

Автоматическое транскрибирование полевых данных: эксперименты и проблемы

Е. Клячко

OPLING-9

Дисклеймер

- я не специалист в области обработки речи
- **цель:**
 - поделиться опытом:
 - что сейчас доступно рядовому пользователю
 - обсудить опыт участников
 - дать мотивацию к собственным экспериментам

Что может потребоваться сделать со звуком, записанным в поле?

- “улучшение” данных (удаление шумов...)
- разметка:
 - диаризация (разделение дикторов) (*speaker diarization*)
 - автоматическое определение языка (*language detection*)
 - выделение ключевых слов (*keyword spotting*)
 - автоматическое выравнивание (*forced alignment*)
 - **автоматическое распознавание (ASR, STT)**
 - транскрипция
 - оформленный текст (с делением на предложения, расстановкой знаков препинания, нормализацией)
- синтез речи (TTS)
- ?....

Цель

≈ 147 часов эвенкийских озвученных словарей
(≈ 95 тысяч фрагментов)

- есть ручная нарезка
- есть переводы
- транскрибировать вручную?

⇒ **дешево и быстро получить черновую транскрипцию для поиска**

- нерасшифрованные эвенкийские тексты

⇒ **получить черновую транскрипцию**

Некоторые термины

набор данных (*dataset*)

выборки: обучающая, валидационная, тестовая

(*train split, validation split, test split*)

обучение (*training*), предобученная (*pretrained*) модель

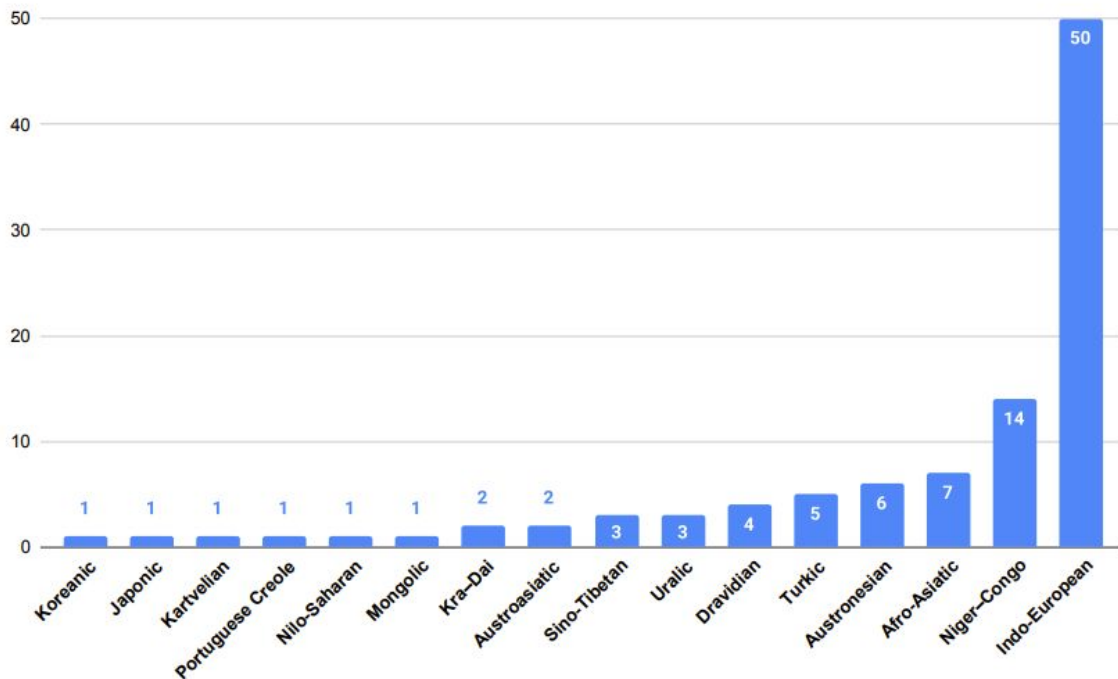
тонкая/точная настройка (*fine-tuning*)

GPU

оценка качества (*evaluation*): WER, CER

бенчмарк (*benchmark*)

Оценка качества: бенчмарки



Conneau, Alexis, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. "Fleurs: Few-shot learning evaluation of universal representations of speech." In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798-805. IEEE, 2023.

Figure 1: *Distributions of language families in FLEURS (y-axis is the count).*

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M. and Baevski, A., 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97), pp.1-52. — **MMS**

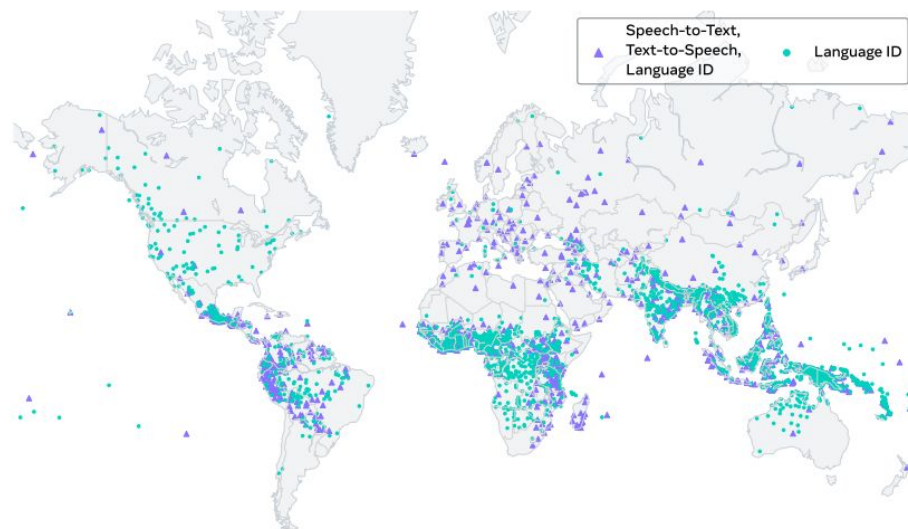


Figure 1: Illustration of where the languages supported by MMS are spoken around the world: MMS models support speech-to-text and text-to-speech for 1,107 languages as well as language identification for 4,017 languages.

- сбор аудиоданных (в основном, озвученные библейские тексты)
- выравнивание
- нормализация текстовых данных
- удаление низкокачественных данных
- обучение модели wav2vec 2.0
- размеченные данные (44 700 часов)
- неразмеченные данные

Эвенкийский ASR: данные

корпус эвенкийского языка

(<https://minlang.iling-ran.ru/corpora/evenki>)

≈ 8 часов звука, ≈ 38 000 словоупотреблений

- спонтанные тексты, записанные от разных дикторов
- разные диалекты
- транскрипция МФА
- запись в эвенкийской орфографии
- текст и звук выровнены по предложениям (файлы .eaf)

Эвенкийский ASR: подготовка данных

- приведение транскрипции к единому виду
- удаление некорректных фрагментов (длины 0)
- подготовка данных в табличном виде
 - строка таблицы имеет вид:
 - файл wav
 - предложение
 - split: train vs test
 - экспорт на huggingface
 - TODO: добавить анонимизированные метаданные

Эвенкийский ASR: данные


<https://huggingface.co/datasets/siberian-lang-lab/evenki-speech>


Dataset Viewer Auto-converted to Parquet </> API Embed Data Studio

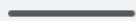
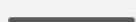
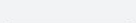
Split (2)
train · 6.5k rows

Split (2)
test · 1.64k rows

Search this dataset

audio
audio · *duration (s)*

0.21 69.2

sentence
string · *lengths*

1 386

▶ 0:00 / 0:05  🔊 ⋮	mohadu: indʲitʃo:w bi do dwadʲsatʲi let əntilnunmi
▶ 0:00 / 0:05  🔊 ⋮	i wəkta nulgiktədʲəŋki:wun nulgimmm nulgidu: indʲitʃo:w
▶ 0:00 / 0:03  🔊 ⋮	simjawun həgdiŋəkəŋtʃə biʲʃo:n ərkə wosʲemʲ kuŋakar biʲʃo:tin

Эвенкийский ASR: [facebook/mms-1b-l1107](https://www.facebook.com/mms-1b-l1107)

- обучалась в основном на библейских текстах
 - результат — в кириллице
- ⇒ сделала транслитератор

Оценка качества ASR: WER

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

WER = ?

Эталон	əŋki:wun	a:rə		gorojə	ta:du:
Предсказание проверяемой модели	əŋki:wun	a:ra	bi:	gorojə	
Результат	✓	S	I	✓	D

Оценка качества ASR: WER

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

WER = ?

Эталон	əŋki:wun	a:rə		gorojə
Предсказание проверяемой модели	əŋki:wun	goro	bi:	gorojə
Результат	✓	S	I	✓

Эвенкийский ASR: facebook/mms-1b-l1107

Образец

bi nənəktə no:wu

bi bultad^ɬam biraldu
d^ɬigali: t^ʰɨktɨ d^ɬami

ənki:wun arə gorojə

WER на тестовой выборке: 1,00

Результат

bi: nonokta – nwo

bi: bultd^ɬim biraldu:
d^ɬegali: t^ʰɨkt d^ɬanni:

ənkiwun ra goroo

CER на тестовой выборке: 0,46

Эвенкийский ASR: обучение

- подготовка данных для обучения и тестирования
- выбор модели
- fine-tuning модели

⇒

<https://huggingface.co/siberian-lang-lab/wav2vec2-large-mms-1b-evenki-colab>

Эвенкийский ASR: обучение

инструкция: https://huggingface.co/blog/mms_adapters (P. von Platen)

- идея: fine-tuning только **слоев адаптера** ⇒ утверждается, что может сработать на небольшом объеме данных
- автоматический выбор размера пачки (`auto_find_batch_size=True`)
- сохраняла **чекпойнты** (потом вручную удаляла менее удачные)

⇒ возобновление с произвольного момента (см. [тетрадь](#))

```
save_total_limit=2,  
push_to_hub=True,  
resume_from_checkpoint=True,  
hub_strategy="all_checkpoints"
```


Изменение WER на тестовой выборке
в зависимости от шага обучения

Видеокарта NVIDIA Tesla T4
≈ 15 часов обучения ⇒ 22 ед.

100 ед. = \$10



Эвенкийский ASR: обученная модель

Образец

bi nənəktə no:wu

bi bultadʲam biraldu
dʲigali: tʲʲikti dʲami

əŋki:wun arə gorojə

WER на тестовой выборке: 0,71

Результат

bi nənəktə no:wəl

bi bultadʲam biraldu:
dʲigali: tʲʲikti dʲamɲi

əŋki:wun ara gorojo

CER на тестовой выборке: 0,2

Эвенкийский ASR: словарь (диктора нет в корпусе)

facebook/mms-1b-l1107

обученная модель

начнётся дождь

тыгдалдн тыгдалдн
тыгдалдн

tigdəldʲan tigdəldʲan
tᶦgdaldʲan

дойти (до дома)

б̄и длви ихим б̄и длви
ихим б̄и длви ихим

bi dʲula:wi əhəm bi
dʲula:wi əhəm bi
dʲula:wi əhəm

самолёт сел

сомолт тэгэрэн
амноннад̄ү сэмэлт
тэгэрэн амнннад̄ү

samolottəgərən
amnonnadu:
samolottəgərən
amnonnadu:

Что дальше?

- Улучшить набор данных (исправить некоторые неконсистентные транскрипции)
- Пополнение датасета?
- Получить черновую транскрипцию для словарей:

на GPU:

тестовая выборка (1,6 ч) → 3 минуты

словари (147 ч) → предположительно, 5 часов

Спасибо за внимание!