Suurten puhemallien ja puheentunnistuksen kehitystä Pohjoismaisille kielille itseohjautuvan opetuksen avulla

– Recent progress in ASR and speech models for Nordic languages by utilizing self-supervised training

Mikko Kurimo **Aalto University**
Dep. Information and Communications Engineering (DICE)

Research questions

1. How much training **data** is needed?
2. How much manually transcribed speech is needed?
3. How much untranscribed speech can help?
4. How much do big pre-trained transformers help?
5. How to measure ASR performance?
6. When ASR fails?

Our case: Developing ASR for spoken **Finnish, Finland-Swedish and Sami languages**

Why to record and process speech data?

To analyse:

- Spoken communication
  - human-machine, human-human
- Spoken conversations
  - interaction, meetings, interviews
- Spoken language
- Spoken information
  - content, topic, intent,
- Speaker information
  - voice, health, proficiency

To develop better tools for:

- Speech recognition
- Speaker diarisation
- Text-to-speech synthesis
- Speech translation
- Spoken information retrieval
- Speaker recognition

Why to record lots of speech data?

- Speech varies a lot by speakers (age, gender, origin, education, dialect)
- Variation based on speaking situation, style, recording
- Spontaneous and colloquial speech differ from text
- Spoken language is captured from speech
- In most languages there are no suitable and large spoken data resources

Aalto ASR research group

Personnel:

- Professor (Mikko Kurimo)
- Research fellow (**Tamas Grosz**) + 2 post docs (Mittul Singh, Guangpu Huang)
- 7 PhD students (**Yaroslav Getman** trained the models presented here!)
- 4 MSc students

Project funding:

- Academy of Finland, Business Finland, Finnish and Nordic foundations

Current collaborations:

- Other groups at Aalto and other universities and companies in Finland
- NTNU, Karolinska Institutet, Univ. Oslo, Univ. Lapland, Uppsala, Tromsö, Greenland

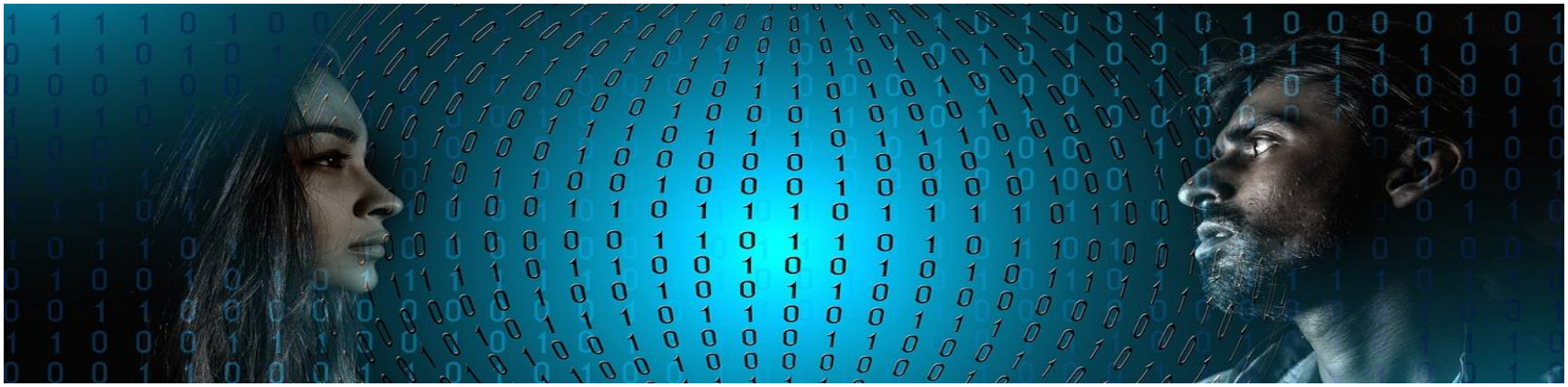
Aalto ASR research group

- Studies deep learning methods in **automatic speech recognition** (ASR) and **language modeling** (LM)
- Challenge: Representation and understanding of real-world spoken conversations

Great variation of speakers, styles and languages

Deep learning methods in ASR and LM

Need for understandable and co-operative AI



Speech processing resources in Language Bank

- Various Finnish speech corpora
- Training and evaluation scripts, e.g. ASR
- Pre-trained speech models, language models, ASR models
- ASR tools, alignment tools

Other big public speech data for FIN/SWE/SAMI

- Yle archives, KAVI, Parliament
- Kotus, Talko (SWE)
- VoxPopuli, Common Voice, Fleurs

EDUSKUNTA
RIKSDAGEN

Parliament sessions 2008 - 2020

Large source of **manually transcribed** spoken **Finnish (and Swedish)**: 4000 hrs and 449 speakers, **videos** and rich demographic **metadata**

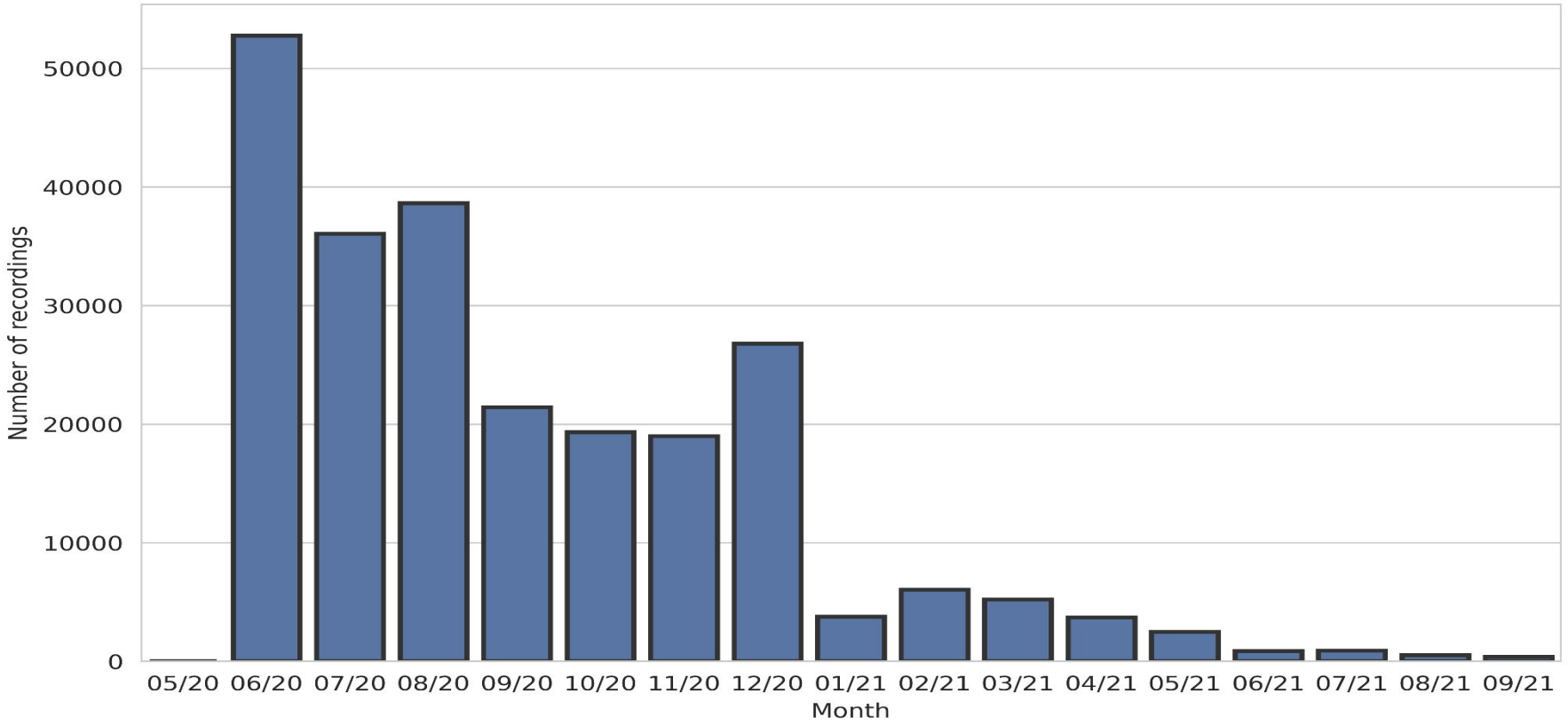
https://www.kielipankki.fi/corpora/fi-parliament-asr/

- automatically segmented and aligned ASR training and test sets
- pipeline for retrieving and processing new recordings and transcripts
- training scripts and models *https://github.com/aalto-speech/fin-parl-models*
- **results:** *Virkkunen, Rouhe, Phan, Kurimo. Finnish Parliament ASR corpus - Analysis, benchmarks and statistics. In: Language resources and evaluation, 2023.* *https://arxiv.org/abs/2203.14876*

Creating ASR resources by collecting new speech data

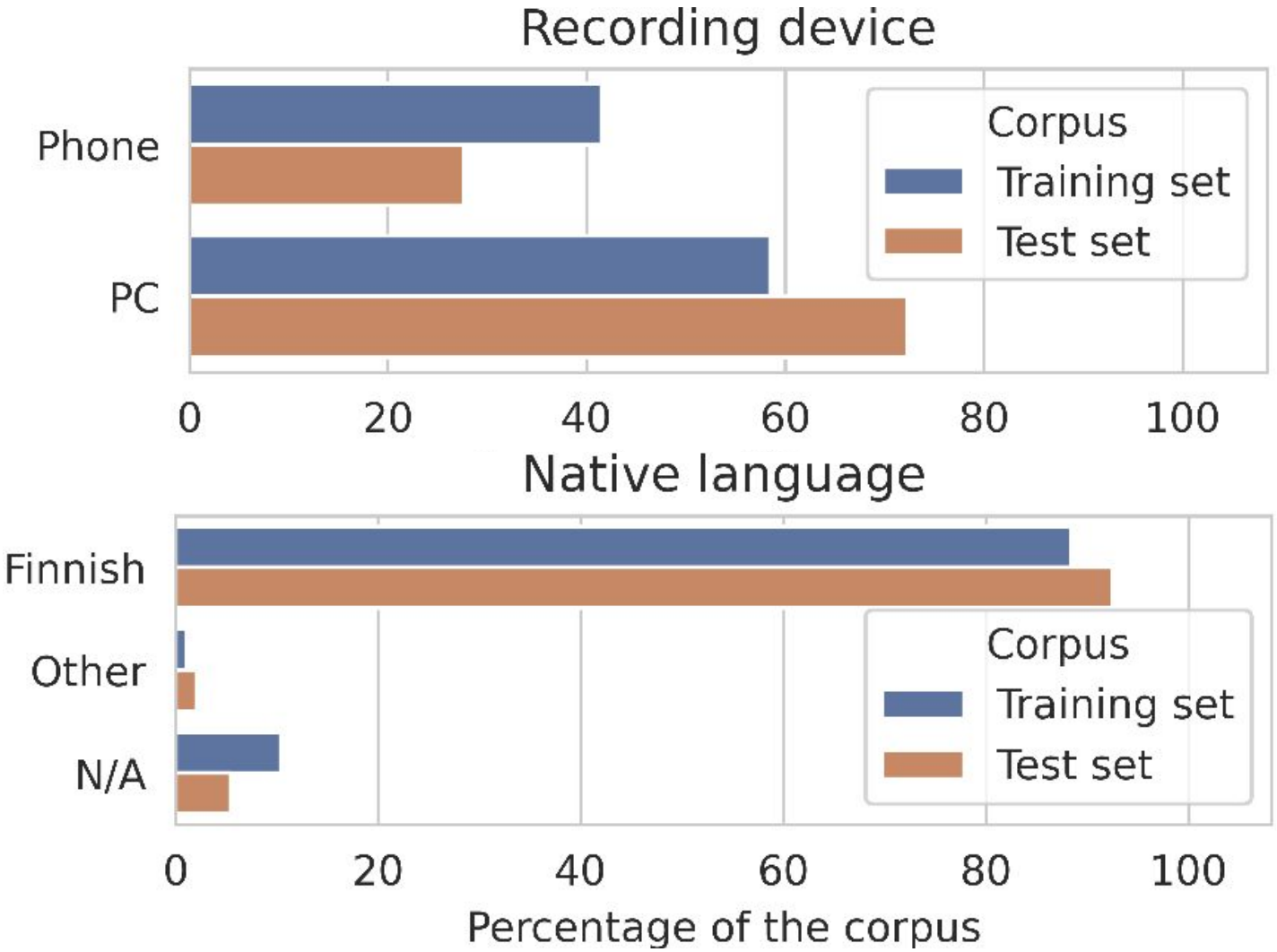
- In 2020 a large-scale **Finnish speech donation** campaign was organized with Yle and FIN-CLARIN (Language Bank of Finland)
- Yle did TV advertising and volunteers donated by recording their speech using the phone app and **lahjoitapuhetta.fi** website
- Speech is personal data protected by GDPR and the collection is based on the legitimate interests of AI research and development
- The target was **to reach out many different speaker groups and variants** of Finnish and **let people speak freely**, e.g. to describe images and videos
- The campaign was awarded by several national prizes and also the best European Digital Audio Project prize by **PRIX EUROPA 2021**

Number of donations in each month



The donated speech data

- More than **3600 hours of spontaneous speech** in 250,000 donations
- Rich metadata: **gender, age, dialect, education, native language, device type, spoken topic** etc.
- After filtering, **1600h was manually transcribed** and 1600h left untranscribed
- With this amount of data we can study how much speech data is needed for decent ASR and try semi- and self-supervised learning
- We can also detect bias in ASR performance between speaker groups and study methods to reduce the bias
- Data, scripts, models etc: https://www.kielipankki.fi/corpora/puhelahjat/
- **Results**: *Moisio, Porjazovski, Rouhe, Getman, Virkkunen, Al-Ghezi, Lennes, Grósz, Lindén, and Kurimo. Lahjoita puhetta – a large-scale corpus of spoken Finnish with some benchmarks. Language resources and evaluation, 2022. https://arxiv.org/abs/2203.12906*
- ASR models (and details): https://github.com/aalto-speech/lahjoita-puhetta-resources


Recording device
Phone
PC
Corpus
Training set
Test set
0
20
40
60
80
100
Native language
Finnish
Other
N/A
Corpus
Training set
Test set
0
20
40
60
80
100
Percentage of the corpus

10 topics

Animal friends (A)

My surroundings (M)

Summer (S)

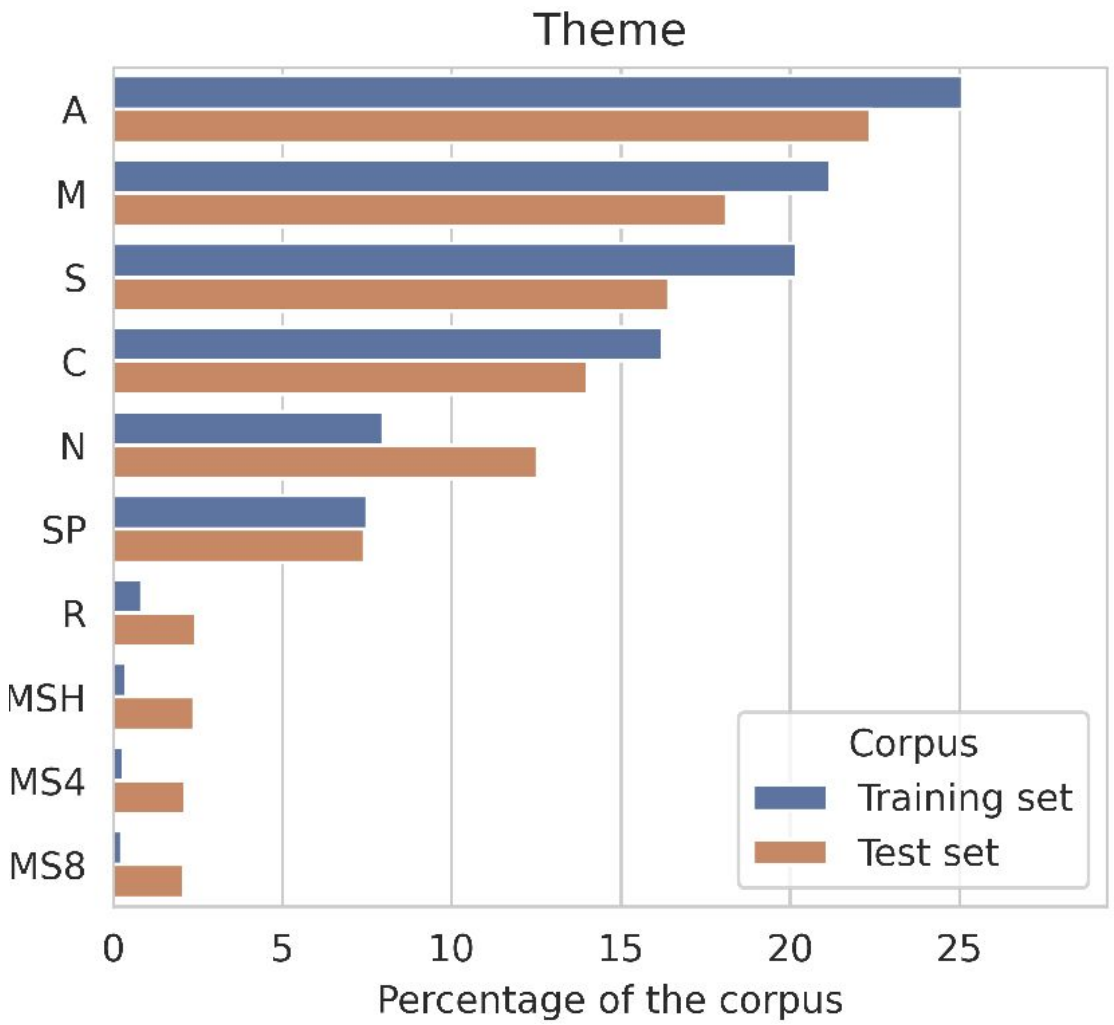
The cursed Covid (C)

Nature (N)

Sports moments (SP)

Rated (R)

Media skills (MS)



9 dialect classes

Savo

Between the SW and Häme (Tran)

Häme

Central and North Ostrobothnia (Pohjanmaa) (CNO)

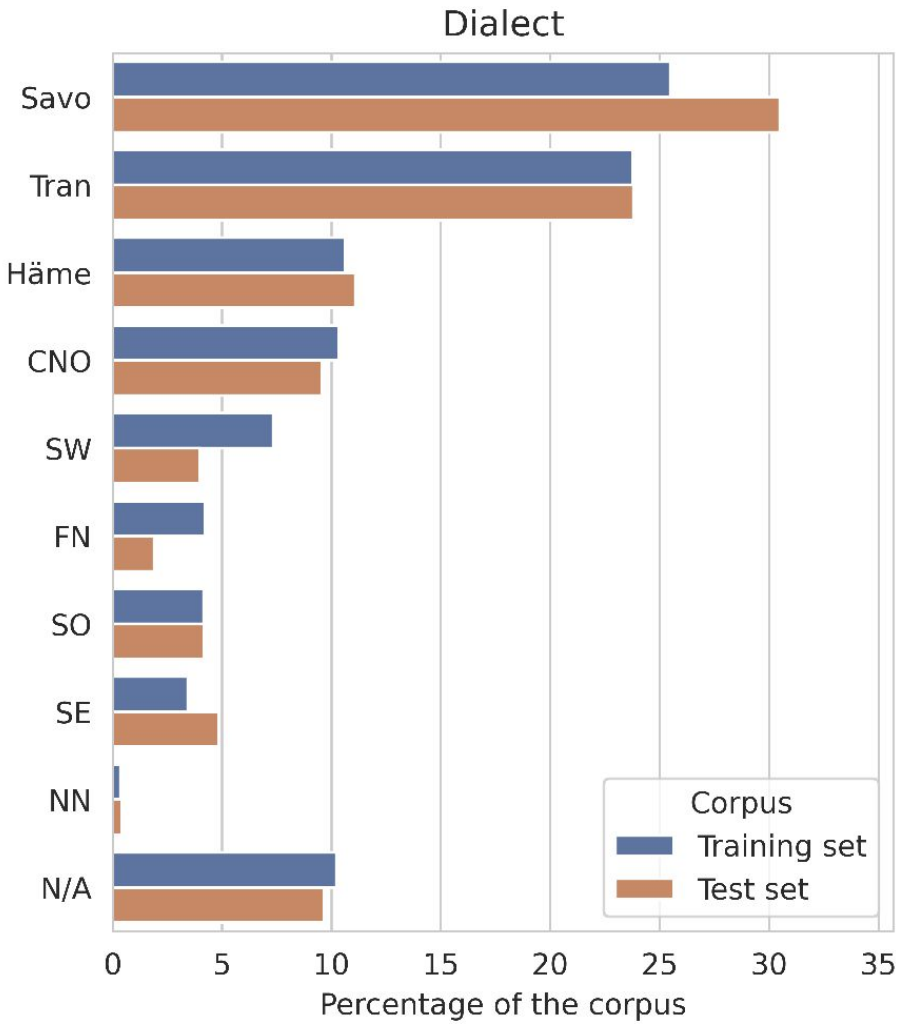
Southwestern dialects (SW)

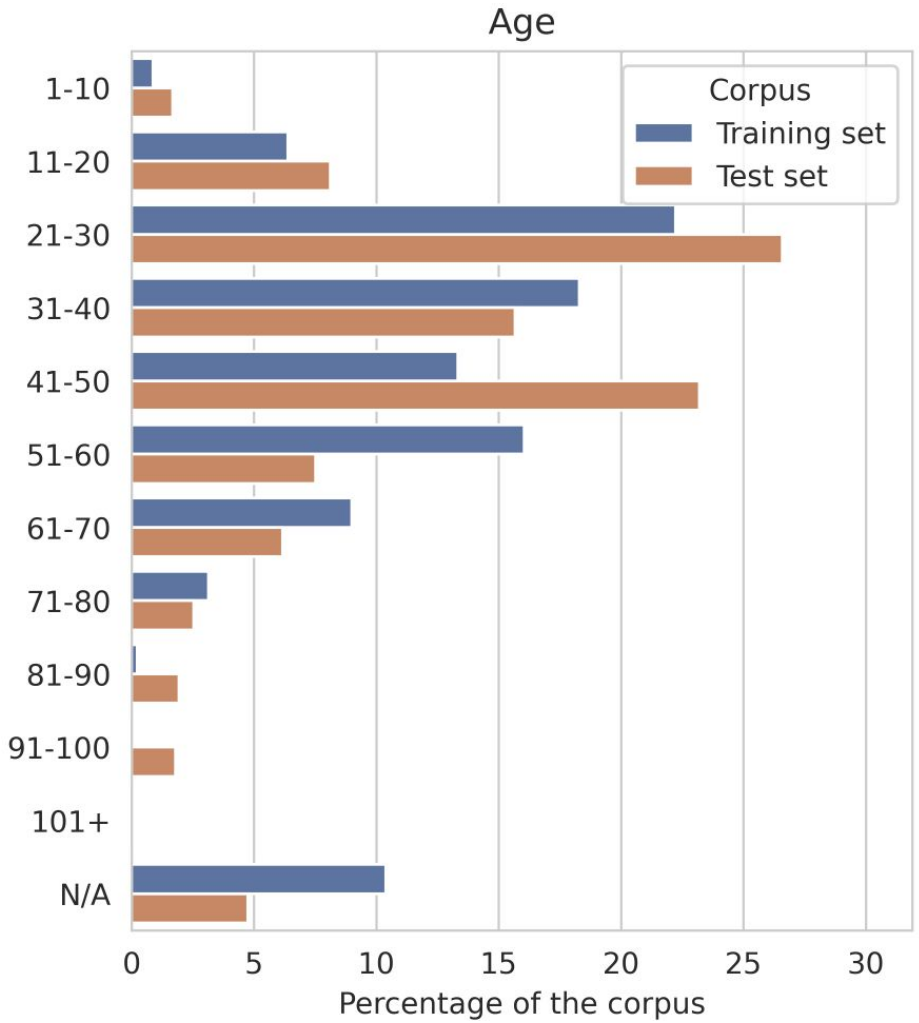
Peräpohjola (the Far North) (FN)

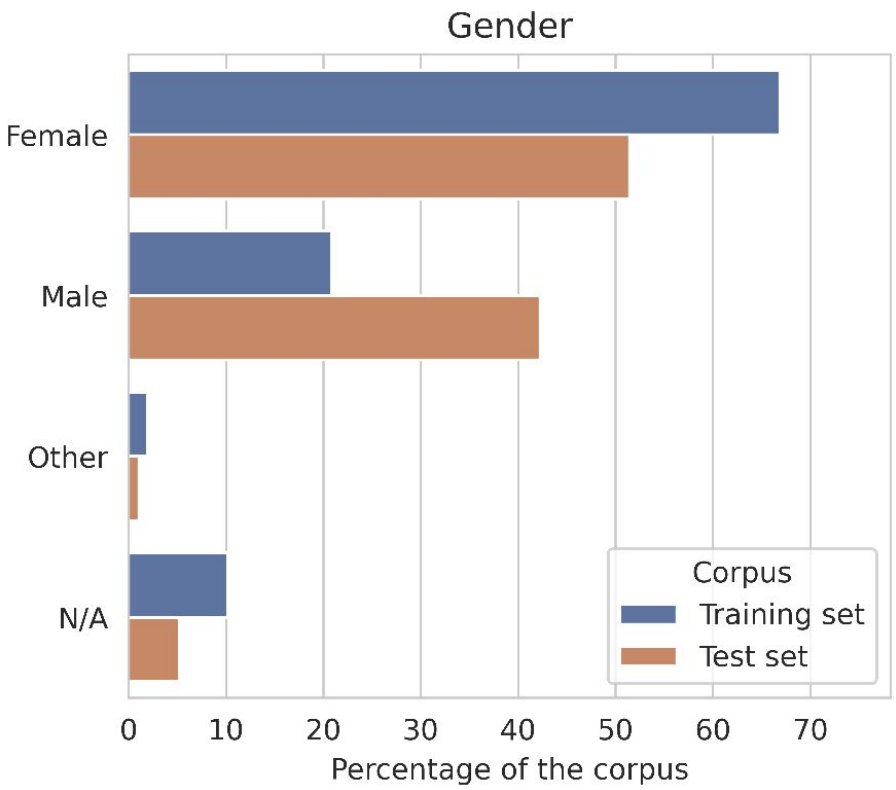
South Ostrobothnia (SO)

Southeastern dialects (SE)

Non-native speakers (NN)



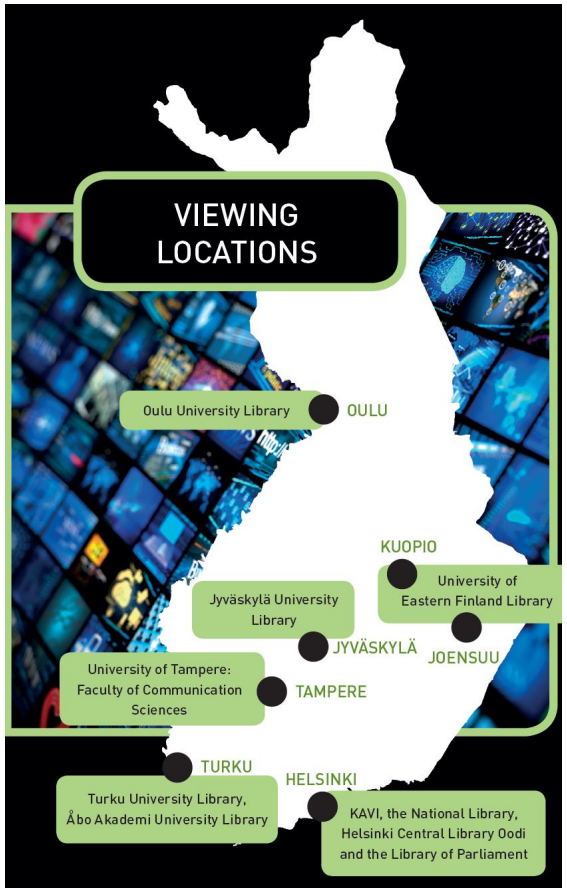

Age
1-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
101+
N/A
Corpus
Training set
Test set
0
5
10
15
20
25
30
Percentage of the corpus


Gender
Female
Male
Other
N/A
Corpus
Training set
Test set
0
10
20
30
40
50
60
70
Percentage of the corpus

KANSALLINEN AUDIOVISUAALINEN INSTITUUTTI
NATIONELLA AUDIOVISUELLA INSTITUTET
NATIONAL AUDIOVISUAL INSTITUTE

Radio & TV Archive (RTVA)

- Founded 2008 as part of the National Audiovisual Institute of Finland
- Operations are based on the Act on the Legal Deposit and Preservation of Cultural Materials (2008)
- Captures live radio and television channels digitally 24/7
- Available for the use of researchers at fixed viewing locations



Aalto University

Ritva database

https://rtva.kavi.fi/

- The entire programme stream and programme information from Finland's main radio and TV channels since 2009, and weekly sample from other channels.
- More than 835 terabytes of programme streams stored: in 2023, approximately 351,000 hours (55 TB) were recorded.
- Hours of programmes in (2023): 4,305,205

Aalto University

KANSALLINEN AUDIOVISUAALINEN INSTITUUTTI
NATIONELLA AUDIOVISUELLA INSTITUTET
NATIONAL AUDIOVISUAL INSTITUTE

LAREINA

Language Resource Infrastructure for AI (LAREINA)

- The Language Resource Infrastructure for AI (LAREINA) project aims to develop a commercially replicable model for building speech interfaces languages spoken in Finland, such as Finnish, Finland-Swedish and the Sámi languages.
- LAREINA is a project funded by Business Finland (2023–2025). The University of Helsinki and Aalto University are collaborating with companies to research, produce, test and pilot speech technology components.
- KAVI joined LAREINA in 2023.
- LAREINA cooperates with the **Alliance for Language Technologies in European Digital Infrastructure Consortium (ALT-EDIC).**

Aalto University

LAREINA

Project partners

- Research partners (staff and budgets)
  - **University of Helsinki**
  - **Aalto University**
- Public organisations (data and/or work)
  - **KAVI**
  - **Social Insurance Institution of Finland (KELA)**
- Commercial organisations (work)
  - **Inscripta Oy**
  - **Lingsoft Oy**
  - **Kielikone Oy**
  - **Solita Oy**
  - **TietoEVRY**

Aalto University

LUMI

LUMI: a supercomputer for research

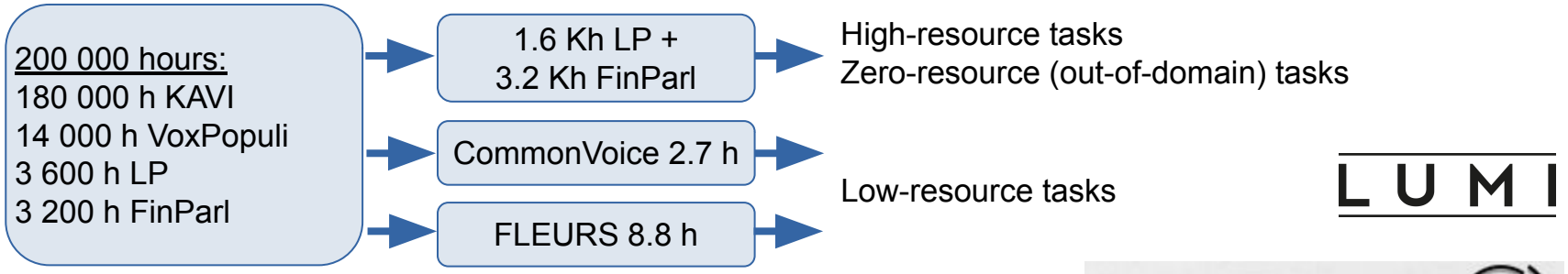
- The Large Unified Modern Infrastructure (LUMI) is **the fastest supercomputer in EUROPE**. It is located in CSC's datacentre in Kajaani
- All **researchers** in Europe can apply for an allocation of the EuroHPC quota (50 %) and in 11 member states also from their country's quota
- I**ndustry and SME** can use 20% of EUROHPC,s quota and 20% Finland's quota

Self-supervised learning: to learn meaningful representations from large untranscribed radio and TV archives (and get ASR models by fine-tune with much smaller transcribed sources)

E.g. Finnish

Unlabeled data for pre-training:

Labeled data for fine-tuning:



“KAVI models”:
Base, Large, X-Large: 1, 5, 10 days by 512 LUMI AMD GPUs

“ASR models”:
Base, Large, X-Large

LUMI

KANSALLINEN AUDIOVISUAALINEN INSTITUUTTI
NATIONELLA AUDIOVISUELLA INSTITUTET
NATIONAL AUDIOVISUAL INSTITUTE

LUMI LAREINA

KANSALLINEN AUDIOVISUAALINEN INSTITUUTTI
NATIONELLA AUDIOVISUELLA INSTITUTET
NATIONAL AUDIOVISUAL INSTITUTE

Training Northern Sami wav2vec2.0 in KAVI

35 Kh radio+TV broadcasts from RTVA archives

VAD-filtering found 22 Kh speech for pre-training

ASR fine-tuned with Sami Parliament 20 h

Tests with 1 h out-domain data

WER/CER	init	train	tune	indom.	outdom.	outdom.	outdom.
	VoxP	KAVI	parl	parl	read	spon	read+spon
		22 Kh	20 h	282 utt.	755 utt.	135 utt.	890 utt.
[Getman24]	XLS-R	Fin	+	–	–	–	47.7/ 15.2
[Hiovain23]	–	Whisper	spon 34h	–	–	–	43.2/ 14.1
KAVI-L.PT	–	+	+	**22.1/ 5.7**	**18.2/ 3.9**	61.7/ 33.0	32.9/ 12.5
KAVI-L.CPT	UralicL	+	+	22.3/ 5.7	19.1/ 4.1	**58.0/ 26.7**	**32.3/ 10.8**

Sharing the Foundation models and ASR models from Aalto

Finnish KAVI models, both non-fine-tuned and fine-tuned for ASR with large Lahjoita Puhetta and Finnish Parliament datasets:
https://huggingface.co/collections/GetmanY1/wav2vec2-fi-150k-66c9d75d18579088974ea37f

Northern Sami KAVI models, both non-fine-tuned and fine-tuned for ASR with the small Sami Parliament dataset:
https://huggingface.co/collections/GetmanY1/wav2vec2-sami-22k-66ead12fe465d6302b63d11b

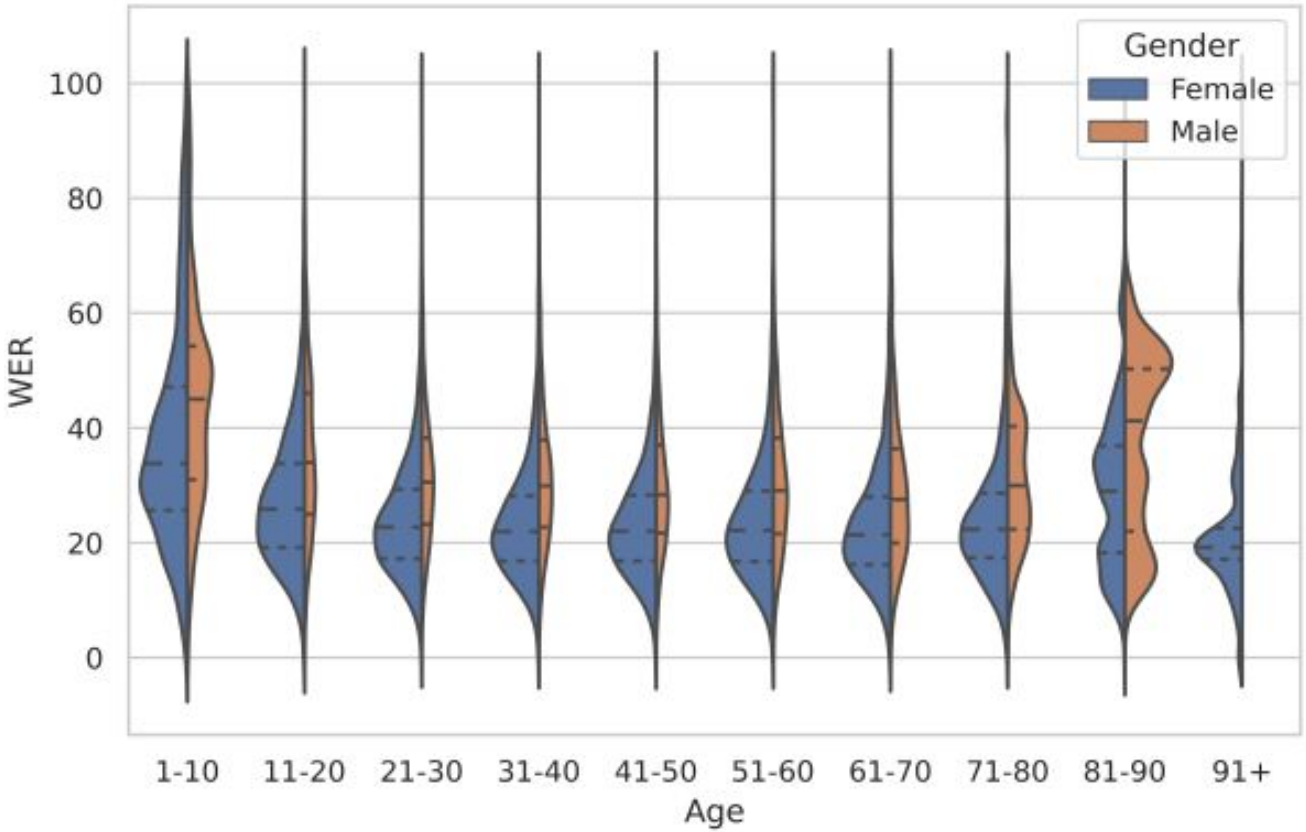
These models are described and evaluated in two papers papers both still in double-blind review for publication, so **DO NOT DISTRIBUTE the models anywhere in public** yet, please.

Variation of ASR accuracy in speaker categories

Largest difference were observed in age and gender.

The system is best for females and worst for children and old males.

Similar observations recently in Finnish Parliament ASR (over 3000h transcribed) data, 50-50 genders



Applications of ASR

Human to human interaction

- Automatic transcription of speech
  - Interviews, meetings, lectures
- Captioning and subtitling
  - Videos, TV and radio broadcasts
- Tools for the hard-of-hearing
  - Captions
  - Hearing devices
- Speech translation and interpretation

Human-computer interaction

- Mobile information services
- Customer support
- Call routing
- Text input
- Chatbots
- Games
- Language learning

Transcription of the speech data

What to transcribe?

- Spoken words w/wo hesitations, , stuttering, repairs, repetitions, mispronunciations
- Names for e.g. anonymisation
- Non-verbal sounds, fillers, vocalized emotions, laughing, coughing
- Speaker information, e.g. name, age, gender, education, occupation, health issues, oral language proficiency
- Noises

Human vs machine? How to evaluate the performance?

Other annotation tasks

- Splitting the recording into sentences or speaker turns
- Aligning the audio and the transcript
- Speaker names and diarization
- Language or dialect identification
- Audio event tagging

Reliability of the annotations? How many annotators? How many per recording?

Collecting speech data and GDPR

- Is it personal data? Is it possible to recognize the speaker?
- What is the justification for the data collection?
- Where the data is stored and who can access it?
- What metadata is collected (contact information, native language, health)?
- Minimising the collected information
- Ethical reviews

For more information

- Contact: *mikko.kurimo@aalto.fi*
- Publications: http://research.aalto.fi
- Home page: (search: *”Aalto asr home”)*
- Software: (search: *”Aalto asr github”)*