Moisio, Anssi; Porjazovski, Dejan; Rouhe, Aku; Getman, Yaroslav; Virkkunen, Anja; AlGhezi, Ragheb; Lennes, Mietta; Grósz, Tamás; Lindén, Krister; Kurimo, Mikko

Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks

Published in:
Language Resources and Evaluation

DOI:
10.1007/s10579-022-09606-3

Published: 01/09/2023

Document Version
Publisher's PDF, also known as Version of record

Language Resources and Evaluation
https://doi.org/10.1007/s10579-022-09606-3

ORIGINAL PAPER

Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks

Anssi Moisio[1] · Dejan Porjazovski[1] · Aku Rouhe[1] · Yaroslav Getman[1] · Anja Virkkunen[1] · Ragheb AlGhezi[1] · Mietta Lennes[2] · Tamás Grósz[1] · Krister Lindén[2] · Mikko Kurimo[1]

Accepted: 8 July 2022

Abstract
The Donate Speech campaign has so far succeeded in gathering approximately 3600 h of ordinary, colloquial Finnish speech into the *Lahjoita puhetta* (*Donate Speech*) corpus. The corpus includes over twenty thousand speakers from all the regions of Finland and from all age brackets. The primary goals of the collection were to create a representative, large-scale resource to study spontaneous spoken Finnish and to accelerate the development of language technology and speech-based services. In this paper, we present the collection process and the collected corpus, and showcase its versatility through multiple use cases. The evaluated use cases include: automatic speech recognition of spontaneous speech, detection of age, gender, dialect and topic and metadata analysis. We provide benchmarks for the use cases, as well downloadable, trained baseline systems with open-source code for reproducibility. One further use case is to verify the metadata and transcripts given in this corpus itself, and to suggest artificial metadata and transcripts for the part of the corpus where it is missing.

Keywords Spoken colloquial language · Speech collection · Automatic speech recognition · Gender, age, dialect and topic recognition

1 Introduction

The preservation of spoken colloquial language is an important task, which requires the collection of relevant materials and their careful curation. The Donate Speech (Lahjoita puhetta) campaign embarked on the quest of preserving the current state of the spoken Finnish language and boosting the development of AI that understands spoken Finnish. To this end, a large collection campaign was initiated that resulted

✉ Anssi Moisio
anssi.moisio@aalto.fi

Extended author information available on the last page of the article

Published online: 09 August 2022

in the creation of a large-scale colloquial Finnish speech corpus. In this paper, we explain how the collection and curation of the data were performed to maximise the amount participants while still ensuring a high quality of the dataset. Furthermore, we will demonstrate with pilot projects and their results how the materials can be used to study and develop new technology and services in the Finnish language.

Currently, there is only one large freely available transcribed Finnish speech corpus, the Finnish Parliament ASR Corpus.[1] It contains over 3000 h of professionally transcribed speech which is rather formal in style and often read from the speaker's notes. However, colloquial, spontaneous Finnish differs significantly from formal Finnish in multiple aspects. Considering phonological features, for instance, durations of phones are longer in read speech than in spontaneous speech (Lennes, 2009). From the morphological and lexical point of view, it is common to truncate or combine words, and to use incorrect word inflections in addition to words not used in written text. Finnish has a near-phonemic orthography: there is usually a one-to-one mapping from letters to phonemes, except for some rare cases such as certain loan words and the "ng" letter pair which is not pronounced as /n/ followed by /g/ (which are the normal phonemes for the letters "n" and "g", respectively) but instead has its own phoneme, /ŋ/. Because of the near-phonemic ortography, the phonological variations can be transcribed mostly unambiguously into text. Since there is no standard transcription style for colloquial speech, the spelling variations of a single word can be numerous [for example, the word "minä" ("I", first person singular) can be written as "mä", "mie" or "mää"], which further increases the distance between the domains of formal and colloquial Finnish.

There are a few smaller corpora that include carefully transcribed spontaneous, colloquial Finnish speech. The SPEECON (Iskra et al., 2002) corpus is a collection of speech for multiple languages, recorded in varying environments. It includes both read and spontaneous speech from 550 speakers. The spontaneous Finnish part includes ten sentences from each speaker, in total about 18.8 h. The FinDialogue[2] part of the FinINTAS (Lennes, 2009) corpus contains 6338 utterances by 22 speakers. The speech is from spontaneous and unmonitored conversations between participants, and includes about 10.4 h of speech in total. The DSPCON[3] corpus consists of free-form conversations between students, recorded at the Aalto University between 2013 and 2016. It includes 5281 spontaneous sentences from 218 different male students and 24 female students, totalling 9.8 h (Enarvi, 2018). Combining these three corpora, there are about 40 h of transcribed spontaneous Finnish speech currently available for research (non-commercial) use,[4] to the best of our knowledge. We note that substantial amounts of Finnish colloquial speech has been collected in the 1960s and 1970s by the National Institute for the languages of Finland as well as some cultural foundations, but that data is not yet available for commercial development use according to the European data protection legislation.

[1] https://urn.fi/urn:nbn:fi:lb-2021051903.
[2] https://urn.fi/urn:nbn:fi:lb-2016041421.
[3] https://urn.fi/urn:nbn:fi:lb-201708251.
[4] While SPEECON is quite expensive, the other two corpora are free.

For major languages like English, large spontaneous and colloquial speech corpora are available for research and commercial use. The Switchboard corpus (Godfrey et al., 1992) consists of about 260 h of telephone conversations among 302 male and 241 female speakers. The Fisher corpus (Cieri et al., 2004) includes approximately 2000 h of colloquial telephone conversations. These two corpora, for example, have been actively used in speech research for many years now, and technologies built for spontaneous English have greatly benefited from the datasets. Even though Finnish has far fewer speakers than the major languages (not even in the top 100), the new *Lahjoita puhetta* corpus covers many more speakers per language (over 20k) than probably any other publicly available spontaneous speech corpus. The *Lahjoita puhetta 2021* release consists of 3600 h of speech out of which about 1600 h have been transcribed. The data covers all regions of Finland and has both male and female, mostly native, speakers in all age brackets. In this work we describe this dataset, how it was curated, and demonstrate its use-cases. Specifically, the contributions of this work are the following:

1. Presenting an open large colloquial speech corpus for Finnish.
2. Describing a successful concept for large-scale speech data curation.
3. Demonstrating the utility of the corpus in speech. recognition and metadata (gender, age, dialect and topic) classification.
4. Defining relevant benchmarks for speech recognition and metadata classification.
5. Providing trained, downloadable baseline systems for the benchmarks, and open-source code for reproducing the systems.

All of the tools and resources described in this work can be accessed online.[5]

2 Data collection

The speech material donated during the campaign is shared by the Language Bank of Finland (Kielipankki),[6] coordinated by the University of Helsinki. Since speech samples may contain personal data, they are protected by European and national data protection legislation, most notably by the General Data Protection Regulation (GDPR).[7] The speech material has been collected based on the legitimate interest of individual researchers, universities, research organisations and private companies to study language or artificial intelligence, to develop AI solutions and to provide higher education in the aforementioned areas. To use legitimate interest as the legal basis for the processing of personal data, it was necessary to accomplish a balance test to ensure that the legitimate interests are not overridden by the interests or fundamental rights and freedoms of the data subjects.

[5] https://github.com/aalto-speech/lahjoita-puhetta-resources.
[6] https://kielipankki.fi.
[7] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

To inform the individuals who donated their speech to the campaign, two essential documents were drafted: a short information page including simple conditions of participation, and a more comprehensive data protection policy. The donors had to acknowledge that they were informed of the conditions of participation before they could start donating.

An ethical review for the data collection was not needed as such a review applies only to the research configurations[8] specified by the Finnish National Board for Research Integrity (TENK). It was also considered that the risks to the rights and freedoms of natural persons were rather low, but to be sure, a data protection impact assessment (DPIA) was made. For a more detailed description of the campaign and its legal background documents, see Lindén et al. (2022).

The goal of the campaign was not merely to collect a vast amount of any kind of speech, but to reach out to as many different groups of Finnish speakers and to as many individuals as possible. In marketing the campaign to citizens, it was emphasised that all variants of spoken Finnish are welcome, including speech from second language Finnish learners. However, in order to understand the privacy notice and the instructions, a certain level of language proficiency was required from the speech donors.

Key issues and challenges for the design of the user interface were in determining elicitation methods that entice a person to speak freely, gaining the trust of the speaker, making him feel comfortable while also satisfying legal constraints for presenting enough required information in an easy to understand format, as well as more technical choices of supported platforms, presentation forms, visual and auditory feedback of the on-going recording or its quality. After some ideas for themes had been formulated and tested, Yle (the Finnish Broadcasting Company) settled on the fail-safe recurring functions of showing a video, a picture or some textual content enticing a person to speak with an easy-to-use one-button starting and stopping of the recording.

Cooperating with Yle was crucial for the marketing of the campaign and for attracting the attention of the Citizens of Finland for the campaign. In the end, Yle developed around 40 straightforward topics, within ten different themes, for stimulating the collecting of speech data. As part of the campaign, Yle made comical infomercials with requests to the general public to donate speech. These were broadcast during programme breaks in national radio and TV channels during the summer and autumn of the Covid-19 pandemic in 2020 with some trailing reruns during spring 2021. In 2021 the data collection campaign was awarded the best European Digital Audio Project prize by PRIX EUROPA, which was founded by the European Parliament, the European Commission and the European Cultural Foundation in 1987.

To illustrate the campaign results with regard to collection speed, the number of recordings received each month during the campaign is shown in Fig. 1. The peaks in the beginning and at the end of 2020 reflect the effects of the increased public advertising activity.

[8] https://tenk.fi/en/ethical-review/ethical-review-human-sciences.

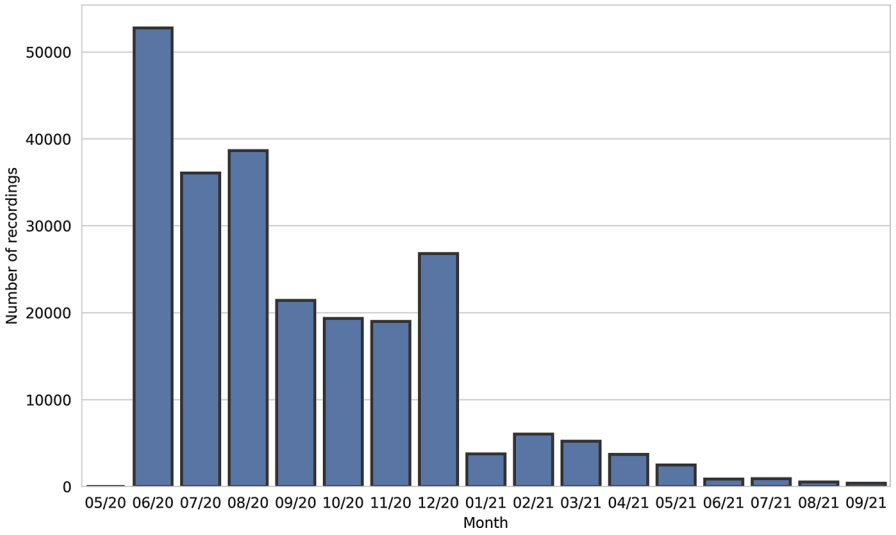

Fig. 1 The number of recordings received in each month during the campaign

2.1 Metadata complementing the speech corpus

Identities of speakers were not collected explicitly, but we assume that one application client identity (of the browser or smart phone application used for recording) corresponds to one speaker. This assumption is not watertight since one person may use multiple application clients, or multiple persons may use one client, but the correspondence generally holds. Assuming this, the number of speakers is well over 20k, which means quite a good sample of all Finnish speakers, which are fewer than six million in total.

Opening the Lahjoita puhetta website or phone app, the user is offered a few different themes to choose from. To focus the campaign, all of the themes are not always available on the website. The complete list of themes, and their English translations and abbreviations used in this text, is the following:

– “Eläinystävät” (“Animal friends”, A)
– “Urheiluhetket” (“Sports moments”, SP)
– “K-18” (“Rated R”, R)
– “Luonto, sää ja mää” (“Nature”, N)
– “Lähelläni juuri nyt” (“My surroundings”, M)
– “Mediataidot 4–6 lk.” (“Media skills—grade 4–6”, MS4)
– “Mediataidot 8–9 lk.” (“Media skills—grade 8–9”, MS8)
– “Mediataidot lukio” (“Media skills—high school”, MSH)
– “Kirottu korona” (“The cursed covid”, C)
– “Sukella kesään” (“Summer”, S)

Each theme includes up to eight different topics that ask a question or in some other way invites the user to speak about the topic. Each recording therefore pertains to some general theme, as well as to a certain topic within that theme. The theme and topic are metadata which can be used to categorise the recordings.

Between the recording prompts, the participant is asked multiple questions about his or her background. These metadata questions include dialect background, gender, native language, age, place of residence, birthplace, occupation and education. In this paper, we focus on the first four of these metadata types.

The dialect background question offers 20 options to choose from. In order to have fewer classes, we clustered these dialect regions into eight larger dialect groups, based on the information provided by The Institute for the Languages of Finland.[9] The dialect groups and their abbreviations used in this paper are:

1. The Southwestern dialects (SW)
   - Varsinais-Suomi
   - Ahvenanmaa
2. The transitional dialects between the Southwestern and Häme dialects (TRAN)
   - Uusimaa
   - Satakunta
3. The Häme (Tavastian) dialects (HÄME)
   - Pirkanmaa
   - Häme
4. The dialects of South Ostrobothnia (Pohjanmaa) (SO)
   - Etelä-Pohjanmaa
   - Pohjanmaa
5. The dialects of Central and North Ostrobothnia (Pohjanmaa) (CNO)
   - Keski-Pohjanmaa
   - Pohjois-Pohjanmaa
6. The dialects of Peräpohjola (the Far North) (FN)
   - Lappi
7. The Savo dialects (SAVO)
   - Pohjois-Savo
   - Etelä-Savo
   - Kainuu
   - Keski-Suomi
   - Pohjois-Karjala

[9] https://kotus.fi/en/on_language/dialects/finnish_dialects_7541.

Table 1 The sizes of the corpus and its subsets

Subset	# of speakers	# of recordings	# of hours
Total original	20,890	218,146	3604.8
Total usable	20,269	205,962	3229.8
Train transcribed	17,821	98,606	1601.5
Train untranscribed	18,825	105,380	1597.1
Train transcribed 100 h	1129	6229	103.5
Dev	103	703	10.5
Test	103	690	10.4
Test multi-transcriber	57	58	1.0
Test multi-transcriber speakers	57	583	10.2

  - Kymenlaakso
  - Päijät-Häme

8. The Southeastern dialects and a few transitional dialects bordering on them (SE)
   - Etelä-Karjala
9. Non-native Finnish speakers (NN)

2.2 Corpus statistics

In the *Lahjoita puhetta 2021* release, there are about 3600 h of recordings in total, and over 20k different speakers. The median speaker donated eight recordings, while the top donor donated 1039 recordings. The median duration of a recording is about 40 s and the longest are about 10 min.

Silent parts were trimmed from the beginnings and endings of the recordings using the `silence` effect of SoX,[10] with a threshold of 0.5% and duration of 0.05 s. After trimming, 3270 h remained, and the randomly selected recordings were sent to human transcribers. When we received the transcribed subset, there were 512 recordings that had empty transcriptions. Some of these were silent audio and some were left empty by mistake by the transcribers, but all 512 were discarded at this point. To verify the quality of the human transcriptions, we generated ASR transcriptions with a hybrid HMM/DNN (hidden Markov model/deep neural network) system trained on the previously existing colloquial Finnish speech data: DSPCON, FinDialogue, SPEECON (see Sect. 1). The average WER (word error rate) was around 38% and CER (character error rate) about 15%. We then filtered out recordings for which both the WER and the CER were over 94% in order to mitigate the chance of having low-quality samples in the ASR training corpus. From the set of about 100k transcribed recordings, 392 had WER and CER over the threshold and

[10] https://sox.sourceforge.net.

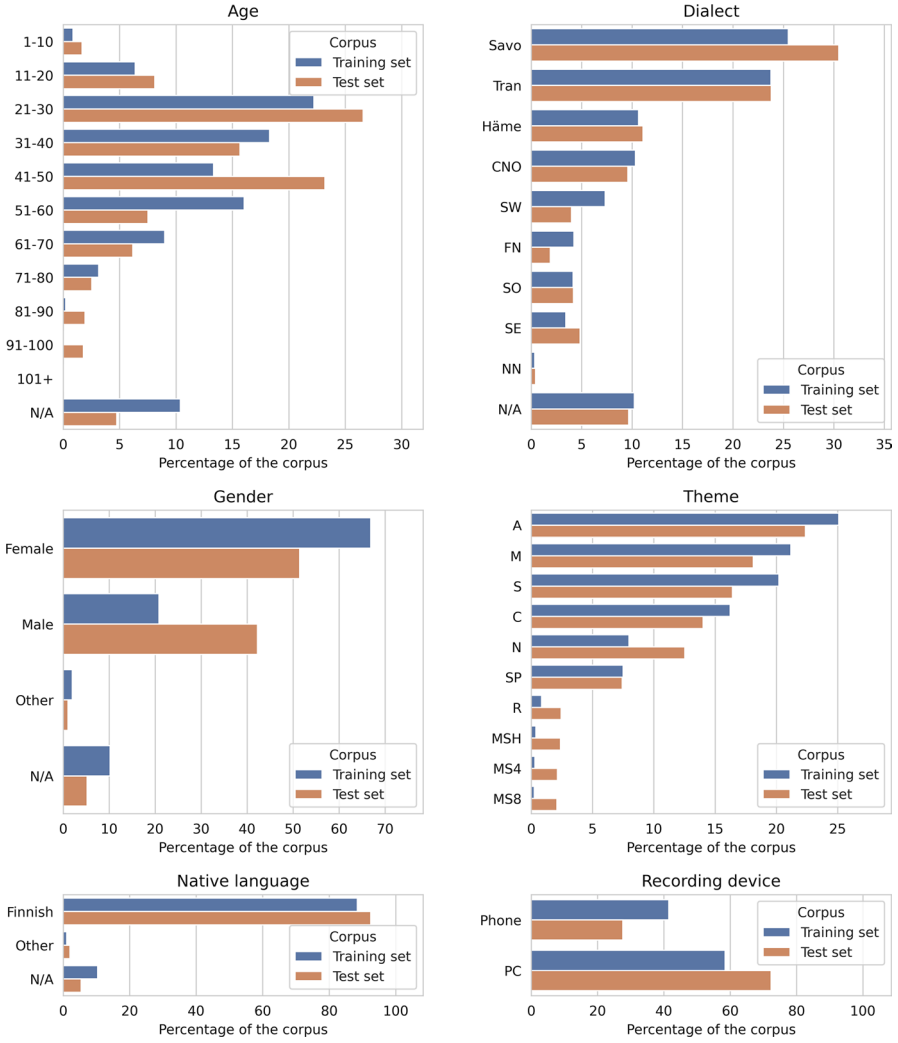

Fig. 2 The distribution of the speaker metadata in the corpus. The "training set" includes both the "train transcribed" and "train untranscribed" described in Table 1. "N/A" means the user has not answered to the question about his or her background, or has given multiple contradicting answers

were excluded. Combined with the 512 empty-transcript recordings, these excluded 904 recordings were about 9.1 h in duration.

We sampled a 10-h test set and 10-h development set from the transcribed speech data, each including at least 10 min of speech for each metadata class in each of the five metadata domains. We also modified the speaker gender ratio of the test and dev sets, so that they have over 40% male speakers although the training set has just over 20%. As a second test dataset, we used a 1-h set that was transcribed by four different transcribers, which includes 58 recordings from 57 speakers. If we add all

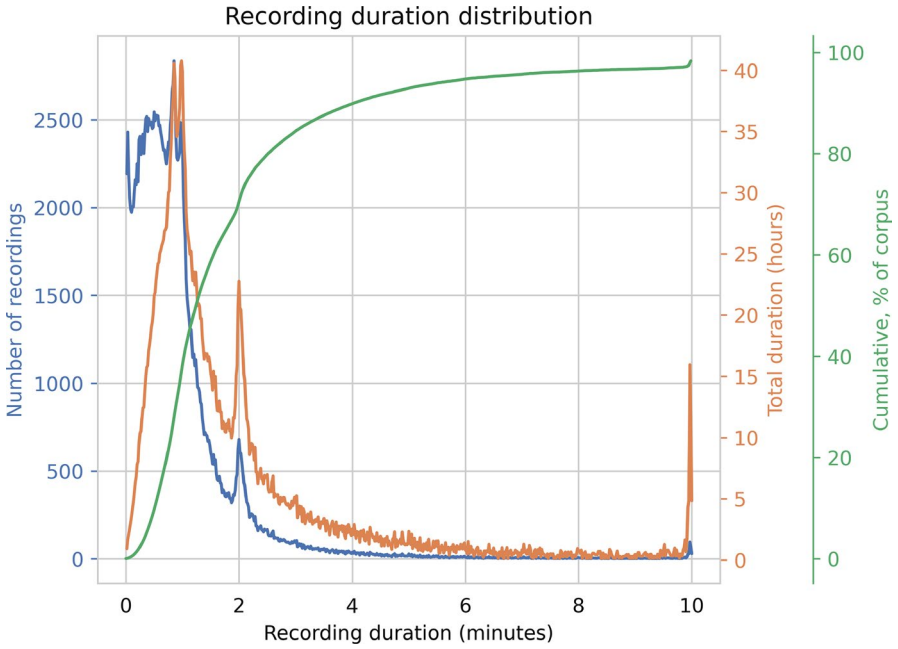

Fig. 3 The recording length distribution. The recording durations are pooled to 1-s bins to generate this figure

recordings by those 57 speakers to this subset, we get a 10-h test set, that we call "test multi-transcriber speakers" in Table 1. The rest of the transcribed speech is used as training data. The train, dev and test sets have no overlap of speakers. There are still recordings that are by the speakers of the dev or test sets but which are not transcribed. These are left unused, leaving about 3230 h in the complete dataset that we use. Table 1 lists the sizes of the corpus subsets.

Figure 2 presents the amount of speech for each metadata type as a portion of the whole training set (both transcribed and untranscribed pooled together) and the 10-h main test set. As the transcribed training set is a 1600-h random sample of the whole dataset, which has about 3199 h out of the complete 3230 h, the training data accurately represents the overall distribution of the whole dataset. We can note that the corpus has varying amounts of speech from the different metadata classes. Younger than 11-year old children have donated some but a relatively small amount, as have older than 80-year-old people. Of the dialects, Savo and Tran have most data, roughly a quarter each. Women have donated significantly more than men: over three times as much. Investigating the reason for this male-female imbalance is beyond the scope of the present work, but we may speculate, for example, that the campaign might have been advertised between TV/radio shows whose audience is predominantly female, although unfortunately we do not know the demographics of the TV or radio audience to which the campaign was advertised. Another factor could be that women might be more likely to answer to surveys in general (Smith, 2008); although speech donation is not exactly a survey, it is similar enough that

the trends found in survey response rate could give some clues to why our dataset is imbalanced. Four themes seem to have a low amount of speech: Rated R and the three Media skills themes, as they were added only when the official marketing campaign had already ended. In all metadata domains, the test set was smoothed to have at least 10 min of speech from each metadata class, visible in the figure.

Figure 3 displays the distribution of the recording lengths. The majority of the recordings are less than 2 min, but longer recordings are not uncommon. There are spikes at the 2-min mark and the 10-min mark. The spike at 10 min was effected by the limit of the duration of recordings: those that would have spoken for longer were cut at 10 min. The other spike, at 2 min, corresponds to the duration of a video clip that was played for the user in one topic. The theme was "Summer", and in this topic the user was asked to describe what is happening in the video clip to an alien while the video displayed sceneries of Finnish summer pastime activities.

3 Annotation procedure

Because a high-quality manual transcription of 1600 h of spontaneous speech is a significant investment, we made an effort to develop a careful process described in detail in this section. The aim was exact transcription, which included not only the verbal content of the speech but also full words, repetitions, hesitations, partially pronounced or only partially audible words, and non-verbal communication such as laughs, growls, and coughs. The guidelines that were given to the transcribers are reproduced in Appendix.

3.1 First phase: annotator selection

To choose the best transcriber companies, we ran a pilot transcription competition, where we shared a 20-h subset of the data with all candidates along with the carefully constructed annotation instructions. The datasets consisted of 19 h of randomly selected data per participant mixed with a common 1-h evaluation set (the composition of the data was not disclosed to the companies). After the competitors submitted their transcripts, we evaluated them automatically and manually using the overlapping 1-h set to determine the quality of their work as well as an hour of random samples from the non-overlapping parts to verify the automated comparisons manually. During the evaluation process, we had no information about individual annotators, so we treated each company as a single transcriber. Our primary goal was to validate that they could produce high-quality transcripts for the collected data.

The automatic evaluation focused on comparing the transcripts of different annotators with each other and with multiple ASR systems. Our goal was to select companies who can produce high quality transcripts that met the standards of the Language bank (Kielipankki). First, standard ASR metrics like word error rate (WER) and character error rate (CER) were used to estimate the inter-annotator agreement. Specifically, one annotators transcript was used to calculate edit distances from the others, treating them as speech recognition systems. This allowed us to create a

Table 2 Pairwise comparison between transcribers (T) using the word and character level edit distances

Transcriber	Word level comparisons			Character level comparisons		
	T2	T3	T4	T2	T3	T4
T1	19.5%	19.8%	20.5%	6.3%	5.8%	6.1%
T2	–	13.6%	15.6%	–	4.7%	5.4%
T3	–	–	16.0%	–	–	4.9%

Table 3 Pairwise comparison between transcribers (T) and ASR systems using the word and character level edit distances

Transcriber	Word level comparisons		Character level comparisons	
	Hybrid (%)	E2E (%)	Hybrid (%)	E2E (%)
T1	33.56	33.65	11.95	10.12
T2	28.02	27.33	10.14	8.59
T3	29.04	28.83	10.69	8.89
T4	29.87	29.87	10.93	9.15

preference order from the perspective of one of the annotators. Repeating this process for all transcriber companies gave us multiple rankings, and we tried to identify outliers by aggregating these preference rankings. In case of an outlier, we could verify that its transcription is of lower quality than the others by manually checking the transcripts with the most differences to the other transcribers. During these analyses, we ignored the non-word symbols, as they were annotated with considerable discrepancies by different annotators.

The inter-annotator disagreements in terms of WER and CER were generally high due to the nature of the data, see Table 2. Still, we can observe considerable differences. These metrics allowed us to create rankings per annotator. Fortunately, we only wanted to ensure the high quality of the transliteration, so we did not have to use complex methods (like the Borda count etc.) to produce a complete order. In the end, we opted for a straightforward scheme to aggregate the individual preference orders by simply eliminating the worst in each round until we get the desired number of annotators.

Looking at the values in Table 2, we can see that T1 had the highest disagreement with the others, both in terms of WER and CER. The transcription quality was also substantiated by manually inspecting 1-h random samples from each candidate. Thus T1 was the first to be eliminated. Of the remaining annotators, T2 and T3 disagree most with T4. Nevertheless, the differences between these three annotators were relatively small, so in the end, we opted to accept all three in this round of selection.

Next, we repeated the experiments, but this time, we compared the transcripts with ASR outputs. Two models were selected for this purpose, a hybrid HMM/DNN, and a Wav2Vec2-based (Baevski et al., 2020) end-to-end network. The hybrid HMM/DNN system was trained on the existing spontaneous colloquial Finnish

speech datasets: DSPCON, FINDialogue and SPEECON (spontaneous part), totalling about 37 h. The 1st pass n-gram LM and 2nd pass RNN LM are trained on the WEBCON (Enarvi, 2018) corpus and the speech transcripts, in total about 76 million words. For the end-to-end model, we decided to utilise the publicly available multilingual *Wav2Vec2 Large* model pre-trained on 100K h of the VoxPopuli dataset (Wang et al., 2021). The model was fine-tuned on the same 37-h colloquial Finnish corpus used to train the hybrid system.

Comparing with ASR models reaffirmed our previous findings (Table 3). We can see that comparing the ASR models with T1 leads to the highest error rates. An interesting observation is that both models seem to favour T2, yielding the lowest error rates, followed by T3 and T4.

Lastly, we also validated the conclusions of all automatic experiments by manually checking the utterances with the largest differences (revealed by the previous examinations). The manual inspection revealed that T4 had transcribed files mostly correctly, but they often used the formally correct spelling instead of writing the verbatim spoken version. This resulted in slightly higher error rates compared with T2 and T3. Comparing T2 and T3 we saw that the latter skipped the extremely noisy part of an utterance, resulting in T2 being selected as the most diligent annotator.

Combining all observations, we concluded that companies T2, T3, and T4 are all capable of creating sufficiently high-quality transcripts, so we continued to work with them to transliterate a large portion of the collected corpus.

3.2 Second phase: quality control

After the initial selection phase, we continued to utilise our ASR models to perform automatic quality control checks. Our goal was to highlight recordings with unusual error rates (WER $\geq$ 94%) for manual inspection. In practice, once we received the transcriptions from the companies, we applied the same ASR models as in the phase one to get the WER and CER for each utterance. To avoid unnecessary checks, we only selected files with a high WER and CER compared with both models.

Our manual examinations revealed several problems that we could address during the annotation process. One of the primary issues that we managed to identify was a mismatch between the transcription and the audio files (approx. 20 transcripts had been assigned to the wrong recording). Naturally, with the help of the annotators, we could fix this problem quickly. The second source of the high ASR error rates was the presence of extreme noises, which made it hard for the ASR systems to recognise the speech. We kept these noisy recordings in the corpus to enable the building of noise-robust models.

Figure 4 depicts the error rates of the hybrid ASR model for each transcriber company. Note that due to legal constraints, we were unable to match the transcribing organizations' ids used here to those in the first phase. Thus we could not analyse how their performance changed on the large dataset. Overall, we can see that the distributions are quite similar, meaning that from the ASR model's viewpoint, they were equally good at providing the gold standard texts. We can see that there is a considerable amount of utterances with more than 100% WER, but overall, the

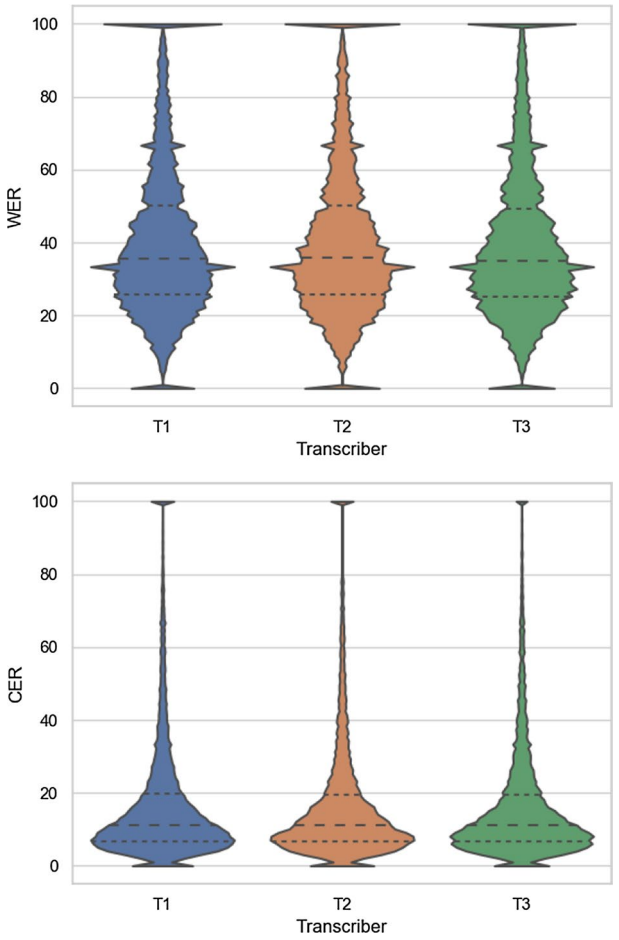

Fig. 4 The distribution of word-level (top) and character-level (bottom) error rates per annotators on the transcribed dataset. *Note* Utterances with more than 100% errors were pooled together for this visualisation. Note also that the transcribers' ids of the second phase do not match to the first phase

vast majority of recordings are recognisable with less than 50% error. The CER statistics further reassured us that the transcription is high quality; more than 75% of the utterances had a CER below 20%. The high errors could be explained by the discovered problems (noise, low volume, speaking far from the microphone).

4 ASR experiments and results

In this section several ASR experiments with various architectures are presented. The goal of the ASR experiments is first to establish that the transcribed Lahjoita puhetta data is useful for creating ASR systems, and then to provide baseline results

and recipe starting points for a few different ASR techniques. The trained ASR systems are also used to provide both time alignments of the manually transcribed part, as well as ASR decoding outputs for the untranscribed part, which can later be used for indexing, searching, or statistical studies on the data, as attested by for example Carrive et al. (2021).

One initial difficulty in using the transcribed Lahjoita puhetta data for ASR is that many of the recordings are longer in duration than is ideal for many speech recognition methods. Bootstrapping alignments for long recordings is more difficult. Long recordings exacerbate the vanishing gradient problem and they also present practical issues related to memory consumption (Narayanan et al., 2019). In these experiments, we are able to bootstrap alignments and create shorter segmentations for different systems by starting from simple monophone HMM/GMM (Gaussian mixture model) systems trained on the shortest utterances.

It is good to note that as Finnish is an agglutinative language, the WER results are not directly comparable to those of, say, English. Hirsimäki et al. (2006) found that as one long Finnish word corresponds to several English words the WER becomes multiplied. For this reason, we report also the CER results, which do not have this problem. Furthermore, some previous works (Enarvi et al., 2017) have used normalisation of colloquial Finnish words in order to mitigate the effect of various spelling variations on the WER results. However, this method is partly manual and thus not easily scalable to large corpora, and we did not use such normalisation. Additionally, the transcripts contain special markers (e.g. for noise and pauses) and some decisions should be made about them in speech recognition: either to predict them, or to simply discard them. We opted for the latter. Before calculating the WER and CER, we removed all the special tokens, such as ".laugh" as well as the dash symbols "-" that indicate dysfluencies in speech, for example false starts ("predi- presidentti" was changed to "predi presidentti").

4.1 Hybrid HMM/DNN ASR systems

The HMM/GMM approach and, later, the hybrid HMM/DNN approach have been popular in speech recognition for the last couple of decades. Although they are now outperformed by newer approaches (mainly end-to-end systems; see the next two Subsects. 4.2 and 4.3), they are still useful since they require relatively small training corpora, and versatile toolkits have been built around these approaches. Namely, the Kaldi (Povey et al., 2011) toolkit provides optimised "recipes" to train and apply ASR systems, which we used to train baseline systems with our data. We then use the best baseline system to align the text and audio and segment the speech. Since large end-to-end systems cannot handle long segments of speech (see Sect. 4.2), segmentation was necessary before we could train the end-to-end models.

In the first phase, we trained two models using mostly standard Kaldi recipes without hyperparameter tuning, one with a 100-h subset (denoted as initial-100 h-TDNN) and another with the complete transcribed training corpus (initial-1600 h-TDNN). To train the HMM/GMM system for monophones and triphones, we used the Kaldi WSJ recipe. This recipe trains the initial monophone model on the shortest

Table 4 Sizes of the language models and their training corpora

Corpus	# of tokens	# of types	# of n-grams
LP 100 h transcriptions	898,700	20,700	400,700
LP 1600 h transcriptions	14,216,500	36,200	4,267,400
WEBCON + DSPCON + LP100 h	126,078,900	42,700	26,005,100
WEBCON + DSPCON + LP1600 h	138,991,000	45,600	28,853,000

The number of n-grams refers to the numbers of unigrams, bigrams, trigrams and 4-grams summed together

Table 5 Error rates of various ASR systems

Training set	System details	Dev set		Test set	
		WER (%)	CER (%)	WER (%)	CER (%)
100 h	Initial-GMM	44.80	18.43	48.94	20.94
	Initial-TDNN	29.16	9.01	32.58	11.04
	+ ext. LM data	26.88	8.46	30.48	10.49
	Semisup-TDNN + ext. LM data	25.37	7.89	28.16	9.78
	Wav2Vec2 + CTC (no LM)	22.50	6.08	24.03	7.02
	Wav2Vec2 + CTC + ext. LM data	20.34	5.83	21.75	6.80
1600 h	Initial-GMM	37.08	15.61	40.87	17.19
	Large-GMM	35.36	14.34	38.99	16.33
	Initial-TDNN	22.09	6.52	24.00	7.64
	TDNN (large-GMM alignments)	21.98	6.47	23.88	7.59
	+ ext. LM data	21.77	6.40	23.82	7.52
	AED	28.80	12.15	34.87	17.04

Larger LMs were trained on external LM data, namely the WEBCON corpus and the DSPCON transcriptions, in addition to the AM training set (either 100 h or 1600 h) transcriptions. All HMM/DNN system LMs are subword-based 4-gram models. The Wav2Vec2 + CTC system uses a word-based 4-gram language model trained on the 1600 h LP transcripts and the external data

utterances in the data, which helps in bootstrapping the alignments. As a deviation from the standard WSJ recipe, we trained the final triphone system using the discriminative, MMI (Bahl et al., 1986) training criterion, which is available as an optional addition in the WSJ recipe. The time-delay neural network (TDNN) (Peddinti et al., 2015; Waibel et al., 1989) models were trained using the HMM/GMM alignments. The TDNN architecture and other hyperparameters were adopted from the Switchboard recipe, since this trains a larger neural network, more suitable for the large training corpus. The TDNN has 15 layers with a dimension of 1536 and a bottleneck dimension of 160. In total the TDNN has about 17M parameters.

Using the SRILM (Stolcke, 2002) toolkit, we trained 4-gram language models (LMs) on the Lahjoita puhetta (LP) 100-h training corpus transcripts, the whole 1600 h training corpus transcriptions, as well as on a corpus of the LP transcripts

pooled with other available colloquial Finnish text corpora, namely the WEBCON corpus and the DSPCON transcriptions. The systems that utilised the external language modelling data are marked with "ext. LM data" in Table 5. We used the Morfessor (Creutz & Lagus, 2002, 2007) toolkit to segment words into subword units. We trained the morfessor model using the same LP transcripts appended with the WEBCON and DSPCON corpora as for the large LMs, with a corpus weight of 0.05. The resulting sizes of the LMs and their training corpora are listed in Table 4. We also trained LMs with a word vocabulary, but subword units yielded better results. For example, the word-based initial-1600 h-TDNN system got a WER of 25.12% on the test set, compared with 24.00% using subword units, so we opted to use subword units in the remaining experiments. The sizes of the training corpora are listed in Table 4. For more details about the language models, see the published recipes.

We used the initial-1600 h-TDNN to segment the training data, so the data could be used for training the E2E ASR systems. The initial-100 h-TDNN with large LM was used to generate transcriptions for the rest of the training corpus, which we then used for training the topic and dialect classification systems (see Sect. 5).

After training the initial ASR systems, we made some simple hyperparameter tuning for the HMM/GMM system to get an idea of how much room for improvement there is, compared with the Kaldi WSJ recipe. The tuning experiments focused mainly on increasing the number of parameters of the GMMs. By increasing the number of Gaussians from 4200 (in the WSJ recipe) to 14,000, and the number of leaves per Gaussian from 40,000 to 200,000, the penultimate, speaker-adaptive triphone system WER on the development set decreased from 42.86% to 39.71%. Training the MMI triphone system on top of the alignments from these systems, the WERs decreased to 37.08% and 35.36%, respectively for the smaller and larger GMM/HMM system. Finally, training the TDNN system on top of these MMI triphone models, the word error rates dropped to 22.09% (smaller GMM/HMM) and 21.98% (larger GMM/HMM) for the dev set and 24.00%/23.88% for the test set.

Decoding with a large language model trained on external data brings additional improvement compared with the LM trained on 100 h transcriptions (see the second and third row in Table 5). However, the 1600 h transcriptions seem to be enough to train a decent language model, and adding external data only brings a small improvement in WER and CER results (see the last two rows in Table 5). It is good to note, however, that the external text data is not exactly in the same domain as the test corpus, although it is colloquial in style.

Additionally, we wanted to demonstrate that the sizeable untranscribed portion of the corpus can be leveraged via semi-supervised training. For this experiment, we choose the approach presented in (Manohar et al., 2018). To demonstrate that the recordings without annotations could be used for improving the ASR systems, we started the semi-supervised training by generating transcriptions of the additional data with the *initial-100 h-TDNN*. Afterwards, we pooled the self-supervised portion (approx. 1587 h) and the 100 h set for the model training. The resulting model (*semisup-100 h-model*) had the same architecture as the *initial-100 h-TDNN* to ensure a fair comparison. From the achieved results (see Table 5), we can conclude that the additional unsupervised data is indeed

valuable, the error rates dropped significantly. On the other hand, we can also see that having more, accurately transcribed data is far more beneficial. The *initial-1600 h-TDNN* outperforms the semi-supervised system by a large margin, and the hyperparameter tuning offers some additional improvements.

4.2 AED ASR systems

Various end-to-end ASR approaches, such as Connectionist Temporal Classification (CTC) (Graves et al., 2006), the Recurrent Neural Network Transducer (RNN-T) (Graves et al., 2013), and Attention-based Encoder–Decoder (AED) (Bahdanau et al., 2016; Chan et al., 2016) models became popular in the 2010s, both in research as well as industrial applications. We train AED models on the transcribed data to serve as end-to-end baselines. Our AED models are trained with the SpeechBrain toolkit (Ravanelli et al., 2021). They consist of a stack of convolution, recurrent, and feed-forward layers in the encoder, a location-and-content aware attention mechanism, and recurrent layers in the decoder with altogether ≈ 28M parameters. The inputs are log-Mel-filterbank-energies and for each output step the network computes a distribution over a vocabulary of 1750 SentencePiece subword units. We trained with dynamic batching, targeting 50 s of audio per batch altogether, for 100 nominal epochs of 10,000 updates each. For the first 20 nominal epochs the encoder learning was aided by using an additional multi-task CTC loss (Kim et al., 2017). We do not use any external language with our AED system, making it fully end-to-end. For further details we refer to the published recipe.

End-to-end models seem to have difficulties with long-form speech, both in learning as well as in generalising (Chiu et al., 2019; Narayanan et al., 2019). Our preliminary experiments with AED systems showed similar issues. Models would not converge with full length utterances. Via segmentations produced with the HMM-based ASR systems, we split the data into shorter utterances. Training converges well on short (up to 10 s) segments and slightly slower on medium length (up to 50 s) segments. Decoding an ad-hoc segmented version of the development set yields a WER of ≈ 22% on both models. However, on the official development set, which has longer utterances, both models have pathological behaviour on a minority of utterances, which increased the error rate considerably. Similar to reports by Keung et al. (2020), our models produce echographic output, i.e. the model repeats a single token or in some cases a long sequence of tokens. The model trained on medium length segments suffers less, so we choose it as our final baseline. Additionally, we implement a simple post-processing filter where we allow repetitions to produce in total a maximum of five tokens. On the development set, this modifies 70 transcripts. This reduces the WER from 45.82 to 28.80%—echographic transcripts account for a significant amount of errors. Listening to the utterances which produced echographic output reveals that these utterances are long, in some cases noisy, and in some cases contain long pauses. Despite the post-processing, our AED baselines fall behind

their HMM/DNN counterparts in performance in Table 5. Due to the initial difficulties with long-form speech, we did not make a system for the 100 h subset.

4.3 Pre-trained Wav2Vec2 fine-tuned with CTC

In recent years, large machine learning models [also called *foundation models* (Bommasani et al., 2021)] that are pre-trained on vast numbers of general-domain, unlabelled data and fine-tuned on downstream tasks with labelled data have achieved state-of-the-art results especially in language processing tasks [for example, Brown et al. (2020), Devlin et al. (2019)]. This *transfer learning* method has recently been shown to be useful in speech recognition as well: models such as *Wav2Vec2* (Baevski et al., 2020) and *HuBERT* (Hsu et al., 2021) are currently used in many systems that achieve SOTA accuracy in speech recognition benchmarks[11]. In this section, we describe how the new dataset can be used to fine-tune a pre-trained *Wav2Vec2* model to create an ASR system that outperforms our other systems.

Wav2Vec2 is a self-supervised framework which learns deep acoustic representations by leveraging large amounts of unlabelled acoustic data. After pre-training on untranscribed speech, the model can be fine-tuned on labelled acoustic data for a downstream task, such as ASR. Fine-tuning for the ASR starts with adding a randomly initialised classification layer on top of the model with classes representing the characters of the target language alphabet and a word boundary token. The model is then optimised with a CTC loss.

In this work, we experimented with a *Wav2Vec2 Large* model (317M parameters) pre-trained on the multilingual VoxPopuli (Wang et al., 2021) corpus. The corpus is composed of 100K h of untranscribed European Parliament plenary session recordings in 23 languages, including 4.4K h of Finnish speech. We fine-tuned this model with CTC on the 100-h subset for 80 epochs with an effective batch size of 48 and a learning rate of 5e−4. We used full length utterances with durations up to 50 s and the segmented recordings for the rest of the training data. We also tried to fine-tune the model on the 1600h set, but it took too much time on our hardware, so we left fine-tuning on the full training set to future work.

The fine-tuned model [see *Wav2Vec2+CTC (no LM)* in Table 5] achieved WER of 22.50% and 24.03% and CER of 6.08% and 7.02% on the development and the test set, respectively. We also incorporated an external language model in order to further improve the model performance. The LM was trained on the 1600 h LP transcriptions and external (WEBCON and DSPCON) data. The dataset included about 84M word tokens and 2.6M word types, and the LM included 3.5M n-grams. With a word-level 4-gram LM (see *Wav2Vec2+CTC + ext. LM data* in Table 5), the word and the character error rates dropped to 20.34/5.83% on the development set and 21.75/6.80% on the test set. In addition, we plan to incorporate the subword-based LM in future experiments, since it provided an improvement in WER compared to word-based LM for some HMM/DNN ASR systems.

[11] For a collection of speech recognition benchmarks, see https://paperswithcode.com/task/speech-recognition.

Table 6 Pearson product-moment correlation coefficients between the WER and the total duration of speech in the training corpus

Metadata type	Correlation coefficient	N	p value
Age	− 0.685	11	0.020
Gender	− 0.618	4	0.382
Dialect	− 0.267	21	0.255
Theme	− 0.829	8	0.011
Device	1	2	–

The classes for age and gender are those specified in Fig. 2, including the N/A classes. For the themes, the "Media Skills" classes were combined as one class. For the dialects, we used the 21 original classes for these calculations

4.4 Analysis of ASR accuracy w.r.t speaker metadata

The rich metadata of Lahjoita puhetta (see for example Fig. 2) allows us to examine differences in the ASR accuracy between speech from different groups of people, different recording devices and on different topics. In recent years, a new research area has emerged that investigates the discriminatory performance of AI systems and its causes (Garnerin et al., 2021; Hovy & Spruit, 2016). In the ASR field, traditional metrics like the aggregated WER and CER are used to measure the overall performance of the models. As we will note in this section, these metrics can hide biases that a model develops during training. To build an excellent general ASR system, we ought to mitigate the risk of the system having a systematically worse recognition rate for any speaker category (for example, gender, age, or dialect). The results we analyse in this section can point us to the weaknesses of the (hybrid HMM/DNN) ASR system trained on the Lahjoita Puhetta dataset, and aid us in future de-biasing efforts.

Dividing the 10-h test corpus into each metadata class yields quite small subsets. To get a larger test corpus for each metadata class, we decoded the rest of the transcribed dataset using the initial-100 h-TDNN ASR system with external LM data (see Sect. 4.1). In this case there is overlap between training and test corpus speakers, although no overlap between the recordings. The average WER for this large set was 26.13% which is a little better than for the fully independent 10-h test set (30.48%, as listed in Table 5).

A basic assumption is that the more training data there are from a specific group the better the speech recognition results are for this group. This means the correlation between the number of training data and WER result should be negative. Table 6 lists the Pearson product-moment correlation coefficients between the WER and the amount of training data, for each metadata type. All of the metadata types have quite a small number of data points (classes, N) to calculate the correlation, and only age and theme have $p < 0.05$. In general, this low N values will not give us very reliable correlation results, but we provide these numbers for some rough indication of how much the metadata class affects speech recognition. Age, gender and theme have the expected results, with quite a strong negative correlation. Dialect

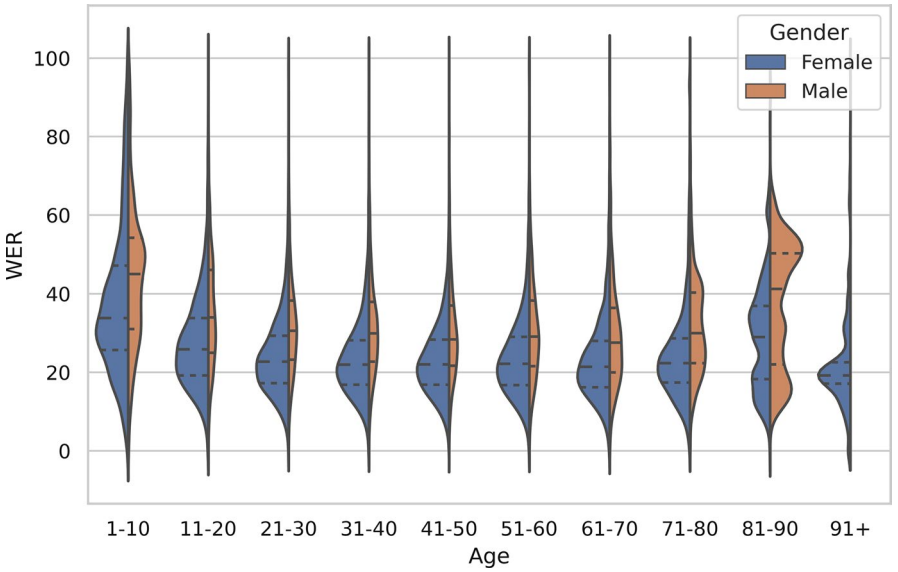

Fig. 5 The distribution of WERs in the test set w.r.t. the age and gender of the speaker

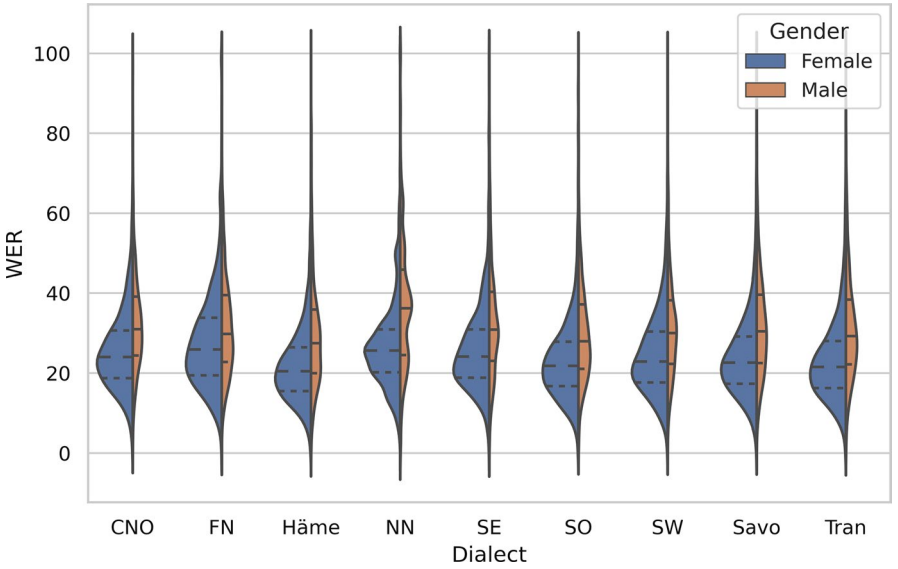

Fig. 6 The distribution of WERs in the test set w.r.t. the dialect and gender of the speaker

has a weak negative correlation, if any, and for the recording device (phone vs. PC), WER correlates positively with the amount of training data.

Figures 5 and 6 enable a more detailed analysis of the results for the metadata groups of gender, age and dialect. The difference between the number of males

and females in the training data is large, which presumably affects the ASR results, although the p-value for the correlation is high because of the low N. The average WER for females, 24.12%, is below the overall average (26.13%) while for males the average WER is well above: 31.78%. Similarly, the number of recordings on a theme in the training corpus correlates with the speech recognition accuracy.

The dialect does not seem to have much of an effect on the speech recognition results. The large dialect groups, Savo and Tran, do not have significantly better results than the average: the WERs are 25.99% and 25.13% respectively. This can be seen also from the correlation coefficient in Table 6, which is not significant. The exception is the group of non-native Finnish speakers, which has a high WER of 30.97%.

From these experiments, the reasons for the differences in the WER results are not entirely clear, where differences exist at all. In general, a larger relative share in the training corpus results in better ASR performance, if there is systematic variation between groups, but other factors presumably affect the results too. For example, there is fewer training data for the speech of young children which might be one reason for the relatively poor ASR performance, but children probably speak less clear Finnish than adults, which also makes speech recognition more difficult. This could apply also to other groups, such as non-native Finnish speakers. Furthermore, speech recorded on smart phones has better ASR accuracy (WER: 24.57%) than speech recorded on a computer (WER: 27.27%) even though there are fewer phone recordings than computer recordings. We speculate that the reason for this is that phones are better than computers, on average, at recording speech.

We should also question whether a user's background is audible in the speech at all, and whether the users have answered the metadata questions accurately. For example, the variability of the speakers' dialects may not be captured by the labels that we used. The users are asked "What dialectal region has affected the most the way you speak?". Even if the dialectal background of a user is very varied (e.g. because they have lived in multiple dialect regions during their lives), they might answer just one region to this question, which can over-simplify this metadata. Similarly, the age bracket of a speaker is not very easy to identify from the voice even for human listeners, at least for certain age groups around middle-age. Furthermore, the metadata classes, given to the user to select an answer from, do not necessarily represent variation in the speech styles of the speakers in the best possible way (a better categorisation might be found e.g. by using classes learned by a machine learning model), but we assume the categories do tell us something about the style, content or the acoustic properties of the speech. In the next section we test this assumption: if the metadata categories are audible in speech, we should be able to build automatic classifiers that recognise the metadata categories that a speech recording belongs to.

5 Gender, age, dialect, and topic classification

Using the metadata, we can build various metadata classifiers, which can later be used in different applications, such as: filling the missing metadata, verifying the correctness of the available metadata, enhancing the speech processing applications

Table 7 Hyperparameters of the audio encoder

Layer	Input size	Output size	Context	Dilation
TDNN 1	40	512	5	1
TDNN 2	512	512	3	2
TDNN 3	512	512	3	3
TDNN 4	512	512	1	1
TDNN 5	512	1500	1	1
Statistical pooling	1500	3000	/	/
Linear	3000	512	/	/

Table 8 Accuracy of the models on the gender classification task

Model	Test	Multi-transcriber test
3 s model	90.03	99.59
50 s model	92.65	99.59

with speaker information, and bias detection. For that purpose, we built and benchmarked baseline models for gender, age, dialect, and topic classification.

The models are built using a 5-layer TDNN with dilated connections, followed by statistical pooling and two linear layers. This is similar to the x-vector models (Snyder et al., 2018). We will call this part audio encoder. For the dialect and topic classification tasks, besides the models trained on audio-only, we additionally trained models that utilise the available transcripts. We did that using an additional text encoder. In the text encoder, word embeddings are extracted using the FinBERT model (Virtanen et al., 2019) and processed through a bi-directional long short-term memory (BLSTM) network (Hochreiter & Schmidhuber, 1997). In the last stage, the outputs of the audio and text encoders are concatenated and passed through a softmax function which produces class probabilities.

As input features, we extracted logarithmic-Mel-filterbanks with 40 filters, using 25 ms window with a stride of 10 ms. To improve the signal-to-noise ratio, we applied mean normalisation to each sample.

The hyperparameters for the audio encoder are given in Table 7. The text encoder is a 2-layer BLSTM with an input size of 768 and an output of 512. As optimiser, we used Adam (Kingma & Ba, 2014), with a learning rate of 1e−4 and a cross entropy loss.

5.1 Gender classification

Gender information plays an important role in many applications, from speech processing (Abdulla et al., 2001) to bias detection (Park et al., 2018). Thus, having a good gender classifier can help us enhance the speech processing models, as well as aid us in detecting the biases related to gender, that those models may contain. For that purpose, we built two gender classifiers, using different segment lengths.

Table 9 Accuracy of the models on the age classification task

Model	Accuracy		Relaxed accuracy	
	Test	Multi-transcriber test	Test	Multi-transcriber test
3 s model	33.59	40.16	79.28	78.48
50 s model	42.39	52.66	86.34	89.55

The audio samples used to train the first model are cut to 50 s. The reason for not using the whole audio samples is that some of them might be too long to process. Additionally, the 50 s limit of the audio should contain a sufficient amount of information for the model to learn the task.

The gender classification models usually work with small (few seconds) audio segments, whereas the average length of the audio samples in our dataset is about 40 s. To make our model more comparable to the others, we constructed another model that uses audio segments up to 3 s. This choice is expected to degrade the performance of the model but will make it more reusable to other applications, where long segments are not available.

In Table 8, we can see how both models performed in terms of accuracy on the test sets (see Table 1 for set descriptions). From the results, we can observe that on the test set, the model using up to 50 s segments performs slightly better than the one using 3 s segments. This is expected, considering that longer segments contain more information. On the multi-transcriber test set, on the other hand, both models perform equally well, achieving almost perfect accuracy score. The significant difference in performance between both test sets could be attributed to the disproportion between male and female speakers. The multi-transcriber test set has many more female speakers than male, and as we will see later, the system is better at detecting the female speakers.

5.2 Age classification

Like gender, age information can also be beneficial in many areas. The age of the speaker can have a large impact on the performance of the ASR system (Wilpon & Jacobsen, 1996). Having a good age classifier can help us find which age group the ASR system struggles with the most, allowing us to improve the model on that end. Additionally, the age information can provide us with clues related to age biases that the model might contain.

The age classification is a challenging task since there is no clear boundary that separates one age class from its neighbouring classes. For example, it is almost impossible to find a difference in speech between a 38-year-old person (age group 31–40) and a 41-year-old person (age group 41–50). Due to that, besides the standard accuracy metric, we also used relaxed accuracy, where the neighbouring classes are also considered as correct predictions.

Table 10 Accuracy of the models on the dialect classification task

Model	Test	Multi-transcriber test
Whole audio	40.74	35.14
Audio + transcripts	32.66	29.20
Audio + ASR transcripts	29.19	30.97
Audio subset	39.73	38.83

For this task, we also developed two models similarly as we did in the gender task. One operating on up to 3-s segments, and another operating on longer, up to 50-s segments. This will give as a clue about what segment lengths are sufficient for learning the task.

In Table 9, we can see the performance of both models on the test sets, using the standard and the relaxed accuracy. From the results, we can see that the model using 50 s segments performs significantly better, which indicates that more information is required for the model to learn this task. Additionally, by using the relaxed accuracy, we gained a large improvement, which suggests that most of the mistakes happen by confusing the actual class with one of the neighbouring classes.

5.3 Dialect classification

The participants in the Lahjoita puhetta campaign were encouraged to use their dialect and provide that information when recording the audio. Automatic dialect classification for Finnish is a challenging and underexplored task. The only previous attempt of combining audio and text modalities for Finnish dialect classification is a system combining FinBERT embeddings and a pre-trained Wav2Vec2 model, achieving good results (Hämäläinen et al., 2021).

Since traces of dialect do not occur in every word (or even sentence), we used longer segments for the dialect classification task. We limited the samples to up to 50 s. The reason that we did not use the whole audio is that some samples can be multiple minutes long, which makes them hard to process.

Besides the acoustic information, for this task, we additionally experimented with enriching the input with morphological information by utilising the transcripts. To utilise both the audio and the transcript information, we used only the audio files that have corresponding transcripts. In this experiment, instead of cutting the audio to 50 s, we discarded the samples that are longer than that. We did so in order for the transcripts to match the audio.

To see if adding the transcripts has any benefit, we trained an audio-only model on the same samples as the model using audio and transcripts (we will refer to this as "audio subset").

Lastly, instead of using the original transcripts, we experimented with the decoded transcripts from the initial-100 h TDNN model (see Sect. 4.1). This model is trained on the same data as the model utilising audio and transcripts, except the 100 h used for training the ASR model. This will give us an opportunity to

Table 11 Accuracy of the models on the topic classification task

Model	Test	Multi-transcriber test
whole audio	65.65	73.93
transcripts	82.06	88.65
ASR transcripts	82.06	86.52
Audio + transcripts	81.34	87.94
Audio + ASR transcripts	80.14	86.52
Audio subset	57.10	69.47

investigate how much the performance differs on ASR-generated transcripts and whether it is a good idea to decode the untranscribed part of the data and train the model on the whole audio and the ASR-generated transcripts.

The accuracy of the models is given in Table 10. Looking at the results, we can observe that the model trained on all the audio performs better than the one trained on the audio and the available transcripts. This could indicate that the dialect information is predominant in the audio since the transcripts are not able to capture information such as pronunciation and accent. Additionally, we can observe that using the ASR transcripts degrades the performance on the test set, but it improves it slightly on the multi-transcriber test set, in comparison to using the original transcripts. This could mean that the words affected by the dialect are also difficult for the ASR model, resulting in incorrect transcriptions. Further, the audio subset model performs better than its counterpart that additionally uses the transcripts. This could indicate that instead of providing additional information, the transcripts introduce noise to the model.

Generally, the accuracy of the models is relatively low in comparison to the other metadata classification tasks. This indicates that the dialect classification in this dataset is a very difficult task and more advanced methods might be required in order to get optimal results. In general, the use of dialects in Lahjoita puhetta may be weaker and less frequent than in datasets where the particular focus on dialects may have affected the choice of participants and collection methods.

5.4 Topic classification

During the collection of Lahjoita puhetta, the participants were asked to choose a theme and then talk about topics within the theme. Due to the large number of topics, we used the themes (listed in Sect. 2.1) as labels, with the only difference being that we combined the three "Media skills" themes into one.

Similar to the dialect classification, for this task we also cut the audio segments to 50 s and trained an audio-only model on the whole data.

Topic classification is often done on text. For that purpose, we trained a text-only model on the samples that are 50 s or less. Additionally, we tried utilising the acoustic and the morphological information by processing the audio and the transcripts together, just like in the dialect classification task. Furthermore, we investigated the

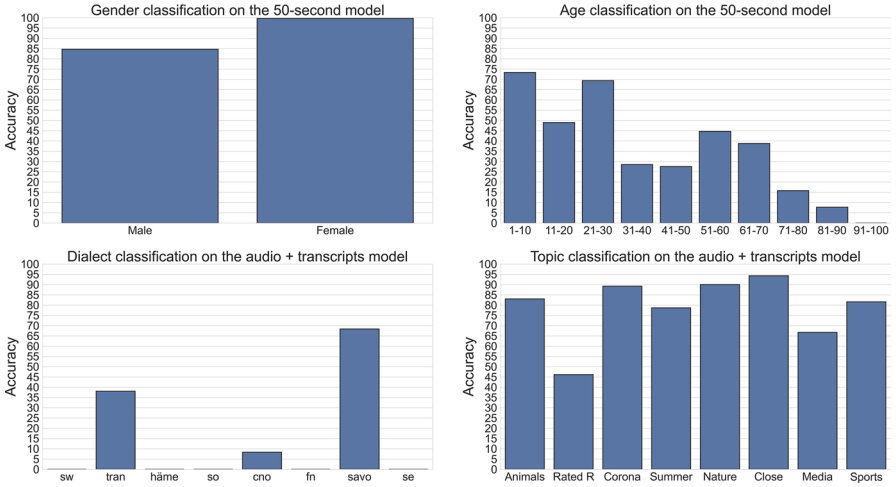

Fig. 7 Metadata accuracy per class on the test set

performance of the model, when provided with ASR decoded transcripts, instead of the gold-standard ones. The ASR transcripts are generated using the same initial-100 h TDNN model as the one in the dialect classification task.

Lastly, we investigated the effect of the audio on the topic classification task. For that purpose, we developed an audio-only model that is trained on the same data as the models using the original transcripts.

The results for the topic classification task are given in Table 11. From the table, we can observe that the model that uses the original transcripts achieves slightly better results than the one using the ASR-generated transcripts on the multi-transcriber test set, whereas on the test set, they perform identically. Additionally, the models using only the transcripts achieve significantly better results than the model using the whole audio, even though the audio-only model was trained on far more data. When jointly using the audio and the transcript information, we can see that there is a small degradation in comparison to using only the transcripts. This could indicate that the audio does not provide any additional information that would help the model. Another thing to consider is that the audio encoder that we are using is quite small, so a bigger model might be necessary if we want to benefit more from the acoustic information. When we combined the audio and the ASR-generated transcripts, we observed only a small degradation in the performance, in comparison to using the audio with the original transcripts. This could indicate that certain keywords affect the topic classification and the ASR system is good at detecting them. Using this knowledge, in future experiments we can generate transcripts for the untranscribed part of the data and use them in addition to the audio, to train a big model that utilises audio and transcript information. From the results obtained on the model trained on the subset of the audio, we can see that there is a significant degradation in the results in comparison to the model that uses only the transcripts. This confirms that the textual information content is sufficiently dense for this task.

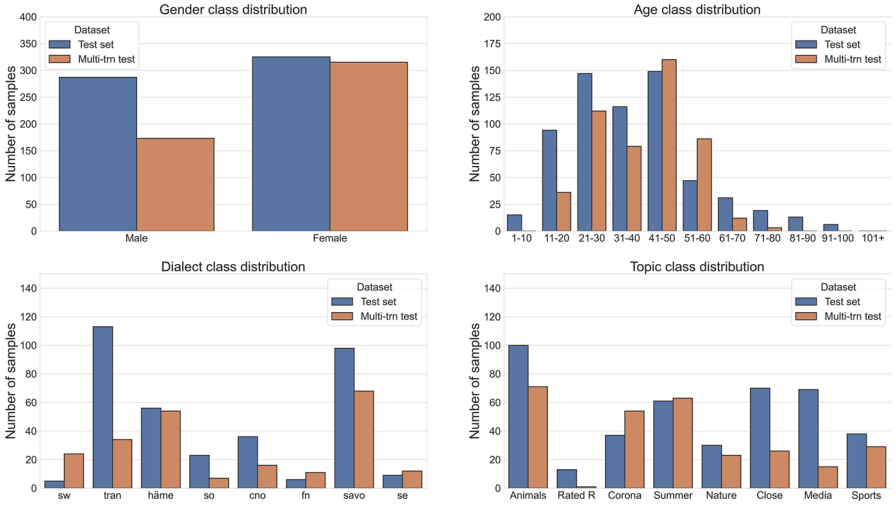

Fig. 8 Metadata class distribution for the test sets

Generally, the models were able to learn the task relatively well, while still leaving some space for improvement, especially on the audio side.

5.5 Analysis of metadata classification errors

To further investigate which classes are challenging for the metadata classification models, we evaluated them on each class individually. The results of the analysis are given in Fig. 7. Additionally, in Fig. 8 we can observe the number of samples per class that were used during the evaluation.

In the gender classification case, we can see that the model performs significantly better on the female examples. The reason could be that there is a high disproportion between male and female samples in the training set.

On the age classification task, we can observe that the model performs better on the lower age groups and struggles with the elderly, especially the ones in the 91–100 age group, where the model misclassified all the samples.

On the dialect classification plot, we can see that the model misclassified all the samples from several dialect groups. This is not surprising, considering that many of those dialect groups have only a couple of samples and the general accuracy of the model is low. To further investigate the mistakes that the model made on this task, we plotted a confusion matrix, presented in Fig. 9. From the matrix, we can see that the HÄME dialects are mostly confused with TRAN and SAVO, which are neighbouring dialects in our dialect grouping. Similar observations can be made with the CNO dialect group, which is mostly confused with its neighbouring SAVO group.

On the topic classification plot, we can see that the model is performing well on almost all the classes. The weakest one seems to be the Rated R class, which generally has a low number of samples in the training and testing sets.

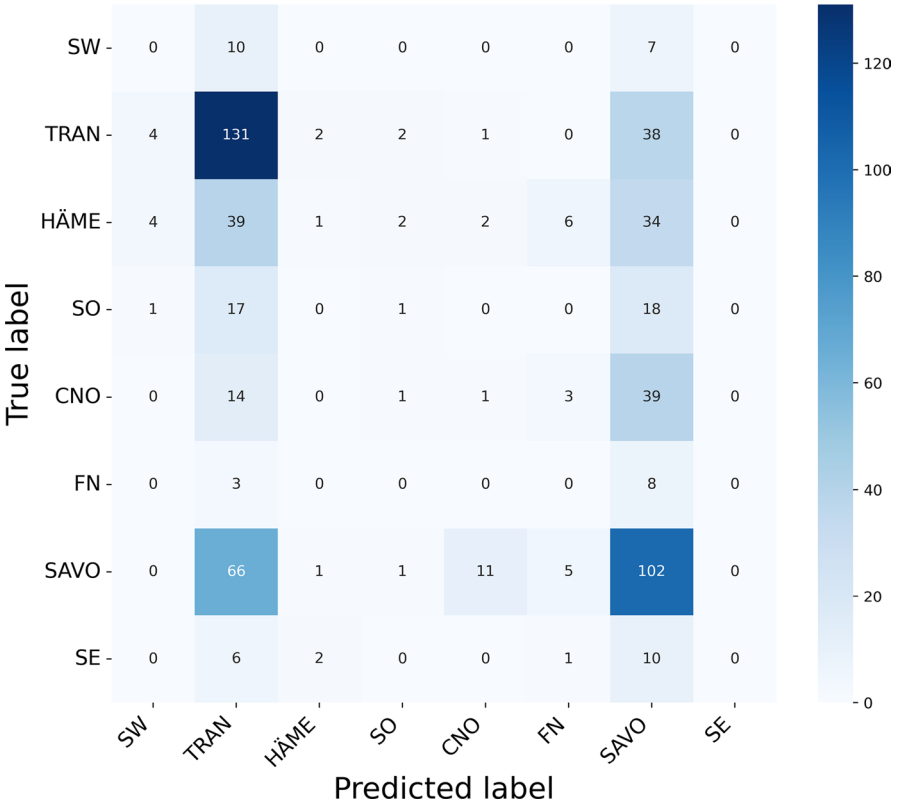

Fig. 9 Confusion matrix for the dialect classification model on the test set

6 Possible future directions

Although we have demonstrated with multiple use-cases the usefulness of the collected corpus in this article, there are still numerous possibilities to utilise the dataset. Out of those possibilities, we plan to realise a few in the near future. Perhaps the most evident utilisation of the corpus is a speaker recognition system. The large number of speakers of various ages speaking different dialects would enable us to build a robust and accurate model for Finnish data. The carefully transcribed portion of the data would make an interesting resource for colloquial Finnish text to speech (TTS) systems. We hypothesise that the marked non-speech parts and disfluencies could be leveraged to create a more natural TTS that can hesitate, restart words, and make non-speech sounds at the appropriate places. The AED ASR experiments uncover clear difficulties with the long-form recordings in this dataset. The results on an ad-hoc segmented version of the development data portion were on par with the HMM/DNN systems' results on the official data, which suggests that solving these technical difficulties would make AED systems viable approach for this data. The last future direction that we wish to mention concern the untranscribed part of

the dataset. We have already demonstrated that it can be used for semi-supervised learning, and we plan to investigate its usefulness with other self-supervised and unsupervised methods. Specifically, we intend to build a truly Finnish Wav2Vec2 model, which would be pre-trained on purely Finnish speech and fine-tuned with the large transcribed part of the corpus.

A similar effort for large-scale collection of donated speech for other languages such as the second national language of Finland, i.e. the variety of Swedish spoken in Finland is already on-going. Efforts for applying this collection concept and tools for collecting minority languages spoken in Finland are also planned.

7 Conclusions

In this paper, we presented a new, large-scale, conversational Finnish speech corpus. The 3600 h, out of which 1600 are transcribed, include over twenty thousand speakers from all age groups and from all the regions of Finland. To ensure the high quality of the transcripts, the transcribers were evaluated using manual and automatic techniques. The techniques for data collection and annotation applied in this paper provide a resource for future similar attempts at collecting large-scale data.

To establish that the *Lahjoita puhetta* dataset is useful for training ASR systems, we built several hybrid HMM/DNN and end-to-end baseline models and made them publicly available. The varied ASR experiments, with the best system achieving 21.75% WER, showed that the dataset is suitable for building such systems. This should also be compared with the initial word-level inter-annotator disagreement ranging between 13 and 20% on this data type. Furthermore, the large untranscribed part of the corpus can be utilised for unsupervised and semi-supervised training. The rich metadata provided by the participants allowed us to successfully train various metadata classification models, demonstrating further use-cases for the dataset. The benchmark metadata classification models are publicly released together with the ASR models.

The large and diverse *Lahjoita puhetta* dataset will be freely available for research purposes, and for commercial use at a cost price. We hope this encourages researchers and companies to further develop language technologies and bridge the gap between research and commercial use.

Appendix

Transcriber instructions

The transcribers were given the following guidelines:

Speech produced continuously is written on the same line without line breaks. A line break is marked at each point where the speaker clearly pauses. In addition to line breaks, breaks or their durations are not marked or separated in any other way. The Finnish alphabet (a-zåäö) is used to record the verbal content of speech. Normal spelling punctuation, such as periods, commas, or question marks, is not included

in the transcription. Numbers are not used either. Words are written as accurately as possible in the exact form in which the speaker produced them, e.g., write "pitsaa", not "pizzaa". Do not attempt to correct any mistakes made by the speaker. Do not add any additional comments to the text. All words can be written in lower case. If the word is clearly a proper noun, a capital letter may be used. However, words beginning an utterance or a sentence are not capitalised.

Hyphens or periods are used for punctuation only in the following special cases: A hyphen indicates, for example, a missed or "incorrectly started" word, e.g. "predi-president" or a word from which only the remainder can be heard, or from which the speaker speaks only the remainder: "-sident." In the case of a compound whose suffix ends in the same vowel in which the suffix begins, a hyphen may be used between the parts: "tila-autolla" ("with a minivan").

When there is a point in the speech where the speaker makes vague fill-in or hesitation sounds, pouts, coughs, laughs, yawns, or sighs so that the sound is clearly heard, and the speaker does not produce the speech at the same time, the sound can be marked with, for example: *.fp* (filled pause) can mark a complex or ambiguous fill or hesitation sound that is not sufficient to describe "mm", "aa" or "öö", *.ct* (clear throat), *.cough* (coughing), *.laugh* (laughing), *.yawn* (yawning), *.sigh* (sigh, loud inhalation and exhalation), *.br* (breath, single clearly audible in- or exhalation sound).

However, if the speaker, for example, laughs or yawns while speaking, do not try to include the laughter or yawn in the transliteration of the speech (for example, using the letters h). In such situations, precisely transcribing is not useful for the purpose of the material. The aim is to transcribe only the verbal content of the speech and, if necessary, the sounds to be heard between the words.

Acknowledgements This work was partly funded by Academy of Finland (Grant Numbers 337073, 329267, 322625 and 345790). The computational resources were provided by Aalto ScienceIT.

Funding Open Access funding provided by Aalto University.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abdulla, W., Kasabov, N., & Zealand, D. N. (2001). Improving speech recognition performance through gender separation. *Changes, 9,* 10.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 12449–12460). Curran Associates Inc.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877–1901.

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher corpus: A resource for the next generations of speech-to-text. *LREC, 4,* 69–71.

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP), 4*(1), 1–34.

Enarvi, S., Smit, P., Virpioja, S., & Kurimo, M. (2017). Automatic speech recognition with very large conversational Finnish and Estonian vocabularies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(11), 2085–2097.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., & Pylkkönen, J. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language, 20*(4), 515–541.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451–3460.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*(3), 328–339.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4945–4949). https://doi.org/10.1109/ICASSP.2016.7472618

Bahl, L., Brown, P., De Souza, P., & Mercer, R. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *ICASSP'86. IEEE international conference on acoustics, speech, and signal processing* (Vol. 11, pp 49–52). IEEE.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint. http://arxiv.org/abs/2108.07258

Carrive, J., Beloued, A., Goetschel, P., Heiden, S., Laurent, A., Lisena, P., Mazuet, F., Meignier, S., Pincemin, B., Poels, G., & Troncy, R. (2021). Transdisciplinary analysis of a corpus of French newsreels: The ANTRACT Project. *Digital Humanities Quarterly, 15*(1). https://hal.archives-ouvertes.fr/hal-03166755, editors: Taylor Arnold, Jasmijn van Gorp, Stefania Scagliola, and Lauren Tilton.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4960–4964). https://doi.org/10.1109/ICASSP.2016.7472621

Chiu, C. C., Han, W., Zhang, Y., Pang, R., Kishchenko, S., Nguyen, P., Narayanan, A., Liao, H., Zhang, S., Kannan, A., Prabhavalkar, R., Chen, Z., Sainath, T., & Wu, Y. (2019). A comparison of end-to-end models for long-form speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)* (pp. 889–896). https://doi.org/10.1109/ASRU46091.2019.9003854

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on morphological and phonological learning, association for computational linguistics* (pp. 21–30). https://doi.org/10.3115/1118647.1118650

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies* (Vol. 1 (Long and Short Papers), pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Enarvi, S. (2018). *Modeling conversational Finnish for automatic speech recognition*. PhD Thesis, Aalto University.

Garnerin, M., Rossato, S., & Besacier, L. (2021). Investigating the impact of gender representation in ASR training data: A case study on librispeech. In *Proceedings of the 3rd workshop on gender bias*

in natural language processing (pp. 86–92). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gebnlp-1.10

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, IEEE international conference on, IEEE computer society* (Vol. 1, pp. 517–520).

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on machine learning* (pp. 369–376). ACM.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649).

Hämäläinen, M., Alnajjar, K., Partanen, N., & Rueter, J. (2021). Finnish dialect identification: The effect of audio and text. arXiv preprint. http://arxiv.org/abs/2111.03800

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th annual meeting of the association for computational linguistics*(Vol. 2: short papers, pp. 591–598). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., & Kiessling, A. (2002). SPEECON—Speech databases for consumer devices: Database specification and validation. In *Proceedings of the third international conference on language resources and evaluation (LREC'02)* (pp. 329–333). European Language Resources Association (ELRA).

Keung, P., Niu, W., Lu, Y., Salazar, J., & Bhardwaj, V. (2020). Attentional speech recognition models misbehave on out-of-domain utterances. arXiv preprint. http://arxiv.org/abs/2002.05150

Kim, S., Hori, T., & Watanabe, S. (2017). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4835–4839). https://doi.org/10.1109/ICASSP.2017.7953075

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint. http://arxiv.org/abs/1412.6980

Lennes, M. (2009). Segmental features in spontaneous and read-aloud Finnish. In: deSilva, V. & Ullakonoja, R., (Eds.), *Phonetics of Russian and Finnish* (pp. 145–166). Frankfurt amMain: Peter Lang.

Lindén, K., Jauhiainen, T., Lennes, M., Kurimo, M., Rossi, A., Kurki, T., & Pitkänen, O. (2022). Donate speech: Collecting and sharing a large-scale speech database for social sciences, humanities and artificial intelligence research and innovation. In Witt, A., & Fisher, D. (Eds.), *Chapter 3.6—CLARIN book*. DeGruyter.

Manohar, V., Hadian, H., Povey, D., & Khudanpur, S. (2018). Semi-supervised training of acoustic models using lattice-free mmi. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4844–4848).

Narayanan, A., Prabhavalkar, R., Chiu, C. C., Rybach, D., Sainath, T. N., & Strohman, T. (2019). Recognizing long-form speech using streaming end-to-end models. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)* (pp. 920–927). https://doi.org/10.1109/ASRU46091.2019.9003913

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. arXiv preprint. http://arxiv.org/abs/1808.07231

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J. C., Yeh, S. L., Fu, S. W., Liao, C. F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., … Bengio, Y. (2021). Speechbrain: A general-purpose speech toolkit. http://arxiv.org/abs/2106.04624

Smith, G. (2008). Does gender influence online survey participation?: A record-linkage analysis of university faculty online survey response behavior. ERIC Document Reproduction Service No ED 501717.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5329–5333). IEEE.

Stolcke, A. (2002). Srilm—An extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. arXiv preprint. http://arxiv.org/abs/1912.07076

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (Vol. 1: Long Papers, pp. 993–1003). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.80

Wilpon, J. G., & Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings* (Vol. 1, pp. 349–352). IEEE.

Authors and Affiliations

Anssi Moisio[1] · Dejan Porjazovski[1] · Aku Rouhe[1] · Yaroslav Getman[1] · Anja Virkkunen[1] · Ragheb AlGhezi[1] · Mietta Lennes[2] · Tamás Grósz[1] · Krister Lindén[2] · Mikko Kurimo[1]

Dejan Porjazovski
dejan.porjazovski@aalto.fi

Aku Rouhe
aku.rouhe@aalto.fi

Yaroslav Getman
yaroslav.getman@aalto.fi

Anja Virkkunen
anja.virkkunen@aalto.fi

Ragheb AlGhezi
ragheb.alGhezi@aalto.fi

Mietta Lennes
mietta.lennes@helsinki.fi

Tamás Grósz
tamas.grosz@aalto.fi

Krister Lindén
krister.linden@helsinki.fi

Mikko Kurimo
mikko.kurimo@aalto.fi

[1] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

[2] Department of Digital Humanities, University of Helsinki, Helsinki, Finland