the prompts from the other categories are single word utterances. In the prompts list about 30% of the prompts are single words, the remaining 70% of the prompts contain 2 to 13 words. In the recorded speech corpus 40% of the utterances are single words and 60% multi word utterances. On average the multi word utterances contain 5.3 words.

2.2. LCC Text Corpus

The *Leipzig Corpora Collection (LCC)* (Goldhahn et al., 2012) contains an Icelandic text corpus, mainly collected from news sites. For this project we have access to 10 million sentences or 180 million tokens from this corpus. The sentences are scrambled, i.e. each sentence is isolated from its original context, and duplicate sentences have been removed.

2.3. Pronunciation Dictionary

During the *Hjal* project (see Section 1.) a pronunciation dictionary with almost 50,000 word forms was developed. The vocabulary selection was done using a frequency list especially compiled for the task, from news papers and other sources. It was transcribed using the SAMPA phonetic alphabet, but the dictionary has since then been extended to about 60,000 unique word forms with 65,000 entries. Furthermore, IPA transcriptions have been added.

3. Resource Processing

3.1. Speech Corpus Training and Test Sets

The largest part of the speech corpus is needed to train the speech recognition system. But a fraction of the data will have to be extracted to create test and development sets used for decoding experiments. A speech corpus created in the same way as Málrómur is expected to contain errors, i.e. not all recordings will exactly match the corresponding text (Hughes et al., 2010; Steingrímsson et al., 2017). By sufficient amount of data this is not necessarily a problem for the acoustic training set. For a test set, however, it is very important that the utterances match the corresponding texts to a high degree. An attempt was made to select high quality recordings for test and development sets. The aim was to collect utterances from reliable speakers for each of the two sets. This was done by listening to 3,000 random recordings from the corpus. Four evaluators (two researchers, two students) used the Eyra application (Pétursson et al., 2016; Guðnason et al., 2017) to rate the recordings and rate them from 1 (=bad) to 4 (=very good). The kappa (Fleiss) value for the evaluation results of the four evaluators was 0.598, which is at the lower limit for substantial agreement (Landis and Koch, 1977). Unifying grades 1 and 2 to one grade (bad) and grades 3 and 4 to one grade (good), a kappa value of 0.64 was computed. The results of the evaluation were used to identify speakers whose recordings generally match the prompts. From those, 40 speakers of each gender were randomly selected, having at least 100 sentence utterances and 50 one word utterances in the corpus. Two constraints were defined for the test and development sets: they should not contain utterances from the phonotactical sentence list and no URLs. The primary role of the phonotactical sentences is to collect acoustic material on as many valid di- and triphones as possible. The sentences, however, often contain very rare

words and combinations and are thus not likely to be predicted by a general language model. Another reason not to include phonotactical sentences in the test sets is that readers are more likely to make errors. As for the URLs, during this first experiment they were not normalized, making it impossible for the ASR system to decode them correctly. The final test and development sets consisted of 6,000 utterances from 20 female and 20 male speakers each. All utterances from the remaining speakers in the Málrómur corpus built the general acoustic training set, 88,285 utterances or approx. 109 hours of speech. Additionally, two experimental training sets were compiled from the general training set: one set were all sentences from the phonotactical sentence list were deleted and another set of the same size where the phonotactical sentences were kept but the corresponding number of news utterances were deleted. These sets contain 78,529 utterances each. The training set with the phonotactical sentences has approx. 95 hours of speech whereas the other set has about 93 hours. The phonotactical sentences are often longer than the news paper sentences, resulting in this difference in duration.

3.2. Text Normalization

The Icelandic LCC corpus was used for language modeling for the ASR system. When using a raw, unrestricted corpus for language modeling, one has to be aware of non-standard words, symbols and punctuation. In line with (Sproat et al., 2001) a taxonomy of non-standard words in Icelandic was developed to prepare text normalization. The normalization of Icelandic texts is a work in progress. Normalization tasks for non-standard words have been defined and for Icelandic, decompounding should also be considered (Adda-Decker and Adda, 2000; Ordelman et al., 2003). In this first setup of the ASR system only basic normalization was performed: all non-relevant symbols and punctuation was removed, common acronyms and abbreviations were replaced, and some corpus specific processing made to remove web page navigation tokens. The text was tokenized, thus separating punctuation from words, and lowercased. Next step would be to adapt the normalization model developed for ASR of Icelandic parliament speeches to replace digits with their spelled out forms (Helgadóttir et al., 2017). This is a non-trivial problem in Icelandic, since cardinal and ordinal numbers from one to four (and all larger numbers ending with these numbers) behave much like adjectives: they are inflected by case and gender and some even by number.

3.3. Extending the Pronunciation Dictionary

For a highly inflective language like Icelandic, a lexicon of 60,000 words is not sufficient for an LVSR system. The Database of Modern Icelandic Inflection contains almost 280,000 inflection paradigms with 2.8 million distinct word forms[4] (Bjarnadóttir, 2012). The text representation of the Málrómur corpus contains 435,000 tokens. The out-of-vocabulary rate (OOV) of this corpus when using the pronunciation dictionary is 19.5%, and of the 32,765 types 17,676 or 54% are contained in the dictionary. Some of the missing types are numbers, acronyms, URLs, foreign words or words with spelling errors. Nevertheless, an extension of the lexicon was necessary in order to reduce the OOV rate and increase lexical coverage. To select words to add to the lexicon two sources were used: a) the LCC text corpus and b) the text representation of the acoustic training set from Málrómur. The 200,000 most frequent words from LCC and all words occurring at least three times in the Málrómur training set were extracted and heuristics used to compile an extended lexicon. These include filtering out words with characters not contained in the Icelandic alphabet (*c, w, ä,* etc.), validation based on upper and lower case, analysis of common n-grams, etc. All words in the original lexicon not occurring in this new word list were deleted. Furthermore, as a general rule, for words that can be written either as one or two words, the one word version was deleted, e.g. *hinsvegar* vs. *hins vegar*. The resulting version of the lexicon contains 134,866 words, 46,209 from the original pronunciation dictionary and 88,657 new words.
For the transcriptions a fixed set of 58 IPA symbols was used. The transcriptions in the original pronunciation dictionary are somewhat inconsistent regarding this set of symbols, and thus some automatic correction procedures were run. Furthermore, some dialect variations were removed, these would need to be added again when training an ASR system where recordings from dialect speakers are included[5]. Overall, transcriptions of about 7,200 words were changed in some way, corrected and/or variations deleted. These revised transcriptions were used to train a grapheme-to-phoneme model using the Sequitur G2P converter[6] (Bisani and Ney, 2008). The extended lexicon was then automatically transcribed using this model. The pronunciation dictionary with the core content from the (corrected) manually transcribed dictionary plus the automatically transcribed word list generated from LCC and the Málrómur training corpus now contains 136,082 entries. This reduces the OOV rate for the whole Málrómur corpus from 19.5% to 8.9% and 76.2% of the types are now included in the pronunciation dictionary compared to 54% before.

4. Training a Baseline ASR System

DNNs are a current standard in acoustic modeling for ASR and represent a very active research field within the ASR community. We used the Kaldi ASR toolkit for the development of our baseline DNN system (Povey et al., 2011) and aligned the acoustic model training to a recipe developed for the Switchboard corpus[7]. The first step in the acoustic model training (after extracting the necessary Mel-Cepstrum-Coefficients (MFCCs)), is to train a hidden Markov model with Gaussian mixture models (HMM-GMM) and to apply Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) to the acoustic features. Then a sequence model based on time delay deep neural network (TD-DNN) (Waibel et al., 1989) layers and long-short term memory (LSTM) network (Sak

[4] http://bin.arnastofnun.is/DMII/

[5] Icelandic does, however, not have highly diverse dialects

[6] https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

[7] https://github.com/kaldi-asr/kaldi/blob/master/egs/swbd/s5c/local/chain/tuning/run_tdnn_lstm_1e.sh

et al., 2014) is trained (Povey et al., 2016). The network takes 40 dimensional LDA feature vectors and a 100 dimensional ivector as input and it is composed of ten hidden layers, of which seven are TDNN layers and three are LSTM layers. Two language models were trained using the MITLM toolkit (Hsu and Glass, 2008). Both models are a trigram model with modified Kneser-Ney smoothing (Chen and Goodman, 1996). One model uses 1 million tokens from the LCC corpus and is used in the first pass decoding in combination with the acoustic model. The second one is used for LM rescoring and uses 10 million tokens from LCC. The extended Icelandic pronunciation dictionary used in the system was described in Section 3.

5. Experimental setup

We trained three ASR models using the training setup as described in the previous Section. The only difference between the models is due to the speech corpus training set: a) ALL-DATA: all training data, 88,285 utterances or approx. 109 hours of speech b) NO-PHONOTACT.: no utterances from the set of phonotactical sentences, and c) PHONOTACT.: a training set of the same size as NO-PHONOTACT. including all phonotactical utterances, but with corresponding reduction in utterances from the news category. In the original test set the OOV-rate is 5.7%, and 13.2% of the types are not contained in the lexicon of the ASR system. An additional test set only with known words was created, and each ASR model used to decode each of the both test sets.

6. Evaluation

Table 2 shows the results of the decoding experiments.

Training data	Full test set	Sub test set
ALL-DATA (109h)	15.72% WER	7.99% WER
NO-PHONOT.. (93h)	16.27% WER	8.60% WER
PHONOTACT. (95h)	16.17% WER	8.41% WER

Table 2: Test results of the Open ASR for Icelandic. Best word-error-rate (WER) is reached when using the largest available acoustic training set (ALL-DATA), approx. 109 hours of speech. The NO-PHONOT. training data set is a slightly smaller set were all phonotactical sentences have been deleted, and the PHONOTACT. set is similar in size to NO-PHONOT., but contains all these specially selected sentences.

The results of the system trained on the whole training set show 15.72% word-error rate (WER) on the full test set. Removing 14-16 hours from the training set increases the WER to 16.17% and 16.27%. Getting more training data would therefore most likely improve the results of the current system.

There is a large gap between the results for the full test set and the test set without unknown words. Recall that the OOV rate of the full test set is only 5.7%, compared to an absolute difference in WER of 7.67% - 7.76% between the full test set and the no OOV test set results. This leads to the assumption that an error caused by an unknown word causes more errors in the succeeding words, as has been shown to be the case for other languages (see e.g. (Salimbajevs and Strigins, 2015)).

One question to be answered by the experiment, was if the carefully selected sentences in the training set have an impact on decoding results. For the full test set and especially for the no OOV test set, the results of the PHONOTACT. set (containing the specially selected sentences) are clearly better than for the set without these sentences. The PHONOTACT. set, however, is slightly larger in duration, approx. 95 hours of speech compared to approx. 93 hours of speech in the NO-PHONOTACT. set. Further experiments are needed to explore this impact. Tests with still smaller training sets, where the phonotactical sentences comprise a larger part of the PHONOTACT. set and where both sets are still closer in duration would be needed. It would be a valuable insight to be able to quantify this impact better. If the impact is substantial, then it is worth the effort to create such a carefully crafted list if working with a low-resourced language. If, on the other hand, the impact is minimal, the work invested in the creation of the list can be skipped, plus that these sentences, at least in the case of the Icelandic sentences at hand, are difficult to read because of very rare and often long and complicated words and unusual word combinations. This diminishes the quality of the speech corpus, measured in match between spoken utterances and their corresponding prompts.

7. Conclusion and Future Work

We described the preparation of language resources for use in an LVSR system for Icelandic. The speech recognition system was implemented in Kaldi, using deep neural networks. The results of this baseline setup show the need for more acoustic training data to improve overall WER, and a larger pronunciation dictionary to be able to deal with open vocabulary decoding.

Experiments with different content of the acoustic training data indicate that careful selection of phonotactical data could be advantageous when developing a speech corpus of limited size. However, due to the extra effort needed to collect such data and the possible loss in recording quality due to reading mistakes, a clear positive impact of special phonotactical data should be evident. More experiments are needed to make a definite statement in this direction.

In addition to the extension of the speech corpus and dictionary, the ongoing work on text normalization and language modeling will be continued, as well as experiments with acoustic model architectures.

8. Acknowledgements

The project Open ASR for Icelandic was supported by the Icelandic Language Technology Fund (ILTF).

9. Bibliographical References

Adda-Decker, M. and Adda, G. (2000). Morphological decomposition for ASR in German. In *Proceedings of the Workshop on Phonetics and Phonology in Automatic Speech Recognition*, volume 5, pages 129–143.

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

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the Workshop on Language Technology for Normalization of Less-Resourced Languages - SaLTMiL 8 - AfLaT2012*, pages 13–18.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *LREC*, pages 759–765.

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 *SLTU*, pages 80–83.

Guðnason, J., Pétursson, M., Kjaran, R., Klüpfel, S., and Nikulásdóttir, A. (2017). Building ASR corpora using Eyra. In *Proceedings of Interspeech*, pages 2173–2177.

Helgadóttir, I., Kjaran, R., Nikulásdóttir, A., and Guðnason, J. (2017). Building an ASR corpus using Althingi's parliamentary speeches. In *Proceedings of Interspeech*, pages 2163–2167.

Hsu, B.-J. and Glass, J. (2008). Iterative language model estimation: Efficient data structure & algorithms. In *Proceedings of Interspeech*, pages 841–844.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., and LeBeu, M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. In *Interspeech*, pages 1914–1917.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ordelman, R., van Hessen, A., and de Jong, F. (2003). Compound decomposition in Dutch large vocabulary speech recognition. In *Interspeech*, pages 225–228.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, B., Hannemann, M., Molticek, P., Quian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Povey, D., Peddinti, V., Galvez, D., Gharhmani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Proceedings of Interspeech*, pages 2751–2755.

Pétursson, M., Klüpfel, S., and Guðnason, J. (2016). Eyra - speech data acquisition system for many languages. In *Proceedings of STLU*.

Rögnvaldsson, E., (2003). *The Icelandic speech recognition project Hjal*, pages 239–242.

Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

Salimbajevs, A. and Strigins, J. (2015). Error analysis and improving speech recognition for Latvian language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 563–569.

Sproat, R., Black, A., and et al., S. C. (2001). Normalization of non-standard words. *Computer Speech & Language*, 15:287–333.

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 of Computational Linguistics*, pages 237–240.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339.