LowResourceEval-2021: a shared task on speech processing for low-resource languages

Elena Klyachko	Daniil Grebenkin	Daria Nosenko	Oleg Serikov
HSE University	NSU	NSU	HSE University
RAS Iling	NSU SDAML	NSU SDAML	DeepPavlov, MIPT

Abstract

This paper describes the results of the first shared task on speech processing for low-resource languages of Russia. Speech processing tasks are notoriously data-consuming. The aim of the shared task was to evaluate the performance of state-of-the-art models on low-resource language data as well as draw the attention of experts to field linguistics data (using Lingovodoc project data). The tasks included language identification and IPA transcription, with three teams participating in them. The paper also provides a description for the datasets as well as an analysis of the participants' solutions. The datasets created as a result of the shared task can be used in other tasks to enhance speech processing and help develop modern NLP tools for both speech communities and field linguists. **Keywords:** automatic speech recognition, language identification, minority languages, low-resource languages

DOI: 10.28995/2075-7182-2021-20-391-402

LowResourceEval-2021: дорожка по обработке речи для малоресурсных языков

Е. Клячко	Д. Гребенкин	Д. Носенко	О. Сериков
НИУ ВШЭ	НГУ	НГУ	НИУ ВШЭ
ИЯз РАН	ЛАПДиМО	ЛАПДиМО	DeepPavlov МФТИ

Аннотация

В статье описываются результаты первого соревнования по обработке речи для малоресурсных языков России. Задания по обработке речи, как правило, требуют больших объемов данных. Задачей соревнования было оценить качество работы современных моделей на данных малоресурсных языков, а также привлечь внимание экспертов к полевым данным (на примере данных проекта Lingvodoc). Задачи соревнования включали идентификацию языка и транскрипцию в МФА. В соревновании участвовали три команды. В статье описываются наборы данных, подготовленные в рамках соревнования, а также анализируются решения участников. Наборы данных могут переиспользоваться для улучшения обработки речи и развития инструментов NLP для языковых сообществ и лингвистов. **Ключевые слова:** автоматическая обработка речи, идентификация языка, малые языки, малоресурсные языки

1 Introduction

The paper describes the results of the first shared task on speech processing for low-resources languages of Russia.

Speech processing tasks are notoriously data-consuming. However, for most of the world's languages little spoken data such as news collections or audiobooks is available, not to speak of manually curated collections. Nevertheless, there are so-called field linguistics datasets, e. g. Paradisec[1], DOBES[2], ELAR[3], Lingvodoc[4]. These datasets have been created primarily for the purpose of language documentation and are often used in linguistic typology or for dialectological and historical linguistics studies. A

[1] https://www.paradisec.org.au

[2] https://dobes.mpi.nl

[3] https://elar.soas.ac.uk

[4] http://lingvodoc.ispras.ru

Shared task	Language	Training set (hours)	Test set (hours)	Number of teams
Interspeech-2018	Tamil	45	4.2	14
Interspeech-2018	Telugu	45	4.2	18
Interspeech-2018	Gujarati	45	5	18
GermEval-2020	Swiss German	70	4	3
Interspeech-2020	non-native English	≈51	≈2.5	7—9 (different tracks)
Sigtyp-2021	16 languages	≈5.5 per language	≈0.7 per language	3

Table 1: Other shared tasks

number of issues, namely, unstable recording quality (with background noise such as dogs barking or cars passing by) as well as vague licensing conditions and non-standard annotation, make field recordings largely unknown and unpopular within the speech processing community. However, these resources are often rich in dialect, age and gender variation and thoroughly annotated by language experts. We can therefore hope that they can be used by the NLP community, too. It is also worth noting that language documentation tasks, being crucial for both theorizing about languages and language revitalization, involve a lot of tedious annotation effort, making it even more important to develop automatic annotation tools. We organized a shared task on low resource speech processing[5], which was active from January to March 2021 and had the following goals:

1. evaluate the quality of modern speech processing methods on field data collections;
2. create a field recording dataset for ASR;
3. promote field data collections and low-resource language data among speech processing experts.

2 Related work

2.1 Other shared tasks on low-resource speech processing

Several competitions in various speech processing tasks have been organized recently in low-resource settings. However, the definition of what "low-resource" means varies from task to task. In 2018, Microsoft organized a shared task in low-resource automatic speech recognition ([6]), releasing data for Telugu, Tamil, and Gujarati, which was provided by Speechocean and Microsoft itself. In 2020, the participants of the GermEval-2020 shared task([13]) had to build a speech-to-text model for Swiss German, using recordings made in the parliament of Bern. Another low-resource speech processing task of 2020 was a challenging automatic speech recognition task for non-native children's speech ([15]), where records of Italian students speaking English were used.

In 2021, SIGTYP has organized a shared task on predicting language IDs (name, genus, and family) from speech[6], which is described in detail in [18]. In contrast with our task, SIGTYP-2021 involved a greater number of typologically diverse languages coming from a greater number of families, though limiting the participants to the language identification task only. They mostly used CMU Wilderness data, which is based on the sounding bible collection ([4]), for the training data as well as Common Voice[7] and OpenSLR[8] for the validation and test data. Moreover, some field data from the Paradisec is also used. The diversity of the data sources is meant to check the robustness of the participants' models. Three teams took part in SIGTYP-2021, with two of them performing better than the baseline. The winning team (Lipsia, [5]) transformed the MFCCs distributed by the organizers into spectrograms and then applied a ResNet-50 CNN based model to them. Another team was NTR ([2]), which used a solution similar to the one submitted to our shared task (see the description for the NTR system below). Finally, the Anlirika ([1]) system combines convolutional and LSTM layers in their approach

The details for the above-mentioned low-resource tasks are summarized in the table below (table 1).

[5] https://lowresource-lang-eval.github.io/content/shared_tasks/asr2021_en.html

[6] https://sigtyp.github.io/st2021.html

[7] https://commonvoice.mozilla.org

[8] https://openslr.org

A low-resource end-to-end speech translation task has been organized in 2021, focusing on two Swahili varieties and French and English[9]. However, its results will only be available later this year.

2.2 Using field linguistic data in speech processing tasks

Most papers on speech recognition for field linguistics datasets are aimed at facilitating language documentation itself, e. g. [12] (for Japhug) or [16](for Samoyedic languages). Moreover, tools for training speech recognition on field datasets have been developed, for instance Persephone ([21]) followed by Elpis ([19]). In [10], Gina-Anne Levow endorses using endangered language data in shared tasks. In [11], the authors show an exemplar case of using field data from the ELAR archive to create datasets for the speaker diarization and identification tasks. The paper also deals with the dual use of linguistic data and its potential consequences.

3 Shared task description

We offered three tasks: number of speakers detection, language identification, and automatic transcription. However, only two latter tasks were actually completed by three teams. The small number of teams is actually comparable to the other low-resource ASR shared tasks, which is perhaps due to the task difficulty and absence of public datasets. The shared task was hosted at the CodaLab automatic scoring platform[10]. The evaluation script is available online[11]

3.1 Number of speakers detection

The track was not completed due to the lack of participants. The aim of the track was to identify the number of speakers in a short recording. The track is crucial for field data processing as recordings often contain dialogues between a linguist and a language consultant. The dataset recordings were thus annotated with a corresponding number of speakers. The participants were to predict the number.

3.2 Language identification task

The shared task participants had to identify the language spoken, its genus, and the language family. The test dataset included surprise languages belonging to the same genera as the previously seen languages. The participants had to classify them as “unknown” (X). We scored the accuracy of classification across all the fields.

3.3 Automatic IPA transcription

The participants were to automatically transcribe speech using IPA. As in the language identification task, the test set included both previously seen and previously unseen data. We scored the length-normalized CER of the transcriptions.

4 Evaluation datasets

The datasets were based on the Lingvodoc platform ([7]), developed at the Institute of System Programming, RAS. The voiced data was compiled by linguists from Russian scientific institutions and processed in a unified way. The project focuses on collecting wordlists and corpora in various dialects of (predominantly) Uralic and Altaic languages, which are usually used for dialectological and historical linguistics studies.

4.1 Dataset preparation

Dataset preparation involved both scraping Lingvodoc and additionally annotating it. Lingvodoc is a joint effort of multiple teams so it is not surprising that the data can suffer from variation in annotation approaches, which has to deal with how wordlists are collected in language documentation projects. It is often a case that a linguist first pronounces the stimulus (a word or a phrase) in an auxiliary language

[9] https://iwslt.org/2021/low-resource

[10] https://competitions.codalab.org/competitions/30008

[11] https://github.com/lowresource-lang-eval/asr_evaluation_scripts

(Russian in case of Lingvodoc). The language consultant then pronounces the translation of the stimulus in their native tongue, sometimes repeating it. These repetitions are sometimes accompanied by a linguist asking the native speaker to pronounce the word again. When uploaded to Lingvodoc, these repetitions are not always split. The stimuli pronounced by the linguist are also not cut off in some cases. Both decisions (whether to split the recording into several repetitions and whether to include the stimuli or not) are usually made by the particular team uploading data to Lingvodoc. This variation in approaches is not crucial for manual processing but can hamper automatic processing. We therefore decided to specify whether there can potentially be repetitions or Russian stimuli in the data. The annotation is also available in the test dataset as two additional columns. We also checked if the transcriptions were IPA-valid using `ipapy`[12], excluding non-IPA transcriptions and normalizing some standard ways of transcribing which are not genuinely IPA, e.g. ɛ́:-ʏ-ak became ɛ́ːɣak

The datasets in the converted format can be found online[13].

4.2 Dataset statistics

Dataset was split into two subsamples, *Train* and *Test* respectively. The tasks were evaluated on the *Test* subsample, which contained both previously seen and surprise languages. The surpise languages were chosen on the following grounds: all the families and genera from the datasets had to be represented.

While there is some difference in recordings lengths across the language groups 1, no specific handling has been applied regarding these distributions.

5 Participants and results

Three teams took part in the shared task, choosing either the classification track (team NTR/TSU)(see [3]in this volume) or the transcription track (teams DG and DN).

5.1 System description

5.1.1 NTR/TSU

NTR/TSU uses a convolutional neural network with a self-attentive pooling layer for the classification task. The input for the network are mel-frequency spectral coefficients calculated from the original audio files. The architecture of the network is QuartzNet ASR. They also used several augmentation techniques, namely, shifting samples in range (-5ms; +5ms), SpecAugment, and adding background noise to the audio files.

5.1.2 Team DG

Daniil Grebenkin, one of the authors of this paper, contributed to this model. The experiment included the following stages:

1. Data preprocessing. The audio files had different sample rates and numbers of channels. Therefore, DG converted the files into mono and set them to the same sample rate (16000 kHz) using sox[14].
2. Creating a multilingual acoustic model for getting the transcriptions from lattices' files. DG used Kaldi ASR ([9]), which is a modular system allowing to add new data to either the language model or the acoustic model at a time without changing the other model. They applied a multilingual model trained on VoxForge[15] corpus to make transcriptions of the train dataset. Then, they trained a language module with these new words and new transcriptions. Finally, they created a new version of a multilingual model with an upgraded language module.
3. Getting the phoneme sequences of the competition train set utterances with the multilingual model. At this stage, they decoded the competition test set to get a transcription for each utterance. They used the epitran tool ([14]) to make a dictionary for the language module. The multilingual VoxForge corpus contains various languages and is rich in phonetic variation. IPA makes it possible to show differences in pronunciation for every language from VoxForge data whereas epitran has support for

[12] `https://github.com/pettarin/ipapy`

[13] `https://lowresource-lang-eval.github.io/content/data/index_data_asr_en.html`

[14] `http://sox.sourceforge.net`

[15] `http://www.voxforge.org`

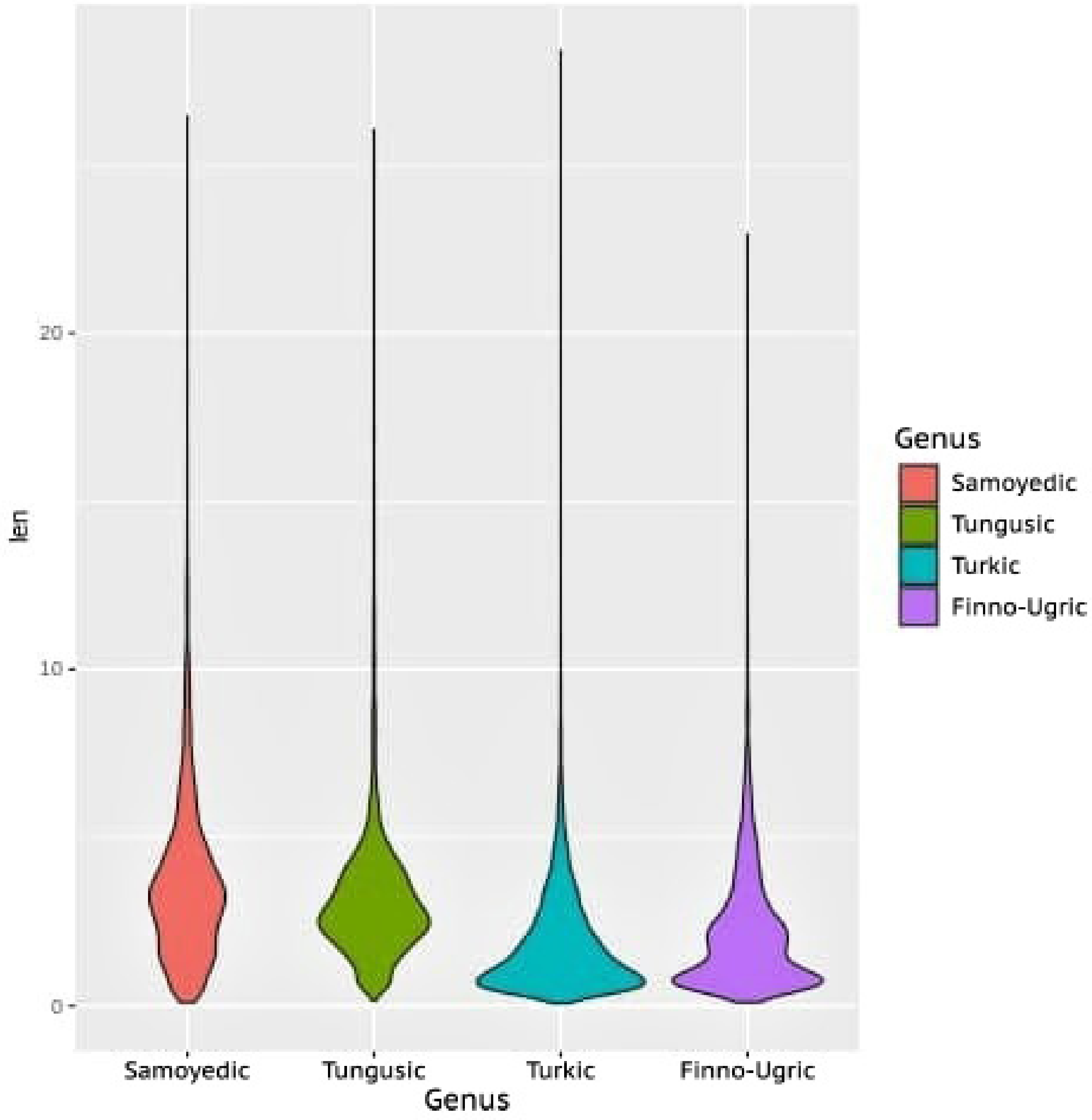

Figure 1: Recordings lenghts (in seconds) distribution across language groups

almost every language of this corpus. The result of the decoding process were lattices which helped get the phoneme sequences of each utterance. The algorithm was as follows:

- compute the best path through lattices and write them out as FSTs;
- store the lattices in archives containing transcriptions, alignments, and acoustic and LM costs;
- convert model-level alignments to phoneme sequences
- convert the phoneme sequences from the X-SAMPA format ([20]) to IPA format with an open-source converter xsampa [16]

4. Creating a new version of the multilingual model. The new transcriptions and the new words were added to the existing dictionary and lexicon files. Afterwards, they created a new language module and a new multilingual model with Kaldi tools.
5. Processing the test set. At this stage, they decoded the test set, getting phoneme sequences, which they then converted to the IPA format.

Some of the audio samples were not recognized by the model, there were $4,2\%$ of lines completed with «none» in final .tsv file which was used for evaluation.

5.1.3 Team DN

Daria Nosenko, one of the authors of this paper, contributed to this model. Team DN solution is founded on an end-to-end neural model for speech recognition QuartzNet([17]) based on Jasper([8]). QuartzNet model uses separable convolutions and is smaller than all other competing models. Team DN used TensorFlow-based NVIDIA OpenSeq2Seq toolkit[17] for their experiments with QuartzNet. The main features of this framework are modular architecture that allows assembling of new models from available components and fast Horovod-based distributed training supporting both multi-GPU and multi-node modes. They chose the multilingual VoxForge corpus for model training. The model training was performed using Horovod on 3 GPUs. The experiment included the following stages:

1. Train dataset preprocessing. All audio files with their annotations were combined into a single multilingual dataset. The .csv file for model training was generated using that dataset. Its rows have the following format: audio file name, audio file size, annotation. The file name contains its absolute path, the speaker folder name and the audio file name, which contains the language tag. If the file name did not contain the language tag, then it was added. The vocabulary for model training was generated from annotations of the entire dataset using Python "set" function. Then VoxForge audio files were converted into mono and set to the same sample rate (16000 kHz) using SoundFile[18]. The multilingual corpus was split into train, validation and test in 80:10:10 ratio in such a way that subsets did not overlap by speakers (i.e., one speaker should not be included in any two subsets at the same time).
2. Training QuartzNet model on the VoxForge dataset. The model was trained on 70 epochs.
3. Predicting annotations for the Dialog test dataset using the model from the previous step.

Finally, there were 412 audio files that were not recognized by the model ($3,9\%$ of the total number).

6 System results

6.1 Language Identification task

The only team to submit the LId task results was the NTR team. Their submission accuracy is outlined in the table 2 along with the random baseline scores. Overall submission confusion matrix is attached in the Appendix A .

While for frequent languages the accuracy is slightly better, there is no significant correlation between how much a language is represented in the data in the data and its identification accuracy. While showing some maybe interesting granular patterns, the confusion matrix is hard to typologically analyze. Unseen language identification is shown to be especially hard.

[16] https://github.com/dohliam/xsampa

[17] https://nvidia.github.io/OpenSeq2Seq/html/index.html

[18] `https://pysoundfile.readthedocs.io/en/latest`

Team	LId	GId	FId
NTR	**0.06**	**0.34**	0.61
baseline	0.01	0.22	**0.82**

Table 2: Results for Language Identification task which consisted of Language Identification (**LId**), language Group Identification (**GId**), language Family Identification (**FId**)

Team	Total test set (files)	Not recognized (files)	normalized CER
DG	10445	438	1.0828
DN	10445	412	1.572267

Table 3: Results for Task2

6.2 ASR task

Despite beating the baseline system, both submissions tend to provide much longer sequences of phonemes than expected. A closely-read analysis discovers that systems prediction performance drops while going from the beginning to the end of the recording with first phonemes usually being nearly guessed. The results of the both teams are summarized in the table below (3).

While the task was formulated using the IPA alphabet, both submissions' alphabets were different due to system design. This complicated the analysis of systems and resulted in significant loss of the evaluation metric.

7 Conclusion

In this paper, we present the results of the first shared task on ASR and speech-based language identification and categorization for the languages of Russia. As a result of the shared task, we prepared several datasets for language classification, transcription, and speaker number detection, for the first time for the languages in question. The participating teams experimented in performing various speech processing tasks for the languages which lack modern ASR technology tools, using state-of-the-art models. When analyzing the results, we also explored the limitations of the systems, which can help improve them

Acknowledgements

We would like to thank the authors of Lingvodoc for allowing us to use their priceless data, and the participants of the shared task for their comment and suggestions.

The work of Elena Klyachko was supported by a grant of the Russian Science Foundation, Project 20-012-00520 (Dynamics of the development of the language situation in local groups of indigenous peoples of Siberia and Russian Far East based on linguistic biographies).

References

[1] Anlirika: An LSTM–CNN Flow Twister for Spoken Language Identification / Andreas Scherbakov, Liam Whittle, Ritesh Kumar et al. // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021. — P. 145–148.

[2] Bedyakin Roman, Mikhaylovskiy Nikolay. Language ID Prediction from Speech Using Self-Attentive Pooling and 1D-Convolutions // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021.

[3] Bedyakin Roman, Mikhaylovskiy Nikolay. Low-Resource Spoken Language Identification Using Self-Attentive Pooling and Deep 1D Time-Channel Separable Convolutions // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021", Moscow, June 16–19, 2021. — 2021.

[4] Black Alan W. CMU Wilderness Multilingual Speech Dataset // ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — 2019. — P. 5971–5975.

[5] Celano Giuseppe GA. A ResNet-50-based Convolutional Neural Network Model for Language ID Identification from Speech Recordings // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021.

[6] Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages / Brij Mohan Lal Srivastava, Sunayana Sitaram, Kalika Bali et al. // SLTU. — 2018. — August.

[7] J.V. Normanskaya O.D. Borisenko. Dictionaries on Samoyedic languages and LingvoDoc software system for collaborative work on dictionaries and online publishing // NORDSCI 2018 Conference Proceeedings. — Vol. 1. — 2018. — P. 313–337.

[8] Jasper: An End-to-End Convolutional Neural Acoustic Model / Jason Li, Vitaly Lavrukhin, Boris Ginsburg et al. // Interspeech 2019. — 2019. — Sep. — Access mode: `http://dx.doi.org/10.21437/interspeech.2019-1819`.

[9] The Kaldi Speech Recognition Toolkit / Daniel Povey, Arnab Ghoshal, Gilles Boulianne et al. // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. — Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society, 2011. — Dec. — IEEE Catalog No.: CFP11SRW-USB.

[10] Levow Gina-Anne. Promoting Language Technology for Endangered Languages with Shared Tasks // Proceedings of the 1st International Conference on Language Technologies for All. — Paris, France : European Language Resources Association (ELRA), 2019. — December. — P. 116–119. — Access mode: `https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.30.pdf`.

[11] Levow Gina-Anne, Ahn Emily P, Bender Emily M. Developing a Shared Task for Speech Processing on Endangered Languages // Proceedings of the Workshop on Computational Methods for Endangered Languages. — Vol. 1. — 2021. — P. 96–106.

[12] Macaire Cécile. Alignement temporel entre transcriptions et audio de données de langue japhug // 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT) / CNRS. — 2020. — P. 9–22. — Access mode: `https://hal.archives-ouvertes.fr/hal-03066031/document#page=15`.

[13] Michel Plüss Lukas Neukom Manfred Vogel. GermEval 2020 Task 4: Low-Resource Speech-to-Text // CEUR-WS.org. — 2020. — Access mode: `http://ceur-ws.org/Vol-2624/germeval-task4-paper1.pdf`.

[14] Mortensen David R, Dalmia Siddharth, Littell Patrick. Epitran: Precision G2P for many languages // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.

[15] Overview of the Interspeech TLT2020 Shared Task on ASR for Non-Native Children's Speech / Roberto Gretter, Marco Matassoni, Daniele Falavigna et al. // Proc. Interspeech 2020. — 2020. — P. 245–249.

[16] Partanen Niko, Hämäläinen Mika, Klooster Tiina. Speech Recognition for Endangered and Extinct Samoyedic languages // arXiv preprint arXiv:2012.05331. — 2020. — Access mode: `https://arxiv.org/ftp/arxiv/papers/2012/2012.05331.pdf`.

[17] Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions / Samuel Kriman, Stanislav Beliaev, Boris Ginsburg et al. // ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — 2020. — May. — Access mode: `http://dx.doi.org/10.1109/ICASSP40776.2020.9053889`.

[18] SIGTYP 2021 shared task: Robust spoken language identification / Elizabeth Salesky, Badr M Abdullah, Sabrina Mielke et al. // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021. — P. 122–129.

[19] User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis / Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer et al. // Proceedings of the 4th Workshop on Computational Methods for Endangered Languages. — 2021.

[20] Wells John C. Computer-coding the IPA: a proposed extension of SAMPA // Revised draft. — 1995. — Vol. 4, no. 28. — P. 1995.

[21] Wisniewski Guillaume, Michaud Alexis, Guillaume Séverine. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? // 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop / European Language Resources Association (ELRA). — 2020. — P. 306–315. — Access mode: `https://arxiv.org/ftp/arxiv/papers/2012/2012.05331.pdf`.

A Appendix

Family	Group	Language	Count
Altaic	Manchu-Tungus	evn	2529
Altaic	Manchu-Tungus	oac	1011
Altaic	Manchu-Tungus	ude	696
Altaic	Manchu-Tungus	ulc	576
Altaic	Turkic	alt-tub	2519
Altaic	Turkic	bak	1044
Altaic	Turkic	sah	388
Altaic	Turkic	tat	273
Altaic	Turkic	tyv	5961
Uralic	Samoyedic	enh	930
Uralic	Samoyedic	nio	725
Uralic	Samoyedic	sel	3022
Uralic	Samoyedic	yrk-for	76
Uralic	Finno-Ugric	kca	661
Uralic	Finno-Ugric	koi-yzv	461
Uralic	Finno-Ugric	kom	318
Uralic	Finno-Ugric	krl	771
Uralic	Finno-Ugric	mdf	103
Uralic	Finno-Ugric	mhr	93
Uralic	Finno-Ugric	mns	775
Uralic	Finno-Ugric	mrj	330
Uralic	Finno-Ugric	sjd	220
Uralic	Finno-Ugric	sms	549

Table 1: **Train** subsample statistics, number of utterances for each language is counted

Family	Group	Language	Count
Altaic	Manchu-Tungus	evn	19
Altaic	Manchu-Tungus	gld	390
Altaic	Manchu-Tungus	neg	360
Altaic	Manchu-Tungus	ude	2060
Altaic	Manchu-Tungus	ulc	17
Altaic	Turkic	alt	671
Altaic	Turkic	alt-tel	21
Altaic	Turkic	alt-tlg	16
Altaic	Turkic	atv-c	12
Altaic	Turkic	bak	296
Altaic	Turkic	chv	1826
Altaic	Turkic	cjs	25
Altaic	Turkic	clw	364
Altaic	Turkic	dlg	19
Altaic	Turkic	kim	229
Altaic	Turkic	kum	3
Altaic	Turkic	sah	483
Altaic	Turkic	tat	34
Altaic	Turkic	tyv	1269
Altaic	Turkic	uig	531
Uralic	Samoyedic	enf	442
Uralic	Samoyedic	enh	46
Uralic	Samoyedic	nio	47
Uralic	Samoyedic	yrk-ntu	93
Uralic	Finno-Ugric	fin	97
Uralic	Finno-Ugric	koi	265
Uralic	Finno-Ugric	kom	9
Uralic	Finno-Ugric	krl	264
Uralic	Finno-Ugric	mhr	15
Uralic	Finno-Ugric	mrj	139
Uralic	Finno-Ugric	myv	15
Uralic	Finno-Ugric	sms	4
Uralic	Finno-Ugric	udm	132
Uralic	Finno-Ugric	vot	232

Table 2: **Test** subsample statistics, number of utterances for each language is counted

Family	Group	Language	Count
Altaic	Manchu-Tungus	evn	2548
Altaic	Manchu-Tungus	gld	390
Altaic	Manchu-Tungus	neg	360
Altaic	Manchu-Tungus	oac	1011
Altaic	Manchu-Tungus	ude	2756
Altaic	Manchu-Tungus	ulc	593
Altaic	Turkic	alt	671
Altaic	Turkic	alt-tel	21
Altaic	Turkic	alt-tlg	16
Altaic	Turkic	alt-tub	2519
Altaic	Turkic	atv-c	12
Altaic	Turkic	bak	1340
Altaic	Turkic	chv	1826
Altaic	Turkic	cjs	25
Altaic	Turkic	clw	364
Altaic	Turkic	dlg	19
Altaic	Turkic	kim	229
Altaic	Turkic	kum	3
Altaic	Turkic	sah	871
Altaic	Turkic	tat	307
Altaic	Turkic	tyv	7230
Altaic	Turkic	uig	531
Uralic	Samoyedic	enf	442
Uralic	Samoyedic	enh	976
Uralic	Samoyedic	nio	772
Uralic	Samoyedic	sel	3022
Uralic	Samoyedic	yrk-for	76
Uralic	Samoyedic	yrk-ntu	93
Uralic	Finno-Ugric	fin	97
Uralic	Finno-Ugric	kca	661
Uralic	Finno-Ugric	koi	265
Uralic	Finno-Ugric	koi-yzv	461
Uralic	Finno-Ugric	kom	327
Uralic	Finno-Ugric	krl	1035
Uralic	Finno-Ugric	mdf	103
Uralic	Finno-Ugric	mhr	108
Uralic	Finno-Ugric	mns	775
Uralic	Finno-Ugric	mrj	469
Uralic	Finno-Ugric	myv	15
Uralic	Finno-Ugric	sjd	220
Uralic	Finno-Ugric	sms	553
Uralic	Finno-Ugric	udm	132
Uralic	Finno-Ugric	vot	232

Table 3: **Overall** dataset statistics, number of utterances for each language is counted

				prediction																								portion
														Altaic							Uralic							
														Manchu-Tungus			Turkic				Samoyedic		Finno-Ugric					
				alt-tub	kca	koi-yzv	mdf	mns	nan	oac	sel	sjd	yrk-for	evn	ude	ulc	bak	sah	tat	tyv	enh	nio	kom	krl	mhr	mrj	sms	
gold	Altaic	Manchu-Tungu	evn	0.11	0	0	0	0.11	0	0	0.11	0	0	***0***	0.26	0	0	0.11	0	0.05	0.05	0	0	0.16	0	0	0.05	0
			gld	0.1	0.01	0.01	0.01	0.06	0	0.02	0.04	0.01	0	0.03	0.15	0.03	0.13	0.08	0.01	0.2	0.02	0.03	0.02	0.05	0	0.02	0.01	0.04
			neg	0.12	0.04	0.01	0.01	0.06	0	0.01	0.05	0	0	0.02	0.16	0.01	0.09	0.07	0.01	0.22	0.03	0.02	0.02	0.03	0	0.01	0.01	0.03
			ude	0.14	0.03	0.01	0.01	0.06	0	0.01	0.04	0	0	0.02	***0.17***	0.01	0.08	0.05	0.01	0.19	0.03	0.02	0.01	0.05	0	0.02	0.01	0.2
			ulc	0.24	0.06	0	0	0.12	0	0	0.12	0	0	0	0.06	***0***	0.06	0	0	0.18	0	0.06	0	0.06	0	0.06	0	0
		Turkic	alt	0.13	0.02	0.01	0	0.07	0	0.02	0.04	0	0	0.03	0.18	0.01	0.08	0.05	0.01	0.2	0.02	0.03	0.02	0.04	0	0.02	0	0.06
			alt-tel	0.1	0.05	0	0	0.1	0.19	0.05	0	0.05	0	0	0.1	0	0.1	0.05	0	0.1	0	0.05	0	0.05	0	0.05	0	0
			alt-tlg	0.13	0.13	0	0	0.19	0	0	0.13	0	0	0	0.06	0	0.06	0.06	0	0.19	0	0	0	0.06	0	0	0	0
			atv-c	0.25	0	0	0	0.08	0	0.08	0	0	0	0	0.08	0	0.25	0	0	0	0	0.08	0.08	0.08	0	0	0	0
			bak	0.11	0.04	0.01	0	0.07	0	0.01	0.04	0.01	0	0.02	0.19	0.02	***0.1***	0.06	0.01	0.2	0.03	0.02	0.02	0.03	0	0.01	0.01	0.03
			chv	0.14	0.02	0.01	0	0.08	0	0.01	0.04	0	0	0.02	0.16	0.01	0.09	0.05	0.01	0.19	0.02	0.02	0.02	0.06	0	0.02	0	0.17
			cjs	0.04	0	0	0	0.04	0	0.04	0.08	0	0	0.04	0.28	0	0.04	0.12	0	0.2	0.04	0.08	0	0	0	0	0	0
			clw	0.13	0.04	0.02	0	0.12	0	0.01	0.04	0.01	0	0.02	0.14	0.01	0.07	0.05	0	0.19	0.04	0.01	0.02	0.04	0.01	0.02	0	0.03
			dlg	0.16	0	0	0	0	0	0	0.05	0	0	0.05	0.05	0	0.11	0.11	0	0.11	0.11	0.11	0	0.16	0	0	0	0
			kim	0.16	0.04	0.01	0	0.09	0	0	0.05	0	0	0	0.15	0	0.08	0.06	0.01	0.22	0.03	0.01	0.03	0.04	0	0.01	0	0.02
			kum	0	0	0	0	0	0	0	0.33	0	0	0	0.33	0	0	0	0.33	0	0	0	0	0	0	0	0	0
			sah	0.14	0.02	0.01	0	0.08	0	0.01	0.05	0	0	0.01	0.17	0.01	0.06	***0.05***	0	0.22	0.03	0.02	0.01	0.07	0.01	0.02	0	0.05
			tat	0.18	0.03	0.06	0	0.06	0	0	0.03	0	0	0	0.09	0	0.12	0.12	***0***	0.15	0.06	0.03	0	0.06	0	0	0.03	0
			tyv	0.16	0.03	0.02	0	0.05	0	0.01	0.04	0	0	0.02	0.16	0.02	0.08	0.06	0.01	***0.19***	0.02	0.03	0.02	0.06	0	0.02	0	0.12
			uig	0.12	0.03	0.02	0	0.07	0	0.02	0.06	0	0	0.02	0.18	0.02	0.09	0.05	0.01	0.19	0.02	0.03	0	0.04	0.01	0.02	0	0.05
	Uralic	Samoyedic	enf	0.13	0.03	0.02	0.01	0.07	0	0.01	0.04	0	0	0.02	0.18	0.01	0.09	0.06	0.01	0.18	0.03	0.02	0.02	0.04	0	0.02	0.01	0.04
			enh	0.02	0.02	0.04	0	0.15	0	0	0.02	0	0	0	0.17	0	0.11	0.07	0.02	0.22	***0***	0	0.02	0.07	0.02	0.04	0	0
			nio	0.15	0.04	0.02	0	0.04	0	0.02	0.04	0.02	0	0.02	0.21	0	0.09	0.09	0.02	0.15	0.04	***0.02***	0	0.02	0	0	0	0
			yrk-ntu	0.16	0.01	0.02	0	0.06	0	0	0.08	0	0.02	0.01	0.14	0.01	0.03	0.09	0.01	0.19	0.03	0.02	0	0.06	0.01	0.03	0	0.01
		Finno-Ugric	fin	0.13	0.01	0	0	0.1	0	0	0.03	0	0	0.02	0.19	0.02	0.07	0.09	0.02	0.11	0.04	0.02	0.06	0.05	0	0.02	0	0.01
			koi	0.12	0.02	0.01	0.01	0.08	0	0.01	0.03	0	0	0.01	0.19	0	0.12	0.05	0.01	0.18	0.03	0.05	0.02	0.05	0	0.02	0	0.03
			kom	0	0	0	0	0	0	0	0	0	0	0	0.33	0	0.11	0.22	0	0.22	0	0	***0.11***	0	0	0	0	0
			krl	0.12	0.04	0.02	0	0.07	0	0.03	0.05	0.01	0	0.01	0.16	0	0.11	0.06	0.01	0.16	0.02	0.03	0.01	***0.05***	0	0.03	0	0.03
			mhr	0.13	0.07	0	0	0.07	0	0	0.07	0	0	0	0.07	0	0	0	0	0.4	0	0	0	0.07	***0***	0.13	0	0
			mrj	0.12	0.03	0.01	0	0.09	0	0	0.04	0	0	0.05	0.17	0.01	0.07	0.06	0.01	0.25	0.01	0.01	0.01	0.05	0	***0.01***	0.01	0.01
			myv	0.13	0	0	0.07	0.07	0	0	0	0	0	0	0.27	0	0	0	0	0.2	0.13	0	0	0.13	0	0	0	0
			sms	0	0	0	0	0	0	0.25	0	0	0	0	0.5	0	0	0	0	0.25	0	0	0	0	0	0	***0***	0
			udm	0.11	0.04	0.01	0	0.05	0	0.02	0.03	0	0	0.02	0.25	0.02	0.08	0.07	0.02	0.19	0.03	0.01	0.02	0.03	0	0.02	0	0.01
			vot	0.14	0.04	0.01	0	0.08	0	0.01	0.06	0.01	0	0.02	0.14	0.02	0.07	0.08	0.01	0.19	0.01	0.01	0.02	0.06	0	0.02	0	0.02

Figure 1: confusion matrix of the team **NTR** submission