Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 316–320
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Manual Speech Synthesis Data Acquisition - From Script Design to Recording Speech

Atli Þór Sigurgeirsson, Gunnar Thor Örnólfsson, Jón Guðnason
Reykjavik University, The Árni Magnússon Institute of Icelandic Studies, Reykjavik University
Menntavegur 1 - Reykjavik Iceland, Laugavegur 13 - Reykjavik Iceland, Menntavegur 1 - Reykjavik Iceland
atlithors@ru.is, gunnarthor@hi.is, jg@ru.is

Abstract

In this paper we present the work of collecting a large amount of high quality speech synthesis data for Icelandic. 8 speakers will be recorded for 20 hours each. A script design strategy is proposed and three scripts have been generated to maximize diphone coverage, varying in length. The largest reading script contains 14,400 prompts and includes 87.3% of all Icelandic diphones at least once and 81% of all Icelandic diphones at least twenty times. A recording client was developed to facilitate recording sessions. The client supports easily importing scripts and maintaining multiple collections in parallel. The recorded data can be downloaded straight from the client. Recording sessions are carried out in a professional studio under supervision and started October of 2019. As of writing, 58.7 hours of high quality speech data has been collected. The scripts, the recording software and the speech data will later be released under a CC-BY 4.0 license.

Keywords: Corpus, Acquisition, Tools

1. Introduction

High quality speech data is imperative to the development of good speech synthesis systems. This fact is often a hurdle for under-resourced languages, like Icelandic, since the acquisition of quality speech data is both labor-intensive and a costly process. Meanwhile, as the language technology (LT) community has grown, the development cost of Text-to-speech (TTS) and Automatic Speech recognition (ASR) systems has decreased.

The availability of Icelandic ASR data has increased tremendously in recent years. The Althingi corpus (Helgadóttir et al., 2017) contains over 500 hours of transcribed parliament speeches, the Málrómur corpus (Steingrímsson et al., 2017) includes 152 hours of recorded speech from 563 participants and the Almannarómur corpus which was collected from 563 participants that provided 219 read sentences on average each (Guðnason et al., 2012). All of these datasets are recorded by multiple speakers under different recording environments which is a benefit when training ASR models while it is a hindrance when training natural-sounding TTS models. Speech synthesis datasets for Icelandic remain sparse.

1.1. Related Work

Typically, speech synthesis datasets are recorded professionally by a single speaker in a controlled environment under supervision. The amount of data required for TTS depends on the chosen model. The CMU arctic corpus consists of 1150 phonetically balanced sentences and was designed for unit selection TTS (Kominek et al., 2003). The Merlin toolkit was used to train a statistical parametric speech synthesis (SPSS) model based on neural networks with 2400 training utterances (Wu et al., 2016). Deep voice (Arik et al., 2017), an end-to-end TTS based entirely on neural networks was trained on approximately 20 hours of speech and reached a mean opinion score (MOS) of 3.94 ± 0.26.

The first development of an Icelandic TTS started around the turn of the century. At least three TTS systems for Icelandic exist today and are in use. Most recently, in 2010, the Icelandic association of the visually impaired hired the Polish software company Ivona to develop a unit selection TTS system. These three systems have all had mixed successes. An important downside to these developments is the fact that all three voices were carried out by foreign firms and no open and available TTS datasets for Icelandic exist today (Nikulásdóttir et al., 2020).

A voice recording client is necessary to facilitate the recording sessions. Common Voice [1] is a well known recording client for crowd sourcing ASR data. Google has used a tool referred to as *Datahound* for collecting and building transcribed speech corpora for many languages (Hughes et al., 2010). A speech data acquisition system made in Iceland referred to as *Eyra* was developed in 2016 (Petursson et al., 2016). Eyra was developed as a crowd sourcing tool and was later used to collect about 35 hours ASR data (Guðnason et al., 2017).

1.2. Overview

This paper presents an overview of the speech data acquisition process for a new Icelandic TTS system. The system is being developed as a part of the Icelandic language technology programme (Nikulásdóttir et al., 2020). The programme spans 4 years and many different projects in LT. This paper presents two of the main goals of the TTS project:

- To generate 3 scripts that maximize a diphone coverage. They should be designed for 1 hour, 10 hour and 20 hour collections.
- To record unit selection TTS data from 4 female speakers and 4 male speakers. 20 hours should be collected from each speaker.

[1] https://voice.mozilla.org/

The 3 scripts should be suitable for TTS recipe development on a varying scale, from unit-selection models to end-to-end neural speech synthesis models. This work started in autumn 2019 and as of writing, the scripts have been finalized. The twenty hour script contains 14400 unique sentences. The list contains at least one occurrence of 87.3% of all possible diphones and 81% of them appear at least 20 times. Speech recording is an ongoing process and we have collected approximately 59 hours of data as of date. Once all speakers have been recorded, the dataset will be published under a CC-BY 4.0 license.

2. Script Design

Before designing the script, 500,000 sentences were extracted from Risamálheild (Steingrímsson et al., 2018), a large Icelandic text corpus containing more than one billion word tokens. All of these sentences had to pass a naive preprocessing step. To pass, the sentence must:

- be at least 10 letters
- be between 5 and 15 words
- only contain characters from the Icelandic alphabet or any of the Icelandic punctuation symbols
- start with a capital letter
- end with a punctuation symbol
- appear in the database of modern Icelandic inflection (Bjarnadóttir, 2012)

Since Icelandic is a highly inflected language, simply checking if all words in a sentence appear in a dictionary would greatly limit the number of sentences that would pass this preprocessing step. Checking if all words appear in the inflection list does not guarantee grammatical correctness however. The length constraints were enforced to minimize prosodic difference between recordings, which can be an issue for very short sentences (Kominek et al., 2003), and to limit the number of mispronunciations while recording the data.

The phones in a randomly sampled list of sentences will follow an uneven distribution where a small number of phones will appear very frequently. Such a list will therefore likely not contain more than one occurrence of a substantial amount of the possible phonetic combinations in the language. This poses a problem for gathering speech synthesis data since it is critical to train a TTS on most phonetic combinations more than once to generate natural-sounding results. TTS scripts are therefore most often designed to maximize some phonetic coverage. A lot of different metrics have been used for measuring such a coverage. It varies both in terms of the phonetic unit used, e.g. diphones (Kominek et al., 2003) or triphones (Ursin, 2002), and also in terms of the context each unit appears in, where in the sentence the unit appears or where in a word it appears and so on. We decided to maximize diphone coverage while limiting sentence length.

A Sequitur grapheme-to-phoneme (G2P) model (Bisani and Ney, 2008) was trained on the Icelandic Pronunciation Dictionary (IPD) (Nikulásdóttir et al., 2018) To acquire predicted phonetization of the source text. This is needed to analyze the phonetic content of the source text. Icelandic is spoken with six rather similar dialects and IPD contains variants for four of those dialects. A standard dialect in the IPD is used to phonetically transcribe the training data in this work. The training set consists of approximately 40,000 verified word and phonetization pairs. The complete list of Icelandic phones in SAMPA format is given below

```
A,ay,ay:,au,au:,A:,c,c0,ey,ey:,f,h,i,
i:,j,k,k0,l,l0,m,m0,n,n0,ou,ou:,p,p0,r,
r0,s,t,t0,U,U:,v,x,C,D,N,N0,9,9y,9y:,
9:,O,oy,O:,E,E:,G,I,I:,J,J0,Y,yy,Y:,T
```

The trained model achieves a phone error rate (PER) of 3.4%. Using this G2P model, the phonetization for each source sentence was predicted. A special symbol was additionally prepended and appended to each phonetization to denote the start and end of sentences. Using this, a list of diphones was generated for each sentence.

A greedy algorithm was used to order the list by a reward function, R, which was constructed to reward sentences that both improve the phonetic coverage and are short. The final script is initialized as the empty set. Given the large list of sentences, all the sentences are scored by R at every time step and sorted. The sentence with the highest reward is inserted into the final script. The reward function is given by

$$R(s) = \frac{1}{|s|} \sum_{i=1}^{n} \frac{1}{\max(1, [d_i \in \mathbf{D}])}$$

$$s = s_1, \ldots, s_m \quad d(s) = d_1, \ldots, d_n$$

Here, s is the sentence, $d(s)$ is the grapheme-to-diphone mapping of s and $\mathbf{D}$ is the set of all diphones in the script already. We define a complete coverage to include at least 20 occurrences of each possible diphone. After that, a diphone does not count towards the reward. The algorithm runs until complete coverage is achieved or 25,000 sentences have been added to the script.

The coverage at every insertion step is shown In Figure 1. The blue curve shows the actual coverage, that is the coverage with regards to all diphones. It is important to point out that not all diphones are valid diphones in Icelandic and never appear. The red curve shows the coverage with regards to the diphones that appear in the source that the algorithm runs on.

After about 6000 insertions the algorithm reaches the maximum possible coverage. The resulting script contains at least one occurrence of 87.3% of all possible diphones and 81% of them appear at least 20 times.

Figure 2 shows phone-to-phone heat maps of the list generated by the proposed method and the same number of randomly sampled sentences. The heat map for the proposed script demonstrates much greater coverage than that of the randomly sampled list. This underlines the issue of randomly sampling sentences.

After sorting the list by the reward, a number of sentences were added to the script in different categories:

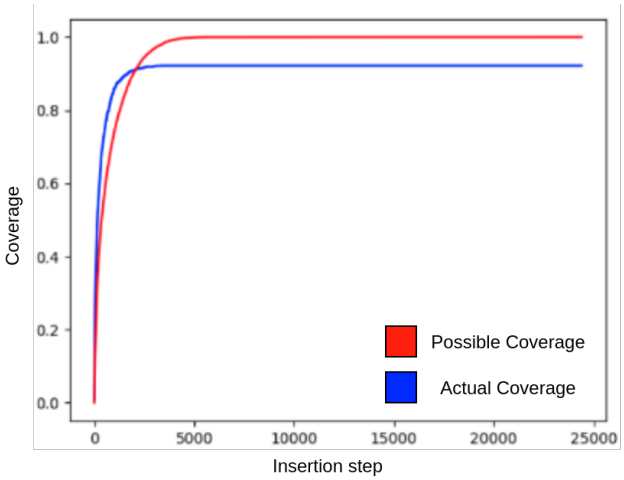
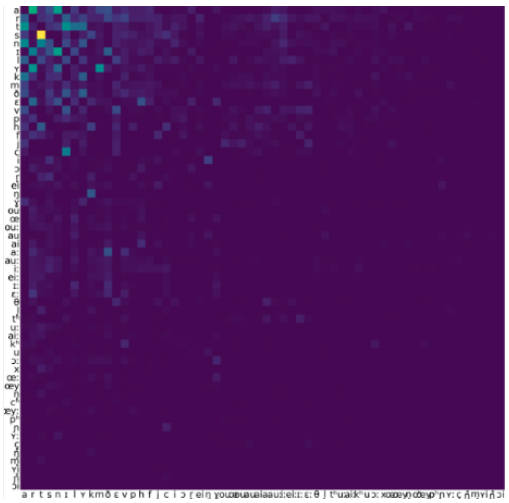
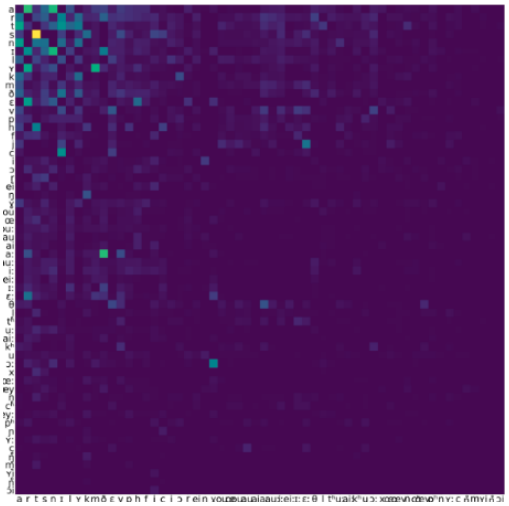


Figure 1: Phonetical coverage at every insertion step of the greedy algorithm.



(a) Randomly sampled script



(b) The proposed script

Figure 2: Heat maps that visualize the diphone distribution of a randomly sampled script and the proposed script. Both axis include all possible phonemes in the same order.

- 2000 sentences between 15-25 words that could be used for learning acoustic alignment for longer sentences.
- 100 sentences containing only digits in written form
- 30 sentences that contain only one word each

To generate the 20 hour script we take the first k sentences from the final list ordered by the reward such that $k = 20 \times 3600\text{s}/5\text{s} = 14,400$ where we estimate that it takes on average 5 seconds to read a single sentence. The 1 hour and 10 hour scripts are generated in a similar manner.

3. Recording Client

A recording client was developed to facilitate voice recording sessions and we call it *Lobe*. Previously, a speech data acquisition tool called *Eyra* (e. Ear) (Petursson et al., 2016) was developed at Reykjavik University. Eyra was used successfully for gathering ASR data (Guðnason et al., 2017). The ASR focused nature of Eyra did not fit the TTS data acquisition task which prompted the development of Lobe. Lobe is at the core a Python package with a Flask [2] web client. It is hosted on a Reykjavik University server and accessible in the browser. The data is stored in a PostgreSQL database with a weekly backup schedule.
Lobe assigns roles to users, either an administration role or a basic user role. A basic user could be

- A speaker whose voice will be recorded
- A person who controls the prompts while recording speech (*prompt manager*)
- Anyone else that wants to access the data

An administrator starts by creating a collection through Lobe. After selecting a collection name and perhaps assigning a speaker to the collection, the prompts are uploaded through Lobe. Lobe accepts multiple file uploads where each line in a file is treated as a single prompt. As Lobe was designed with the script design in mind it also accepts prompts that include the following in a tab-separated format:

- The prompt itself.
- The source of the prompt (e.g. a certain newspaper).
- An order score. If a score is higher it appears earlier in the prompts.
- The phonetization of the prompt.

In this way, we can start by recording phonetically rich sentences as determined by the reward function.
Next, the administrator creates a new user for the speaker through Lobe. Lobe stores user information such as age, sex and dialect. After that, recording sessions can be carried out. Each recording session contains 50 prompts. The prompt manager presses a key to initialize audio capture and a visual sign prompts the speaker to start speaking. After the speaker reads the prompt the prompt manager

[2] https://flask.palletsprojects.com/

presses another key to stop audio capture. At that point the prompt manager can go to the next prompt or download the current audio capture. Since the scripts are not guaranteed to be grammatically correct, the prompt manager also has the option to skip the current prompt and the sentence will be marked as faulty in the database. That sentence will not appear further as a prompt. Lobe has a simple quality con-

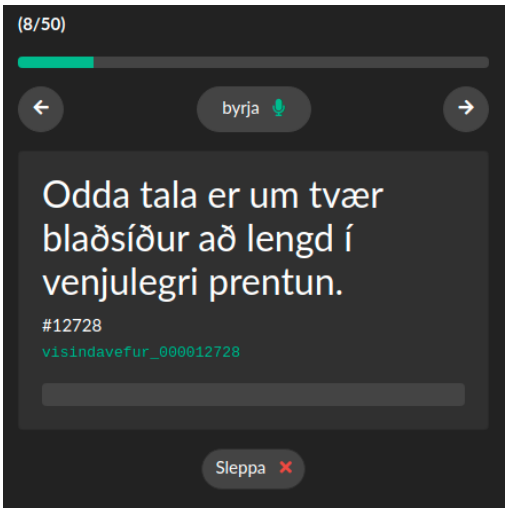


Figure 3: The prompt screen shown to both the speaker and the prompt manager

trol check that runs after each recording. It will prompt the manager if the recording is either too quiet or too loud. For further inspection, the recording can be downloaded and analyzed in any available audio software. We use the MediaRecorder [3] interface to record the audio. It is sampled at 41KHz with a 24 bit depth.

After a session is finished, the prompt manager can start a new recording session or log out of Lobe. At any time, the collection can be downloaded as a separate dataset through Lobe. The client creates an archive that includes all prompts and recordings as well as information about each recording session, the speaker and the collection itself.

Since we are using Merlin (Wu et al., 2016) for generating SPSS voice recipes, we made sure that the lobe dataset exports could be easily imported into Merlin.

4. Recording Speech

Eight speakers will be recorded, 20 hours each. As of date we have started recording four of those. One of our goals is to attain diversity in age, dialect and overall speaking style. Four female speakers will be recorded and four male speakers. The first four speakers are in the age range of 49-71 years as shown in Table 1. They also all speak in the same standard dialect. It is therefore important to select the next four speakers with this fact and the goal of attaining diversity in mind. A voice sample of five sentences is recorded and analyzed before a speaker is added to the dataset. We evaluate the speech rate, volume and the overall pleasantness of the voice. Once the speaker has

[3] https://developer.mozilla.org/en-US/docs/Web/API/MediaRecorder

Speaker ID	Age	Sex	Amount recorded
M1	70	Male	20 hours
F1	59	Female	17.3 hours
M2	49	Male	9.8 hours
F2	71	Female	11.6 hours

Table 1: The recording progress for the first 4 speakers

been approved, the speaker is assigned a recording schedule with a prompt manager. A voice recording test is carried out during the first session. This is done to determine the external sound card level that ensures that the recording stays between -18dB and -12dB in playback with 0dB as the distortion threshold. The sound card level is recorded for future reference.

Recording sessions are carried out in a studio at the national broadcaster of Iceland. The studio is separated into a recording space and a monitoring space. The recording space is sound proof and designed to limit resonance. Both prompt managers and the speakers monitor the distance from the pop filter attached to the microphone at the start of each recording session as the distance could affect the recorded results. The speakers are also told not to bring anything else into the recording space and limit movement. The prompt manager sits in the monitor space and communicates with the speaker using a talkback system in the studio. Before starting, the prompt manager starts a voice

Figure 4: The recording environment shown from inside the recording space.

recording test. A single sentence is recorded and then analyzed. If the monitor value is not within the (-18dB;-12dB) range, the prompt manager changes the sound card level accordingly and records a new sound card level. At this point a session can start.

Each session is configured to go through 50 prompts. The speaker never records for more than two hours each day to reduce the risk of vocal strain. Typically eight to twelve such sessions can be completed in a two hour span. The session duration varies between speakers but is normally between seven to 13 minutes with an average duration of about 9 minutes.

5. Conclusions and Future Work

We aim to finish recording 20 hours from the eight speakers each by the 1st of October 2020. The dataset will thereafter be made available. Parallel to this, work on unit selection TTS and SPSS model recipes will be carried out and trained for the speakers that have reached the 20 hour goal.

Work to improve Lobe is ongoing. Most importantly is the work on expanding the built-in quality control. We additionally aim to make Lobe more configurable with regards to the number of prompts in a session, sample rate, bit depth and so on. More features will also soon be added to Lobe to facilitate different types of data collections. Firstly, support for multi-speaker collections will be added. This is necessary as part of the Icelandic language technology programme is to collect 2 hours from 40 speakers each for voice mixing synthesis projects. Secondly, support for video capture will be added to facilitate audio-visual speech recognition data acquisition.

6. Bibliographical References

Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., et al. (2017). Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 195–204. JMLR. org.

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 (beygingarlỳsing íslensks nútímamáls). *Language Technology for Normalisation of Less-Resourced Languages*, page 13.

Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsson, H. H., Loftsson, H., Helgadóttir, S., Jóhannsdóttir, K. M., and Rögnvaldsson, E. (2012). Almannaromur: An open icelandic speech corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

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

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.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. J., and LeBeau, M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. In *Eleventh Annual Conference of the International Speech Communication Association*.

Kominek, J., Black, A. W., and Ver, V. (2003). CMU ARCTIC databases for speech synthesis.

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

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

Petursson, M., Klüpfel, S., and Gudnason, J. (2016). Eyra-speech data acquisition system for many languages. *Procedia Computer Science*, 81:53–60.

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 Guðnason, J. (2018). Risamálheild: A very large icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ursin, M. (2002). Triphone clustering in finnish continuous speech recognition. *Diplomityö, Teknillinen korkeakoulu*.

Wu, Z., Watts, O., and King, S. (2016). Merlin: An open source neural network speech synthesis system. In *SSW*, pages 202–207.