МІНІСТЭРСТВА АДУКАЦЫІ РЭСПУБЛІКІ БЕЛАРУСЬ
БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ
ФАКУЛЬТЭТ ПРЫКЛАДНОЙ МАТЭМАТЫКІ І ІНФАРМАТЫКІ
Кафедра дыскрэтнай матэматыкі і алгарытмікі

ТРАФІМАЎ Аляксандр Сяргеевіч

АЎТАМАТЫЧНАЕ ПЕРАЎТВАРЭННЕ БЕЛАРУСКАГА МАЎЛЕННЯ Ў ТЭКСТ МЕТАДАМІ ГЛЫБОКАГА НАВУЧАННЯ

Магістарская дысертацыя

спецыяльнасць _____ _____

шыфр *назва*

Навуковы кіраўнік
Гецэвіч Юрый Станіслававіч
кандыдат тэхнічных навук,
загадчык лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі

Дапушчана да абароны
«____» _____2022 г.
Загадчык кафедры дыскрэтнай матэматыкі і алгарытмікі
Котаў Уладзімір Міхайлавіч
доктар фізіка-матэматычных навук, прафесар

Менск, 2022

ЗМЕСТ

УМОЎНЫЯ АЗНАЧЭННІ

- **STT** – Speech-to-text – пераўтварэнне маўлення ў тэкст
- **TTS** – Text-to-speech – пераўтварэнне тэкста ў маўленне, сінтэз маўлення
- **ASR** – automatic speech recognition, аўтаматычнае распазнаванне маўлення
- **MFCC** – mel-frequency cepstrum, мел-частасныя кепстральныя каэфіцыэнты
- **LPC** – linear predictive coding, кадаванне з лінейным прэдыктарам
- **GMM** – gaussian mixture model, мадэль сумесі гаўсавых размеркаванняў
- **HMM** – hidden Markov models, схаваныя маркаўскія мадэлі
- **CTC** – Connectionist temporal classification
- **RNN** – Recurrent neural network, рэкурэнтная нейронная сетка
- **WER** – word error rate, доля памылковых словаў

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА РАБОТЫ

Магістарская дысертацыя, 52 старонкі, 13 выяваў, 5 табліцаў, 3 формулы, 27 крыніцаў.

Ключавыя словы: РАСПАЗНАВАННЕ МАЎЛЕННЯ, ПЕРАЎТВАРЭННЕ МАЎЛЕННЯ Ў ТЭКСТ, НЕЙРОННЫЯ СЕТКІ, ГЛЫБОКІЯ НЕЙРОННЫЯ СЕТКІ, ГЛЫБОКАЕ НАВУЧАННЕ, WAV2VEC2, БЕЛАРУСКАЯ МОВА.

Мэта работы – пабудова якаснай мадэлі распазнавання агульнага беларускага маўлення, а таксама збор неабходных для гэтага дадзеных.

Аб'ект даследавання – аўтаматычнае распазнаванне маўлення.

Прадмет даследавання – аўтаматычнае распазнаванне агульнага беларускага маўлення метадамі глыбокага навучання.

У магістарскай дысертацыі былі прааналізаваныя асноўныя падыходы да распазнавання маўлення а таксама папярэднія эксперыменты па пабудове сістэм распазнавання беларускага маўлення.

Быў сабраны вялікі корпус начытаных беларускіх тэкстаў пры ўдзеле вялікай колькасці дыктараў.

Была паспяхова навучаная сістэма распазнавання маўлення на аснове нейроннай архітэктуры Wav2Vec2 з выкарыстаннем n-грамнай моўнай мадэлі. Атрыманая сістэма апублікаваная на платформе Hugging Face. Таксама створаны і апублікаваны на платформе Hugging Face дэманстрацыйны вэб-інтэрфэйс, які дазваляе распазнаваць беларускае маўленне наўпрост у браўзеры.

Атрыманую мадэль можна выкарыстоўваць для стварэння сістэмы галасавога ўводу тэксту, пабудовы галасавых дапаможнікаў, стварэння сістэм аўтаматызаванага навучання беларускай мове.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Магистерская диссертация, 52 страницы, 13 рисунков, 5 таблиц, 3 формулы, 27 источников.

Ключевые слова: РАСПОЗНАВАНИЕ РЕЧИ, ПРЕОБРАЗОВАНИЕ РЕЧИ В ТЕКСТ, НЕЙРОННЫЕ СЕТИ, ГЛУБОКИЕ НЕЙРОННЫЕ СЕТИ, ГЛУБОКОЕ ОБУЧЕНИЕ, WAV2VEC2, БЕЛОРУССКИЙ ЯЗЫК.

Цель работы – построение качественной модели распознавания общей белорусской речи, а также сбор необходимых для этого данных.

Объект исследования – является автоматическое распознавание речи.

Предмет исследования – автоматическое распознавание общей белорусской речи методами глубокого обучения.

В магистерской работе были проанализированы основные подходы к распознаванию речи а также предыдущие эксперименты по построению систем распознавания белорусской речи.

Был собран большой корпус начитанных белорусских текстов при участии большого числа дикторов.

Была успешно обучена система распознования речи на основе архитектуры Wav2Vec2 з использованием n-грамной лингвистической модели. Полученная система опубликована на платформе Hugging Face. Также создан и опубликован на платформе Hugging Face демонстрационный веб-интерфейс, который позволяет распознавать белорусскую речь прямо через браузер.

Полученую модель можно использовать для создания системы голосового ввода текста, построения голосовых помощников, создания систем автоматизированного обучения белорусскому языку.

SUMMARY

Master thesis, 52 pages, 13 figures, 5 tables, 3 formulas, 27 resources.

Keywords: SPEECH RECOGNITION, SPEECH-TO-TEXT, SPEECH TO TEXT, NEURAL NETWORKS, DEEP NEURAL NETWORKS, DEEP LEARNING, WAV2VEC2, BELARUSIAN LANGUAGE.

The objective – to build a good speech recognition model for general Belarusian language and to gather a required dataset for this task.

The object of research – automated speech recognition.

Research focus – automated speech recognition for general Belarusian language using deep learning methods.

In this master's thesis main speech recognition approaches were analyzed as well as previous experiments to build speech recognition models for Belarusian language.

A large dataset of voiced texts was gathered with a participation of huge amount of people.

A Wav2Vec2-based speech recognition system was successfully built using additional n-gram Language model. The system was published on Hugging Face platform. A demo web-interface was built and published on Hugging Face platform that allows to recognize Belarusian speech directly from a web-browser.

Results of this research can be applied to develop voice input systems, build voice assistants, create a system to teach Belarusian language in an automated way.

УВОДЗІНЫ

Нягледзячы на істотныя поспехі для такіх задач апрацоўкі вялікіх неструктураваных дадзеных як аналіз відарысаў і апрацоўка натуральных моваў, сістэмы аўтаматычнага распазнавання маўлення не дасягнулі параўнаных вышыняў.

Адна з прычын палягае ў адметнасці задачы. Традыцыйным сістэмам цяжка распазнаваць маўленне ў зашумленых умовах. Таксама няпростым з'яўляецца пабудова традыцыйных сістэм, здольных аднолькава якасна распазнаваць маўленне розных дыктараў (speaker independent сістэмы).

Сучасныя тэхналогіі, заснаваныя на метадах глыбокага навучання, дазволілі зрабіць істотны прагрэс у задачы распазнавання маўлення. Але для вялікай колькасці моваў пабудова такіх сістэмаў з'яўляецца цяжкай задачай з прычыны адсутнасці вялікіх адкрытых набораў агучаных тэкстаў.

Апошнім часам адбываецца дэмакратызацыя патрабаванняў да пабудовы якасных сістэмаў распазнавання маўлення. Напрыклад становіцца магчымым зменшыць колькасць анатаваных дадзеных, патрэбных для навучання.

У межах дадзенай работы вядзецца пабудова сістэмы распазнавання адвольнага (без абмежавання слоўнікам) беларускага маўлення з выкарыстаннем сучасных архітэктур глыбокага навучання. Выбар сучасных мадэляў абумоўлены іх добрымі рэзультатамі па распазнаванні маўлення, зменшанай патрэбай адносна анатаваных датасэтаў у параўнанні з папярэднімі мадэлямі, а таксама адсутнасцю для беларускай мовы вялікіх анатаваных корпусаў агучаных тэкстаў.

Актуальнасць задачы тлумачыцца адсутнасцю для беларускай мовы якасных сістэм распазнавання адвольнага маўлення. Пабудова такой сістэмы дазволіць перайсці да пабудовы больш складаных моўных тэхналогіяў і пачаць інтэграваць іх у штодзённае жыццё.

Работа складаецца з наступных частак. У раздзеле 1 прыводзіцца агляд работы сістэм распазнавання маўлення, аналізуюцца рэзультаты папярэдніх даследаванняў па распазнаванні беларускага маўлення і фармулюецца задача гэтай работы. У раздзеле 2 праводзіцца аналіз патрабаванняў да датасэта для навучання мадэлі, апісваецца працэс збору дадзеных і аналіз атрыманага датасэта. У раздзеле 3 апісваецца пабудова і ацэнка якасці выніковай сістэмы распазнавання маўлення.

РАЗДЗЕЛ 1
АСАБЛІВАСЦІ ЗАДАЧЫ

1.1 Апісанне задачы

Задача пераўтварэння маўлення ў тэкст або задача распазнавання маўлення (**Speech-to-Text, STT** або **Automatic Speech Recognition, ASR**) – гэта адна з асноўных задачаў апрацоўкі натуральнага маўлення. Яе мэтай з'яўляецца аўтаматычная пабудова тэкставай транскрыпцыі для аўдыяфайла з запісаным маўленнем аднаго ці некалькіх людзей.

Прыкладам выкарыстання такой тэхналогіі з'яўляецца галасавы набор тэкста або пабудова галасавых інтэрфэйсаў для кіравання прыладамі, у тым ліку стварэнне галасавых дапаможнікаў. У апошнім выпадку тэкставая транскрыпцыя перадаецца на ўваход іншым мадэлям, якія ставяць у адпаведнасць распазнанаму тэксту набор дзеянняў, якія неабходна выканаць. Пасля гэтага галасавы дапаможнік мусіць паведаміць карыстальніку пра распазнаныя дзеянні і абвесціць пра іх далейшае выкананне. Зваротная камунікацыя з карыстальнікам можа ажыццяўляцца ў тэкставай форме. Але найбольш натуральным з'яўляецца адказ карыстальніку сінтэзаваным голасам: па згенераваным тэкставым адказе асобная мадэль сінтэзу маўлення (**Text-to-Speech, TTS**) стварае аўдыяфайл, які імкнецца найбольш якасна зымітаваць вымаўленне адказа чалавекам. У прыведзеным прыкладзе мадэль пераўтварэння маўлення ў тэкст з'яўляецца першым складнікам складанага ансамбля мадэляў, і пры яе адсутнасці немагчымая пабудова ўсёй сістэмы.

Фармальна, уваходны аўдыясігнал $X = (x_i)_{i=1}^{N}$ мадэль пераводзіць у набор словаў $W = (w_j)_{j=1}^{M}$. Прычым атрыманая транскрыпцыя будуецца без уліку пунктуацыі – атрыманыя транскрыпцыі складаюцца толькі са словаў. Даданне пунктуацыі з'яўляецца асобнай задачай апрацоўкі мовы.

Памер уваходнага сігналу N вызначаецца працягласцю запіса (аналагавага сігналу) і **частасцю дыскрэтызацыі**. Максімальнае значэнне выбаркі (фрэйма) x_i вызначаецца **шырынёю** – колькасцю бітаў, якія адводзяцца пад захаванне кожнай выбаркі (фрэйма). Аўдыясігнал можа быць захаваны ў адзіным **канале** (мона) або ў некалькіх (стэрэа). У апошнім выпадку аўдыясігнал можна ўявіць як $X = (X_j) = ((x_{j,i})_{i=1}^{N})_{j=1}^{K}$, дзе K – колькасць каналаў у запісе.

У залежнасці ад тыпу мадэлі на распазнаныя словы w_j можа як накладацца абмежаванне на прыналежнасць да вызначанага слоўніка V, так і не.

Для навучання мадэлі распазнавання маўлення неабходна сабраць адмысловы набор дадзеных (**датасэт)**. Ён мусіць складацца з: (1) набору гукавых файлаў з запісамі маўлення і (2) тэкставых транскрыпцыяў, якія ў дакладнасці адпавядаюць вымаўленаму на запісах тэксту. Каб пабудаваць устойлівую мадэль, здольную якасна распазнаваць маўленне ў разнастайных умовах (наяўнасць фонавага шуму, рэха, недасканаласць мікрафонаў), неабходна збіраць датасэт адпаведным чынам – аўдыязапісы мусіць мець дастатковую варыятыўнасць. Таксама варта памятаць пра наяўнасць для некаторых моваў розных дыялектаў. Калі для аднаго і таго ж слова існуюць некалькі дапушчальных варыянтаў вымаўлення, наш датасэт мусіць па магчымасці змяшчаць іх усе.

1.2 Актуальнасць

Найбольш даступнымі тэхналогіямі распазнавання маўлення з’яўляюцца прадукты ад вялікіх кампаніяў: Google, Microsoft, Amazon, Apple, Яндэкс. Іх агульным недахопам з’яўляецца абмежаваная колькасць моваў, якія яны падтрымліваюць. Так, ніводны з прыведзеных сервісаў не падтрымлівае распазнаванне беларускай мовы. Гэта значыць, што ў беларусаў няма магчымасці карыстацца галасавым наборам, галасавымі дапаможнікамі, галасавым перакладам па-беларуску.

Распрацоўка тэхналогіі распазнавання маўлення для беларускай мовы дазволіць не толькі спрасціць беларусам узаемадзеянне з камп’ютарамі і прыладамі, але і паспрыяе развіццю і захаванню мовы праз стварэнне сучасных і тэхналагічных адукацыйных рэсурсаў, зможа натхніць большую колькасць людзей на вывучэнне мовы.

1.3 Традыцыйныя падыходы да распазнавання маўлення. Асноўныя складнікі такіх мадэляў.

З прычыны таго, што ў некаторых сферах дзейнасці чалавека (кіраванне электроннымі прыладамі, біяметрычная ідэнтыфікацыя чалавека, інш.) задача распазнавання маўлення з'яўляецца надзвычай важнай, рашаць яе пачалі даволі даўно. Падыходы да яе рашэння змяняліся разам з развіццём інфармацыйных тэхналогіяў.

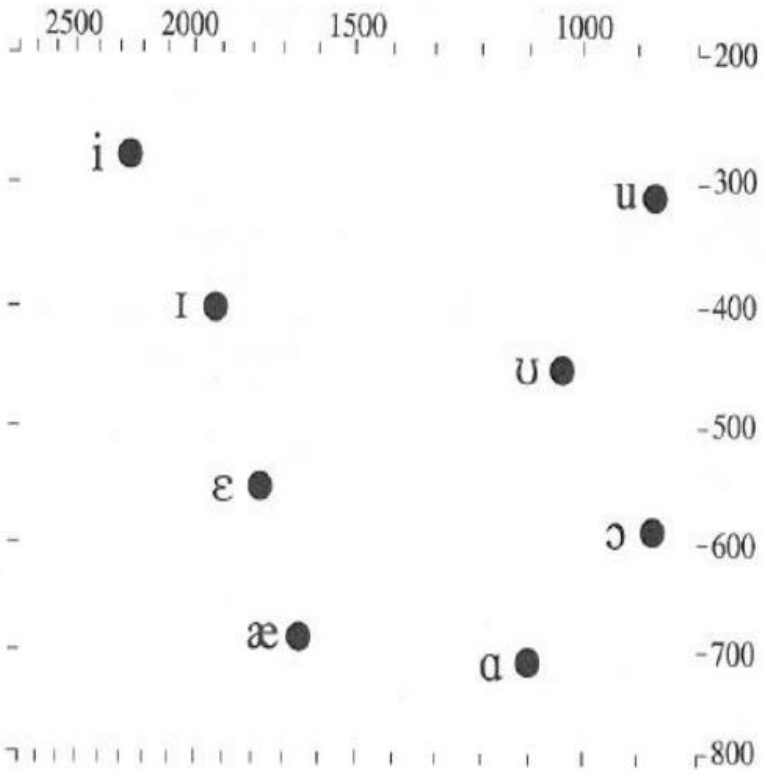
Для традыцыйнага падыходу (conventional ASR) уласціва стварэнне мадэлі распазнавання маўлення, якая складаецца з асобных кампанентаў, пабудаваных і навучаных незалежна адзін ад аднаго [16]. Найчасцей такімі кампанентамі з'яўляюцца: **блок вылучэння прыкмет** (feature extractor), **акустычная мадэль** (acoustic model), **фанетычны працэсар** (phonetical processor) ды **моўная мадэль** (language model).

1.3.1 Блок вылучэння прыкметаў

У традыцыйных мадэлях распазнавання маўлення метады вылучэння прыкметаў па ўваходным аўдыясігнале найчасцей праэктуюцца ўручную, самімі даследнікамі. Для гэтага патрабуюцца веды асаблівасцяў акустычнай прыроды маўлення.

Напрыклад вядома, што фанемы – асноўныя складнікі маўлення – адрозніваюцца між сабою ў частаснай прасторы [10]. На выяве 1 прыводзіцца дыяграма рассейвання для амерыканскіх ангельскіх галосных фанемаў, размешчаных у двухмернай прасторы першай і другой **фармант** F_1, F_2 – першых двух пікаў у спектральным дыяпазоне пры іх вымаўленні. Заўважна, што фанемы знаходзяцца адасоблена адна ад адной.

Для аўтаматызацыі вылучэння спектральных прыкметаў звычайна выкарыстоўваюць пабудову мел-частасных кепстральных каэфіцыэнтаў (MFCC) і метад кадавання з лінейным прэдыктарам (LPC) – яны дазваляюць апісаць галасавы сігнал у спектральнай прасторы ў кожны момант часу.



Выява 1. Дыяграма рассейвання для амерыканскіх ангельскіх галосных. Вертыкальная вось – фарманта F_1, гарызантальная вось – фарманта F_2 (герцы, Hz) [10]

1.3.2 Фанетычны працэсар

Фанетычны працэсар ставіць у адпаведнасць слову спосаб яго вымаўлення – набор фанемаў. Праініцыялізаваць такую мадэль можна з дапамогай падрыхтаваных загадзя фанемных слоўнікаў. Такія слоўнікі могуць змяшчаць па некалькі варыянтаў вымаўлення аднаго і таго ж слова, што тлумачыцца магчымай варыятыўнасцю ва ўнармаваным вымаўленні або наяўнасцю ў мове дыялектаў.

Каб фанетычны працэсар умеў працаваць з адсутнымі ў фанетычным слоўніку словамі, дадаткова будуецца мадэль пераводу графем у фанемы – статыстычная мадэль, якая на аснове правілаў мовы можа пабудаваць магчымы спосаб вымаўлення невядомага слова.

Атрыманы фанетычны працэсар выкарыстоўваюць падчас навучання акустычнай мадэлі для пераводу наяўных у корпусе тэкставых транскрыпцыяў у набор фанемаў.

1.3.3 Акустычная мадэль

Акустычная мадэль уяўляе сабою блок сістэмы распазнавання маўлення, які па вылучаных з уваходнага аўдыясігнала прыкметах будуе паслядоўнасць фанемаў, вымаўленых з найбольшай імавернасцю.

Так акустычная мадэль можа спярша вылучыць часавыя прамежкі з адносна нязменнымі значэннямі спектральных прыкметаў, а пасля для кожнага знойдзенага прамежка адшукаць фанему з найбольш падобнымі спектральнымі ўласцівасцямі гучання.

Гэта можна зрабіць, напрыклад, з выкарыстаннем мадэлі сумесі гаўсавых размеркаванняў (GMM), якая падчас навучання падбірае найбольш верагодныя наборы спектральных прыкметаў для кожнай фанемы, і схаванай маркаўскай мадэлі (HMM), якая выкарыстоўвае пабудаваную GMM і дазваляе падабраць найбольш верагодную паслядоўнасць фанемаў для ўваходнага аўдыясігналу [10].

1.3.4 Моўная мадэль

Моўная мадэль патрэбная для пераводу атрыманага па ўваходным аўдыязапісе набора фанемаў у набор словаў – выніковую транскрыпцыю. Для гэтага выкарыстоўваецца статыстычная мадэль, якая вызначае імавернасць сустрэчы новага слова ў залежнасці ад папярэдняга кантэкста. Такая мадэль будуецца аўтаматычна (без настаўніка) па вялікім корпусе тэкстаў [17].

Прыкладам моўнай мадэлі з’яўляецца **n-грамная мадэль**. N-грама – гэта паслядоўнасць n словаў. Для n-грамнай моўнай мадэлі імавернасць сустрэчы паслядоўнасці словаў $W = (w_i)_{i=1}^{m}$ набліжаецца наступнай формулай:

$$p(W) \approx \prod_{i=1}^{m} p(w_i | w_{max\{1,\ i-n+1\}}, \ldots, w_{i-1}) \qquad (1.1)$$

дзе

$$p(w_i|w_{i-n+1},\ldots,w_{i-1}) = \frac{C(w_{i-n+1},\ldots,w_{i-1},w_i)}{C(w_{i-n+1},\ldots,w_{i-1})} \quad (1.2)$$

а c(w) – частасць сустрэчы слова або паслядоўнасці словаў *w* у навучальным корпусе [8].

1.4 Метады з выкарыстаннем глыбокага навучання

1.4.1 Матывацыя выкарыстання метадаў глыбокага навучання

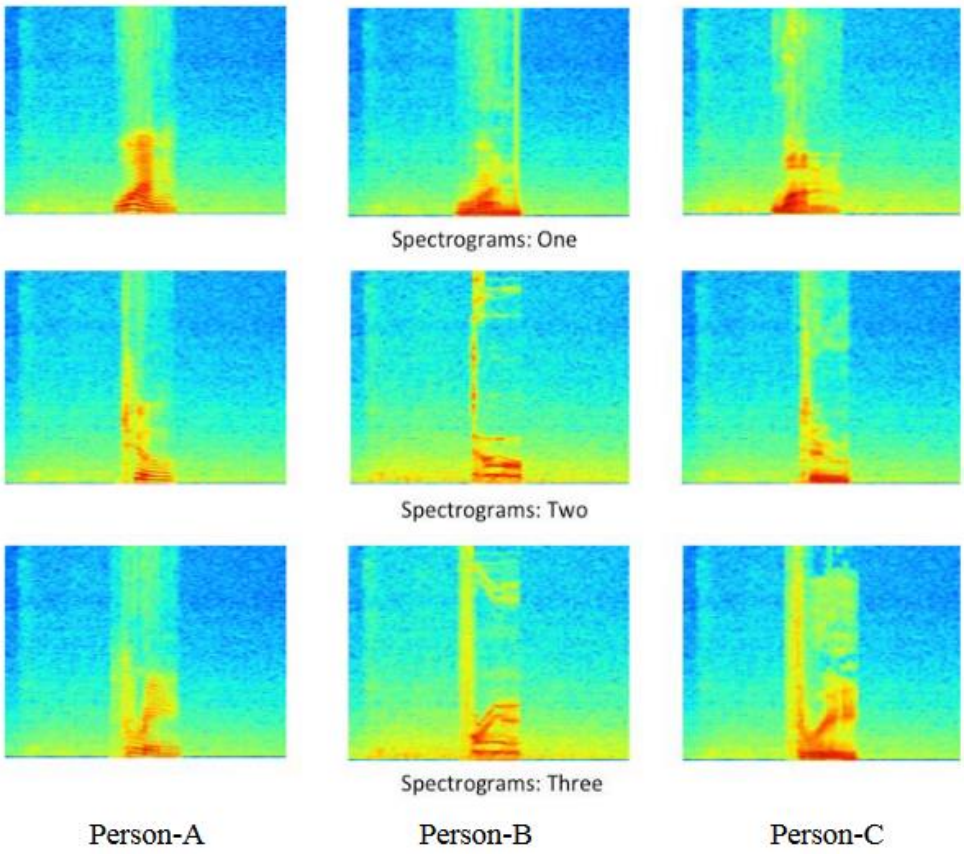
Асноўнай праблемай традыцыйных метадаў распазнавання маўлення з'яўляецца неабходнасць ручнога канструявання пайплайнаў для вылучэння прыкметаў для апісання галасавога сігналу. Праблема ўскладняецца магчымай наяўнасцю на гукавым запісе фонавых шумоў ды рэха, недасканаласцю мікрафонаў а таксама вялікай варыятыўнасцю ў вымаўленні аднолькавых словаў паміж рознымі людзьмі і дыялектамі адной мовы. Для забеспячэння ўстойлівасці мадэлі ў згаданых умовах даследнікам неабходна прыкладаць многа намаганняў і канструяваць дадатковыя складнікі мадэлі.

Метады глыбокага навучання дазваляюць істотна аблегчыць задачу. Іх здольнасць аўтаматычна вылучаць прыкметы, заканамернасці і шаблоны, карыстаючы вялікія аб'ёмы дадзеных, дазваляюць пазбавіцца ад канструявання складаных кампанентаў уручную.

На выяве 2 прыводзяцца спектраграмы вымаўлення трох словаў "one", "two", "three" трыма рознымі людзьмі. Бачна, што спектраграмы для адных словаў падобныя паміж сабою. Пры карыстанні метадаў глыбокага навучання знікае патрэба ручнога пошуку інварыянтных прыкметаў для адрознення фанемаў або цэлых словаў па наяўнай спектраграме – гэты працэс аўтаматызуецца нейроннай сеткай.

Нейрасеткавыя мадэлі выкарыстоўваць альбо як замену некаторым складнікамі (напрыклад толькі для вылучэння прыкметаў), альбо як самастойныя end-to-end сістэмы. **End-to-end** сістэмамі прынята называць [16] сістэмы, якія будуюць наўпроставую адпаведнасць паміж уваходнымі дадзенымі і выходнымі, прычым усе кампаненты такой сістэмы вучацца адначасова, а пры навучанні

аптымізуецца канчатковая метрыка (напрыклад WER). Прыкладам такой сістэмы з’яўляецца энкодэрна-дэкодэрная архітэктура глыбокага навучання. Энкодэрная частка такіх мадэляў стварае ўнутранае прадстаўленне для ўваходных дадзеных (напрыклад аўдыяфайл), якое з дапамогай дэкодэрнай часткі ператвараецца ў дадзеныя выходнага фармата (напрыклад літары). Апошнім часам end-to-end сістэмы паказваюць найлепшыя рэзультаты для розных задачаў машыновага навучання [6].



Выява 2. Спектраграмы вымаўлення словаў "one", "two", "three" (па вертыкалі) трыма рознымі дыктарамі (па гарызанталі) [10]

1.4.2 Выраўноўванне тэкставых транскрыпцыяў па часе. CTC.

Важнай праблемай пры пабудове нейрасеткавых мадэляў з’яўляецца адсутнасць выраўнаваных па часе тэкставых транскрыпцыяў. Калі для невялікіх

датасэтаў яшчэ можна пабудаваць анатацыі, у які момант часу вымаўляецца якая фанема або слова, то для вялікіх датасэтаў гэта становіцца непрактычным.

Калі б у нас былі такія транскрыпцыі, то мы маглі бы разбіць уваходны гукавы сігнал на кавалкі, а пасля для кожнага з іх рашыць задачу класіфікацыі на фанемы або словы. З іншага боку, нават пры наяўных выраўнаваных па часе транскрыпцыях, навучаную мадэль было бы немагчымы скарыстаць для стварэння транскрыпцыі для новага аўдыяфайла, для якога транскрыпцыя адсутнічае ўвогуле. Падобная праблема паўстае таксама ў іншых задачах: распазнавання падзеяў на відэа, распазнавання почырку па відарысах або паслядоўнасці рухаў мышы/стылуса і іншых.

Рашыць праблему дазваляе алгарытм Connectionist temporal classification (**CTC**) [6]. Ён дазваляе знайсці патрэбнае пераўтварэнне ўваходнай паслядоўнасці X (прыкметы, вылучаныя з аўдыяфайла) у выходную паслядоўнасць Y (тэкставая транскрыпцыя). Набліжанае рашэнне $\hat{Y}$ шукаецца як $\hat{Y} = arg\max_{Y} p(Y|X)$.

CTC працуе з выхадам мадэлі, якая генеруе размеркаванне імавернасцяў кожнага сімвала на кожны часавы прамежак. Па атрыманым размеркаванні абіраецца набор найбольш імаверных паслядоўнасцей, напрыклад алгарытмам Beam search – промневага пошуку.

Прыклад работы алгарытма прыводзіцца ў табліцы 1.

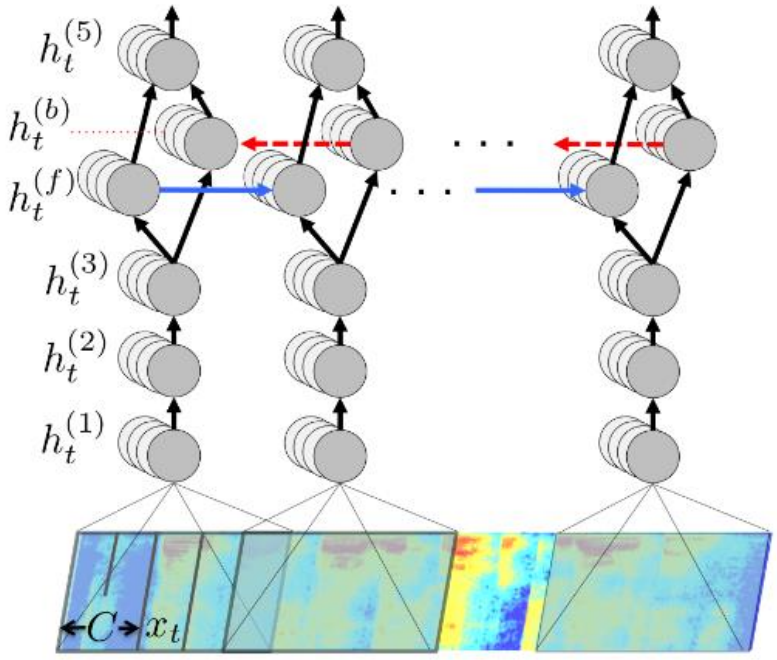
Табліца 1. Прыклад работы алгарытма CTC

Уваход, X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Прамежкавы выхад	м	м	я	к	к	є	к	і	і
Выхад, $\hat{Y}$	м		я	к			к	і	

Кожнаму элементу ўваходнай паслядоўнасці алгарытм ставіць у адпаведнасць сімвал з алфавіту Σ — літараў мовы разам з адмысловым сімвалам “є”. Выходная паслядоўнасць затым сціскаецца — аднолькавыя сімвалы замяняюцца на адзін. Сімвал “є” карыстаецца, каб перадухіліць сціскання аднолькавых сімвалаў, якія адпавядаюць двум паслядоўным аднолькавым літарам у сапраўднай транскрыпцыі («кк» у слове «мяккі»).

1.4.3 Прыклад end-to-end архітэктуры. DeepSpeech

Першай нейрасеткавай end-to-end мадэлью распазнавання маўлення стала **DeepSpeech** [3]. Яна значна палепшыла рэзультаты распазнавання маўлення (у тым ліку для аўдыязапісаў, створаных у шумных умовах) у параўнанні з папярэднімі мадэлямі, заснаванымі на традыцыйных падыходах. На ўваход архітэктуры падаецца спектраграма аўдыяфайла $x_{t,p}$, дзе t адпавядае моманту часа, а p – індэксу прамежка частасцяў, на якія спектраграма разбіваецца падчас дыскрэтызацыі. На выхадзе, замест генерацыі паслядоўнасці фанемаў, архітэктура дазваляе адразу генераваць паслядоўнасць сімвалаў з алфавіта мовы. Па аналогіі з традыцыйнымі метадамі распазнавання маўлення такую нейронную архітэктуру часам называюць **акустычнай мадэлью** [16]



Выява 3. Архітэктура DeepSpeech [3]

Схема архітэктуры прыводзіцца на выяве 3. Яна ўяўляе сабою рэкурэнтную нейронную сетку (RNN), якая складаецца з 5 унутраных слаёў. Уваходнымі значэннямі для першага слою $h_t^{(1)}$ у момант часу t з'яўляецца фрэйм спектраграмы x_t разам з C суседнімі фрэймамі з абодвух бакоў. Першыя 3 слаі не з'яўляюцца рэкурэнтнымі – яны выконваюць ролю блока вылучэння прыкметаў.

Чацвёрты слой $h_t^{(4)} = (h_t^{(f)}, h_t^{(b)})$ з'яўляецца двухскіраванам рэкурэнтным слоем (bi-directional recurrent layer). Ён складаецца з двух унутраных блокаў:

прамога $h_t^{(f)}$ і адваротнага $h_t^{(b)}$ распаўсюду сігналу (forward and backward recurrence). Кожны з гэтых унутраных блокаў змяшчае інфармацыю пра ўвесь левы (або правы) кантэкст бягучага фрэйма. Пяты ўнутраны слой архітэктуры аб'ядноўвае выхады ўнутраных блокаў на чацвёртым слаі і скіроўвае свой выхад праз softmax-слой. Выхадам апошняга слоя з'яўляецца $\hat{y}_{t,k}$ – вектар імавернасці сустрэчы ў фрэйме x_t сімвала c_t з алфавіта мовы.

Мадэль вучыцца з выкарыстаннем CTC як функцыі стратаў [6]. Для рэгулярызацыі выкарыстоўваецца dropout [14], аднак толькі для першых трох (нерэкурэнтных) слаёў.

Архітэктура DeepSpeech дазволіла істотна палепшыць распазнаванне маўленне, аднак для яе навучання аўтарам спатрэбілася каласальная колькасць анатаваных дадзеных: блізу 5'000 гадзін маўлення ад 9'600 дыктараў. Большую частку дадзеных аўтары збіралі самастойна, а сабраны імі датасэт з'яўляецца прапрыетарным (закрытым для вольнага выкарыстання).

Вялікія аб'ёмы дадзеных, патрэбных для навучання DeepSpeech "з нуля" (то бок без выкарыстання трансфернага навучання), а таксама адсутнасць у вольным доступе вялікіх датасэтаў (асабліва для непашыраных моваў, кшталту беларускай) ускладняюць выкарыстанне дадзенай мадэлі.

1.4.4 Моўная мадэль у end-to-end архітэктурах

Нейронная акустычная мадэль, навучаная на дастаткова вялікіх аб'ёмах дадзеных, здольная ствараць чытэльныя і адносна якасныя тэкставыя транскрыпцыі. Аднак такая мадэль дапускае граматычныя ды арфаграфічныя памылкі, або ўвогуле прадказвае неіснуючыя ў мове словы. Падобныя праблемы часта бываюць выкліканы **словамі-амафонамі** – словамі, якія вымаўляюцца аднолькава, але маюць рознае напісанне – або блізкімі па гучанні фразамі.

Таксама мадэль можа памыляцца на словах, адсутных у навучальнай выбарцы. Амаль немагчыма сабраць для навучання такі агучаны датасэт, які бы ўбіраў у сябе ўсе словы мовы з усімі магчымымі формамі ўжывання. Да таго ж збор такога датасэту быў бы непрактычны, бо для кожнай мовы вымагаў бы велізарных намаганняў па зборы дадзеных.

Без дадатковых ведаў пра граматыку мовы, без слоўніка з пералікам дапушчальных словаў, а таксама без разумення кантэксту акустычнай мадэлі

няпроста выправіць свае памылкі. Аднак варта заўважыць, што сучасныя нейрасеткавыя мадэлі, напрыклад заснаваныя на архітэктуры увагі (attention [8]), здольныя ўлічваць кантэкст усёй уваходнай фразы пра пабудове транскрыпцыі, што змяншае колькасць памылак [6].

Для выпраўлення памылак акустычнай мадэлі, яе выхад (размеркаванне імавернасцяў сімвалаў з мовы для кожнага моманта часу) звычайна падаюць на ўваход **моўнай мадэлі**, якая ўвасабляе ў сабе правілы мовы, а працэс стварэння выніковых транскрыпцыяў называюць **дэкодынгам** [3]. Як і ў выпадку з традыцыйнымі сістэмамі распазнавання маўлення, моўнай мадэлью можа з'яўляцца, напрыклад, n-грамная мадэль, пабудаваная па тэкставым корпусе у рэжыме без настаўніка. Даданне моўнай мадэлі істотна паляпшае выніковую якасць сістэм распазнавання маўлення.

У табліцы 2 прыведзены прыклады транскрыпцыяў, пабудаваных мадэлью DeepSpeech. Першы слупок змяшчае транскрыпцыі, атрыманыя акустычнай мадэлью, другі – транскрыпцыі, выпраўленыя з дапамогай моўнай мадэлі. Заўважна, што акустычная мадэль памыляецца: (1, 3 радкі) у выпадку калі вымаўленне слова не супадае з напісаннем ("bostin" vs "boston"), (2 радок) на рэдкім выразе ("Narendra Modi" – уласнае імя), (4 радок) у выпадку блізкіх па гучанні словаў ("rose" і "roast" не з'яўляюцца поўнымі амафонамі, але гучаць падобна). Моўная мадэль дапамагае выправіць гэтыя памылкі.

Табліца 2. Прыклад выпраўлення моўнай мадэлью прадказанняў акустычнай мадэлі

Выхад акустычнай мадэлі	Рэзультат выпраўлення моўнай мадэлью
what is the weather like in bostin right now	what is the weather like in boston right now
prime miniter nerenr modi	prime minister narendra modi
arther n tickets for the game	are there any tickets for the game
rose beef looming before us	roast beef looming before us

Функцыя стратаў для аб'яднанай акустычнай і моўнай мадэлі можа выглядаць наступным чынам [3]:

$$Q(c) = log(P(c|x)) + \alpha\ log(P_{lm}(c)) + \beta\ word\_count(c)$$

дзе $P(c|x)$ – ацэнка імавернасці атрыманай транскрыпцыі c пры наяўнасці ўваходнага сігналу x (будуецца акустычнай мадэллю); $P_{lm}(c)$ – імавернасць сустрэчы атрыманай транскрыпцыі моўнай мадэллю; α ды β – гіперпараметры мадэлі, якія дазваляюць кантраляваць унёсак акустычнай мадэлі, моўнай мадэлі і агульнай працягласці транскрыпцыі ($word\_count$) у выніковае значэнне функцыі стратаў. Гіперпараметры (α, β) падбіраюцца, напрыклад, з дапамогай крос-валідацыі для кожнай задачы індывідуальна. Аптымізацыя такой функцыі стратаў праводзіцца, напрыклад, алгарытмам промневага пошука (beam search).

Важна заўважыць, што для пабудовы якаснай моўнай мадэлі неабходна выкарыстоўваць вялікі тэкставы корпус, які мусіць пакрываць вялікую колькасць словаў з мовы і ўтрымліваць разнастайныя формы іх ужывання.

1.4.5 Прыклад end-to-end архітэктуры. Wav2Vec2

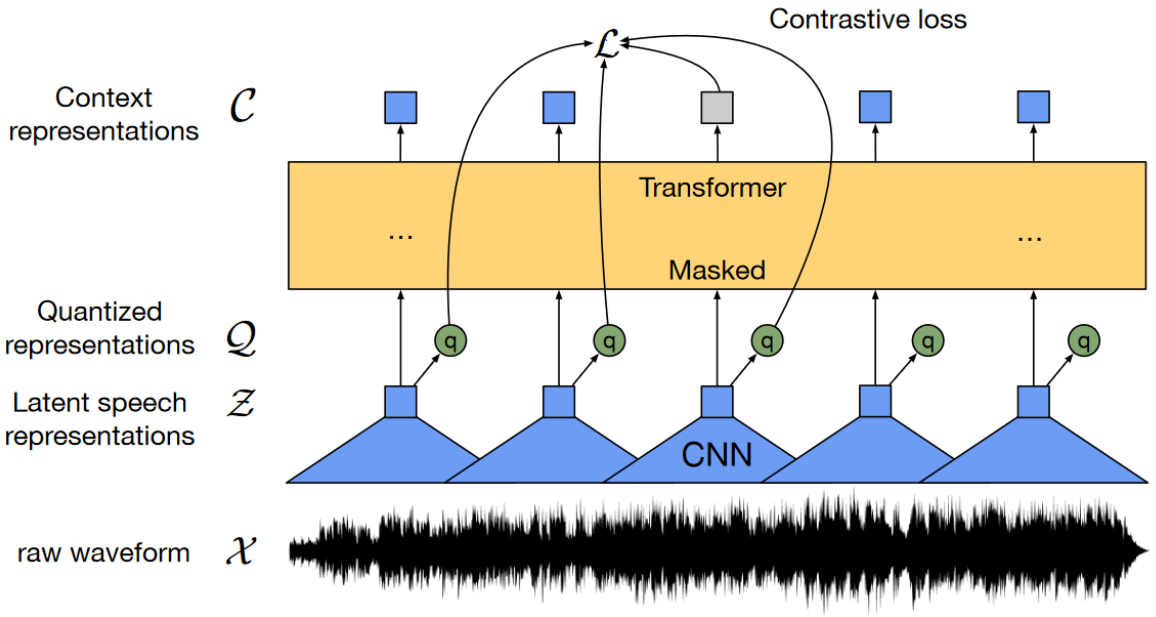
Прыкладам больш сучаснай end-to-end архітэктуры ў параўнанні з DeepSpeech з'яўляецца **wav2vec2** [6]. У яе аснове палягае ідэя пабудовы мадэлі, здольнай пераднавучвацца на корпусе неанатаваных дадзеных (аўдыяфайлы з запісаным маўленнем, але з адсутнымі транскрыпцыямі) без настаўніка. Атрыманая пераднавучаная мадэль здольная вылучаць з уваходнага аўдыязапіса прыкметы, якія выкарыстоўваюцца ў далейшых падзадачах — такіх як пераўтварэнне маўлення ў тэкст.

З прычыны таго, што пераднавучанне мадэлі адбываецца ў рэжыме без настаўніка, для гэтай задачы становіцца магчымым выкарыстанне вялікай колькасці неанатаваных аўдыязапісаў: радыёэфіраў, аўдыякніг, падкастаў, запісаў спектакляў, тэлешоў, выступаў, канферэнцыяў – для вялікай колькасці з прыведзеных крыніц аўдыязапісаў маўлення няма транскрыпцыяў, асабліва для не пашыраных у свеце мовах (такіх як беларуская, чэшская, шведская і інш).

Для выкарыстання мадэлі ў падзадачы (напрыклад для распазнавання маўлення) пераднавучаную мадэль неабходна давучыць на анатаваным датасэце (трансфернае навучанне). Займальна тое, што аўтарамі было паказана [6], што мадэль дасягае добрых рэзультатаў на падзадачы нават пры невялікім памеры анатаванага датасэту. Гэта робіць магчымым выкарыстанне архітэктуры wav2vec2 для задачы распазнавання маўлення для моваў з адсутнымі вялікімі карпусамі анатаваных дадзеных (такіх як беларуская).

Так, у адным з эксперыментаў [6] аўтары wav2vec2 спярша пераднавучылі мадэль на 960 гадзінах неанатаванага ангельскага маўлення, далей давучылі мадэль толькі на 10 хвілінах анатаванага маўлення і атрымалі прымальную якасць мадэлі (base-мадэль, 4-грамная моўная мадэль – 9.1 test clean WER, 15.6 test other WER). Такі эксперымент даказвае магчымасць пабудовы якасных сістэмаў распазнавання маўлення пры адсутнасці вялікіх анатаваных датасэтаў (зрэшты для пераднавучання патрабуецца корпус неанатаваных дадзеных).

Схема архітэктуры wav2vec2 прыведзеная на выяве 4.



Выява 4. Архітэктура wav2vec2. [6]

Напачатку ўваходны сігнал апрацоўваецца наборам **згорткавых нейронных сетак (CNN)** для вылучэння набору прыкметаў (Z на выяве), кожная з якіх захоўвае інфармацыю пра сігнал пэўнай працягласці (напрыклад 2 мс). Па аналогіі з задачай **MLM (Masked language modeling)** частка атрыманых прыкметаў замяняецца на **токен-маску**. Далей прыкметы падаюцца на ўваход **трансформернай сетцы** [8] для пабудовы кантэкстуалізаваных прыкметаў (C на выяве). Яшчэ адным крокам з'яўляецца **квантаванне** прыкметаў вылучаных згорткавымі сеткамі, то бок іх перавод з непарыўнага ў концае мноства мноства значэнняў (з Z у Q на выяве).

Задача пераднавучання мадэлі фармулюецца як **кантрастная задача** пошуку: для кожнай замаскаванай кантэкстуалізавай прыкметы c_t шукаецца найбольш падобная да яе квантаваная прыкмета q_t. Такім чынам атрымліваецца дасягнуць добрай якасці выніковых прыкметаў, вылучаных мадэлью.

1.5 Ацэнка якасці мадэлі распазнавання маўлення. WER

Для ацэнкі якасці мадэлі ў задачы распазнавання маўлення звычайна карыстаецца метрыка **WER (Word Error Rate)** — доля памылковых словаў [14]. Метрыка грунтуецца на **рэдактарскай адлегласці** паміж радкамі *S* (source, зыходны тэкст) ды *T* (target, мэтавы тэкст).

Фармальна, для пары *S, T* з дапамогай алгарытму пошуку рэдактарскай адлегласці (рэалізуецца дынамічным праграмаваннем) знаходзіцца колькасць правільных словаў (П) і колькасць кожнага з тыпу выпраўленняў, якія неабходна выканаць над мэтавым тэкстам, каб прывесці яго да зыходнага. Тыпамі выпраўленняў з'яўляюцца: замена слова (З), выдаленне лова (В), устаўка слова (У).

Так, значэнне метрыкі вызначаецца наступнай формулай:

$$WER = \frac{З + В + У}{З + В + П} \tag{1.3}$$

Разгледзім наступны прыклад падліку метрыкі:

- *T* = «раз два тры чатыры».
- *S* = «раз раз чатыры»
- Паслядоўнасць аперацыяў на мэтавым тэкстам, каб прывесці яго да зыходнага: П, З, В, П.
- П = 2, З = 1, В = 1, У = 0
- $WER = \frac{1+1+0}{2+1+1} = 50\%$

1.6 Папярэднія работы па распазнаванні беларускага маўлення і пастаноўка задачы

Эксперыменты па распазнаванні беларускага маўлення праводзіліся і раней. Так, у [1] праводзіўся эксперымент па распрацоўцы галасавога інтэрфэйса кіравання робатам з абмежаваным наборам камандаў. Таксама эксперымент па распазнаванні абмежаванага набору словаў быў праведзены ў [21].

Першым даследаваннем па распрацоўцы мадэлі распазнавання адвольнага (без абмежавання на распазнаваемыя словы) беларускага маўлення стала работа [2]. Распрацаваная мадэль уяўляе сабою аб'яднанне традыцыйных падыходаў да распазнавання маўлення разам з выкарыстаннем глыбокага навучання. Так, мадэль пераводу графем у фанемы (g2p-seq2seq) рэалізоўвалася з дапамогай рэкурэнтных нейронных сетак (RNN). Для рэалізацыі быў абраны фрэймворк CMU Sphinx.

Аўтарамі была праведзеная грунтоўная работа па зборы і апрацоўцы датасэту. Ім атрымалася сабраць корпус беларускіх начытаных тэкстаў агульнай працягласцю 18.5 гадзін. Ён складаўся з 16 гадзін начытаных дыктарамі тэкстаў з мастацкіх твораў і 2.5 гадзін начытаных звычайнымі людзьмі тэкстаў з часопісаў ды газет на беларускай мове. Агулам у запісе прынялі удзел 23 дыктары.

Аднак у даследаванні былі наступныя недахопы:

- "Усе тэксты былі начытаны ў асобным кабінеце, спецыяльна абсталяваным для запісу голасу з адсутнасцю знешніх шумоў, але з прысутнасцю натуральнага шумавога фону". Навучанай на такіх дадзеных мадэлі будзе цяжка распазнаваць маўленне ў звычайных умовах (напрыклад маўленне, запісанае на вуліцы на мікрафон тэлефона).
- "Мастацкія тэксты былі начытаны прафесійнымі дыктарамі". Маўленне дыктараў адрозніваецца выразнай, а часам і залішне артыстычнай дыкцыяй. Праз гэта мадэлі можа быць цяжка распазнаваць маўленне звычайных людзей. Аднак аўтары спрабавалі ўлічыць гэта ў зборы дадзеных і акрамя дыктараў запрасілі да агучвання звычайных людзей.
- "Дыктары не выбіраліся з пункту гледжання разнастайнасці дыялекта". Пры адсутнасці ў датасэце дыялектнай варыятыўнасці ў вымаўленні словаў, мадэлі будзе цяжка разумець некаторых дыктараў.
- У даследаванні адсутнічаюць формулы для метрык, выкарыстаных для ацэнкі якасці мадэляў. Акрамя гэтага самі значэнні метрык супярэчлівыя: "Падлікі

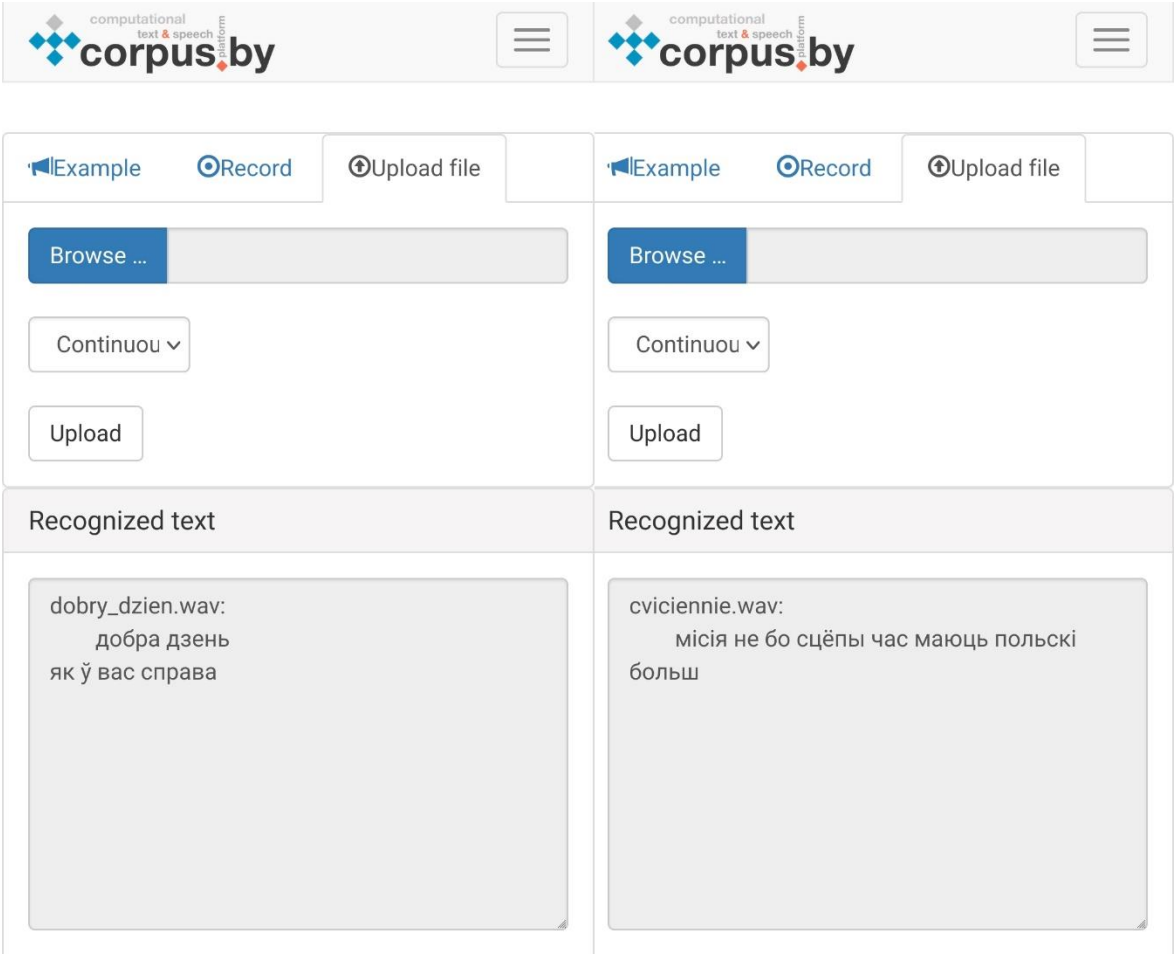
памылак былі наступнымі: для распазнавання сказаў – 18,6 % (49/264), асобных слоў – 5,8 % (170/2917)". Наступная цытата: "Вынікі тэставання натрэніраванай мадэлі для беларускай мовы паказалі *дакладнасць* пры распазнаванні сказаў 67,3 %, а пры распазнаванні асобных слоў – 14,7 %.". Няясна, што разумеецца пад словамі "памылкі" і "дакладнасць".

- Малы памер і варыятыўнасць датасэта, на якім ацэньвалася мадэль, не дазваляе ацаніць яе якасць ва ўмовах, набліжаных да жыцця.
- Моўная мадэль будавалася толькі на тэкстах, якія ўвайшлі ў датасэт (транскрыпцыі) – то бок на невялікім корпусе тэкстаў. Звычайна для пабудовы моўнай мадэлі патрабуецца вялікі корпус тэкстаў – большы, чым корпус транскрыпцыяў. Таксама неабходна імкнуцца да збору тэкстаў з самых розных крыніцаў, каб перадаць як мага большую варыятыўнасць моўных канструкцыяў.

Аўтары даследавання загрузілі атрыманую мадэль распазнавання маўлення ў вольны доступ [5]. Было вырашана праверыць якасць распазнавання адвольнага маўлення на прыкладзе наступных фразаў: (1) *"Добры дзень. Як у вас справы?"* і (2) *"Цвіценне працягваецца з мая да позняй восені"*. На выяве 5 прыводзяцца рэзультаты распазнавання. Першая фраза была часткова распазнаная, аднак пабудаваная транскрыпцыя для другой фразы, на жаль, далёкая ад вымаўленага тэксту.

Такім чынам паставім перад сабою **наступныя мэты** у межах дадзенай магістарскай работы:

1. Пабудаваць мадэль для распазнавання адвольнага (без абмежавання на распазнаваемыя словы) беларускага маўлення добрай якасці. Мадэль будзем будаваць на аснове нейрасеткавай end-to-end архітэктуры wav2vec2 з прычыны таго, што мадэль з'яўляецца найбольш сучаснай, а таксама праз нізкія патрабаванні да колькасці анатаваных дадзеных.
2. Сабраць неабходны для гэтага корпус начытаных тэкстаў на беларускай мове дастатковага аб'ёму з як мага большай варыятыўнасцю дыктараў і ўмоваў запісу.



Выява 5. Прыклад работы мадэлі [5] для распазнавання адвольнага маўлення. На прыкладзе злева была начытана фраза “Добры дзень. Як у вас справы?”, справа - “Цвіценне працягваецца з мая да позняй восені”.

Высновы

У дадзеным раздзеле былі прааналізаваныя метады распазнавання маўлення: як традыцыйныя, так і сучасныя на аснове нейронных end-to-end архітэктур. Былі апісаныя асноўныя складнікі традыцыйных мадэляў, апісаныя нейронныя архітэктуры DeepSpeech, wav2vec2, апісаны алгарытм выраўноўвання транскрыпцыяў CTC і метрыка Word Error Rate.

Былі даследаваны папярэднія работы па распазнаванні беларускага маўлення, вызначаны іх недахопы і пастаўленыя задачы для дадзенай работы: (1) пабудаваць мадэль распазнавання адвольнага беларускага маўлення, (2) сабраць неабходны для гэтага датасэт.

РАЗДЗЕЛ 2
ЗБОР І АНАЛІЗ САБРАНЫХ ДАДЗЕНЫХ

2.1 Патрабаванні да датасэту

Для выканання пастаўленай у рабоце задачы – навучання мадэлі wav2vec2 для распазнавання беларускага маўлення – было неабходна сабраць анатаваны датасэт, які мусіць складацца з аўдыяфайлаў з беларускім маўленнем і адпаведных ім тэкставых транскрыпцыяў. Датасэт павінен быць дастатковага памеру. Эксперыменты ў [6] паказваюць, што для атрымання прымальнай якасці мадэлі распазнавання, аўтарам спатрэбілася 960 гадзін неанатаванага маўлення для пераднавучання мадэлі і мінімум 10 хвілін анатаванага маўлення для давучвання мадэлі. Але гэты эксперымент ставіўся адной і той жа мовы: і пераднавучанне мадэлі, і яе давучванне праводзіліся для ангельскай мовы.

Было вырашана сабраць анатаваны корпус як мага большага памеру. Агульным фактам у глыбокім навучанні з’яўляецца паляпшэнне якасці мадэлі пры павелічэнні аб’ёмаў дадзеных для навучання, і эксперыменты з [6] гэта пацвярджаюць. Да таго ж, калі атрымаецца сабраць напраўду вялікі датасэт, то можна будзе не толькі давучыць на беларускім датасэце пераднавучаную на іншай мове мадэль, але паспрабаваць таксама правесці само пераднавучанне wav2vec2-мадэлі з нуля на сабраным беларускім датасэце. Пачатковай мэтай ставілася сабраць **мінімум 100 гадзін** анатаванага беларускага маўлення. У дадатак сабраны датасэт мусіць мець як мага большую **варыятыўнасць дыктараў** (пол, узрост, тэмп маўлення, узровень валодання беларускай мовай) **і ўмоваў запісу** (фонавы шум, рэха, розныя мікрафоны), каб забяспечыць якасную работу пабудаванай сістэмы ў розных умовах.

Важна адзначыць, што значная частка беларусаў дапускае памылкі ў літаратурных фанетычных нормах беларускай мовы. Напрыклад згодна з літаратурнай нормай слова “усміхаешся” неабходна чытаць як [ус'м'іхајэс:'а] – адбываецца асіміляцыя шыпячага “ш” да свісцячага “с”. Аднак многа хто вымаўляе гэтае слова як [ус'м'іхајэшс'а] – без аніякай асіміляцыі. Каб дазволіць мадэлі разумець як мага большую колькасць беларусаў, мы мусім улічваць падобныя разыходжанні літаратурнай мовы з узусам. Адпаведна датасэт мусіць змяшчаць абодва прыведзеныя вымаўленні слова “ўсміхаешся” і для ўсіх падобных словаў.

2.2 Выбар падыходу да збору дадзеных

Каб сабраць датасэт патрэбнага памеру было вырашана выкарыстаць **краўдсорсынгавыя** (ад ангельскага слова “crowdsourcing”) платформы – анлайн-платформы для калектыўнага збору дадзеных пры ўдзеле вялікай колькасці ўдзельнікаў.

Патрэбны для работы датасэт можна сабраць двума рознымі падыходамі:

1. Сабраць набор тэкстаў для іх далейшага агучвання (начыткі)
2. Сабраць базу аўдыязапісаў з беларускім маўленнем і альбо атрымаць ужо даступныя для іх тэкставыя транскрыпцыі (напрыклад для аўдыякніг), альбо пабудаваць такія транскрыпцыі самім

Для пачатку было вырашана засяродзіцца на першым падыходзе з прычыны таго, што ён дазваляе кантраляваць працэс начытання тэкстаў. Так, да ўдзелу ў праэкце можна прыцягнуць вялікую колькасць людзей, і кожны з іх будзе павялічваць агульную варыятыўнасць галасоў, акцэнтаў, вымаўленняў. У дадатак, калі весці збор дадзеных не ў адной студыі, а ў розных месцах, гэта забяспечыць варыятыўнасць акустычных умоваў запісу.

У другім падыходзе для пабудовы тэкставых транскрыпцыяў кожны з сабраных аўдыяфайлаў неабходна разбіць на неперасечныя кавалкі і размеркаваць іх па ўдзельніках працэса збору дадзеных. У выпадку наяўнасці тэкставых транскрыпцыяў, неабходна нарэзаць на неперасечныя кавалкі як працяглыя аўдыяфайлы, так і адпаведныя ім транскрыпцыі. Для гэтага можна выкарыстаць існуючыя прылады, напрыклад з фрэймворка Kaldi [18]. Недахопам такога падыхода з'яўляецца тое, што даўгія аўдыязапісы часта бываюць начытаныя адным чалавекам (напрыклад аўдыякнігі, радыёэфіры). Гэта стварае дызбаланс па размеркаванні дыктараў у датасэце: агульная працягласць запісаў адных людзей будзе складаць некалькі дзясяткаў хвілін або нават гадзін, у той час як працягласць гучання некаторых іншых людзей можа скласці ўсяго толькі пару секунд.

2.3 Платформы для збору дадзеных

Такім чынам было вырашана спярша сабраць корпус тэкстаў на беларускай мове, а пасля іх агучыць. Але перш чым распачаць працэс збору тэкставага корпусу

неабходна было вызначыцца з платформай для далейшай начыткі тэкстаў і азнаёміцца з умовамі яе выкарыстання.

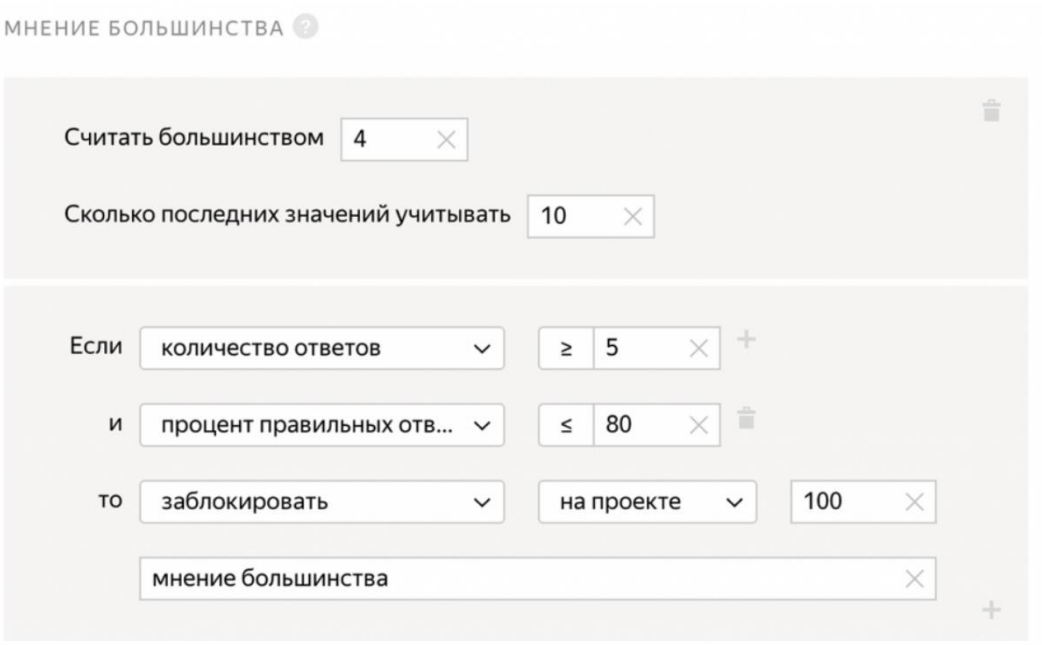
Былі разгледжаныя наступныя вядомыя краўдсорсынгавыя платформы для начытання сабраных тэкстаў:

- Amazon Mechanical Turk
- Яндэкс.Талака
- Mozilla Common Voice

Яндэкс.Талака и **Amazon Mechanical Turk** з'яўляюцца прыкладамі краўдсорсынгавых анлайн-платформаў, дзе людзям прапаноўвацца выконваць простыя задачы за грашовую аплату. Так людзям могуць прапанаваць вылучаць машыны і мінакоў на фотаздымках вуліц, ацэньваць таксічнасць каментароў і начытваць прапанаваныя тэксты. Найчасцей такія платформы выкарыстоўваюць для стварэння датасэтаў для задачаў машыновага навучання. Так, выкарыстанне платформы Mechanical Turk для стварэння датасэту агучаных тэкстаў апісваецца ў [7].

Беларускае слова "талака" ў назве платформы "Яндэкс.Талака" элегантна падкрэслівае краўдсорсынгавы характар работы сервіса. Адно са значэнняў слова "талака" адпавядае калектыўнай дапамозе пры выкананні сельскагаспадарчых работ, а ў другім значэнні слова азначае проста групу людзей.

І Яндэкс.Талака, і Mechanical Turk дазваляюць кантраляваць якасць выканання заданняў кожным з выканаўцаў, напрыклад з дапамогай **перакрыцця**, калі адное і тое ж заданне выконваецца некалькімі выканаўцамі, іх рэзультаты параўноўваюцца між сабою і абіраецца нейкі адзін канчатковы. Напрыклад для задачы стварэння набора анатацыяў для задачы класіфікацыі, а ў якасці "правільнага" абіраецца найбольш часты адказ. Таксама ёсць магчымасць дадаваць да спісу неразмечаных аб'ектаў **кантрольныя аб'екты** (пытанні) – аб'екты, пазнакі для якіх вядомыя. На такіх аб'ектах праводзіцца кантроль якасці асобных выканаўцаў. Акрамя гэтага можна падлічваць **каэфіцыэнт узгодненасці** паміж рознымі выканаўцамі (для задачы вылучэння аб'ектаў на фота такой мерай можа быць сярэдняе значэнне меры Жакара – Intersection over Union, IoU). Такія веды дазволяць адфільтроўваць адказы выканаўцаў, якія сільна адрозніваюцца ад астатніх або маюць нізкае значэнне якасці на кантрольных пытаннях. Прыклад інтэрфэйса стварэння задання для Яндэкс.Талакі прыводзіцца на выяве 6.



Выява 6. Інтэрфэйс стварэння задання для Яндэкс.Талакі

Аднак агульным недахопам платформаў Amazon Mechanical Turk і Яндэкс.Талакі з’яўляецца іх платны характар. Гэты фактар сур’ёзна ўскладняе збор вялікай колькасці дадзеных у некамерцыйных праектах.

Насупраць, краўдсорсынгавая анлайн-платформа **Mozilla Common Voice** з’яўляецца адкрытай і бясплатнай. У дадатак яна спецыялізуецца менавіта на зборы датасэтаў начытанага маўлення. Праз гэтыя і шэраг іншых прычынаў (апісаныя далей) было вырашана выкарыстоўваць для збору дадзеных менавіта гэту платформу.

2.4 Mozilla Common Voice

Пералічым асноўныя адметнасці краўдсорсынгавай анлайн-платформы **Mozilla Common Voice** [5], [20] для збору корпусаў начытаных тэкстаў:

- Платформа з’яўляецца бясплатнай да выкарыстання
- Common Voice была адмыслова распрацаваная для збору адкрытых датасэтаў з агучаным маўленнем. На момант 19.01.2022 на платформе сабраныя і даступныя для спампоўкі датасэты для **87** моваў агульнай працягласцю больш за 18’000 гадзін
- Выніковыя датасэты, публікуемыя платформай, знаходзяцца ў публічнай прасторы, а менавіта маюць ліцэнзію **CC-0**. Гэта значыць, дадзеныя можна вольна і бясплатна выкарыстоўваць у любых мэтах (нават у камерцыйных) без абавязковай атрыбуцыі
- Платформа скіраваная на збор такіх дадзеных, з якімі сістэмам распазнавання маўлення давядзецца сутыкнуцца ў жыцці. Кожны з удзельнікаў начытвае тэксты ў звычайных і не ў студыйных умовах
- Прадуманая і наладжаная працэдура збору ды валідацыі дадзеных
- Датасэты публікуюцца з пэўнай рэгулярнасцю (1-2 разы на год)
- Сабраныя дадзеныя выкарыстоўваюцца як даследнікамі, так і тэхналагічнымі кампаніямі (у тым ліку буйнымі карпарацыямі) з усяго свету
- Дадзеныя публікуюцца з зафіксаванай разбіўкай на выбаркі: навучальную, валідацыйную, тэставую (train, dev, test). Гэта дазваляе выкарыстоўваць іх у якасці бенчмарка для параўнання розных мадэляў між сабою
- Зручны інтэрфэйс, які зніжае бар’еры да выкарыстання платформы людзям з розным узроўнем валодання камп’ютарамі
- Лёгкасць доступу. Для ўдзелу чалавеку патрабуецца толькі вэб-браўзер або мабільны дадатак, мікрафон і трошку вольнага часу
- Падчас збору дадзеных прысутнічае элемент спаборніцтва. Як паміж рознымі мовамі (можна паглядзець на колькасць сабраных для кожнай мовы дадзеных), так і паміж асобнымі ўдзельнікамі праэкта
- Зыходны код платформы знаходзіцца ў адкрытым доступе
- Платформа мае актыўную суполку мадэратараў і ўдзельнікаў, якія могуць пракансультаваць і дапамагчы з рэалізацыяй розных ідэяў

Збор дадзеных на платформе складаецца з 3 асноўных этапаў: (1) папярэдні збор тэкстаў і іх запампоўка на платформу; (2) агучванне сабраных тэкстаў на

платформе любымі ахвочымі людзьмі; (3) праверка сабраных дадзеных любымі ахвочымі людзьмі з выкарыстаннем перакрыцця.

Платформа дазваляе дасягнуць вялікай варыятыўнасці ў сабраных дадзеных. Няма абмежавання на ўдзельнікаў праэкта: дыктарам можа стаць абсалютна любы чалавек, нават той, для каго мова з'яўляецца не роднай а вывучанай. Запіс агучвання вядзецца ў разнастайных умовах: з карыстаннем розных мікрафонаў і з рознымі акустычнымі ўмовамі (фонавы шум, рэха, інш.)

2.4.1 Умовы выкарыстання платформы

Датасэты, сабраныя на гэтай платформе, публікуюцца пад самай вольнай з усіх магчымых ліцэнзіяй – **CC-0**. Гэта значыць, атрыманыя дадзеныя можна карыстаць як у камерцыйных, так і некамерцыйных праэктах, а таксама распаўсюджваць дадзеныя без атрыбуцыі першакрыніцы. Такі падыход да ліцэнзавання істотна спрашчае карыстанне датасэтамі і спрыяе развіццю моўных тэхналогіяў. Але ў той жа час ліцэнзія CC-0 накладае абмежаванні на тэксты, якія загружаюцца на платформу з мэтай агучвання – тэксты мусяць таксама мець ліцэнзію CC-0 (знаходзіцца ў публічным дамене).

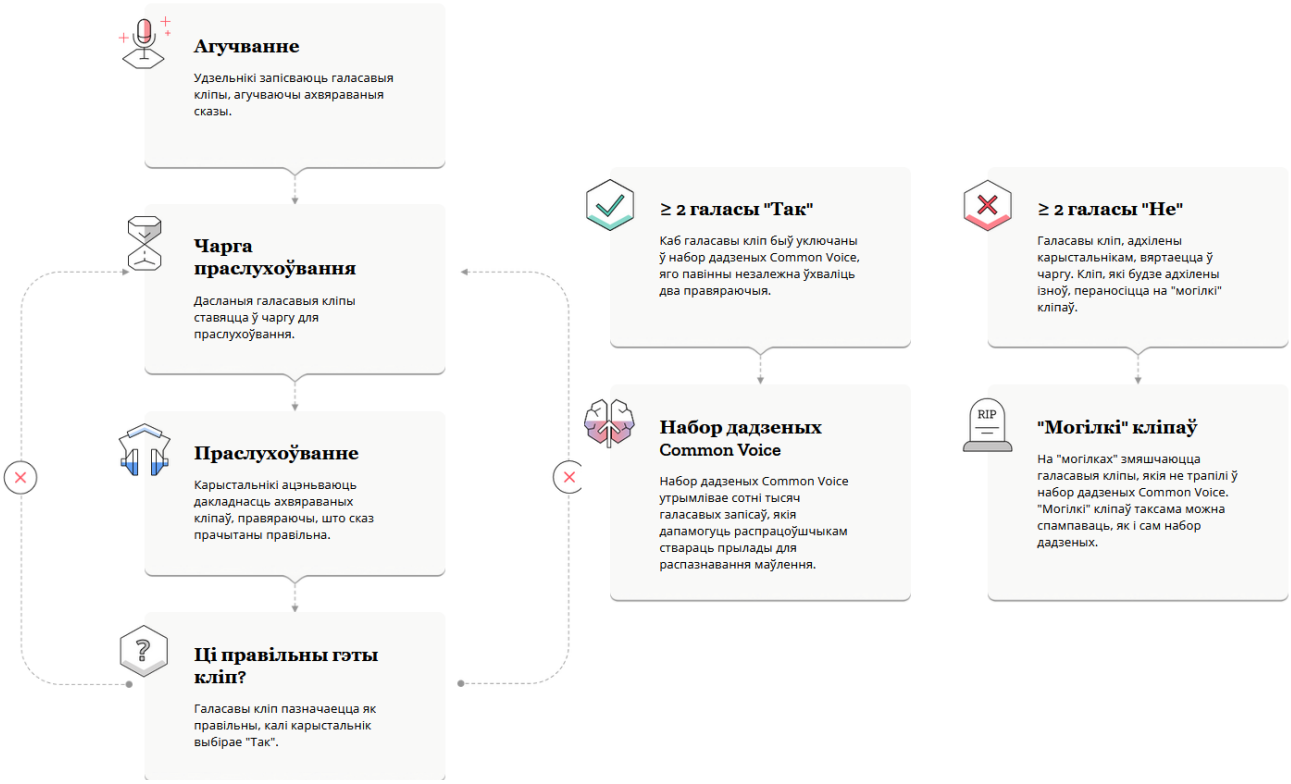
2.4.2 Механізм работы платформы

Першапачаткова беларускай мова на Common Voice была ў неактыўным стане – гэта значыць, збіраць агучванні было немагчыма. Таму было вырашана зрабіць усё неабходнае для актывацыі беларускай мовы і запуску агучвання. Ніжэй прыводзіцца паслядоўнасць крокаў, патрэбных для дадання і актывацыі новай мовы на платформе Common Voice:

1. Запыт мовы. Нехта з удзельнікаў запытвае адміністратараў платформы дадаць новую мову (было зроблена да нас)
2. Лакалізацыя сайта на беларускую мовы (было зроблена да нас)
3. Збор пачатковага набору сказаў (≥ 5'000) для агучвання. Пасля гэтага кроку адбываецца актывацыя мовы на платформе
4. Любыя ахвочыя людзі заходзяць на платформу праз інтэрнэт і пачынаюць агучваць сказы (транскрыпцыі)

5. Людзі правяраюць агучванні адно аднаго. Такім чынам фільтруюцца памылковыя запісы і сказы. Праверaныя запісы дадаюцца да выніковага датасэту разам з транскрыпцыямі.
6. Платформа публікуе дадзеныя з пэўнай перыядычнасцю (раз на паўгады)

Праверка дадзеных з перакрыццём рэалізаваная наступным чынам. Агучванні, якія набралі 2 і больш ухвальныя ацэнкі ад іншых людзей, трапляюць у выніковы датасэт. Іншыя агучванні таксама даступныя да спампоўкі і карыстання, аднак яны не трапляюць у афіцыйную разбіўку на train-, dev-, test-выбаркі. Схема агучвання, праслухоўвання і праверкі аўдыякліпаў прыводзіцца на выяве 7.



Выява 7. Схема апрацоўкі дадзеных, запісаных на Common Voice.

2.5 Збор і фільтрацыя тэкставых дадзеных

Згодна з умовамі выкарыстання, тэксты, запампаваныя на платформу, мусяць таксама знаходзіцца ў публічнай прасторы (мець ліцэнзію CC-0). Гэта звужае кола магчымых крыніцаў. Так, тэкст альбо ад моманту публікацыі мусіць знаходзіцца ў

публічнай прасторы, альбо яго праваўладальнік мусіць яўным чынам перадаць свае тэксты ў публічную прастору.

Паміж Вікіпедыяй ды ўладальнікамі платформы Mozilla Common Voice была падпісаная дамова пра магчымасць выкарыстання на платформе тэкстаў з Вікіпедыі (большасць тэкстаў Вікіпедыі маюць ліцэнзію CC BY-SA – адрозную ад CC-0). Аднак такая дамова накладае абмежаванні на выкарыстанне не болей за 3 сказы з кожнага артыкула.

Такім чынам пачатковы корпус тэкстаў для беларускай мовы быў створаны з тэкстаў артыкулаў беларускай Вікіпедыі. Пазней тэкставы корпус быў дапоўнены беларускімі мастацкімі творамі, напісанымі не пазней за 70 гадоў таму, што аўтаматычна пераводзіць іх у публічную прастору згодна з беларускім заканадаўствам.

Выгружаныя тэксты перад запампоўваннем на Common Voice апрацоўваліся і фільтраваліся. Пакідаліся толькі сказы працягласцю не больш за 14 словаў (каб не ўскладняць дыктарам задачу начыткі), якія складаюцца з літараў беларускага алфавіта (у тым ліку апострафа ‘), прагалаў, асноўных сімвалаў пунктуацыі (,.!?) і розных варыянтаў злучкоў (-, –, інш.). Лічбы і літары з іншых алфавітаў не дазваляліся.

Выдаляліся сказы, якія змяшчалі такія словы, што: (1) яны адсутнічаюць у граматычнай базе беларускай мовы і таксама (2) яны сустракаюцца ў беларускай Вікіпедыі ≤ 60 разоў (каб улічыць адсутнасць у граматычнай базе некаторых распаўсюджаных у мове словаў і іх формаў).

Дадаткова выдаляліся ўсе сказы з уласнымі імёнамі, каб аблегчыць навучанне мадэлі (Вікіпедыя змяшчае вялікую колькасць разнастайных уласных імёнаў, якія паходзяць з самых розных моваў).

Сабраны тэкставы корпус налічвае 347’010 унікальных сказаў. Пасля загрузкі на платформу Common Voice першых 5’000 сказаў, беларуская мова набыла статус актыўнай на платформе, што зрабіла даступным агучванне загружаных сказаў (згодна з правіламі платформы).

2.5.1 Прыклады сабраных сказаў

Прыклады сабраных 347’010 сказаў прыводзяцца ў табліцы 3.

Табліца 3. Прыклады сказаў з сабранага тэкставага корпуса

№	Сказ
1	У цэнтры другога паверха быў зроблены балкон, аздоблены каванай агароджай.
2	Да гэтага ж часу адносіцца пачатак яго выкладчыцкай дзейнасці.
3	Як змяніўся горад за гэты час?
4	Але гэта і ёсць нараджэнне дэмакратыі.
5	Там адна з каманд пакажа спектакль на актуальную тэму.
6	Для атрымання камерцыйнай выгады трэба, каб быў вялікі паток заказаў.
7	У магістратуры і аспірантуры займаўся археалогіяй.
8	Але з гадамі светапогляд мяняецца.
9	Магу запэўніць, што з часу яго львоўскай біяграфіі юбіляр вонкава амаль не змяніўся.
10	Улетку ўсе будуць займацца гэтым конкурсам.
11	А калі ў цябе ёсць ідэя, то неабходна рашыцца ўзяць і зрабіць.
12	Можна павольней ехаць, можна хутчэй.
13	Ля карэтнай майстэрні выступалі індзі-гурты.
14	Пакуль пад'язджалі да гарадка, з шчырасці расказаў пра сябе.
15	Таксама пісаў пейзажы і нацюрморты.

2.6 Агучванне

Платформа Common Voice дазваляе любому ахвочаму агучваць загружаныя на яе сказы. Для гэтага патрабуецца толькі доступ да інтэрнэту, браўзер і мікрафон. На платформе можна зарэгістравацца, але гэта не з’яўляецца абавязковым для ўдзелу. Рэгістрацыя дазваляе адсочваць свой прагрэс колькасці агучаных і праверaных аўдыязапісаў, задаваць асабістыя мэты (напрыклад агучваць 15 сказаў штодня), а таксама робіцца магчымым пазначыць свой пол і прыблізны ўзрост (у тым ліку, каб можна было збалансаваць выбаркі ў выніковым датасэце). З прычыны таго што агучваць сказы можа любы чалавек, якасць запісаных аўдыязапісаў правяраецца іншымі людзьмі. Каб аўдыязапіс патрапіў у выніковы датасэт, ён мусіць назбіраць не менш за 2 ухвальныя адзнакі.

Агучванне дадзеных вялося на працягу паловы года. Пасля актывацыі на беларускай мовы на платформе Common Voice, у сацыяльных сетках было абвешчана пра з’яўленне магчымасці агучваць тэксты на беларускай мове. Да ўдзелу ў праэкце атрымалася прыцягнуць вялікую колькасць людзей: агульная колькасць дыктараў склала 6’160 чалавек. Найбольшая актыўнасць агучвання і праверкі аўдыязапісаў была дасягнутая падчас правядзення двух двухтыднёвых марафонаў агучвання з розніцай паміж імі прыблізна ў паўгады.

2.7 Аналіз сабраных дадзеных

2.7.1 Колькасць дадзеных і іх фармат

На момант красавіка 2022 года беларуская мова патрапіла ў два рэлізы датасэту Common Voice: рэліз версіі 7.0 ад 21.07.2021 – 356 агучаных гадзін (275 з іх праверана); і рэліз версіі 8.0 ад 19.01.2022 – 987 агучаных гадзін (903 з іх праверана).

Статыстыка беларускага датасэту Common Voice 8.0 (ад 19.01.2022):

- Агульная колькасць сабраных гадзін – **987**
- Колькасць праверaных гадзін – **903**
- Агульная колькасць дыктараў – **6’160** чалавек
- Колькасць унікальных сказаў у датасэце – **347’010**
- Колькасць праверaных аўдыязапісаў – **677’936**
- Фармат захавання аўдыязапісаў – **mp3**
- Частасць дыскрэтызацыі – **32кГц**

Датасэт Common Voice 8.0 (ад 19.01.2022) быў прааналізаваны на колькасць даных, якія ўвайшлі ў кожную з зафіксаваных выбарак, а таксама праверaны на разнастайнасць (варыятыўнасць) сабраных дадзеных. Рэзультаты прыводзяцца ў секцыях ніжэй.

2.7.2 Разбіўка на выбаркі

Перад публікацыяй датасэту на платформе Common Voice, усе сабраныя дадзеныя для кожнай мовы падзяляюцца на выбаркі. Сярод іх 3 асноўныя: навучальная (train), валідацыйная (dev), тэставая (test). І 4 дадатковыя: праверaныя (validated), памылковыя (invalidated), астатнія (other), адрынутыя (reported).

Падбыварка праверaных кліпаў змяшчае набор усіх праверaных агучаных сказаў разам з іх транскрыпцыямі. Менавіта на яе аснове далей ствараюцца навучальная, валідацыйная і тэставая выбаркі. У кожную з іх дадзеныя адбіраюцца так, каб знутры адной падбываркі сказы не паўтараліся (адпаведна дазваляецца не больш за 1 агучка для кожнага са сказаў у выбарцы). Таксама згаданыя выбаркі не павінны мець агульныя сказы і агульных дыктараў, каб перадухіліць магчымыя выцёкі дадзеных (data leakage). Падчас разбіцця датасэту памер навучальнай выбаркі максімізуецца, а памер валідацыйнай і тэставай выбарак абіраюцца такімі, каб яны былі статыстычна значнымі (99% давяральны інтэрвал з хібай ≤1%) ў параўнанні з памерам навучальнай выбаркі [21].

Памылковая выбарка змяшчае аўдыязапісы, якія не былі прынятыя падчас праверкі з перакрыццём. Выбарка “астатнія” змяшчае аўдыязапісы, якія не назбіралі дастатковую колькасць ацэнак, каб перайсці ў статус праверaных або памылковых. Выбарка “адрынутыя” складаецца са сказаў, якія былі адрынутыя падчас агучвання. Такія сказы напрыклад змяшчаюць арфаграфічныя або грубыя граматычныя памылкі.

На выяве 8 прыводзіцца колькасць агульных сказаў для кожнай магчымай пары выбарак. Бачна, што выбаркі train, dev, test не маюць агульных сказаў, каб перадухіліць магчымыя выцёкі дадзеных.

	df_train	df_dev	df_test	df_validated	df_invalidated	df_other	df_reported
df_train	314305	0	0	314305	19163	26535	1756
df_dev	0	15803	0	15803	677	1292	206
df_test	0	0	15801	15801	701	1489	198
df_validated	314305	15803	15801	345909	20541	29316	2160
df_invalidated	19163	677	701	20541	21564	1524	408
df_other	26535	1292	1489	29316	1524	29687	266
df_reported	1756	206	198	2160	408	266	2199

Выява 8. Перасячэнне выбарак па сказах

З прычыны таго, што кожная з train, dev, test выбарак не змяшчае дублікатаў па сказах, значэнні на дыяганалі для адпаведных выбарак таксама азначаюць колькасць аўдыязапісаў, якія ўвайшлі ў адпаведную выбарку.

Для выбаркі validated (праверанай), агульная колькасць аўдыязапісаў складае 677’936, колькасць унікальных сказаў – 345’909. Бачна, што амаль палова з усіх запісаных і праверaных аўдыязапісаў не патрапіла ані ў навучальную, ані ў валідацыйную, ані ў тэставую выбаркі. Гэта азначае, што **пры выкарыстанні стандартных (зафіксаваных пры стварэнні датасэта) train, dev, test выбарак не выкарыстоўваецца амаль палова з запісаных аўдыякліпаў**.

Калі дазволіць уваходжанне ў навучальную, валідацыйную, тэставую выбаркі розных агучванняў адных і тых сказаў, то колькасць аўдыязапісаў у кожнай выбарцы павялічваецца да 613’160, 32’391, 32’385 адпаведна.

Зафіксаваныя ў датасэце навучальная, валідацыйная, тэставая выбаркі дазваляюць выкарыстоўваць датасэт у якасці бенчмарку для параўнання між сабою рэзультатаў розных мадэляў. Але схема разбіцця дадзеных на выбаркі ў датасэце Common Voice не дазваляе выкарыстоўваць амаль палову з даступных дадзеных. Для навучання найбольш якасных мадэляў распазнавання маўлення неабходна павялічваць зафіксаваныя train, dev, test выбаркі за кошт аўдыязапісаў з выбаркі validated.

На выяве 9 прыводзіцца колькасць агульных дыктараў для кожнай магчымай пары выбарак. Нагадаем, што агульная колькасць унікальных дыктараў склала 6’160. Бачна, што train, dev, test выбаркі не маюць агульных дыктараў, каб перадухіліць магчымыя выцёкі дадзеных.

	df_train	df_dev	df_test	df_validated	df_invalidated	df_other
df_train	2222	0	0	2222	1804	556
df_dev	0	1017	0	1017	528	81
df_test	0	0	2779	2779	998	103
df_validated	2222	1017	2779	6018	3330	740
df_invalidated	1804	528	998	3330	3384	559
df_other	556	81	103	740	559	831

Выява 9. Перасячэнне выбарак па дыктарах

2.7.3 Аналіз варыятыўнасці дадзеных

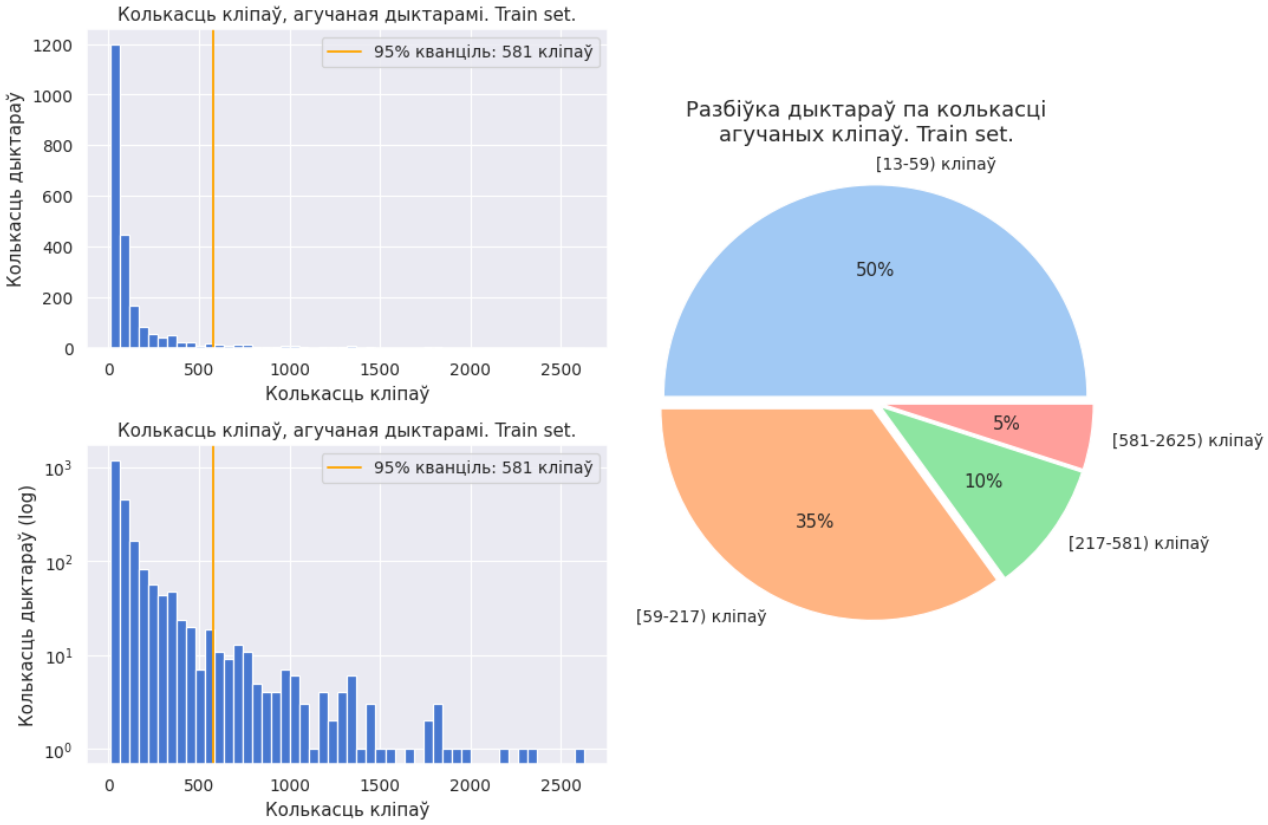
Нагадаем, што мы хацелі бы сабраць такі датасэт агучаных тэкстаў, каб забяспечыць максімальную варыятыўнасць у дадзеных: розныя дыктары (пол, узрост, узровень валодання беларускай мовай, дыкцыя, інш.), розныя акустычныя ўмовы запісу (фонавы шум, якасць мікрафону, інш).

Праслухоўванне асобных аўдыязапісаў пацвярджае той факт, што сабраны датасэт атрымаўся разнастайны. Прысутнічаюць галасы людзей рознага полу, узросту, з разнастайнымі тэмбрамі, тэмпам маўлення. Акустычныя ўмовы таксама адрозніваюцца: некаторыя запісы ствараліся ў добрых акустычных умовах, з мінімальнай прысутнасцю фонавага шума, вымаўленне словаў яснае, разборлівае, гучнае; а на некаторых запісах наадварот прысутнічае большая колькасць фонавага шума, словы вымаўляюцца цішэй, у запісе могуць прысутнічаць пабочныя гукі – тым не менш чалавек можа распазнаць маўленне на такіх запісах, а значыць і камп’ютары мусяць навучыцца. Аўдыязапісы з яшчэ горшымі ўмовамі запісу (больш істотны шум, памылкі, запінкі падчас вымаўлення, залішне ціхае вымаўленне, інш.) пазначаліся людзьмі як памылковыя і ў train, dev, test падвыбаркі не патрапілі.

Платформа Common Voice дазваляе зарэгістравацца і стварыць асабісты профіль, і пазначыць у ім інфармацыю пра дыктара: пол і прыблізны узрост. Рэгістрацыя і стварэнне асабістага профіля не з’яўляюцца абавязковымі, таму такія дадзеныя даступныя не для ўсіх дыктараў. На старонцы апублікаванага датасэту прыводзіцца наступныя **статыстыка пола дыктараў**: жанчыны – 14% дыктараў, мужчыны – 10% дыктараў, астатнія дыктары пол не пазначылі. На той жа старонцы прыводзіцца наступная **статыстыка прыблізнага ўзроста дыктараў**: узрост да 19 гадоў – 1% дыктараў, узрост ад 19 да 29 гадоў – 6% дыктараў, узрост ад 30 да 39 гадоў – 8% дыктараў, узрост ад 40 да 49 гадоў – 7% дыктараў, астатнія дыктары не пазначылі свой узрост. Калі адкінуць з разгляду дыктараў з невядомымі полам і ўзростам, то можна зрабіць наступныя высновы: (1) колькасць жанчын сярод дыктараў прыблізна ў 1.5 разы болей за мужчын; (2) узрост дыктараў размеркаваны на прамежку ад 19 да 49 гадоў прыблізна раўнамерна, а самай частай стала ўзроставая група ад 30 да 39 гадоў.

Пры аналізе колькасці запісаных кожным з дыктараў аўдыязапісаў (сярод тых, што патрапілі ў навучальную выбарку) выявілася, што размеркаванне з’яўляецца сільна скошаным і нагадвае размеркаванне Пуасона (выява 10). Большасць

дыктараў агучыла невялікую колькасць сказаў, а самыя актыўныя ўдзельнікі агучылі значна большую колькасць сказаў. Так, палова дыктараў запісала не больш за 59 кліпаў, 85% – не больш за 217 кліпаў, 95% – не больш за 581 кліп, і толькі 5% самых актыўных дыктараў запісалі ад 581 да 2625 кліпаў. Гістаграма размеркавання разам з кругавой дыяграмай прыведзеныя на выяве 10.



Выява 10. Размеркаванне колькасці агучаных дыктарамі сказаў для навучальнай выбаркі. Злева – гістаграмы размеркавання (у тым ліку з лагарыфмічнай восью), справа – кругавая дыяграма

Каб ацаніць варыятыўнасць аўдыясігналаў былі прааналізаваныя аўдыязапісы з валідацыйнай выбаркі (з прычыны яе меншага памеру ў параўнанні з навучальнай выбаркай). Рэзультаты аналізу можна абагуліць і на ўсю навучальную выбарку, бо памеры валідацыйнай і тэставай выбаркі з'яўляюцца статыстычна значнымі адносна памераў навучальнай выбаркі (99% давяральны інтэрвал з хібай $\leq$1%) [21].

Кожны аўдыязапіс з валідацыйнай выбаркі апісваўся наступнымі прыкметамі:

- 95% працэнціль модуля значэнняў аўдыязапіса. Дазваляе ацаніць агульны ўзровень гучнасці аўдыязапіса і не ўлічваць пікі (у адрозненне ад проста максімальнага па модулі значэння), якія могуць быць абумоўленыя шумам, выбухнымі зычнымі (п, к, т) і г.д.
- Тэмпам маўлення, які падлічваўся як *колькасць выбарак (фрэймаў) аўдыясігнала, падзеленая на колькасць літараў у сказе*.

Для кожнага дыктара далей знаходзілася сярэдняе значэнне пабудаваных статыстык па агучаных ім сказах. Дыяграма рассейвання дыктараў з валідацыйнай выбаркі па выніковых статыстыках прыводзіцца на выяве 11. Па выяве бачна, што дыктарамі пакрываецца шырокі спектр значэнняў адпаведных статыстык, што сведчыць пра высокую варыятыўнасць сабраных дадзеных.

Выява 11. Візуалізацыя варыятыўнасці дыктараў.
Па гарызанталі – сярэдняе 95% працэнціля модуля значэнняў аўдыязапісаў,
па вертыкалі – сярэдняе значэнне тэмпу маўлення.

2.8 Высновы

Для навучання нейрасеткавай мадэлі распазнавання агульнага беларускага маўлення было неабходна сабраць аб’ёмны і варыятыўны датасэт. У якасці пачатковай мэты было вырашана сабраць мінімум 100 гадзін анатаванага маўлення. Пастаўленыя задачы былі выкананыя: атрымалася сабраць варыятыўны датасэт (з разнастайнымі дыктарамі ды разнастайнымі ўмовамі запісу) агульнай працягласцю ў 987 гадзін (903 з іх праверaны з дапамогай перакрыцця, калі некалькі людзей ацэньваюць якасць агучанага іншым чалавека сказа). Колькасць праверaных аўдыязапісаў склала 677’936, колькасць сказаў, якія ўвайшлі ў датасэт склала 347’010. Да ўдзелу ў агучванні і праверцы агучаных сказаў атрымалася прыцягнуць вялікую колькасць беларусаў. Колькасць дыктараў склала 6’160 чалавек. Статыстыка прыводзіцца для датасэта Common Voice 8.0 ад 19.01.2022.

Сабраныя дадзеныя былі прааналізаваныя на варыятыўнасць. Узрост дыктараў прыблізна раўнамерна размеркаваны на прамежку ад 19 да 49 гадоў, жанчын-дыктараў прыблізна ў паўтары разы больш за мужчын (статыстыка на аснове даступных дадзеных пра дыктараў). Акрамя гэтага, для кожнага дыктара з валідацыйнай выбаркі былі пабудаваныя статыстыкі, якія сведчаць пра наяўнасць варыятыўнасці ў сабраных дадзеных. Атрыманыя рэзультаты пераносяцца і на навучальную выбарку з прычыны статыстычна значнага памера валідацыйнай выбаркі. Праслухоўванне асобных аўдыязапісаў таксама сведчыць пра вялікую разнастайнасць дадзеных.

Вялікі аб’ём сабраных дадзеных дазваляе будаваць сістэмы распазнавання агульнага маўлення з выкарыстаннем метадаў глыбокага навучання, а высокая варыятыўнасць сабраных дадзеных дазваляе мадэляваць умовы, з якімі сістэмы распазнавання маўлення сутыкнуцца пры іх выкарыстанні ў звычайным жыцці.

Датасэт апублікаваны пад свабоднай ліцэнзіяй CC-0, што робіць яго даступным для выкарыстання для розных задач (у тым ліку камерцыйных праэктаў).

РАЗДЗЕЛ 3
ПАБУДОВА СІСТЭМЫ РАСПАЗНАВАННЯ МАЎЛЕННЯ

Як было апісана ў Раздзеле 1, сучасныя нейрасеткавыя end-to-end архітэктуры распазнавання маўлення паказваюць найлепшыя рэзультаты, нават пры ўмове адсутнасці вялікіх карпусоў анатаваных дадзеных. У адной з задач дадзенай работы было вырашана навучыць менавіта такую мадэль на аснове архітэктуры wav2vec2 [6]. Памер сабранага ў Раздзеле 2 корпуса дадзеных дазваляе як давучыць пераднавучаную мадэль на анатаваным корпусе беларускага маўлення, так і паспрабаваць пераднавучыць wav2vec2 “з нуля” на частцы сабраных дадзеных, а пасля давучыць мадэль для задачы распазнавання маўлення на рэшце дадзеных. Каб дасягнуць найлепшых рэзультатаў было вырашана прытрымацца першага варыянта: выкарыстаць пераднавучаную мадэль.

У якасці пераднавучанай мадэлі была абраная мадэль **“facebook/wav2vec2-base”** [3]. Яе пераднавучанне праводзілася на датасэце LibriSpeech [3], які складаецца прыблізна з 1’000 гадзін анатаванага маўлення на ангельскай мове. Пераднавучанне мадэлі праводзілася ў рэжыме без настаўніка – гэта значыць, што анатацыі (транскрыпцыі) не выкарыстоўваліся. У якасці крыніцаў начытаных тэкстаў для гэтага датасэту былі абраныя аўдыякнігі з праэкту LibriVox. Гукавыя файлы захаваныя ў датасэце з частасцю дыскрэтызацыі 16 кГц.

3.1 Пабудова акустычнай мадэлі

3.1.1 Падрыхтоўка дадзеных

Для навучання карысталіся зафіксаваныя train, dev, test выбаркі датасэта Common Voice 8.0. Нагадаем, што згаданыя выбаркі складаюць толькі палову ад усіх праверaных запісаў (каля 350’000 кліпаў з каля 680’000 праверaных) – рэшта патрапіла ў выбарку validated з прычыны таго, што ў кожную з train, dev, test выбарак трапляла не больш за 1 агучка кожнага сказа. Паўторныя агучкі з’явіліся з прычыны вялікай актыўнасці ўдзельнікаў марафонаў агучвання і недастатковай вялікай колькасці запампаваных для агучвання сказаў.

Такім чынам навучальную і праверачныя выбаркі можна было бы павялічыць прыблізна ў 2 разы, а значыць і палепшыць выніковыя мадэлі. Аднак з прычыны вялікага аб’ёму дадзеных, які і так прадстаўлены ў згаданых зафіксаваных

выбарках, і вялікага часу, які сыходзіў на навучанне мадэлі, ад гэтай ідэі было вырашана адмовіцца. Да таго ж, зафіксаваныя Mozilla выбаркі выступаюць у якасці бенчмарку – даследнікам не трэба праводзіць пераразбіццё дадзеных і ўпэўнівацца ў адсутнасці ў ім памылак, каб параўнаць свае метрыкі з іншымі даследнікамі.

Для аўдыязапісаў перадапрацоўка зводзілася да прывядзення іх да агульнага фармата: частасць дыскрэтызацыі 16кГц, 1 канал.

Тэкставыя транскрыпцыі прыводзіліся да ніжняга рэгістру; прыбіраліся ўсе сімвалы акрамя літараў алфавіта і лічбаў; кожная паслядоўнасць сімвалаў-прагалаў (прагал, знак табуляцыі, інш.) замянялася на 1 прагал.

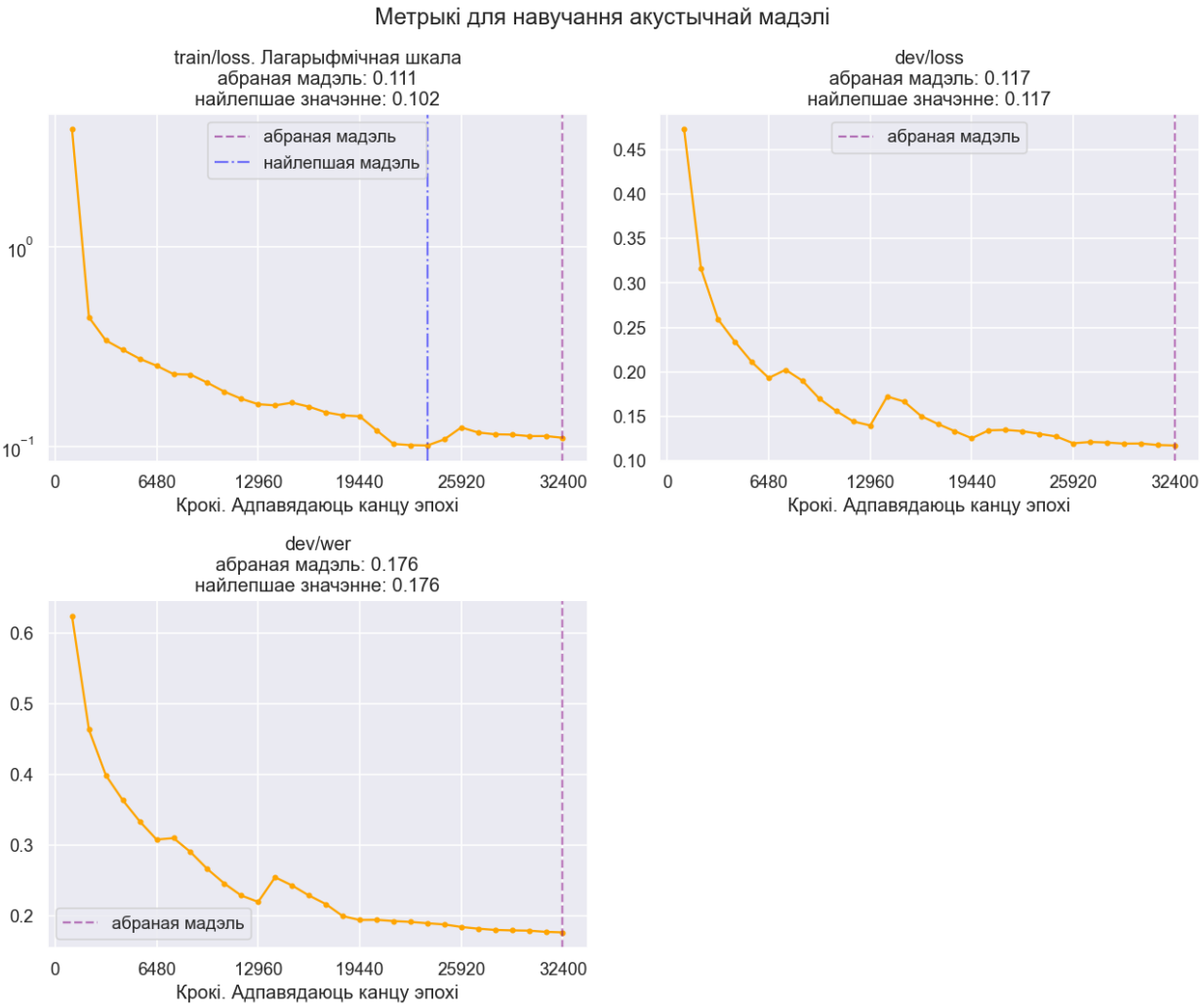
3.1.2 Навучанне

У якасці функцыі стратаў для давучвання мадэлі на задачы распазнавання маўлення выкарыстоўвалася CTC [6]. Аптымізацыя параметраў адбывалася з дапамогай алгарытма AdamW [18] – выпраўленай версіі папулярнага алгарытма аптымізацыі Adam. Для аптымізацыі спажывання памяці таксама выкарыстоўваўся метад Gradient checkpointing [8]. Выбар найлепшай мадэлі праводзіўся з дапамогай метрыкі WER на валідацыйнай выбарцы.

Для праграмнай рэалізацыі навучання акустычнай мадэлі быў абраны папулярны фрэймворк для навучання NLP і ASR мадэляў **HuggingFace** [15], які з’яўляецца абгорткай над іншым фрэймворкам — **PyTorch**. Навучанне праводзілася на серверы з 3 відэакартамі NVIDIA GeForce RTX 2080 Ti. Памер батча для навучання і ацэнкі якасці мадэлі быў роўны 16 (з улікам трох відэакартаў Сярэдні час на эпоху склаў блізу 8 гадзін. У сувязі з гэтым, і з прымальнымі значэннямі метрык, навучанне было спынена пасля 5 эпохаў. Ацэнка якасці (evaluation) і захаванне прамежкавых параметраў (checkpointing) праводзілася некалькі разоў на працягу кожнай эпохі.

Графікі функцый стратаў і метрыкі WER прыводзяцца на выяве 12. Крывыя функцыі стратаў на навучальнай і валідацыйнай выбарках сведчаць, што навучанне ішло добра, перанавучання не адбылося. Бачна, што пасля 5 эпохаў мадэль пачынае збягацца: можна працягваць навучанне мадэлі, але далейшае паляпшэнне метрык будзе нязначным.

Найлепшае значэнне WER на валідацыйнай выбарцы было дасягнута на апошнім чэкпоінце (0.176). Ён і быў абраны ў якасці найлепшай акустычнай мадэлі.



Выява 12. Метрыкі падчас навучання акустычнай мадэлі.

Тэг dev/wer адпавядае метрыцы WER на валідацыйнай выбарцы

Якасць найлепшай акустычнай мадэлі, абранай з дапамогай метрыкі WER на валідацыйнай выбарцы, была праверaная на адкладзенай тэставай выбарцы. Значэнне Test WER склала 0.187.

3.2 Пабудова моўнай мадэлі

Для паляпшэння прадказанняў акустычнай мадэлі была пабудаваная 5-грамная моўная мадэль з выкарыстаннем мадыфікаванага згладжвання Кнэсер-Нэй (modified Kneser-Ney smoothing) [8]. Для пабудовы такой мадэлі была выкарыстаная папулярная бібліятэка KenLM, створаная аўтарамі [8].

Моўная мадэль вучылася на корпусе тэкстаў з датасэта Common Voice 8.0. У корпус увайшлі унікальныя сказы з навучальнай выбаркі (train) і выбаркі validated без сказаў з dev, test выбарак, каб перадухіліць магчымыя выцёкі дадзеных (data leakage). Агульная колькасць сказаў для пабудовы моўнай мадэлі склала 314’676.

Як згадвалася ў раздзеле 1, важна, каб моўная мадэль будавалася на вялікім і разнастайным корпусе тэкстаў. 300 тысяч сказаў не з’яўляюцца шырокім тэкставым корпусам, здольным перадаць усе асаблівасці мовы і змясціць усе словаформы і іх ужыванні. Аднак збор падобнага корпуса тэкстаў з’яўляецца асобнай аб’ёмнай і руплівай задачай. Таму ў межах дадзенай работы было вырашана пабудаваць моўную мадэль толькі на даступных для навучання сказах.

3.3 Ацэнка якасці выніковай сістэмы

У табліцы 4 прыводзяцца значэнні метрыкі WER на валідацыйнай і тэставай выбарках разам з доляй цалкам распазнаных сказаў (без памылак). Метрыкі прыводзяцца асобна для акустычнай мадэлі і для акустычнай мадэлі разам з пабудаванай моўнай мадэлью.

Табліца 4. Метрыкі выніковай мадэлі

Мадэль	dev WER	test WER	Доля распазнаных сказаў, test
Акустычная	0.1761	0.187	36.688%
Акустычная + моўная	0.115	0.124	52.269%

Даданне моўнай мадэлі дазволіла знізіць WER на тэставай выбарцы з 0.187 да 0.124. Канчатковы рэзультат – WER 0.124 (або 12.4%) з’яўляецца даволі добрым для мадэляў распазнавання. Так напрыклад цяперашняе найлепшае значэнне test WER для нямецкага датасэту Common Voice роўнае 5.7%.

У табліцы 5 прыводзяцца прыклады сапраўдных сказаў і пабудаваных з дапамогай найлепшай акустычнай мадэлі разам з моўнай мадэлью транскрыпцыяў. З прыкладаў у 4 радку можна зрабіць выснову, што пры навучанні моўнай мадэлі на больш вялікім корпусе (куды напрыклад будзе ўваходзіць слова “ваеннапалонных”), памылка ў гэтым слове можа знікнуць. Распазнаны тэкст у 5

радку гучыць надзвычай падобным да сапраўднай транскрыпцыі (“па лесе ішоў сабака” і “алеся ішоў сабак”) – гэтая памылка таксама можа знікнуць пры паляпшэнні моўнай мадэлі.

Табліца 5. Прыклады распазнавання выніковай мадэлью

Транскрыпцыя	Распазнаны тэкст	Поспех	WER
цвіценне працягваецца з мая да позняй восені	цвіценне працягваецца з мая да позняй восені	1	0
таксама пройдзе канферэнцыя фестываль і мастацкая выстава	таксама пройдзе канферэнцыя фестываль і мастацкая выстава	1	0
у нас была яе персанальная выстава	у нас была персанальная выстава	0	0.167
далейшае супрацоўніцтва паміж нямецкім і савецкім прыняло форму абмену польскіх ваеннапалонных	далейшы супрацоўніцтва між нямецкім савецкім пыняафораму абмену польскіх дваеннапогам	0	0.54
па лесе ішоў сабака	алеся ішоў у сабак	0	1

Увогуле ад пабудаванай сістэмы ствараецца добрае ўражанне, яна здольная распазнаваць адвольнае беларускае маўленне на дастаткова добрым узроўні.

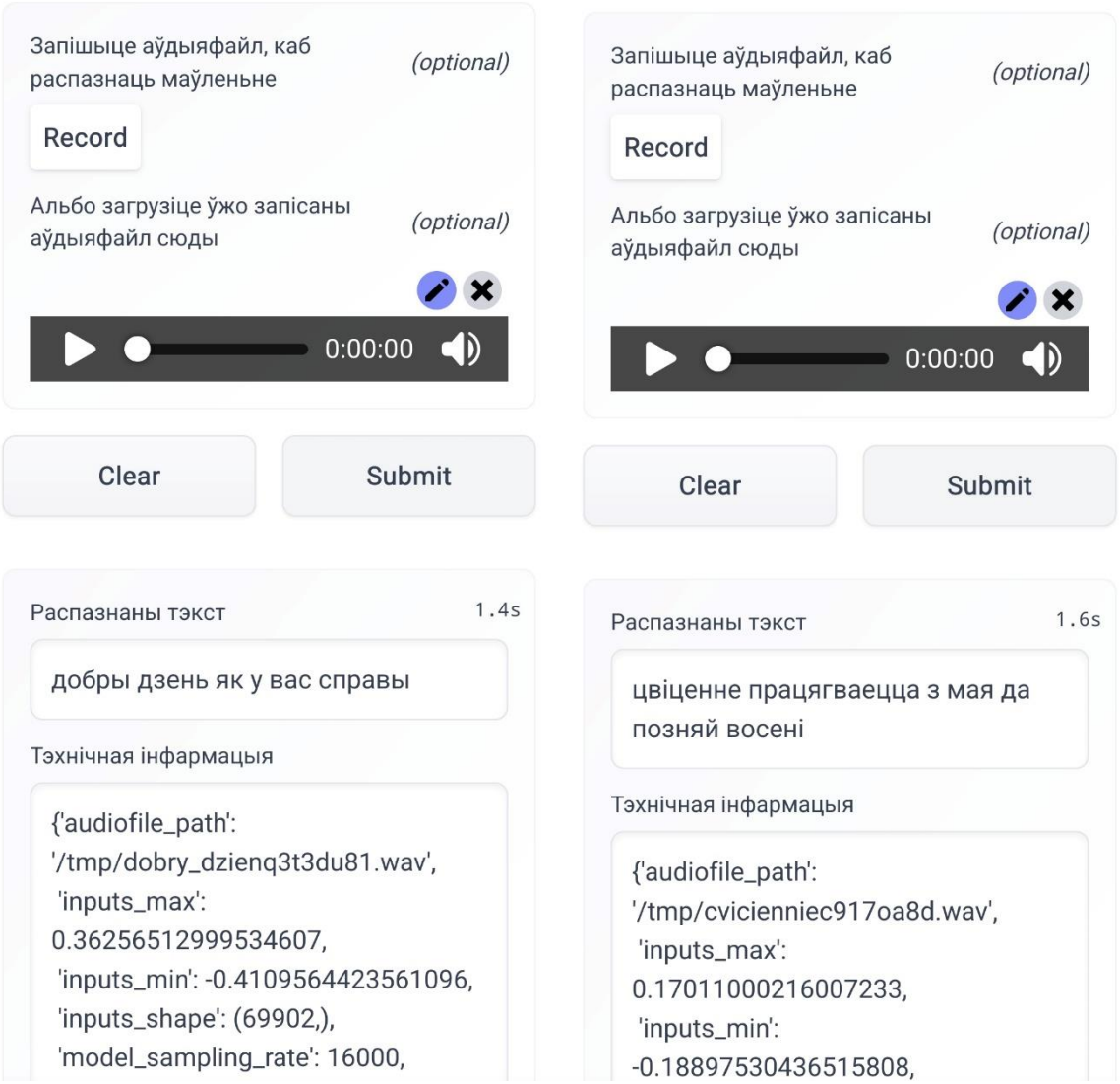
3.4 Стварэнне вэб-інтэрфэйса для дэманстрацыі работы мадэлі

Пабудаваная акустычная мадэль разам з моўнай мадэлью былі загружаная на платформу Hugging Face Hub пад вольным доступам, што дазваляе любому ахвочаму выкарыстоўваць мадэлі далей [16].

У дадатак на платформе Hugging Face Spaces быў рэалізаваны вэб-інтэрфэйс [15] для дэманстрацыі работы навучанай сістэмы распазнавання маўлення, якая складаецца з акустычнай і моўнай мадэляў. Інтэрфэйс дазваляе запісаць аўдыяфайл наўпрост у браўзеры і падаць яго на ўваход навучанай сістэме распазнавання. Таксама падтрымліваецца магчымасць загрузіць існуючы аўдыяфайл.

На выяве 13 прыводзіцца прыклад работы навучанай сістэмы з дапамогай пабудаванага вэб-інтэрфэйсу. На ўваход мадэлі падаваліся тыя самыя аўдыязапісы, што і пры тэставанні папярэдняй сістэмы распазнавання агульнага маўлення [2], [5],

якое праводзілася ў Раздзеле 1 (выява 5). Тады, для першага аўдыяфайла была пабудаваная часткова правільная транскрыпцыя, а для другога аўдыяфайла транскрыпцыя была зусім няправільная. Цяпер жа транскрыпцыі, пабудаваныя навучанай мадэлью, цалкам супадаюць з вымаўленым тэкстам. Гэта яшчэ раз падкрэслівае якасць мадэлі і здольнасць распазнаваць агульнае (не абмежаванае слоўнікам) маўленне.



Выява 13. Распазнаванне адвольнага маўлення навучанай сістэмай [15] праз пабудаваны вэб-інтэрфэйс.

На прыкладзе злева была начытана фраза “Добры дзень. Як у вас справы?”, справа - “Цвіценне працягваецца з мая да позняй восені”.

3.5 Высновы

У дадзеным раздзеле апісваецца працэс пабудовы сістэмы распазнавання беларускага маўлення. Сістэма складаецца з акустычнай і моўнай мадэляў. Акустычная мадэль была навучаная з дапамогай давучвання мадэлі “facebook/wav2vec2-base” на сабраным раней датасэце Common Voice 8. Пасля 5 эпохаў навучання значэнні функцыі стратаў CTC і метрыкі Word Error Rate выйшлі на плато і стабілізаваліся. Найлепшая акустычная мадэль мае значэнні WER 0.176 ды 0.187 на валідацыйнай і тэставых выбарках адпаведна.

Для паляпшэння распазнавання маўлення ў дадатак да акустычнай мадэлі была пабудаваная моўная 5-грамная мадэль з мадыфікаваным згладжваннем Кнэсер-Нэй. Моўная мадэль будавалася на сказах з датасэту Common Voice. У будучыні, пасля збору больш аб’ёмнага тэкставага корпусу, можна пабудаваць больш якасную моўную мадэль. Але выкарыстанне цяперашняй моўнай мадэлі таксама дазволіла істотна зменшыць значэнні WER да 0.115 ды 0.124 на валідацыйнай і тэставых выбарках адпаведна.

Пабудаваная сістэма распазнавання (акустычная + моўная мадэлі) была загружаная пад вольным доступам на платформу Hugging Face Hub і з’яўляецца даступнай для выкарыстання любымі ахвочымі [16].

Для лёгкага азнаямлення з магчымасцямі мадэлі быў створаны дэманстрацыйны вэб-інтэрфэйс, таксама загружаны на платформу Hugging Face Spaces [15]. Ён дазваляе хутка запусціць мадэль распазнавання маўлення на запісаным наўпрост у браўзеры або загружаным з камп’ютара або тэлефона аўдыяфайле.

Зыходны код для прыводзіцца ў [4].

Нізкае значэнне метрыкі Word Error Rate а таксама рэзультаты ручных тэстаў сведчаць пра добры ўзровень распазнавання сістэмай адвольнага беларускага маўлення.

Такім чынам можна лічыць, што пастаўленая задача па навучанні мадэлі распазнавання маўлення выканана.

ВЫСНОВЫ

У выніку праведзенай работы была пабудаваная новая для беларускай мовы сістэма распазнавання адвольнага маўлення высокай якасці (Test WER = 0.124 або 12.4%), заснаваная на end-to-end архітэктуры з выкарыстаннем глыбокага навучання. У параўнанні з мінулай мадэлью, заснаванай на традыцыйных метадах распазнавання маўлення [2], [5], істотна палепшылася якасць распазнавання адвольнага маўлення, у тым ліку ў складаных умовах запісу (прысутнасць фонавага шуму, інш.). Мадэль знаходзіцца ў адкрытым доступе [16], [15], [4], што спрашчае доступ да яе іншых даследнікаў.

Для пабудовы сістэмы распазнавання маўлення быў сабраны вялікі корпус начытаных тэкстаў на беларускай мове з дапамогай платформы Mozilla Common Voice. Агульная працягласць сабраных аўдыязапісаў складае 987 гадзін (з іх 903 праверана), а ў агучванні тэкстаў прыняло ўдзел 6’160 дыктараў. Гэта першы з падобных датасэтаў настолькі вялікага памеру для беларускай мовы. Высокая варыятыўнасць сабраных дадзеных як адносна дыктараў (пол, узрост, тэмп маўлення, іншыя асаблівасці вымаўлення), так і адносна ўмоваў запісаў (розныя мікрафоны, наяўнасць фонавага шуму, інш.) дазваляе навучыць сістэмы распазнавання маўлення працаваць ва ўмовах, набліжаных ды тых, з якімі гэтым сістэмам давядзецца працаваць у штодзённым жыцці.

Такім чынам усе пастаўленыя ў рабоце задачы (навучанне мадэлі распазнавання адвольнага беларускага маўлення і збор неабходнага для гэтага датасэту) выкананы. Наяўнасць якаснай мадэлі распазнавання адвольнага маўлення адкрывае для беларускай мовы перспектывы далейшага развіцця больш складаных моўных тэхналогіяў: галасавы ўвод тэкста, галасавыя дапаможнікі, аўтаматызаванае навучанне беларускай мове і інш.

Вольная ліцэнзія CC-0, пад якой распаўсюджваецца сабраны датасэт, дазваляе свабодна і бясплатна выкарыстоўваць яго як у наступных даследаваннях, так і ў камерцыйных праектах. Збор і публікацыя датасэта на папулярнай платформе Mozilla Common Voice спрыяе лёгкаму доступу да дадзеных з боку буйных тэхналагічных кампаніяў – гэта дазволіць ім на аснове гэтага датасэта будаваць свае сістэма апрацоўкі маўлення і інтэграваць іх у існуючыя і новыя прадукты.

ВЫКАРЫСТАНЫЯ КРЫНІЦЫ

1. Гецэвіч, Ю.С. РАСПРАЦОЎКА КАМПАНЕНТА РАСПАЗНАВАННЯ МАЎЛЕННЯ ДЛЯ НАТУРАЛЬНА МАЎЛЕНЧАГА ІНТЭРФЕЙСУ / Ю.С. Гецэвіч, К.А. Нікалаенка, Л.І. Кайгародава // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2015) : материалы V междунар. науч.-техн. конф. (Минск, 19 – 21 февраля 2015 года) / пад рэд. В. В. Голенков (отв. ред.) [и др.]. Минск : БГУИР, 2015. — С. 507-512.
2. Казлоўская, Н.Д. Распрацоўка алгарытмаў дыктаранезалежнага распазнавання беларускага маўлення / Н.Д. Казлоўская, М.В. Шыбко, А.І. Пратасеня, Ю.С. Гецэвіч // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2017) : доклады XVI Международной конференции, Минск, 16 ноября 2017 г. / ОИПИ НАН Беларуси ; под науч. ред. А.В. Тузиков, Р.Б. Григянец, В.Н. Венгеров. — Минск : ОИПИ НАН Беларуси, 2017. — С. 299-304.
3. Мадэль facebook/wav2vec2-base // [Электронны рэсурс]. – 2022. Рэжым доступу : https://huggingface.co/facebook/wav2vec2-base. – Дата доступу : 01.04.2022.
4. Праграмная рэалізацыя навучання сістэмы распазнавання беларускага маўлення // [Электронны рэсурс]. – 2022. Рэжым доступу : https://github.com/yks72p/stt_be. – Дата доступу : 01.04.2022.
5. Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. "Common voice: A massively-multilingual speech corpus." *arXiv preprint arXiv:1912.06670* (2019).
6. 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.
7. Callison-Burch C., Dredze M. Creating speech and language data with amazon’s mechanical turk //Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk. – 2010. – С. 1-12.
8. Chen T. et al. Training deep nets with sublinear memory cost //arXiv preprint arXiv:1604.06174. – 2016.

9. Corpus.by. ThematicSpeechRecognizer // [Электронны рэсурс]. – 2022. Рэжым доступу : https://www.corpus.by/ThematicSpeechRecognizer/?lang=be. – Дата доступу : 01.04.2022.
10. Dhankar, Abhishek. "Study of deep learning and CMU sphinx in automatic speech recognition." In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2296-2301. IEEE, 2017.
11. Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." In *Proceedings of the 23rd international conference on Machine learning*, pp. 369-376. 2006.
12. Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).
13. Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. "Scalable modified Kneser-Ney language model estimation." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 690-696. 2013.
14. Hinton G. E. et al. Improving neural networks by preventing co-adaptation of feature detectors //arXiv preprint arXiv:1207.0580. – 2012.
15. Hugging Face. Дэманстрацыйны вэб-інтэрфэйс для мадэлі распазнавання беларускага маўлення // [Электронны рэсурс]. – 2022. Рэжым доступу : https://huggingface.co/spaces/ales/wav2vec2-cv-be-lm. – Дата доступу : 01.04.2022.
16. Hugging Face. Мадэль распазнавання беларускага маўлення // [Электронны рэсурс]. – 2022. Рэжым доступу : https://huggingface.co/ales/wav2vec2-cv-be. – Дата доступу : 01.04.2022.
17. Lamere P. et al. The CMU SPHINX-4 speech recognition system //Ieee intl. conf. on acoustics, speech and signal processing (icassp 2003), hong kong. – 2003. – Т. 1. – С. 2-5.
18. Loshchilov I., Hutter F. Decoupled weight decay regularization //arXiv preprint arXiv:1711.05101. – 2017.
19. Morris A. C., Maier V., Green P. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition //Eighth International Conference on Spoken Language Processing. – 2004.

20. Mozilla Common Voice Datasets // [Электронны рэсурс]. – 2022. Рэжым доступу : https://commonvoice.mozilla.org/en/datasets. – Дата доступу : 01.04.2022.
21. Mozilla Common Voice Corpora Creator // [Электронны рэсурс]. – 2022. Рэжым доступу : https://github.com/common-voice/CorporaCreator. – Дата доступу : 01.04.2022.
22. Nikalaenka, K. Training algorithm for speaker-independent voice recognition systems using HTK / K. Nikalaenka, Y. Hetsevich // PRIP ‘2016: Pattern recognition and information processing : proceedings of the 13th International Conference, (3-5 October 2016, Minsk, Belarus). – Minsk : Publishing Center of BSU, 2016. – P. 126-129
23. Panayotov V. et al. Librispeech: an asr corpus based on public domain audio books //2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2015. – С. 5206-5210.
24. Povey, Daniel, et al. "The Kaldi speech recognition toolkit." *IEEE 2011 workshop on automatic speech recognition and understanding*. No. CONF. IEEE Signal Processing Society, 2011.
25. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.
26. Wang S., Li G. Overview of end-to-end speech recognition //Journal of Physics: Conference Series. – IOP Publishing, 2019. – Т. 1187. – №. 5. – С. 052068.
27. Wolf T. et al. Huggingface's transformers: State-of-the-art natural language processing //arXiv preprint arXiv:1910.03771. – 2019.