



Samrómur Milljón: An ASR Corpus of One Million Verified Read Prompts in Icelandic

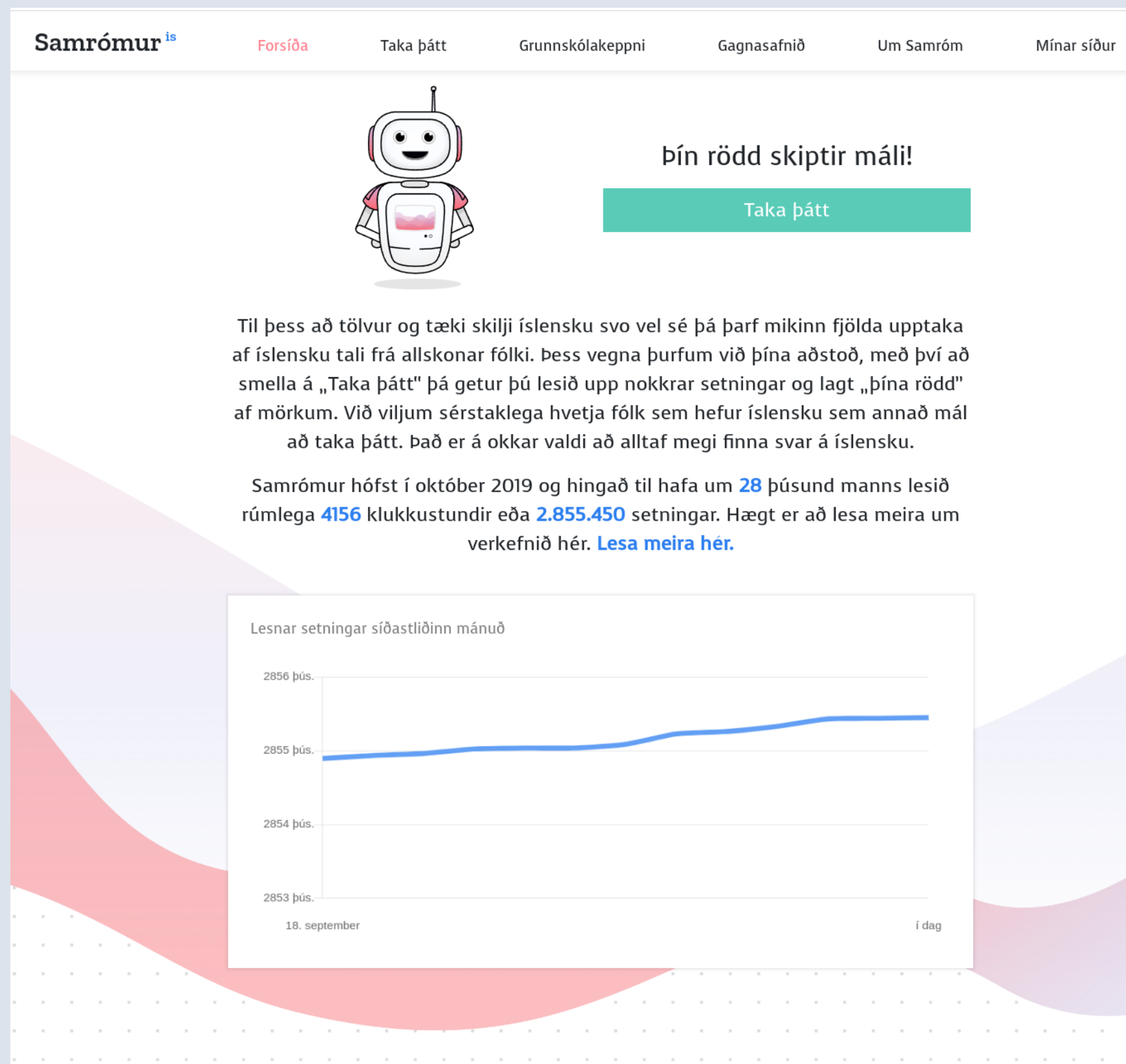
Carlos Mena, Þorsteinn Daði Gunnarsson, Jón Guðnason

Language and Voice Lab, Reykjavík University

{carlosm, thorsteinng, jg}@ru.is



The Platform samromur.is



Samromur Unverified 22.07

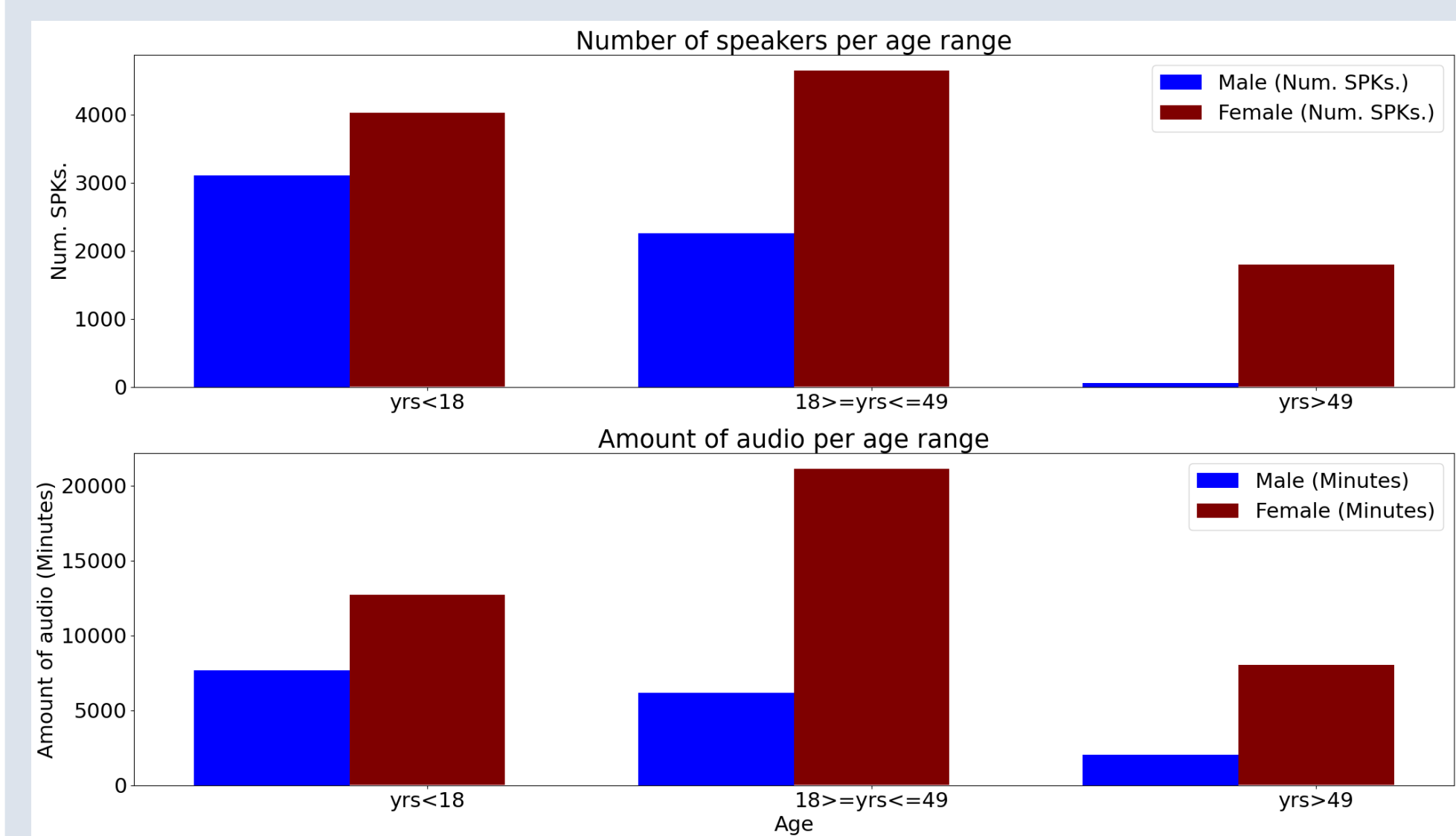
Samromur Unverified 22.07 contains the recordings collected by the platform samromur.is up to July, 2022.

A total of 2,159,314 (2,233 hours) speech recordings in Icelandic that are essentially unverified!



Samrómur Milljón

Gender	Female	Male	Unknown
Duration	697h22m	264h28m	5h16m
Utterances	714,564	282,499	5,094
Speakers	10,447	5,948	209



Data for the Verification

The total duration of the sum of these four datasets is 1000 hours (875h of published data + 125h of unpublished material).

The verification process was performed using one acoustic model at a time running in a GPU Tesla A100.

Corpus	Size
Althingi Corpus	514h
Malrómur Corpus	119h
Samrómur Corpus	114h
Samrómur Children Corpus	127h
Unpublished Material	125h

Verification with NeMo

The verification model used to perform the verification with NeMo was:

carlosdanielhernandezmena/stt_is_quartznet15x5_ft_ep56_875h

- The verification model was trained with 875 hours
- NeMo verified 2 million recordings in 7 hours.
- No language model used.
- 348,295 of perfect matches.

Verification with Wav2Vec2

The verification model used to perform the verification with Wav2Vec2 was:

carlosdanielhernandezmena/wav2vec2-large-xlsr-53-icelandic-ep10-1000h

- The verification model was trained with 1,000 hours
- Wav2Vec2 verified 2 million recordings in 7 days.
- No language model used.
- 1,002,218 of perfect matches.

Verification with Whisper

The verification model used to perform the verification with Whisper was:

carlosdanielhernandezmena/whisper-large-icelandic-30k-steps-1000h

- The verification model was trained with 1,000 hours
- Whisper verified 2 million recordings in 36 days (138,992 transcription in 2.5 days).
- **We were using a corrupted model!** This issue motivated the creation of the Faster-Whisper Model
- 18,839 of perfect matches.

Verif. with Faster-Whisper

The verification model used to perform the verification with Faster-Whisper was:

carlosdanielhernandezmena/whisper-large-icelandic-30k-steps-1000h-ct2

- This model was created from the Whisper model shown in the previous slide.
- 863,220 of perfect matches.

This model was created with 2 lines of code!

Verification Results

V=Wav2Vec2, N=NeMo, W=Whisper, F=Whisper-Fast

Sys.	Matches	Percent.
V+N+F	325,713	32.50%
V+N+W	4,449	0.44%
V+F	537,453	53.62%
V+N	18,072	1.80%
V+W	14,390	1.43%
V	102,080	10.18%

More than 80% of the data was verified by at least 2 systems!

Wav2Vec2 Model

V.M.=Verification Model; S.M.=Samrómur Milljón Model.

Dataset	V. M.	S. M.
Samrómur (Test)	9.84%	7.69%
Samrómur (Dev)	8.73%	6.78%
Samrómur Children (Test)	9.39%	6.46%
Samrómur Children (Dev)	6.05%	4.23%
Malrómur (Test)	5.64%	6.63%
Malrómur (Dev)	6.15%	5.83%
Althingi (Test)	11.43%	17.90%
Althingi (Dev)	11.09%	17.93%

Whisper Model

V.M.=Verification Model; S.M.=Samrómur Milljón Model.

Dataset	V. M.	S. M.
Samrómur (Test)	8.47%	7.76%
Samrómur (Dev)	7.29%	7.03%
Samrómur Children (Test)	7.74%	7.04%
Samrómur Children (Dev)	4.59%	4.42%
Malrómur (Test)	5.11%	11.51%
Malrómur (Dev)	5.28%	11.00%
Althingi (Test)	8.25%	16.18%
Althingi (Dev)	7.99%	16.00%

Conclusions & Further Work

- Corpus Samrómur Milljón.
- Automatic Verification of 1,002,157 speech recordings (967 hours).
- Acoustic Models in Wav2Vec2, Whisper and Faster-Whisper trained with Samrómur Milljón.
- Further Work: Verify the remaining 1 million recordings of "Samrómur Unverified 22.07".