Málrómur: A Manually Verified Corpus of Recorded Icelandic Speech

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
steinst@hi.is

Jón Guðnason
Reykjavik University
jg@ru.is

Sigrún Helgadóttir
The Árni Magnússon Institute for Icelandic Studies
sigruhel@hi.is

Eiríkur Rögnvaldsson
Department of Icelandic
University of Iceland
eirikur@hi.is

Abstract

This paper describes the Málrómur corpus, an open, manually verified, Icelandic speech corpus. The recordings were collected in 2011–2012 by Reykjavik University and the Icelandic Center for Language Technology in cooperation with Google. 152 hours of speech were recorded from 563 participants. The recordings were subsequently manually inspected by evaluators listening to all the segments, determining whether any given segment contains the utterance the participant was supposed to read, and nothing else. Out of 127,286 recorded segments 108,568 were approved and 18,718 deemed unsatisfactory.

1 Introduction

A common way to gather speech corpora for automatic speech recognition is to build a large list of sentences that are then read and recorded by a large number of people. Ideally, the sentence list provides a good coverage of the language structure and the number of people is large enough to capture the acoustic and phonetic variation present in the spoken language. Each recording in the corpus is accompanied by its text transcription and possibly additional metadata, such as the speaker's gender, age, or recording conditions. The recordings are then commonly used in a supervised learning algorithm to produce an acoustic model for automatic speech recognition systems.

The quality of the trained model depends both on the quality of the recordings, and on the correctness of their transcriptions. Errors in speech corpora will lead to degraded acoustic models. Verifying the correctness of the data in a speech corpus increases its quality.

Almannarómur, a free Icelandic speech corpus, was created in 2011–2012 (Guðnason et al., 2012). The data were recorded in cooperation with Google, on Android G1 phones using Datahound (Hughes et al., 2010). The main aim of the project was to create a database of spoken sentences to aid development of automatic speech recognition for Icelandic. However, the database can be used for many other types of spoken language technologies.

We created and executed a procedure to verify the recordings in Almannarómur. It was evident that a proportion of the Almannarómur recordings was flawed. The most prominent errors occur when the participants read the prompts only in part, read them incorrectly or say something completely different. Recordings are also sometimes incomplete, starting too late or stopping too early. The purpose of our verification process was to create a subset of a raw speech corpus that is as close to being 100% correct as possible. This was done by manually checking and verifying all the recordings. Manual verification of speech recordings can be a tedious and time consuming task so we designed a simple workflow, which could be used both for verification by an individual and by a group, and requires very little instruction.

All the recordings, along with the results of the verification process and other relevant metadata, are published with a CC BY 4.0 license[1] on a website for Icelandic Language Technology (LT) resources[2], *Málföng* (Helgadóttir and Rögnvaldsson, 2013), under the name *Málrómur*.

2 Evaluating the Speech Corpus

The process starts by pre-processing the recordings. The recordings then enter a two stage verification process and finally the accuracy of the verification is evaluated.

[1] http://www.malfong.is
[2] https://creativecommons.org/

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 237–240,
Gothenburg, Sweden, 23-24 May 2017.

2.1 Pre-Processing the Data

Before the verification process starts, we created new sound files by automatically trimming long periods of silence at the beginning and end of the recordings. By this we achieve two objectives. 1) The manual verification process takes less time. The total duration of the untrimmed files is 151 hours, 55 minutes and 26 seconds. The total duration of the trimmed files is 90 hours, 16 minutes and 4 seconds. The duration of the trimmed files is thus only 59% of the original files' duration. 2) Recordings that are expected to contain no speech can be identified as being completely truncated in this process and removed from the working database for the verification.

To reduce bandwidth and loading times during data verification, the audio is transcoded into smaller files using a lossy codec.

2.2 Data Verification

In the verification process, the recordings are played back to an evaluator who then classifies them according to given criteria. For a recording to be verified as correct it has to be read correctly and clearly and have no additional speech before or after it. Vocal segregates (e.g. uh, um, ah, eh) are an exception, they are allowed when they occur before the prompt is read, if there is a clear pause in between. Background noise is also allowed if it is clearly lower than the read segment.

To make the process as simple as possible for the evaluators, a web-based system was implemented for these tasks, using PyBossa[3], a crowdsourcing environment. Due to the decentralized nature of the setup, evaluators are not bound to a physical workplace and, furthermore, collaboration of many hired evaluators and/or volunteers is easily achievable. No training is required, the guidelines for the evaluators can be explained in less than five minutes.

We set the evaluation up as a two stage process, designed so that we could build up a database of verified recordings as fast as possible. In the first stage the trimmed recordings, as described in Section 2.1, are used. Most of the recordings are expected to be correct so we only give the evaluators two choices. If a recording meets the criteria described above it is to be accepted as correct. If not it should be rejected.

[3] `http://pybossa.com`

In the second stage, recordings that were rejected in stage one are categorized into five categories: 1) Unclear – unclear, inaudible and silent recordings. 2) Incomplete – recording begins or ends during the reading of the prompt. 3) Additional Speech – read correctly, but there is additional speech before or after. 4) Incorrectly Read – but clear and sensible. 5) Correct – truncated file is flawed or the recording was incorrectly marked as flawed in stage one.

As the data pre-processing can generate errors of type 2 – Incomplete, the non-trimmed audio is used for playback in the second stage.

The time spent on the verification was logged. The logs give insight into the workload of the process and make it possible to estimate the duration of future verifications.

The process allows for each recording to be checked multiple times by different evaluators, which would likely reduce verification errors. The process also makes it possible to crowdsource the evaluation process. Using the workload calculations from this project the cost of doing more than one pass of evaluation and the feasibility of trying to crowdsource the work can be estimated.

Four evaluators worked on verifying the speech data in stage one and two. Each recording was only checked once, by one evaluator. Four other evaluators then estimated the accuracy of the evaluation process by listening to a subset of 3000 recordings and classifying them. All the accuracy evaluators listened to all the 3000 files, so each of the 3000 files was checked four times. The accuracy evaluators listened to the original untrimmed recordings and the results were compared to the classification in stage one.

3 Results

Total files recorded were 127,286. Failed recordings were 5,401, thus 121,885 were pre-processed. Out of these, 2,795 were identified as silent by the truncating process. Therefore, 119,090 recorded segments were to be verified.

In stage one, four evaluators listened to the recordings. 100,020 recordings were accepted as correct, and 19,070 were rejected and sent to stage two (see Table 1). Total duration af the segments labelled as correct was 136 hours.

There are three types of utterances in the corpus. Single word utterances, multiword utterances and internet domain names. There were consid-

erably fewer errors for single word utterances and domain names than for multiword utterances, as illustrated in Table 1.

Utterance type	Correct	Total	Correct (%)
Single Word	30,670	35,262	(86.98%)
Multiword	58,053	71,092	(81.66%)
Web Domain	11,297	12,736	(88.70%)
Total	100,020	119,090	(83.99%)

Table 1: Stage one results for different utterance types.

The results obtained by each of the evaluators range from 18.6% to 19.3% error rate. Taking into account that the evaluators did not listen to the same ratio of each of the three utterance categories (see Table 1), this range should not be surprising.

In stage two, two evaluators listened to the original untrimmed 19,070 recordings. The evaluators classified each of the recordings into one of five classes, as described in Section 2.2. In this round 8,548 recordings, or 45% of the recordings previously classified as incorrect, were classified as correct, giving a total of 10,522 incorrect recordings, classified by four types of error:

Class	Count	(%)
Unclear	1,381	(7.24%)
Incomplete	5,526	(28.98%)
Additional Speech	592	(3.10%)
Incorrectly Read	3,023	(15.85%)
Correct	8,548	(44.82%)
Total	19,070	(100.00%)

Table 2: Stage two results.

Average time spent on each recorded segment in stage one was 4.2 sec. In stage two, the two evaluators spent 7.1 sec on average verifying each segment. By multiplying that duration with the total number of recorded segments entering stage one, we can approximate the time saved by verifying the data in two stages to be in the vicinity of 58 hours, compared to using the method in stage two for all the recordings.

In order to evaluate the correctness of the verification process four new evaluators listened to 3000 recordings as described in section 2.2. Their results were compared to that of the verification process in stage one. Out of the 3000 recordings 1509 had previously been classified as correct and 1491 had been classified as incorrect. Accuracy evaluation is shown in table 3.

Evaluation	Stage One Correct	Stage One Incorrect
Correct	1,499	726
Incorrect	10	765
Agreement	99.34%	51.31%

Table 3: Evaluating accuracy of stage one verification.

The recordings classified as correct in the verification process were classified in the same way in 99.34% of the cases in the correctness evaluation. Recordings classified as incorrect were classified the same way 51.31% of the time by the correctness evaluators. The ratio of segments previously marked as incorrect, but which the correctness evaluators mark as correct is not far from the ratio in stage two of the verification process, as evident by comparing tables 2 and 3. This is expected as the same kind of data was being evaluated, in both cases the original, untrimmed recordings. The trimmed versions of the same recordings were rejected in stage one. This may indicate that in the pre-processing trimming phase the threshold for cutting silent segments was set too low. Further tweaking of the parameters might have resulted in better results.

4 Availability and Use

The final, verified corpus is published on the Icelandic LT website *Málföng* under a permissive license (CC BY 4.0) to promote research and development using this Icelandic language resource. The recordings are made available for download as recorded, in 16 kHz WAV-format, accompanied by all relevant metadata: duration of recording in file, environment conditions, gender of speaker, age of speaker, prompt text and class determined by the verification process described in this paper and listed in table 2.

5 Conclusion and Further Work

We have determined that out of the 121,885 speech segments that were successfully recorded in the Almannarómur project, 108,568 files, or 89%, were verified to be correct.

The accuracy evaluation shows that over 99% of recordings classified as correct in stage one were

verified to be correct. Having a corpus that has such a low ratio of incorrect data will be of great benefit for users of speech corpora.

About 51% of the recordings classified as incorrect were verified as incorrect by the accuracy evaluators (see Table 3). This is in line with the results of stage two of the verification process, where about 55% of the recordings previously classified as incorrect in stage one were verified as such (see Table 2). The reason for this is that in stage one trimmed recordings were used for classification but in the accuracy evaluation and in stage two untrimmed recordings were used. It was important to use the untrimmed recordings for evaluating accuracy of stage one verification to see the accuracy of the data classification rather than just the accuracy of the four stage one evaluators.

We have shown that rather than verifying speech data in one stage with no pre-processing, manual verification of a speech corpus can be done faster by using a two stage verification process after pre-processing. The pre-processing includes trimming the files used for the first stage of verification and removing recordings identified as silent.

Gathering information about different types of errors is important, as analysis of errors in the incorrect data may allow to identify patterns the incorrect recordings exhibit. This can give feedback to adjust the prompt selection or recording setup to improve the correctness of further recordings.

One error class in stage two, type 3 errors – Additional Speech, could be processed further in a third stage. These rejected recordings include correct utterances but they are preceded and/or followed by unwanted speech. The recording could be cropped and the good part of the segment added to the correct recordings. This has not been done as a part of this project, but it would be worthwhile to estimate how much time is needed to crop the recordings in a third stage.

The *Málrómur* corpus is the largest of its kind for Icelandic and is already being used for training an Icelandic speech recognizer. It will also be used to develop tools helping corpus creators to automatically evaluate the correctness of new Icelandic speech data.

References

Jón Guðnason, Oddur Kjartansson, Jökull Jóhannsson, Elín Carstensdóttir, Hannes Högni Vilhjálmsson, Hrafn Loftsson, Sigrún Helgadóttir, Kristín M. Jóhannsdóttir, and Eiríkur Rögnvaldsson. 2012. Almannarómur: An Open Icelandic Speech Corpus. In *Proceedings of SLTU '12, 3rd Workshop on Spoken Languages Technologies for Under-Resourced Languages*, Cape Town, South Africa.

Sigrún Helgadóttir and Eiríkur Rögnvaldsson. 2013. Language Resources for Icelandic. In K. De Smedt, L. Borin, K. Lindén, B. Maegaard, E. Rögnvaldsson, and K. Vider, editors, *Proceedings of the Workshop on Nordic Language Research Infrastructure at NODALIDA 2013*, pages 60–76. NEALT Proceedings Series 20. Linköping Electronic Conference Proceedings, Linköping, Sweden.

Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno, and Mike LeBeau. 2010. Building Transcribed Speech Corpora Quickly and Cheaply for Many Languages. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1914–1917, Makuhari, Chiba, Japan.