Building an ASR corpus using Althingi's Parliamentary Speeches

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, Jón Guðnason
Center for Analysis and Design of Intelligent Agents, Language and Voice Lab (https://lvl.ru.is)
Reykjavik University, Iceland

REYKJAVIK UNIVERSITY

Overview

We developed a speech corpus for Icelandic parliamentary speeches. It is the largest speech corpus publicly available for Icelandic, and has been used to train a preliminary transcription system for the Icelandic parliament.

Althingi speeches

- The data obtained from the Icelandic parliament consists of 6600 recordings of parliament speeches, dating from 2005 to 2016.
- The average speech length is 6 minutes, but the range is around 1-30 minutes.
- Speakers are 197: 105 male and 92 female.

Expansion error rates for different language model's n-gram orders

	%Incorrect		
	2-gram	3-gram	5-gram
Althingi abbreviations	10.6	10.1	9.1
General abbr. + numbers	10.6	10.1	10.2

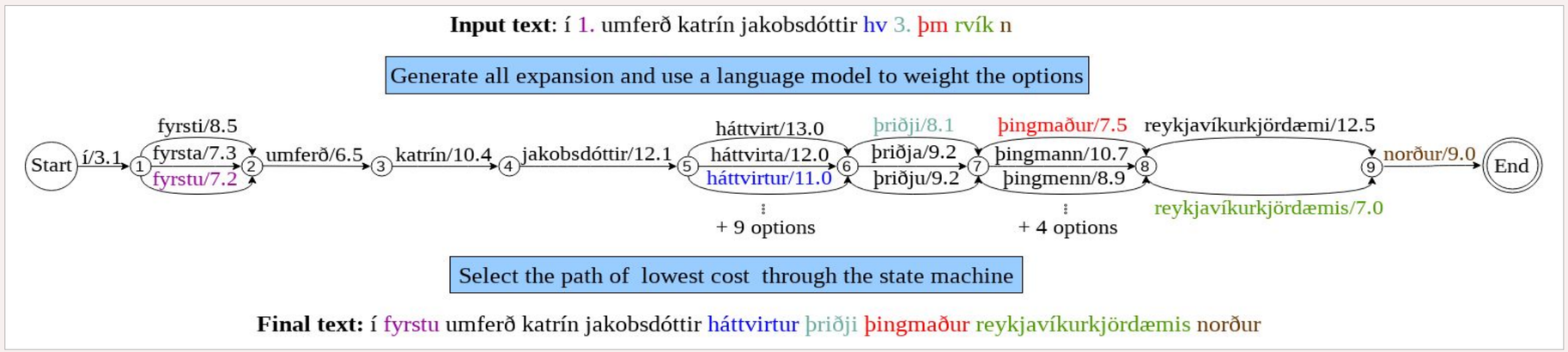
Text processing and normalization

- The transcriptions don't reflect the speech recordings accurately enough to be useful for ASR training, due to numbers and abbreviations.
- Basic text normalization performed.
- Numbers and abbreviations expanded, a non-trivial problem in Icelandic.
- The OpenGrm Thrax Grammar Development Tools[1] were used to generate the expansions.
- A language model used to select the most likely one.

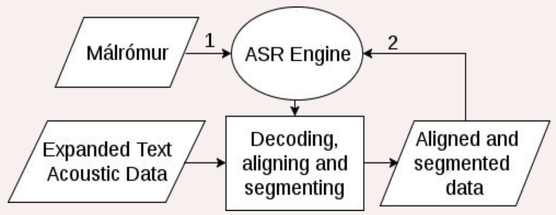
Segmentation and alignment

- The segmentation and alignment of the database was done in two stages.
- An experimental version of an open LV ASR system for Icelandic was used to align 100 hrs of audio and text.
- The aligned data was segmented into segments of maximum 15 seconds of audio.
- A new speech recognizer was trained on 60 hours of the newly aligned parliamentary speeches.
- This intermediate ASR system was used to align and segment the rest of the data.

Text expansion example



The two stage segmentation and alignment



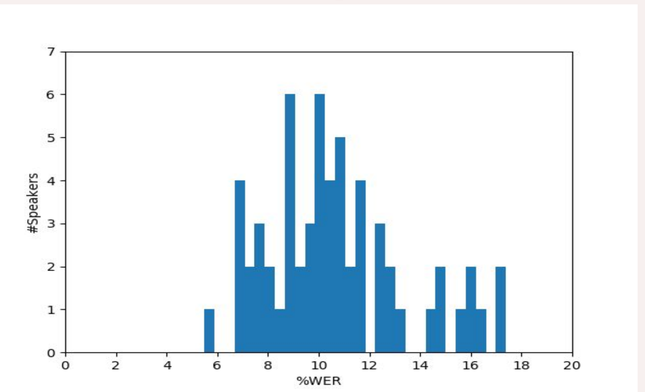
Database and corpus structure

Total duration	**542 hrs and 25 min**
Avg. segment dur.	9.8 sec
# word tokens	4,583,751
Training set size	514.5 hrs, 192 speakers
Dev/eval set sizes	14 hrs, 59 speakers ('16 data)

WER for different acoustic models

Acoustic model	Dev. set [%]	Eval. set [%]
GMM-SAT	22.61	22.24
DNN	17.48	17.28
TD-DNN	16.71	16.38
TD-DNN w/sp	16.44	16.20
LSTM-RNN	15.17	14.76
LSTM-TDNN (LF-MMI) w/sp	10.30	**10.17**

%WER distribution over speakers



Post-processing of the ASR output

- The ASR returns a stream of words with no punctuation.
- Thrax[1] used to denormalize numbers and abbreviations.
- Punctuator 2[3] used to insert periods, commas, question marks and colons in the text.
- Regular expressions used to capitalize the beginning of sentences, collapse spelled out acronyms and rewrite.

Data and sources

- The data is freely available on http://www.malfong.is
  - Aligned and segmented speech corpus
  - The original data
  - Pronunciation dictionary
  - Three language models
- The whole recipe is available on https://github.com/ingarun/kaldi/tree/master/egs/althingi
  - Data normalization, alignment and segmentation.
  - Training of an ASR
  - Postprocessing of the ASR output
  - Applying the ASR system and postprocessing

Kaldi

The Kaldi speech recognition toolkit[3] is used to train the system.

References

[1] http://thrax.opengrm.org/
[2] https://github.com/ottokart/punctuator2
[3] http://kaldi-asr.org