Talrómur: A large Icelandic TTS corpus

Atli Thor Sigurgeirsson
University of Edinburgh

Þorsteinn Daði Gunnarsson, Gunnar Thor Örnólfsson, Eydís Huld Magnúsdótir
Ragnheiður Kr. Þórhallsdóttir, Stefán Gunnlaugur Jónsson, Jón Guðnason
Menntavegur 1
Reykjavík University

Abstract

We present Talrómur[1], a large high-quality Text-To-Speech (TTS) corpus for the Icelandic language. This multi-speaker corpus contains recordings from 4 male speakers and 4 female speakers of a wide range in age and speaking style. The corpus consists of 122,417 single utterance recordings equating to approximately 213 hours of voice data. All speakers read from the same script which has a high coverage of possible Icelandic diphones. Manual analysis of 15,956 utterances indicates that the corpus has a reading mistake rate no higher than 0.25%. We additionally present results from subjective evaluations of the different voices with regards to intelligibility, likeability and trustworthiness.

1 Introduction

All statistical TTS models require some training data to learn the mapping from text to speech. Unit selection TTS models are capable of producing an intelligible voice using less than 2 hours of aligned speech (Conkie, 1999). HMM-based TTS models can produce somewhat natural-sounding speech using less than 500 utterances (Yoshimura et al., 1999). The more recent neural end-to-end models have reached a considerably higher mean opinion score (MOS) in regard to naturalness. However, they require a much larger training corpus; most require tens of thousands of utterances to converge and reach natural sounding synthesis (Wang et al., 2017) (Ren et al., 2019). The widely used LJ Speech corpus consists of 13,100 recordings amounting to approximately 24 hours (Ito and Johnson, 2017).

[1] "Tal" means speech and "rómur" means voice.

To produce high quality synthesised speech with minimal noise, the corpora used for training TTS models are most often captured in a studio under supervision. New approaches have lowered the language-specific expertise needed for high quality TTS but at the cost of requiring larger amounts of training data (Sotelo et al., 2017) (Arik et al., 2017) (Wang et al., 2017) (Ren et al., 2019). The large amount of data needed and the quality of that data limits the ability of many low resource language communities to benefit from these recent advancements in the TTS domain.

The Icelandic language program (ILP) is a 5 year government funded program to make the Icelandic language viable in the digital world (Nikulásdóttir et al., 2020). TTS development for Icelandic is a significant part of the ILP ranging from unit selection voices to multi-speaker TTS models. A prerequisite for all TTS projects of the ILP is a large high quality TTS corpus which up to this point has not been available for open use (Nikulásdóttir et al., 2020).

Previous work in spoken language technology for Icelandic has been more focused on speech recognition, both in terms of data acquisition and acoustic modelling (Helgadóttir et al., 2017) (Guðnason et al., 2012) (Steingrímsson et al., 2017) (Mollberg et al., 2020). Since most of that data is found or crowd-sourced data from multiple speakers it is not ideal for speech synthesis where low background noise and high recording quality is important. An Icelandic pronunciation dictionary for TTS exists as well as a limited text normalisation system (Nikulásdóttir et al., 2018) (Nikulásdóttir and Guðnason, 2019). To address the lack of high quality Icelandic TTS data, Talrómur has been created.

2 The Talrómur Corpus

One of the aims of the Talrómur project is to attain diversity in age, speaking style, dialect and

ID	Name	Gender	Age	# Utterances	Duration	# Characters	# Words	# Unique Words
A	Rósa	F	59	9,899	16h32m12s	556,767	93,002	19,272
B	Bjartur	M	70	12,048	25h43m05s	713,578	118,564	22,617
C	Diljá	F	71	13,443	27h57m33s	843,530	139,636	25,492
D	Búi	M	49	12,357	22h32m58s	766,037	126,814	23,857
E	Ugla	F	26	20,050	31h28m04s	1,298,318	215,176	33,629
F	Álfur	M	35	19,849	29h07m18s	1,284,508	212,979	33,401
G	Salka	F	33	16,886	30h09m38s	1,078,978	178,818	29,966
H	Steinn	M	39	17,637	29h49m01s	1,134,244	187,868	30,977

Table 1: Overview of corpus, outlining key statistics and information for each speaker. The "Name" column contains pseudonyms for the speakers in the corpus

prosody. Voice samples from speaker applicants were analysed and evaluated with this and a subjective evaluation of pleasantness in mind. Each participating speaker got a recording schedule, typically two hours each working day until completion.

Dialect diversity is low in Iceland and six main but rather similar regional variants are listed in the Icelandic pronunciation dictionary (Nikulásdóttir et al., 2018). Speakers A-F all speak in the most frequent standard dialect while speakers G and H speak in the second most frequent regional variant. Speakers A, B and C differ a bit from the rest of the group and their qualities deserve a specific mention.

Speaker A was the first speaker we recorded. At that time the development of the recording client was ongoing and we had limited experience with the studio and equipment. As shown in table 1 that speaker has significantly fewer hours recorded.

Speaker B is a 70 year old man with limited eyesight. This speaker often had issues with reading the prompts fluently. This results in unnatural pauses in the middle of sentences that correspond with where the line is split on the screen. We have looked into using silence detection to remove these silences and current results suggest that this task is easily automated. We release the data in the raw format however, without any trimming.

Speaker C is a female voice actor with a deep, breathy voice. This speaker's recordings are more similar to audio-book recordings in that they have a more animated speaking style when compared to the other speakers.

Technical Details

Each speaker reads single sentence prompts from the same reading list. The reading list was designed to have a high coverage of diphones in the Icelandic language (Sigurgeirsson et al., 2020). The prompts were sourced from Risamálheild, a large Icelandic text corpus consisting of text from many different types of sources (Steingrímsson et al., 2018).

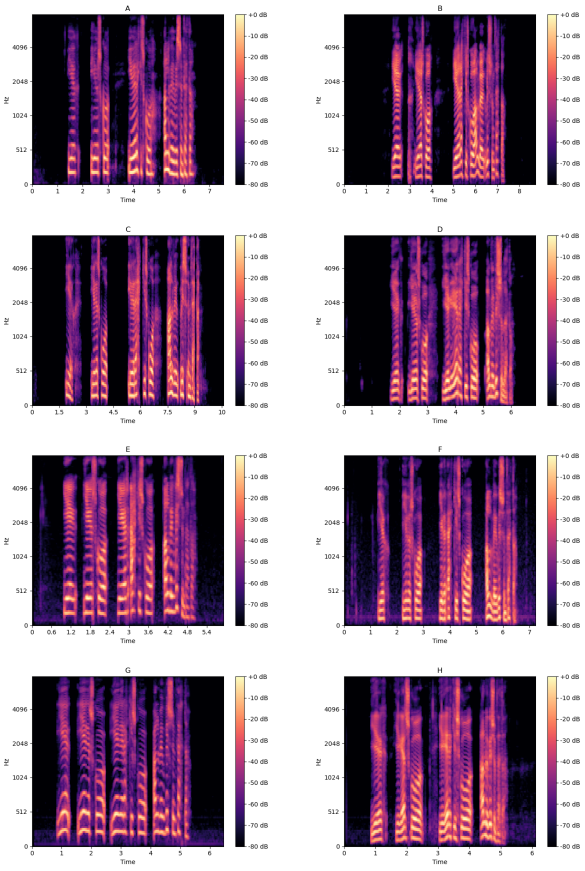

Figure 1: Mel-frequency spectrograms of all speakers saying the same phrase: "*Ég, ég er sko, ég er ekki sko, alveg viss um þetta*".

Recording sessions were carried out in a stu-

dio at the national broadcaster of Iceland. After recording the first 2 speakers, the project was moved to a different studio at the national broadcaster due to restrictions caused by the COVID-19 pandemic. The last two speakers reside in northern Iceland and they were therefore recorded in a third studio. The recordings were captured between November 2019 and September 2020.

Since the speakers read prompts from the same reading list nearly all sentences in the corpus are spoken by multiple speakers. This makes the corpus ideal for multi-speaker TTS development, prosody transfer, voice conversion and other research domains where the speaker identity and linguistic content have to be disentangled by the TTS model (Skerry-Ryan et al., 2018) (Wang et al., 2018).

The same recording hardware was used for all recordings. The hardware consisted of an AKG ULS series condenser microphone equipped with a CK-61 cardioid capsule, an SPL channel one pre-amplifier and a Clarett 2Pre sound card. The recordings are captured using a recording client specifically made for this project (Sigurgeirsson et al., 2020).

We store some information about every recording captured, such as how the text appeared on the monitor to the speaker, the session ID and technical information about the recordings. Most recordings are sampled at 48kHz with a 16 bit depth. Some recordings of speakers A and B are sampled at 44.1kHz. All recordings are single channel.

3 Recording Analysis

Type of error	Occurrence	Rate
Volume too low	8	0.05%
Volume too high	70	0.44%
Audio flaw	347	2.17%
Prompt mismatch	196	1.23%
Actual mismatch	39	0.25%

Table 2: The results of 15,956 recording analyses. The evaluators judge long silences as prompt mismatches resulting in 196 prompt mismatch evaluations. Subtracting those results in a much lower number or 39.

We have analysed a portion of the recordings for quality. Of the approximately 122,417 recordings 15,956 recordings have been analysed. Using a proprietary tool, human evaluators are asked to first listen to a single recording and then indicate whether the recording matches the prompt and whether the recording quality is good. We specifically ask the evaluators to indicate whether the volume is either too high, resulting in pops or distortions, or too low making the recording hard to comprehend or whether any other audio flaws are audible in the recording.

Of the recordings analysed 613 were marked as bad or about 3.8%. Only 1.23% of the recordings were indicated to have a mismatch between the prompt and the recording. Upon further inspection it seems that the evaluators marked recordings with untimely silences as prompt mismatches. Most of those are spoken by speaker B as explained in section 2. After a second pass over the evaluations we are confident that a better estimate of prompt mismatches is no more than 0.25%.

The rate of audio flaws is 2.17% but reviewing the samples in question revealed that a significant portion of these recordings do not have any unwanted artefacts. Upon inspection we believe some of these recordings have a higher than normal volume, making them sound unpleasant when compared to other recordings. This is particularly common for speaker B. The volume of recordings can be too high if the speaker has moved too close to the microphone, the hardware has not been configured correctly or the speaker speaks with more effort than is natural to the speaker. There are however some recordings that do have unwanted artefacts. In most cases this consists of a small pop at a random location in the recording. These pops mostly appear in recordings from speakers A and B and we therefore deduce that the source of this artefact is the hardware configuration in the recording studio for those speakers.

4 Subjective Listening Experiment

To gain further information about which voice would be most suitable for general TTS use, we set up a subjective listening experiment with 50 participants. During the listening experiment, the participants listen to a single recording at a time. They are then asked one of three questions[2]:

Q1: How easy is it to understand this voice?
Q2: How pleasant is this voice?
Q3: How trustworthy is this voice?

[2]In Icelandic: Q1: *Hversu auðskiljanleg þykir þér þessi rödd?* Q2: *Hversu viðkunnanleg þykir þér þessi rödd?* Q3: *Hversu traustverðug þykir þér þessi rödd?*

ID	SR		F0			Duration		
	words / sec	chars / sec	Min	Mean ± SD	Max	Min	Mean ± SD	Max
A	2.34	13.70	147.77	198.82 ± 22.56	246.68	1.30	6.01 ± 1.55	14.38
B	1.73	10.19	76.58	150.71 ± 24.57	215.14	2.22	7.68 ± 1.91	18.68
C	1.89	11.20	107.62	173.61 ± 24.52	331.10	2.71	7.48 ± 1.77	17.76
D	2.24	13.28	79.69	143.69 ± 28.02	210.22	0.91	6.57 ± 1.53	15.97
E	2.94	17.39	128.37	210.74 ± 27.00	294.88	1.86	5.65 ± 1.50	14.46
F	3.26	19.33	102.03	128.10 ± 12.89	165.02	1.78	5.28 ± 1.36	12.96
G	2.39	14.13	154.71	237.08 ± 20.60	271.69	2.26	6.43 ± 1.64	14.82
H	2.60	15.42	98.45	142.69 ± 23.36	213.84	1.44	6.09 ± 1.56	14.57

Table 3: Estimation of speaking rate (SR) and average F0. Pitch was estimated by averaging pitch over voiced segments in the phrase used in figure 1. *ProsodyPro* was used for pitch tracking (Xu, 2016).

The participants then rate the recording on a scale from 1 to 5, e.g. from very untrustworthy to very trustworthy. Before starting the evaluation participants are made aware that the sentences being spoken should not affect their judgement and that they should focus on the voice itself.

Each participant listens to 3 recordings from each speaker for each of the three questions, resulting in 24 evaluations per question and 72 evaluations in total per participant. We used a balanced Latin square experimental design with 24 different recordings tested for each evaluation question (MacKenzie, 2002). This resulted in 1074 Q1 responses, 1074 Q2 responses and 1088 Q3 responses. The number of responses per utterance ranges from 4 to 8.

Results from this experiment are shown in table 4. These scores are relative between the 8 speakers since listeners only listen to recordings from the Talrómur corpus. Due to the fact that the listening test wasn't anchored, the interpretation of the rating scale varied noticeably between listeners. The results we present here are normalised per listener, and the raw scores are higher, particularly for Q1. Voice G is rated as the most intelligible, voice H as the most likable and most trustworthy, although they didn't score significantly higher than the second highest for each question.

ID	Q1	Q2	Q3
A	2.78 ± 0.36	2.84 ± 0.33	2.80 ± 0.32
B	1.82 ± 0.36	1.66 ± 0.30	1.50 ± 0.30
C	2.96 ± 0.37	1.95 ± 0.38	2.14 ± 0.35
D	3.57 ± 0.33	3.02 ± 0.31	3.55 ± 0.29
E	4.13 ± 0.28	2.87 ± 0.32	3.72 ± 0.28
F	3.54 ± 0.31	3.10 ± 0.34	2.87 ± 0.34
G	$\mathbf{4.27} \pm 0.22$	2.91 ± 0.33	3.32 ± 0.30
H	3.97 ± 0.28	$\mathbf{3.15} \pm 0.27$	$\mathbf{3.73} \pm 0.28$

Table 4: Normalised mean opinion score with standard deviation for each speaker and each question. Q1 tested for intelligibility, Q2 for likeability and Q3 for trustworthiness.

5 Summary and Future Work

In this paper we introduce the Talrómur corpus which is the result of the first TTS data acquisition phase of the Icelandic language program. Talrómur is a large, high quality speech corpus designed specifically for TTS. The corpus consists of 8 different voices with a wide range in prosodic effect, speaking style and age. The quality and amount of data in Talrómur matches or exceeds that used in many state-of-the-art end-to-end neural TTS models for the English language. A subjective evaluation indicates which voice users are likely to prefer but we believe most of the voices are good candidates for general TTS use. As with other deliverables belonging to the ILP, the data will be published under open licenses to encourage wide use and adoption of the data. The data has been made available through the CLARIN project[3].

6 Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur [4], is funded by the Icelandic Ministry of Education, Science and Culture.

[3]https://repository.clarin.is/repository/xmlui/handle/20.500.12537/104
[4]https://almannaromur.is/

References

Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. 2017. Deep voice: Real-time neural text-to-speech.

Alistair Conkie. 1999. Robust unit selection system for speech synthesis. In *137th meeting of the Acoustical Society of America*, page 978.

Jón Guðnason, Oddur Kjartansson, Jökull Jóhannsson, Elín Carstensdóttir, Hannes Högni Vilhjálmsson, Hrafn Loftsson, Sigrún Helgadóttir, Kristín M Jóhannsdóttir, and Eiríkur Rögnvaldsson. 2012. Almannaromur: An open Icelandic speech corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2017. Building an asr corpus using althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

I Scott MacKenzie. 2002. Within-subjects vs. between-subjects designs: Which to use? *Human-Computer Interaction: An Empirical Research Perspective*, 7:2005.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jon Gudnason. 2020. Samrómur: Crowd-sourcing data collection for Icelandic speech recognition. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3463–3467.

Anna Björk Nikulásdóttir and Jón Guðnason. 2019. Bootstrapping a Text Normalization System for an Inflected Language. Numbers as a Test Case. In *INTERSPEECH*, pages 4455–4459.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for Icelandic 2019-2023. page 3414–3422.

Anna Björk Nikulásdóttir, Jón Guðnason, and Eiríkur Rögnvaldsson. 2018. An Icelandic pronunciation dictionary for tts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, pages 3171–3180.

Atli Sigurgeirsson, Gunnar Örnólfsson, and Jón Guðnason. 2020. Manual speech synthesis data acquisition-from script design to recording speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 316–320.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. In *In ICLR2017 workshop submission*.

Steinþór Steingrímsson, Jón Guðnason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2017. Málrómur: A manually verified corpus of recorded Icelandic speech. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 237–240.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. pages 4006–4010.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR.

Y Xu. 2016. Prosodypro. a praat script for large-scale systematic analysis of continuous prosodic events. In *In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*.