The Althingi ASR System

Inga R. Helgadóttir, Anna B. Nikulásdóttir, Michal Borský,
Judy Y. Fong, Róbert Kjaran, Jón Guðnason

Reykjavik University
Language and Voice Lab
{ingarun,jg}@ru.is

Abstract

All performed speeches in the Icelandic parliament, Althingi, are transcribed and published. An automatic speech recognition system (ASR) has been developed to reduce the manual work involved. To our knowledge, this is the first open source speech recognizer in use for Icelandic. In this paper the development of the ASR is described. In-lab system performance is evaluated and first results from the users are described. A word error rate (WER) of 7.91% was obtained on our in-lab speech recognition test set using time-delay deep neural network (TDNN) and re-scoring with a bidirectional recurrent neural network language model (RNN-LM). No further processing of the text is included in that number. In-lab F-score for the punctuation model is 80.6 and 61.6 for the paragraph model. The WER of the ASR, including punctuation marks and other post-processing, was $15.0 \pm 6.0\%$, over 625 speeches, when tested in the wild. This is an upper limit since not all mismatches with the reference text are true errors of the ASR. The transcribers of Althingi graded 77% of the speech transcripts as Good. The Althingi corpus and ASR recipe, constitute a valuable resource for further developments within Icelandic language technology.

Index Terms: speech recognition, parliamentary transcription, human-computer interaction, post-processing of ASR output, Icelandic

1. Introduction

The policy of the Icelandic parliament, Althingi, has been to publish all speech recordings and transcripts. First, a human transcriber listens to the audio and transcribes the speech. The transcriber usually enriches the text with minor changes, e.g. fixes grammatical errors and removes repetitions. Next, in the manual editing stage, the transcript is made fit for publication on the Althingi website[1].

The parliament currently employs multiple transcribers and editors to provide this service. The main objective of the project is to replace the first stage, manual transcription, with an ASR system. The output text from the ASR system needs to be correct and readable enough to not slow down the work of the editors of Althingi. Hence, on top of a good speech recognizer, we need a good punctuation insertion model, good grammar which formats and abbreviates the text according to Althingi's conventions, and a paragraph model, so that when the editors open a long speech, they don't see a wall of text, but a text which is structured into smaller units.

The parliament has collaborated with researchers from Reykjavik University since 2016 in order to develop an automatic transcription system. The abundance of data, already existing transcription procedures, clear definition of the application and advances in ASR technologies made this an ideal opportunity for automation. The system entered a test phase in October 2018, and has been in production since January 2019. The first results have already been presented in [1] and some preliminary evaluation has been carried out [2]. This paper presents a comprehensive description of the system and its parts. Our ASR system is open source and freely available on the Language and Voice Lab GitHub page[2]. The training data is available at http://www.malfong.is.

2. System overview

2.1. The speech recognizer

The automatic speech recognition system consists of a speech recognizer and automatic post-processing of the output text. The speech recognizer contains an acoustic model and a language model. In our development we have tried many different neural network architectures for acoustic modelling. Our current acoustic model is based on a recipe developed for the Switchboard corpus[3], using the Kaldi ASR toolkit [3]. It is a sequence trained neural network based on a factorized form of time delay neural networks (f-TDNN) [4]. The network consists of eleven factorized TDNN layers with a semi-orthogonal constraint on the first factor [5]. The layer dimensions are 1280 with a linear bottleneck dimension of 256. Skip connections and $l2$-regularization are applied but dropout is not. The network takes 40 dimensional LDA feature vectors and a 100 dimensional i-vector as input. Before using this acoustic model we used a model based on lattice-free MMI [6], which consisted of seven time delay deep neural network layers and three long-short term memory layers (LF-MMI LSTM-TDNN) [7]. That model gave similar results, but had a decoding real time factor (RTF) of 1.03, while our current model structure trained on the same training set had a decoding RTF of 0.18.

A pruned trigram language model (LM) is used for decoding. Both 5-gram and recurrent neural network LMs have been used to re-score the decoding results. The n-gram language models are trained using the KenLM toolkit [8]. The recurrent neural network language model (RNN-LM) is based on a recipe for the Switchboard corpus[4]. The network consists of three TDNN layers and two LSTM layers with layer dimensions of 1024. A backwards RNN-LM can also be trained to apply on top of the forward RNN-LM. Even though using RNN-LMs for re-scoring the decoding results gives slightly better results, so far at Althingi we have rather used the 5-gram. Re-scoring with

[1] http://www.althingi.is

[2] https://github.com/cadia-lvl/kaldi/tree/master/egs/althingi

[3] https://github.com/kaldi-asr/kaldi/blob/master/egs/swbd/s5c/local/chain/tuning/run_tdnn_7n.sh

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

the RNN-LM takes over 20 times longer than using n-grams and it uses more memory. Over the whole recognition process, using RNN-LMs instead of n-grams, more then doubles the maximum memory usage.

The lexicon is based on the pronunciation dictionary from the Hjal project [9], available at Málföng[5]. Inconsistencies in the pronunciation dictionary were fixed and we added words from the language model training data, which appeared three or more times, with some constraints, resulting in a dictionary containing roughly 270k words. The phonetic transcriptions of the new words were found using the Sequitur G2P toolkit [10].

2.2. Automatic post-processing

The ASR returns a stream of words with no punctuation or formatting. Since the output of the ASR is to be read by the editors of Althingi, human readability needs to be factored into the final transcription. The OpenGrm Thrax Grammar Development tool [11, 12] was used to compile grammars into weighted finite-state transducers, in order to denormalize numbers and abbreviate words, according to parliamentary conventions. Thrax grammar rules are also used to collapse expanded acronyms, as well as for other small formatting, such as of timestamps, the look of regulations, time intervals, and websites. Currently repetitions, except those that most commonly appear correctly in text, are removed as well.

The Punctuator toolkit [13] is used to restore punctuation marks in the text, specifically periods, commas, colons and question marks and to capitalize the start of sentences. Punctuator is a bidirectional recurrent neural network model with an attention mechanism. It can both be trained on punctuation annotated text only, or additionally, take in pause annotated text. Our tests have not been able to confirm the benefit of a two-stage training, hence we use a single-stage text-only training. To improve the readability of longer texts, paragraphs are inserted. An adaptation of the punctuation model is used to predict the paragraph splits.

The following example shows how the ASR output changes throughout the following post-processing steps; **1)** original ASR output, **2)** after applying Thrax grammar to rewrite numbers and abbreviate, **3)** after applying the punctuation model, **4)** final text, after last fixings, e.g. insertion of periods after abbreviations, **5)** reference text.

- **1)** *að verðbólga fari ekki yfir tvö komma níu prósent ákvæði sextugustu og níundu grein laga um almannatryggingar var lögfest með tíundu grein laganna númer hundrað þrjátíu nítján hundruð níutíu og sjö í greinargerð*
- **2)** *að verðbólga fari ekki yfir 2,9% ákvæði 69. gr laga um almannatryggingar var lögfest með 10. gr laganna nr 130/1997 í greinargerð*
- **3)** *að verðbólga fari ekki yfir 2,9 %. Ákvæði 69. gr laga um almannatryggingar var lögfest með 10. gr laganna nr 130/1997. Í greinargerð*
- **4)** *að verðbólga fari ekki yfir 2,9%. Ákvæði 69. gr. laga um almannatryggingar var lögfest með 10. gr. laganna nr. 130/1997. Í greinargerð*
- **5)** *að verðbólga fari ekki yfir 2,9%. <EOP>*
  Ákvæði 69. gr. laga um almannatryggingar var lögfest með 10. gr. laganna nr. 130/1997. Í greinargerð

2.3. Integration with the Althingi system

The ASR connects with the rest of the Althingi servers to retrieve the templated speech XML document and audio segment. Then, the transcription is sent to Documentum repositories for the speech department to finish processing. The connections are enabled via a representational state transfer application programming interface (RESTful API), SLURM workload manager, and Documentum Foundation Classes (DFC). The ASR process starts when the API is sent the ending time of the parliamentarian's speech. Using the ending timestamp, the API queries for the XML metadata and the audio segment. Then two SLURM jobs are queued: 1. The ASR transcript is created and automatic post-processing performed. 2. The API is called with the unique speech ID which compiles the speech contents into a fully-formed XML document. The resulting file is imported into Documentum. Finally, the manual post-processing begins and the edited speeches are posted onto the Althingi website.

Currently, the ASR is housed on a virtual server and interacts with the rest of the Althingi servers through the RESTful API and DFC. The ASR server is built on Centos 7 with 8 CPUs and 8 GB of RAM. The transcription process is allocated up to 7 CPUs and 6 GB of RAM with the number of parallel transcriptions limited by the ratio of CPUs to Gigabytes of RAM available. The current ratio is 1 CPU per 2 GB of RAM. This means that at most 3 speeches can be transcribed in parallel. During parliamentary meetings, speeches tend to be processed in parallel when earlier speeches are at least two times as long as succeeding speeches.

3. Data

The details about the preparation of the ASR training data can be found in [1]. The total speech corpora prepared in that fashion is roughly 6300 hours. The training set used for our current speech recognizer is a subset of the whole set, roughly 1000 hours of parliamentary speeches and corresponding text, which are cleaned further using Kaldi's clean and segment data function[6]. Speed perturbations (sp) are used to generate new acoustic data at 1.1 and 0.9 times the original speed of the original recordings, resulting in extra 2000 hours of data. The development and test sets are based on parliamentary speeches from 2016 and are roughly 11 hours each or 94k words. The language models are trained on the total parliamentary text set, 59M tokens. The data set is split on end-of-sentence markers.

The punctuation model training text set contains roughly 58M words. The development and test set contain 114k and 111k words, respectively.

The paragraph model training and test sets are parliamentary texts, cleaned from symbols not present in denormalized ASR transcripts. The training set contains roughly 300k paragraphs and 52M words. The development and test sets contain around 16.5k paragraphs and 2.9M words each.

4. In-lab results

We found that data quality mattered more than the amount of data for the quality of our acoustic model. We have trained a model using the same architecture with both a 6200 hour data set and a 1500 hour subset and the difference in WER is minuscule. However when we cleaned a subset of the training set

[5] http://www.malfong.is

[6] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/clean_and_segment_data.sh

Table 1: *The word-error-rate and real-time-factor of different acoustic models, trained using different training set sizes and cleanliness of data. The different acoustic model architectures and language models are described in Sec. 2.1*

Acoustic model	Training set	WER [%]	RTF
LSTM-TDNN (LF-MMI) w/sp	1500 hrs	9.07	1.03
f-TDNN w/sp	1500 hrs	9.17	0.18
f-TDNN	6200 hrs	9.05	0.21
f-TDNN w/sp	1000 hrs, re-cleaned	8.52	0.15
f-TDNN w/sp w/RNN-LM	1000 hrs, re-cleaned	8.17	0.15*
f-TDNN w/sp w/bi-RNN-LM	1000 hrs, re-cleaned	**7.91**	0.15**

* RNN-LM re-scoring takes ≈20 times longer than n-gram re-scoring

** Bi-directional RNN-LM re-scoring takes ≈40 times longer than n-gram re-scoring

further, resulting in about 1000 hour training set, we got significant improvement, as shown in Table 1 and Figures 1a and 1b.

Re-scoring with an RNN-LM instead of a 5-gram gives better WER. On our test set, applying a forward or bi-directional RNN-LM improves the WER by 4.1% and 7.2% relative, respectively, compared to 5-gram re-scoring. However, 5-gram re-scoring happens almost instantly. On the test set re-scoring with a forward RNN-LM took 21 times longer and with a bi-directional RNN-LM 40 times longer. However, the re-scoring is only one part of the whole recognition process. For a small set of 30 speeches, of varied length, we recorded the total recognition time, i.e. the time it took to transcribe and automatically post-process the speeches, either re-scoring with a 5-gram or a forward RNN-LM. The rnnlm transcription process took on average 1.25 times longer. The average memory usage was 522 MB when using n-grams and 1.2 GB when using a RNN-LM. However, we could not see the improvement in WER as in our test set. Hence, and because of the higher memory consumption of RNN-LMs, we decided to continue using n-grams for the Altingi ASR system, resulting in a production system with 8.52% WER.

When analysing the kinds of errors the system makes[7] [14], it shows that a substantial portion of the errors is unlikely to affect the meaning of the output. Icelandic is a highly inflected language showing numerous morphological forms for its inflected word classes. Further, it has a very productive compound mechanism, where words are strung together to build new words. These characteristics of the language are often the source for errors in our ASR system.

Table 2 shows that approx. 39% of the substitution errors represent errors where a reference word is substituted with another word form from the same paradigm, i.e. both the reference word and the substitution are morphological forms of the same lemma.

Table 2: *Report on whether the substitutions in the ASR output on our test set belong to the same inflection paradigm as the reference word or not. The percentages missing to reach 100% are either non-inflective words or words missing from the inflection database[15].*

#Substitutions	Same lemma	Different lemma
3916	38.76%	32.71%

Similarly, compound word errors considerably increase WER. However, these errors normally do not affect meaning. Out of 714 occurring compound words in the test set, 347 had been split into two, in 88 utterances they were the only errors. In those cases the meaning of the sentence is still easily understood and the error can be corrected by deleting a single space. However, both a substitution and an insertion error are noted for each split compound word which increases the WER significantly.

The WER of the current ASR, before any post-processing is done, is 8.52% on the test set after 5-gram re-scoring. In real life, when text post processing has been applied, this number is going to be higher, since imperfect punctuation reconstruction will add to the errors in the output.

In-lab punctuation model results on a test set of parliamentary speeches from 2016 can be seen in Table 3. The rules for the use of commas in Icelandic are not very clear, making learning their position difficult, which drags down the overall F-score of the punctuation model.

Table 3: *Punctuation insertion results on the parliamentary punctuation test set.*

Punctuation	Precision	Recall	F-score
Comma	76.9	53.5	63.1
Period	91.0	88.4	89.7
Question mark	89.8	81.5	85.5
Colon	83.3	80.5	81.9
Overall	86.7	75.3	80.6
Err: 2.37%			
SER: 32.0%			

The results in Table 3 are obtained on well structured text. Error rates are higher in the automatically transcribed speeches, since the ASR output reflects spoken language, which has looser sentence structure, contains hesitations and speech errors.

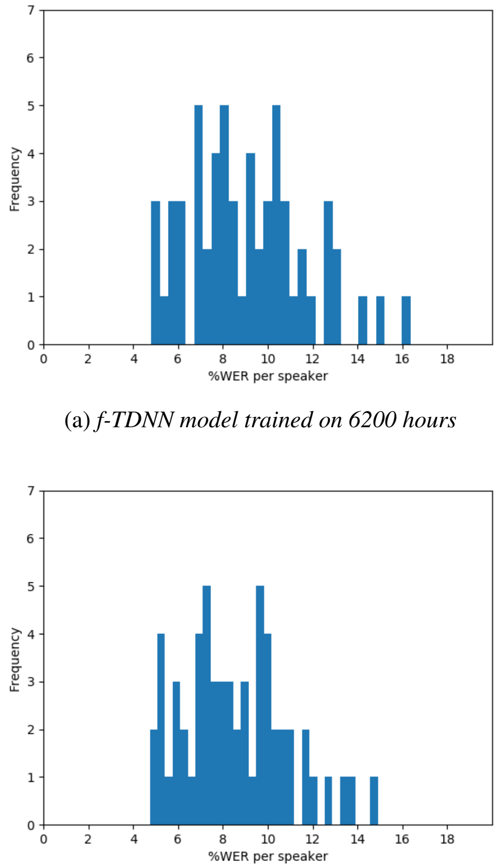
Current results for the paragraph model are in Table 4. The accuracy is not high but the results are nevertheless tolerable, since the main objective is to improve the readability of the text and incorrect paragraph splits are not difficult to fix.

Table 4: *Precision, recall and F-score of paragraph insertion on a parliamentary test set using the paragraph model.*

Precision	Recall	F-score
64.1	59.4	61.6

Due to the morphological richness of the language, the denormalized version of a text can be more correct than the original ASR output. This primarily affects numbers and abbreviations, where e.g. the denormalization of wrong word forms causes the errors to vanish: *ákvæði níunda grein laga* ('the

[7] https://github.com/cadia-lvl/ice-asr



(a) *f-TDNN model trained on 6200 hours*

(b) *Our current ASR: f-TDNN model trained on re-cleaned 1000 hours*

Figure 1: *The average word error rate per speaker for (a) a f-TDNN model trained on 6200 hours, and (b) our current f-TDNN model trained on re-cleaned 1000 hours. Both re-scored by a 5-gram.*

provisions of the ninth article of law'), where the word forms *níunda* and *grein* are incorrect for *níundu* and *greinar*, becomes *ákvæði 9. gr. laga*, i.e. the erroneous word forms are correctly transformed to a numeral and an abbreviation. We have, however, not measured the effect of this on overall WER.

5. Feedback from transcribers and editors

The editors at Althingi have noted that the speech transcripts are now available sooner than before, since the speech recognizer starts transcribing as soon as it sees a new audio file, and the ASR system does not take coffee breaks or go home when office hours end.

In the spring of 2018 we tested the editors' happiness with the ASR we had at the time. That recognizer had a WER of 9.63% on our test set, before any post-processing is done. The raw ASR outputs were compared to transcripts of the corresponding speeches, created by human transcribers. However, the human transcribers fixed more than we would have liked so the WER is higher than what represents true errors. The average WER of the ASR system was 20.1 ± 5.8% with punctuation marks and 14.4 ± 4.8% without them. Out of 234 graded speeches 26 were graded as Bad, 105 as Medium, and 103 as Good. No guidelines were given regarding these grades, they are only based on the evaluators' intuitions.

In those experiments we also received comments regarding what the editors would like changed in the text post-processing. We reacted to their comments and in December 2018, we had a similar test, using the recognizer we described in this paper as our current one. This time we got feedback on 625 speeches. The WERs are calculated between the ASR outputs and the corresponding outputs after a transcriber has corrected the text. We see that in some instances the transcribers did more than just correct the text which increases the WER number, but the closeness is still higher than in the spring test. The average WER of the ASR system was 15.0 ± 6.0% with punctuation marks and 11.3 ± 5.2% without them. 616 of the speeches were graded, 475 as Good, 111 as Medium and 30 as Bad. The grades are qualitative, based on the transcribers' intuitions. However, the average editing time per word for speeches graded as Good was 0.76 sec/word while for speeches with Medium or Bad grades it was 0.94 and 1.09 sec/word, respectively. By comparing the grades between the two tests we can deduce that the users' happiness with the ASR has increased. The percentage of speeches graded as Good has increased from 44% to 77%.

6. Conclusion

We have shown that using a clean 1000 hour subset of our speech corpus gives better ASR transcripts than training on a 6200 hour training set that only goes through basic cleaning steps. We have also discussed the linguistic challenges of creating an automatic speech recognizer for Icelandic and shown with both in-lab and in-the-wild results that it is well possible. Our system is already in use and speeding up the transcription process at the Icelandic parliament. We have established that punctuation insertion models work for Icelandic, even though the accuracy for commas might be better. We have even shown that a good text post-processing can sometimes hide some of the linguistic errors prone to Icelandic ASR. Paragraph splits can also be learned and inserted into the denormalized output text. Further tests are needed to see whether a pause annotated data set created from the re-cleaned acoustic training set can improve the punctuation and paragraph models.

The responses from the staff at Althingi show that they are happy with the transcripts they receive from the ASR and it is considered useful, even for the speakers that are most often badly transcribed. Further work includes looking into whether it is beneficial to use linguistic information to correct the morphological form of words and whether extending the LM with a cache model of the pertinent bills improves recognition.

7. Acknowledgements

This project was made possible through the support of Althingi's information and publications departments. The authors would like to thank Solveig K. Jónsdóttir and Þorbjörg Árnadóttir for their valuable help, Ingvi Stígsson for handling the technical aspects of the test at Althingi, as well as the editors and transcribers, for reading over the ASR transcriptions, correcting them, timing the process and giving valuable comments.

8. References

[1] I. R. Helgadóttir, R. Kjaran, A. B. Nikulásdóttir, and J. Gudnason, "Building an ASR corpus using Althingi's parliamentary speeches," in *Proc. Interspeech 2017*, 2017, pp. 2163–2167. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-903

[2] J. Y. Fong, M. Borsky, I. R. Helgadóttir, and J. Gudnason, "Manual post-editing of automatically transcribed speeches from the icelandic parliament - althingi," 2018.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[4] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018), Hyderabad, India*, 2018.

[5] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[6] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech 2016*, 2016, pp. 2751–2755.

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

[8] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proc. of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011, pp. 187–197.

[9] E. Rögnvaldsson, "The Icelandic speech recognition project Hjal," *Nordisk Sprogteknologi. Årbog*, pp. 239–242, 2003.

[10] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[11] T. Tai, W. Skut, and R. Sproat, "Thrax: An open source grammar compiler built on OpenFst," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[12] B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, and T. Tai, "The OpenGrm open-source finite-state grammar software libraries," in *Proc. of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 61–66.

[13] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration." in *Proc. Interspeech 2016*, 2016, pp. 3047–3051.

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

[15] K. Bjarnadóttir, "The database of modern icelandic inflection (beygingarlỳsing íslensks nútímamáls)," *Language Technology for Normalisation of Less-Resourced Languages*, p. 13, 2012.