

Модели ИИ для распознавания и синтеза речи на кыргызском языке

Исаев Руслан Рамилевич, к. ф-м. н.,
декан факультета Инженерии и информатики

медицина | инженерия | образование | технологии



Бишкек, 2024

Цели и задачи исследования

Цель: Создание ASR и TTS моделей для кыргызского языка

Задачи:

- Сбор наборов данных - языковые корпуса, голосовые наборы данных
- Обработка и чистка данных
- Выбор архитектур нейронных сетей, её настройка и обучение
- Тонкая настройка модели ИИ

Актуальность исследования

Программа развития государственного языка и совершенствования языковой политики в Кыргызской Республике на 2021-2025 годы

23 сентября 2019 года на мероприятии, посвященном 30-летию принятия Закона Кыргызской Республики "О государственном языке Кыргызской Республики", Президент Кыргызской Республики отметил следующее: "Много поколений осталось в истории, а наш великий язык живет. Поэтому наш родной язык - это священный залог, дарованный нашими предками, который мы должны беречь и передать потомкам. Сохранение языка имеет такое же значение, как и защита государства, независимости..."

О государственном языке Кыргызской Республики от 17 июля 2023 года № 140

Данные проект Common Voice (записи голоса)

Download the Dataset

We've made some changes. Delta Segments just contain the most recent clips since the last release. [Read more about this work.](#)

Select the desired language dataset and choose the version you wish to download.

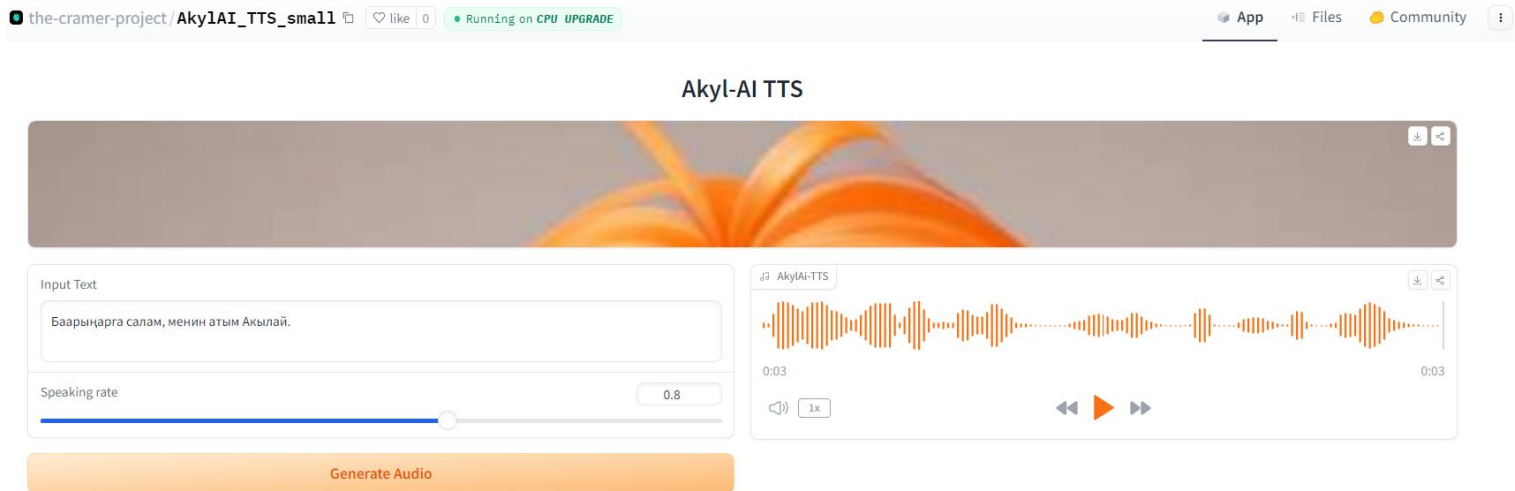
Language
Kyrgyz

Version	Date	Size	Recorded Hours	Validated Hours	License	Number of Voices	Audio Format	Splits (Age and Sex)
✓ Common Voice Corpus 17.0	3/20/2024	1.03 GB	48	39	CC-0	283	MP3	65% 20 - 29 17% < 20 9% 30 - 39 7% No information 1% 40 - 49 0% 50 - 59 0% 60 - 69 0% 70 - 79 0% 80 - 89 0% 90 - 99 54% Male/Masculine 35% Female/Feminine 12% No Information 0% Transgender 0% Non-Binary 0% Don't Wish To Say
Common Voice Delta Segment 17.0	3/20/2024	36.19 MB	2	1	CC-0	11	MP3	
Common Voice Corpus 16.1	1/5/2024	1,015.39 MB	46	39	CC-0	272	MP3	
Common Voice Delta Segment 16.1	1/5/2024	0 byte	1	1	CC-0	3	MP3	
Common Voice Corpus 15.0	9/14/2023	1,016.09 MB	46	39	CC-0	269	MP3	
Common Voice Delta Segment 15.0	9/14/2023	14.48 MB	1	1	CC-0	5	MP3	
Common Voice Corpus 14.0	6/28/2023	1,001.6 MB	46	39	CC-0	264	MP3	

<https://commonvoice.mozilla.org/en/datasets>

Модель Akyl.ai (TTS)

Полная версия кыргызоязычного искусственного интеллекта AkylAI выйдет примерно к весне 2024 года



<https://the-tech.kz/pervyj-ai-assistent-akylai-zagovoril-na-kyrgyzskom-yazyke/>
<https://huggingface.co/the-cramer-project>

Модель AkyL.ai (STT)

the-cramer-project / Kyrgyz-STT-Small

like 0

Running

App


Files

Community


Kyrgyz-STT-Small

Realtime demo for Kyrgyz speech recognition using a fine-tuned Whisper small model.


audio



0:030:03



1x



Clear

Submit

output

Баарыңарга салам менин атын акыл-ай


<https://the-tech.kz/pervyj-ai-assistent-akylai-zagovoril-na-kyrgyzskom-yazyke/>
<https://huggingface.co/the-cramer-project>

Модель Söyle (ASR)

Söyle: Noise-Robust Multilingual Speech Recognition with Long Transcription Featuring the Tatar Speech Corpus

Language:

▶ 0:00 / 0:03 🔊 ⋮

 Powered by
GitHub

<https://github.com/IS2AI/Söyle>

<https://issai.nu.edu.kz/ru/soyle-project-rus/>

https://colab.research.google.com/drive/1b3iF5QrwTFN3_Yr8nYB3-RQRMO6hLwAh?usp=sharing

Сбор аудио-данных

▲ 172.104.249.207/record



Kyrgyz Speech Dataset

Көптөгөн жыныстык жол менен жугуучу инфекциялар, анын ичинде хламидиоз, гонорея, гепатит В, ВИЧ жана сифилис, кош бойлуу жана төрөт учурунда энеден балага жугушу мүмкүн

Sentence: 0/100

▶ 0:00 / 0:00 ———— 🔊 ⋮

Submit Recording

Record Again



Жасалма интеллект

40 members



Notifications



2 saved messages



42 audio files



436 voice messages

Модель Soyle и наши данные

көз бұлшындары денедегі ең активдеу бұлшын болып саналады теория бойынша адамның көзі он миллионға шейін түсті айырмалай алады шындығында ал жүзге ғана жақын түсті айырмалай алады ал емеккесі би түске байланышты адамдар суретшілер дизайнерлер бұл жолменен жүз елу түсті айырмалай ала ашады

көз булчумдары денедеги эң активдүү булчум болуп саналат теория боюнча адамдын көзү он миллиондор чийин түстү айрмалай алат чындыгында ал жүзгө гана жакын түстү айрмалай алат ал эми кесеби түстү байланыштуу адамдар сүрөтчүлөр дизайнерлер болжол менен жүз елүү түстү айрмалай алышат

WER: ~8%



Сбор корпуса по техническим терминам

"Генетический алгоритм": {

"translation": "Генетикалык алгоритм",

"article": "Машина үйрөнүү - жасалма интеллект методдорунун классы, анын мүнөздүү өзгөчөлүгү көйгөйдү түз чечүү эмес, бирок ушул сыяктуу көптөгөн маселелерди чечүү жолдорун колдонуу аркылуу үйрөнүү. Мындай ыкмаларды куруу үчүн математикалык статистиканын инструменттери, сандык методдор, математикалык анализ, оптималдаштыруу методдору, ыктымалдуулук теориясы, график теориясы жана цифралык формадагы маалыматтар менен иштөөнүн ар кандай ыкмалары колдонулат. Тренингдин эки түрү бар: Прецеденттик окутуу, же болбосо индуктивдүү окутуу, аныктоого негизделген эмпирикалык мыйзам ченемдүүлүктөрү - жылы маалыматтар. Дедуктивдүү окутуу эксперттердин билимин формалдаштырууну жана аларды компьютерге билим базалары. Дедуктивдүү окутуу доменге таандык эксперттик системалар. ошондуктан терминдер машинаны үйрөнүү жана прецеденттик окутуу синонимдер деп эсептесе болот. Классикалык статистикалык ыкмаларга альтернатива катары көптөгөн индуктивдүү окутуу ыкмалары иштелип чыккан. Көптөгөн ыкмалар маалымат алуу менен тыгыз байланышта, маалыматтарды казып алуу. == Жана башка дагы == Көптөгөн объектер жана көптөгөн мүмкүн болгон жооптор бар. Жооптор менен объектердин ортосунда кандайдыр бир көз карандылык бар, бирок ал белгисиз. Прецеденттердин чектүү жыйындысы гана белгилүү - "\"объект, жооп\"" жуптары, окутуу топтому деп аталат. Бул маалыматтардын негизинде жашыруун көз карандылыкты калыбына келтирүү, башкача айтканда, ар кандай мүмкүн болгон киргизүү объектинин жетиштүү так классификациялык жоопту чыгарууга жөндөмдүү алгоритмди куруу зарыл. Бул көз карандылык сөзсүз түрдө аналитикалык түрдө туюндурулбайт жана бул жерде нейрон тармактары эмпирикалык түрдө түзүлгөн чечим принцибин ишке ашырат. Бул учурда маанилүү өзгөчөлүк болуп окутуу системасынын жалпылоо, башкача айтканда, учурдагы окутуу үлгүсүнүн чегинен чыккан маалыматтарга адекваттуу жооп берүү жөндөмдүүлүгү саналат. Жооптордун тактыгын өлчөө үчүн сапатты баалоо функциясы киргизилет. Бул билдирүү классикалык маселелерди жалпылоо болуп саналат жакындаштыруу милдеттери. Классикалык жакындаоо маселелеринде объектилер чыныгы сандар жана векторлор болуп саналат. Чыныгы тиркемелерде объектилер жөнүндө маалыматтар толук эмес, так эмес, сандык эмес, гетерогендүү болушу мүмкүн. Бул өзгөчөлүктөр машинаны үйрөнүү методдорунун ар түрдүүлүгүнө алып келет. == Машина үйрөнүү ыкмалары == Машина үйрөнүү бөлүмү, бир жагынан, нейрондук тармактар илиминин тармактарды окутуу ыкмаларына жана алардын архитектурасынын топологияларынын түрлөрүнө бөлүнүүнүн натыйжасында түзүлсө, экинчи жагынан математикалык статистиканын ыкмаларын камтыган. Төмөндө саналып өткөн машинаны үйрөнүү ыкмалары нейрондук тармактарды колдонууга негизделген, бирок окутуу топтомуна негизделген башка методдор бар - мисалы, байкалган статистиканын жалпыланган дисперсиясы жана ковариациясы боюнча иштеген дискриминанттык анализ же Байес классификаторлору. Нейрондук тармактардын негизги типтери, мисалы, перцептрон жана көп катмарлуу перцептрон күчөтүлгөн жана өзүн-өзү уюштуруу менен окутуучу менен да, окутуучусуз да үйрөтүлүшү мүмкүн. Бирок кээ бир нейрон тармактары жана көпчүлүк статистикалык ыкмалар окутуу ыкмаларынын бирине гана таандык болушу мүмкүн. Ошондуктан, эгер сиз окутуу ыкмасына жараша машиналык үйрөнүү ыкмаларын классификациялоо керек болсо, анда нейрон тармактарын белгилүү бир типке классификациялоо туура эмес болуп калат, нейрондук тармакты окутуу алгоритмдерин классификациялоо туура болот; Мугалим менен окутуу - ар бир прецедент үчүн "\"кырдаал, талап кылынган чечим\"" жуп берилет: Жасалма нейрон тармагы Герен үйрөнүү Ката оңдоо ыкмасы Артка жайылтуу ыкмасы Колдоо вектордук машина Көзөмөлсүз окутуу - ар бир прецедент үчүн объектердин жуптук окшоштугу боюнча маалыматтарды пайдалануу менен объектерди кластерлерге топтоо

Планы

Добавление модели в открытую библиотеку для распознавания речи “Vosk”

<https://alphacephei.com/vosk/>

Добавление модели NLP для <https://spacy.io/models>

Сбор корпуса Кыргызского языка для <https://universaldependencies.org/>