УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Богдан Томић

УНАПРЕЂИВАЊЕ МАЛИХ ASR МОДЕЛА ЗА СРПСКИ ЈЕЗИК ПРИМЕНОМ ДЕСТИЛАЦИЈЕ ЗНАЊА

мастер рад

Београд, 2026.

Ментор:

др Младен Николић, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Јована Ковачевић, ванредни професор
Универзитет у Београду, Математички факултет

др Јелена Граовац, ванредни професор
Универзитет у Београду, Математички факултет

Датум одбране: _____

Наслов мастер рада: Унапређивање малих ASR модела за српски језик применом дестилације знања

Резиме: У раду се испитује колико се најмања варијанта архитектуре Whisper може приближити знатно већим системима за препознавање говора на српском језику и које технике обучавања притом дају највећи допринос. Прикупљен је и обрађен корпус од око 200 сати јавно доступног аудио материјала. Обрада је обухватила сегментацију дугих записа, груписање говорника по боји гласа, стратификовану поделу и присилно поравнање, чиме је добијен коначни скуп од 41.082 узорка. Над истим корпусом упоређена су три приступа обучавању, стандардно фино подешавање, дестилација знања уз фино подешен модел `whisper-large-v3` у улози учитеља, и итеративно поновно учење на сопственим грешкама кроз пет рунди.

Мерено подударношћу са референтним транскриптом, фино подешавање и дестилација дају блиске вредности, 76,14% и 74,64%. Мерено стопом грешке на нивоу речи разлика је изразита, 56,80% наспрам 39,80% у просеку. Дестиловани модел не препознаје реченицу просечног квалитета боље од фино подешеног, али готово упола ређе даје потпуно промашен излаз, онај у коме упада у петљу или измишља текст. Дестилација, дакле, не подиже просечан квалитет транскрипције него уклања најгоре случајеве, што је за практичну употребу важније. Ученик притом задржава преко деведесет одсто тачности учитеља уз око четрдесет пута мање параметара.

Итеративни поступак у основној поставци доследно је спуштао валидациону грешку из рунде у рунду, али ниједна рунда није надмашила дестиловани модел од кога је пошла. Узрок је нађен у односу чланова хибридне функције грешке. Са исправљеним односом, две рунде спуштају просечну стопу грешке на 35,71% уз подударност од 76,03%, при чему друга рунда доноси већи добитак од прве.

Кључне речи: аутоматско препознавање говора, дестилација знања, Whisper, фино подешавање, српски језик, стопа грешке на нивоу речи, говорни корпус

Садржај

Глава 1

Увод

Аутоматско препознавање говора (енг. *automatic speech recognition*, ASR) [1] данас је једна од главних области на које се стручњаци фокусирају у вештачкој интелигенцији и обраде природних језика. Технологија која је некада била ограничена на једноставне гласовне команде развила се у системе способне да у реалном времену претварају слободан говор у текст, раде превођење и анализирају смисао изреченог. Значај система за препознавање говора види се у њиховој широкој примени, од паметних виртуелних асистената и диктирања медицинских извештаја, па све до аутоматизације корисничке подршке и алата који помажу особама са инвалидитетом да лакше користе дигиталне услуге. Велики скок у прецизности ових система у последњој деценији директно је повезан са преласком са старих статистичких модела на дубоке неуронске мреже, а нарочито са појавом трансформерских архитектура. Модели који користе механизме пажње омогућили су далеко боље повезивање аудио сигнала и текстуалног контекста. Савремени модели који раде у једном пролазу (енг. *end-to-end*), међу којима се посебно издваја архитектура Whisper, обучени су на стотинама хиљада сати вишејезичног аудио материјала и показују одличну отпорност на различите акценте, позадинску буку и лошији квалитет самог аудио записа [8].

Упркос одличним резултатима, савремени системи за препознавање говора имају један велики проблем, највећа прецизност захтева и најјачи хардвер. Велики модели имају стотине милиона или милијарде параметара, што доводи до огромних захтева над графичком и радном меморијом, као и до кашњења (латенције) у раду. Овакви захтеви онемогућавају локално покретање модела на слабијим уређајима, мобилним телефонима и уграђеним

системима. Поред тога, стално коришћење удаљене инфраструктуре у облаку (енг. *cloud*) и скупих сервера ствара велике финансијске трошкове и захтева много електричне енергије. На крају, све више области попут здравства, правосуђа и банкарства захтева да се обрада звука ради искључиво локално због заштите приватности података, што потпуно искључује слање података на туђе сервере. Због тога се истраживачи све више фокусирају на методе компресије модела и развој мањих, ефикасних архитектура које покушавају да задрже тачност великих модела, али унутар скромнијег хардвера.

Када се проблем оптимизације модела посматра из угла српског језика, изазови постају још већи. У области обраде природних језика, српски се сврстава у језике са ограниченим ресурсима (енг. *low-resource languages*). То значи да за његов развој не постоје јавно доступни, огромни акустички корпуси од неколико десетина хиљада сати потпуно очишћеног материјала за обучавање. Додатно, српски језик има специфичне лингвистичке особине које отежавају рад система за препознавање говора, као што је изузетно богата и флексибилна морфологија, где иста реч може имати десетине различитих облика кроз падеже и глаголске промене. Слободан ред речи у реченици захтева да језички део модела буде веома напредан. Иако српски језик користи и ћирилично и латинично писмо, у овом раду фокус је искључиво на употреби латиничног писма, јер је српски текст у пракси најчешће у латиници. Whisper подржава и ћирилицу, али се ћирилични знакови због специфичности његове токенизације често деле на већи број ситних токена, што компликује стандардизацију текста и непотребно оптерећује декодер модела.

Због недостатка обимних података, мањи модели који се на српски језик директно примене „из кутије”[1] показују изразито високу стопу грешке. Са друге стране, покушај да се овај проблем реши само обичним финим подешавањем (енг. *fine-tuning*) мањих модела често доводи до брзог преприлагођавања (енг. *overfitting*) због мале количине доступних података.

Предмет овог мастер рада је практична примена, развој и детаљна провера методологије дестилације знања (енг. *knowledge distillation*) на малом моделу `whisper-tiny`, са крајњим циљем да се прилагоди за ефикасан рад на српском језику. Цело истраживање је усмерено на пренос знања са великог, прецизног

[1]Примена унапред обученог модела на нови задатак или језик без додатног обучавања или финог подешавања (енг. *out-of-the-box*, *zero-shot*).

и рачунарски захтевног модела у улози учитеља (`whisper-large-v3`) на мањи модел у улози ученика. Овај процес се реализује кроз директно усклађивање одговора, где мали модел учи да опонаша понашање и излазне вероватноће великог модела. То се постиже праћењем и копирањем вероватноћа на нивоу сваког појединачног токена, као и кроз преношење знања на нивоу целих реченица помоћу текстуалних транскрипата које велики модел даје за потребе обучавања.

Главни циљ рада јесте изградња и провера кружног поступка обучавања у коме модел учи на сопственим грешкама. У сваком циклусу мери се где модел греши и какве су природе те грешке, а затим се тим налазом усмерава избор узорака за наредни круг обучавања, тако да тежиште падне на оне врсте грешака које се доучавањем и могу исправити. Други циљ рада јесте изградња и чишћење звучног корпуса од око 200 сати на српском језику, при чему се поступцима поравнања звука и текста из њега уклањају нејасни и лоше исечени снимци, који би иначе унели шум у обучавање и умањили способност учења малог модела.

На основу овако постављеног предмета и циљева истраживања, у раду се дефинише радна хипотеза која тврди да ће дестилација знања, вођена циљаном анализом грешака, омогућити малом моделу знатно поузданије препознавање говора на српском језику у односу на исти модел обучен стандардним финим подешавањем над истом количином података. Поузданост се овде мери стопом грешке на нивоу речи, при чему се претпоставља да добитак неће доћи од побољшања реченице просечног квалитета него од уклањања потпуно промашених излаза, оних у којима модел упада у петљу или измишља текст, док подударност са референтним транскриптом остаје на нивоу који се постиже финим подешавањем.

Структура рада је подељена на осам међусобно повезаних поглавља. Након првог поглавља, које доноси уводне напомене и дефинише предмет, хипотезу и циљ истраживања, друго поглавље даје теоријске основе, архитектуру модела Whisper, појам дестилације знања и преглед постојећих решења за српски језик. Треће поглавље детаљно описује прикупљање података и формирање почетног аудио корпуса, док четврто поглавље обухвата анализу квалитета података кроз груписање узорака по боји гласа и проверу конзистентности транскрипата. Пето поглавље описује стратификацију и присилно поравнање текста и аудио сигнала. Шесто поглавље обухвата фино

подешавање, дестилацију знања и итеративно поновно учење на грешкама. Седмо поглавље је посвећено резултатима и дискусији, обухватајући поређење сва три приступа, промену структуре грешака кроз рунде, анализу осетљивости на однос чланова функције грешке и захтеве над рачунарским ресурсима. Осмо поглавље доноси закључак, у коме се проверава хипотеза, сумирају доприноси рада, наводе његова ограничења и дају смернице за будућа истраживања.

1.1 Напомена о употреби алата вештачке интелигенције

Пошто су алати засновани на великим језичким моделима и сам предмет овог рада, ред је рећи и где су они у његовој изради коришћени. Њихова употреба овде има две јасно раздвојене улоге, једну методолошку и једну помоћну.

Методолошка улога описана је у самом раду и представља део поступка, а не помоћ при писању. Мултимодални модел `gemini-2.5-flash` употребљен је за генерисање референтних транскрипата целог корпуса, на начин детаљно описан у одељку 3.2.1, укључујући и тачан текст упита. Модел `whisper-large-v3`, фино подешен на српском језику, употребљен је као учитељ у дестилацији и као други, независни транскриптор при провери конзистентности текста. Ови избори, њихова ограничења и последице по поузданост референтних ознака размотрени су у одељку 8.3.

Помоћна улога тиче се израде самог рада. За њу је коришћен четбот Claude компаније Anthropic, заснован на великом језичком моделу, и то у две ситуације. Прва је писање програмског кода, где је модел коришћен за предлоге при писању помоћних скрипти, за објашњавање порука о грешкама и за тражење узрока неисправног понашања кода. Друга је рад на тексту, где је модел коришћен за језичко и стилско дотеривање већ написаних пасуса, за проверу правописа и за предлоге сажимања предугачких реченица.

У свим наведеним случајевима алати су имали улогу помоћног средства, а не извора садржаја. Целокупна структура рада, поставка хипотезе, избор и редослед експеримената, тумачење добијених бројева и сви закључци дело су аутора. Ниједан резултат, вредност у табели ни навод из литературе није преузет из излаза језичког модела без провере, сви бројеви потичу из

покренутих експеримената и програмског кода датог у одељку 8.2, а сви наведени радови проверени су у изворнику. Сав програмски код настао уз помоћ ових алата ручно је прегледан, покренут и проверен, а одговорност за садржај рада у целини сноси аутор.

Глава 2

Теоријске основе

Whisper је напредни систем за препознавање говора који је развила компанија OpenAI, а дизајниран је као систем који ради у једном пролазу [8]. То значи да сирови аудио снимак улази директно у мрежу, а из ње излази готов, чист текст, без потребе за компликованим раздвајањем на посебне акустичке и језичке подсистеме, како се то некада радило. За разлику од старијих модела који су обучавани у строгим условима на пажљиво очишћеним и вештачким снимцима, Whisper је обучен на огромном и разноликом скупу од преко 680.000 сати вишејезичног аудио материјала покупљеног са интернета. Захваљујући томе, модел добро подноси услове из стварне употребе и сналази се са различитим дијалектима, позадинском буком, слабијим квалитетом звука и неуобичајеним нагласцима, што је важно за примену на српском говорном подручју.

2.1 Како Whisper прима аудио сигнале

Пошто су улазни подаци за Whisper аудио снимци различитих дужина и формата, модел захтева да се они на самом почетку стандардизују. Систем најпре подешава учестаност одабирања свих снимака на 16 kHz како би сви били истог квалитета. Након тога, аудио запис се сече на једнаке делове у трајању од тачно 30 секунди. Ако је неки снимак краћи, модел му на крај аутоматски додаје тишину (енг. *padding*), док се дужи снимци секу на више повезаних делова од по 30 секунди. На овај начин, подаци на улазу у мрежу увек имају потпуно исте димензије.

Када се звук исече на блокове од 30 секунди, он се претвара у облик који

Whisper може да обрађује. Аудио сигнал се посматра кроз мале временске прозоре ширине 25 милисекунди, који се померају у корацима од по 10 милисекунди дуж целог снимка. На сваки од ових појединачних прозора примењује се краткотрајна Фуријеова трансформација (енг. *short-time Fourier transform*, STFT). Пошто класична Фуријеова трансформација анализира цео сигнал одједном и губи временску компоненту, ова трансформација решава проблем тако што рачуна фреквенцијски садржај локално, унутар сваког појединачног временског прозора. Сам процес превођења сигнала из временског у фреквенцијски домен изводи се коришћењем алгоритма брзе Фуријеове трансформације, који знатно смањује број потребних операција и убрзава израчунавање густине спектралне снаге. Овим поступком за сваки корак добија се информација о јачини појединих фреквенција у том тренутку, чиме се формира класичан спектрограм, дводимензионални приказ који показује који су тонови били најгласнији у ком тренутку.

Следећи корак је пребацивање фреквенција на мел скалу [7], која је подељена на тачно 80 фиксних канала. Ови канали представљају одвојене фреквенцијске опсеге, од нижих ка вишим тоновима. Пошто људи много боље примећују мале промене у нижим тоновима где се налази већина људског говора, мел скала је направљена тако да ниске фреквенције подели детаљније, кроз више канала, док високе тонове сажима у мањи број канала. Процес дељења се изводи тако што се јачина звука из обичног спектрограма групише и распоређује у ових 80 канала, чиме се задржавају само она својства гласа која су важна за препознавање речи. Број канала зависи од варијанте модела. Осамдесет канала користе све варијанте закључно са `whisper-large-v2`, међу њима и `whisper-tiny`, док `whisper-large-v3`, употребљен у овом раду у улози учитеља, ради са 128 канала.

На самом крају, на вредности ових канала примењује се логаритамска функција како би се добио коначни лог-мел спектрограм, јер људско уво и јачину звука доживљава логаритамски. Излаз овог процеса је матрица димензија 80×3000 код варијанти са осамдесет канала, дакле осамдесет фреквенцијских канала и три хиљаде временских корака, што за корак од десет милисекунди тачно одговара блоку од тридесет секунди. Та матрица се као готов пакет прослеђује енкодеру.

2.2 Архитектура Whisper модела

Архитектура модела Whisper [8] заснована је на класичној енкодер-декодер трансформерској структури [9], која је специјално прилагођена за задатке обраде и препознавања говора. Уместо сировог аудио сигнала, модел као улаз користи претходно припремљени лог-мел спектрограм, чиме се комплексни звучни талас преводи у формат погодан за матрично рачунање. Сама мрежа је подељена на две главне компоненте. Оне су аудио енкодер, који обрађује акустичке сигнале, и текстуални декодер, који на основу тих сигнала генерише коначан текст.

Текстуални декодер преузима звучну репрезентацију и претвара је у текст, предвиђајући токен по токен преко механизма унакрсне пажње, при чему истовремено узима у обзир улазне атрибуте из енкодера и већ исписане токене. Декодер подржава вишезадатно учење, па пре почетка исписивања добија специјалне почетне токене који одређују шта треба да ради. Токен `<|startoftranscript|>` означава почетак транскрипције, а на тој позицији модел процењује и вероватноћу да у снимку уопште нема говора, преко токена `<|nospeech|>`. Наредни токен активира језички профил, чиме се речник сужава на изабрани језик, а последњи одређује да ли се ради транскрипција или превод на енглески.

2.3 Дестилација знања

Друга целина теоријске основе овог рада тиче се начина на који се знање једног модела преноси на други. Уобичајено обучавање неуронске мреже ослања се на тачну ознаку, за сваки пример постоји један исправан излаз, а све остало је погрешно. Хинтон, Вињалс и Дин приметили су да велики, већ обучен модел садржи знатно више од тога [2]. Његова излазна расподела показује и колико је модел сигуран у сваку од алтернатива, па се из ње види који су одговори били блиски тачном, а који потпуно промашени. Ту информацију они називају мрачним знањем (енг. *dark knowledge*).

Дестилација знања је поступак у коме мањи модел, ученик, учи да опонаша расподелу већег модела, учитеља. Уместо да види само тачну ознаку, ученик добија цео вектор вероватноћа. Да би разлике међу мањим вероватноћама постале видљиве, логити се пре примене softmax функције деле температуром

T, чиме се расподела омекшава, што је температура већа, разлике су мање изражене, а реп расподеле добија већи значај.

Укупна грешка обично је комбинација два члана, једног који везује ученика за тачну ознаку и другог који га везује за омекшану расподелу учитеља:

$$L = \alpha \cdot L_{\mathrm{CE}} + (1 - \alpha) \cdot T^2 \cdot L_{\mathrm{KL}},$$

где је L_{CE} унакрсна ентропија у односу на тачну ознаку, L_{KL} одступање ученика од омекшане расподеле учитеља, а коефицијент $\alpha \in [0, 1]$ одређује њихов однос. Пошто омекшавање смањује величину градијената дестилационог члана приближно фактором $1/T^2$, тај члан се множи са T^2, да би оба члана остала на упоредивој скали. Тачан облик оба члана, у виду у коме су коришћени у овом раду, дат је у одељку 6.2.2.

Према прегледу области [5], дестилација се дели на три врсте према томе шта се преноси. Дестилација заснована на излазу (енг. *response-based*) преноси излазну расподелу последњег слоја, дестилација заснована на атрибутима (енг. *feature-based*) поред тога користи и међуслојеве, а дестилација заснована на односима (енг. *relation-based*) преноси односе међу примерима. У овом раду примењена је прва врста, јер не захтева да учитељ и ученик имају упоредиву унутрашњу структуру, што је овде важно пошто се два модела разликују и по броју слојева и по димензији репрезентација.

Фурланело и сарадници показали су да ученик не мора бити мањи од учитеља и да поновљена дестилација може дати модел бољи од полазног [4]. Тај налаз је непосредна мотивација за трећи приступ овог рада, у коме се исти поступак понавља кроз више рунди.

За језике са ограниченим ресурсима дестилација је привлачна из практичног разлога, она додаје нов извор надзора, али не тражи ниједан нов означен пример. Иста количина података носи више информације када уз њу иде и расподела великог модела.

2.4 Постојећа решења за српски језик

Whisper подржава српски у оквиру свог вишејезичног скупа, али је удео српског у 680.000 сати обуке мали, па су резултати без додатног прилагођавања слаби, што потврђује и мерење у поглављу 7.

Јавно доступна решења за српски настају скоро искључиво доучавањем великих варијанти модела Whisper на отвореним корпусима. Најбоље резултате пријављују модели заједнице објављени на платформи Hugging Face, модел `whisper-large-v3` доучен на српском делу скупова Common Voice 13 [20] и Google FLEURS [21], са пријављеном стопом грешке на нивоу речи од 5,56% [16], и старија варијанта заснована на моделу `whisper-large-v2`, доучена на скупу Common Voice 11, са пријављених 10,76%. Постоје и модели засновани на архитектури wav2vec 2.0 [23], од којих је један коришћен у овом раду за присилно поравнање, што је описано у одељку 5.2.1.

Корпуси на којима су ти модели обучени углавном садрже читани говор. Common Voice чине снимци добровољаца који читају унапред задате реченице, а FLEURS је вишејезични скуп направљен превођењем и читањем истог материјала. Изузетак је корпус Јужних вести [22], састављен од телевизијских прилога, у трајању од око педесет сати.

Из тога следе две празнине на које се овај рад ослања. Прва је величина модела, пошто су сва наведена решења велике варијанте, чија примена тражи јаку графичку картицу, док за најмању варијанту нема објављених резултата на српском језику. Друга је природа података, јер читани говор не одражава услове у којима се препознавање заиста користи, па је корпус овог рада намерно састављен од спонтаног говора у стварним условима, што је описано у поглављу 3.

Глава 3

Подаци и формирање корпуса

Подаци коришћени у овом раду прикупљени су из јавно доступних, отворених извора са интернета, са примарним фокусом на платформу YouTube, и употребљени су искључиво у некомерцијалне, научноистраживачке сврхе. Коришћење овог садржаја ослања се на законска ограничења ауторског права предвиђена Законом о ауторском и сродним правима Републике Србије. Подаци се користе искључиво за моделовање система општег препознавања говора, без обраде биометријских или приватних личних података говорника, чиме је испоштован и етички оквир приватности субјеката који су садржај свесно пласирали у јавну сферу.

Сакупљени корпус састоји се од укупно 200 сати аудио записа на српском језику, који је структурно подељен на више целина сличних података. Једну трећину корпуса чини артикулисано читање унапред припремљеног текста (аудио књиге), другу трећину телевизијска гостовања, интервјуи и подкасти, а трећу различити дугачки видео записи са интернета попут документараца. Коришћење платформе YouTube као примарног извора оправдано је тиме што је она највећа и најразноврснија доступна база живог, природног говорног српског језика у стварним условима (енг. *in-the-wild*). Зато се овакви корпуси често користе за обучавање и фино подешавање система за препознавање говора.

Управо зато што су подаци преузети из ауторски заштићених извора и што снимак људског гласа истовремено представља податак о личности, пошто омогућава идентификацију говорника, сам корпус се не објављује. Изузетак предвиђен за научно истраживање покрива обраду тог материјала за потребе овог рада, али не и његово даље дељење. Поновљивост рада зато не

почива на скупу него на поступку, прикупљање, сегментација, критеријуми одабира и све мере квалитета описани су у наредним поглављима са тачним вредностима прагова и статистикама добијеног корпуса, тако да се идентичан поступак може спровести над било којим другим материјалом. Обучени модели и програмски код, за разлику од самих података, јавно су доступни, а њихове адресе дате су у одељку 8.2.

Главна разлика између јавно доступних аудио записа са поменуте платформе и локално снимљених аудио података огледа се у квалитету и степену компресије сигнала. Током подизања видео садржаја, алгоритми платформе YouTube примењују компресију звука са губитком (енг. *lossy compression*), претварајући га у Opus или AAC, формате за аудио компресију, са протоком од 128 kbps до 160 kbps [10]. Насупрот томе, сирови, локално снимљени аудио записи задржавају пун квалитет пулсно-кодне модулације (енг. *pulse-code modulation*, PCM), стандардног формата за дигитализацију аналогног звука [11], са протоком од 256 kbps за моно формат при учестаности одабирања од 16 kHz, односно 1411 kbps за стерео формат при учестаности од 44,1 kHz.

Ова компресија директно утиче на спектрална својства сигнала. Алгоритам платформе YouTube примењује нископропусни филтер који агресивно одсеца фреквенције изнад 16 kHz или 18 kHz, у зависности од тога која се од више чуваних копија различитог квалитета преузима, интерпретирајући их као мање приметне за људски слух. Поред тога, примењује се и компресија динамичког опсега која уједначава амплитуду најтиших и најгласнијих делова говора ради конзистентности репродукције. Ове измене не сметају моделу Whisper јер се он свакако фокусира на ниже тонове и аутоматски своди сваки снимак на 16 kHz. Ипак, овакво сечење неизбежно уништава фине детаље у гласу и ствара дигитални шум. То се највише примећује код пискавих сугласника, као што су с, ш, ћ и ч, који због губитка тих високих тонова губе своју оштрину и природан звук.

Ипак, Whisper добро подноси оваква изобличења и дигитални шум. Као што је већ наведено, оригинални модел је обучен методом слабог надгледања (енг. *weak supervision*) на преко 680.000 сати разноврсних снимака са интернета на више језика. Управо због обучавања на толикој количини несавршених података, модел је отпоран на слаб квалитет звука.

3.1 Алгоритам за сегментацију дугих аудио записа

У фази претходне обраде, прикупљени аудио материјал сегментише се на смислене узорке, аудио исечке, чије трајање варира у опсегу од 5 до 30 секунди, са циљаном просечном дужином од 15 секунди, чиме је формиран коначан скуп података прилагођен за ефикасно секвенцијално учитавање и обучавање. Како би се овај процес аутоматизовао над сировим подацима, развијен је наменски алгоритам за динамичко сечење аудио записа дужих од 30 секунди, са кључним задатком да се сигнал открива и пресеца искључиво у тренуцима паузе у говору, чиме је спречено пресецање речи на пола. Алгоритам се састоји из претходне обраде сигнала и главног сечења аудио записа.

Пре саме анализе, аудио сигнал се претвара у монофонски формат и учестаност одабирања му се смањује (енг. *downsampling*) на 16 kHz, што је стандард за обраду говора унутар архитектуре Whisper. За препознавање пауза у говору коришћена је функција за детекцију тишине из библиотеке pydub. Праг осетљивости постављен је на вредност -40 dB, док је минимални непрекидни прозор тишине дефинисан на 400 ms. Тако проналазимо све сегменте који испуњавају ове критеријуме.

Логика динамичког сечења заснива се на прорачуну средишње тачке за сваку детектовану паузу. Временски маркер S_{start} дефинише тренутак када ниво сигнала падне испод -40 dB, док S_{end} означава тренутак када се сигнал врати изнад овог прага, уз стриктан услов да трајање паузе испуњава критеријум $S_{\text{end}} - S_{\text{start}} > 400$ ms. Временски маркер потенцијалног реза M_{mid} рачуна се као средиште тако одређене паузе:

$$M_{\text{mid}} = \frac{S_{\text{start}} + S_{\text{end}}}{2}.$$

Средишња тачка узима се као потенцијална локација реза како би се обезбедила максимална удаљеност од суседних речи и очувао акустички интегритет лексичких целина. Даљи процес сегментације вођен је стриктним временским ограничењима, при чему алгоритам акумулира трајање говора од почетка текућег сегмента до наредне идентификоване тачке реза и доноси одлуке на основу скупа дефинисаних временских прагова.

Ако је акумулирани сегмент краћи од доње границе од 5 секунди, тачка реза се свесно прескаче, а сигнал наставља да се акумулира у оквиру истог блока. У случају да дужина сегмента упадне у оптималну зону између 5 и 15 секунди, алгоритам унапред испитује следећу тачку реза, ако и наредна тачка задржава укупну дужину унутар максималне границе, рез се одлаже ради максимизације дужине узорка ка циљаном просеку, док се у супротном операција пресецања извршава на тренутној позицији. Ако ипак укупна дужина сегмента премаши горњу границу од 15 секунди, алгоритам се враћа корак уназад и форсира рез на претходној валидној тачки тишине, под условом да она задовољава услов минималног трајања.

Коначно, алгоритам обрађује граничне случајеве на самом крају аудио записа, односно ситуације када је преостали део сувише кратак, испод 5 секунди, или предугачак за стандардни временски оквир. Ако је тај крајњи сегмент дужи од 5 секунди, он се уписује као самосталан узорак, док се у супротном аутоматски припаја претходном сегменту ради очувања контекста. За екстремне случајеве где у изворној датотеци уопште нема регистрованих пауза, алгоритам задржава комплетан запис у оригиналном облику како би се избегло насилно пресецање речи усред говора. Таква датотека се накнадно подвргава секундарној обради помоћу измењене верзије алгоритма, специјализоване за сегментацију непрекидних аудио записа велике дужине. То се огледа у томе што основни алгоритам сече говор на сегменте од 5 до 15 секунди, док ова верзија динамички помера границе до максималних 30 секунди, што одговара максималном временском прозору који Whisper може да обради одједном током обучавања.

Након сечења, добијени делови се аутоматски нумеришу и чувају у засебне директоријуме у некомпримованом `wav` формату, чиме се добија чист и акустички уредан говорни корпус за обучавање неуронске мреже, који у укупном збиру броји 51.428 генерисаних узорака, од чега 1.842 узорка траје дуже од 30 секунди.

3.2 Припрема метаподатака за аудио сегменте

Задатак у овој фази је обележавање прикупљених података како би се за сваки аудио узорак дефинисао тачан циљни текст са којим ће се модел поредити током обучавања. Пошто обучавање неуронске мреже захтева

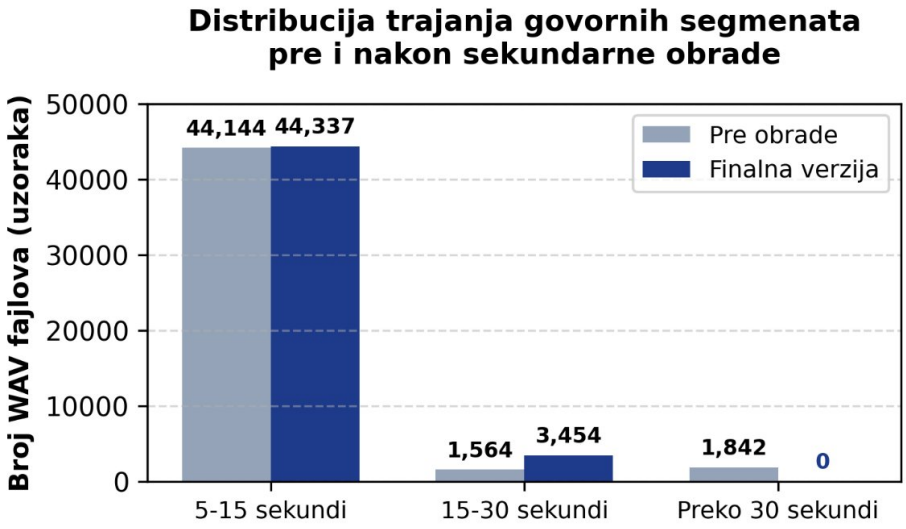
упарене податке, овај поступак аутоматизује генерисање циљног текста и његово спајање са одговарајућим звуком. Крајњи излаз је скуп парова датотека истог назива, где `wav` датотека носи звук, а `json` датотека садржи тачан текст и метаподатке који служе као директна референца за учење модела.

У првој фази сирови аудио сегменти прослеђују се моделу `whisper-large-v3` ради генерисања примарног транскрипта, без икаквог помоћног упита, уз једини задати податак да је језик снимка српски. Поред текста, Whisper за сваки запис даје и оцену сопствене сигурности. Он снимак интерно дели на N мањих временских целина и за сваку рачуна просечну логаритамску вероватноћу изговорених речи, коју овде означавамо са $\overline{\ell}_i$, из чега се добија коначан проценат сигурности c за цео запис:

$$c = \left(\frac{1}{N} \sum_{i=1}^{N} e^{\overline{\ell}_i} \right) \times 100.$$

Пошто је општи Whisper склон изостављању речи и грешкама на српском језику, референтни транскрипт добијен је помоћу мултимодалног модела `gemini-2.5-flash`, који аудио запис анализира непосредно, док му транскрипт модела Whisper служи само као помоћна смерница о теми. Правилима у упиту наложено му је да текст исписује малим словима латинице, без интерпункције, да стране речи и имена пише фонетски и да бројеве исписује словима, а тачан текст упита дат је у одељку 3.2.1. Тако добијени текст представља референцу коју модел треба да предвиди и чува се у засебном пољу, поред сировог излаза модела Whisper, што омогућава њихово касније поређење. Употреба таквих модела за аутоматско обележавање све је чешћа у пракси, јер знатно смањује потребу за спорим ручним означавањем.

Након генерисања текста систем проверава трајање сваког записа, читајући број одбирака и учестаност непосредно из заглавља звучне датотеке. Записи краћи од тридесет секунди добијају одговарајућу ознаку у пратећој `json` датотеци, преко које алгоритам за накнадно сечење препознаје које датотеке треба додатно да обради. Након те обраде добијен је 47.791 узорак краћи од тридесет секунди, чија је расподела по трајању приказана на слици 3.1, где 44.337 узорака траје између пет и петнаест секунди, а 3.454 узорка између петнаест и тридесет.



Слика 3.1: Расподела генерисаних узорака према трајању

Важно је напоменути да је почетни корпус садржао 200 сати сировог аудио материјала. Међутим, током обраде су аутоматски уклоњени сви сегменти на којима није забележен чист глас, већ тишина, музика, позадинска бука или неразговетан говор. Будући да за такве делове није било могуће генерисати поуздан текстуални пар, они су елиминисани из крајњег скупа података. Управо ти сегменти чине разлику између укупног броја исечака добијених сечењем, наведеног у одељку 3.1, и броја узорака приказаног на слици 3.1. На овај начин добијени су јасно дефинисани и филтрирани парови звучних сигнала и припадајућих транскрипата оптималне дужине.

Сваком исечку се на основу његове локације и идентификатора додељују метаподаци матичног записа, оригинални назив, путања, изворна адреса и тип садржаја. Пошто исечци из различитих извора могу носити исто име, при преклапању се на крај имена додаје јединствени суфикс.

3.2.1 Упит за генерисање референтног транскрипта

Упит је постављен тако да модел не преписује понуђени текст него да га користи само као оквир, док звук остаје једини извор истине. Правила о писму, интерпункцији, страним речима и бројевима служе стандардизацији ознака, јер свака таква разлика иначе улази у меру грешке као да је модел погрешно чуо. Захтев да одговор буде у једном реду олакшава аутоматску

обраду, а добијени текст се додатно нормализује, своди се на мала слова и уклањају се сувишни размаци. Поступак се у случају неуспеха понавља највише три пута.

```
POMOĆNI TEKST: '{original_text}'

TVOJ ZADATAK JE DA NAPRAVIŠ SAVRŠENU FONETSKU TRANSKRIPCIJU
AUDIO FAJLA!

Ovo su apsolutna i najstroža pravila koja moraš slepo da
pratiš:

1. AUDIO JE GLAVNI I JEDINI IZVOR ISTINE. Pomoćni tekst služi
samo da otprilike znaš o čemu se radi. Moraš pažljivo odslušati
kako čovek izgovara svaku reč i to verno preneti.

2. SVA SLOVA MORAJU BITI ISKLJUČIVO MALA SLOVA SRPSKE LATINICE
(a, b, c, č, ć, d, dž, đ, e, f, g, h, i, j, k, l, lj, m, n, nj,
o, p, r, s, š, t, u, v, z, ž) i razmaci. Nikakva velika slova
nisu dozvoljena!

3. STRANE REČI PIŠI FONETSKI, onako kako su izgovorene na
srpskom (npr. ako izgovara 'da vinčijev' moras napisati 'da
vinčijev', ako izgovara 'den braun' piši 'den braun', ne 'dan
brown').

4. SVE BROJEVE MORAŠ NAPISATI REČIMA, tačno onako kako su
izgovoreni (npr. 'dvadeset dva'). Nije dozvoljeno korišćenje
numeričkih cifara u odgovoru pod bilo kakvim uslovima!

5. ZABRANJENA INTERPUNKCIJA: Odgovor ne sme da sadrži apsolutno
nikakve znakove interpunkcije (bez tačaka, zareza, upitnika i
slično).

6. VRATI ISKLJUČIVO TRANSKRIBOVAN TEKST U JEDNOM JEDINOM REDU
(BEZ PRELOMA). Bilo kakvi komentari, zaglavlja, objašnjenja ili
prelasci u novi red (Enter/Newline) su najstrože zabranjeni.
```

У упиту је `{original_text}` место на које се уписује транскрипт добијен у првом пролазу. Уз тако састављен упит моделу се прослеђује и сам аудио запис, као низ бајтова, док су остале поставке модела остале подразумеване.

Уз сваки исечак чува се и сирови излаз модела Whisper, са интерпункцијом и великим словима, његова оцена поузданости, коначни стандардизовани

транскрипт који служи као референца за учење, ознака да ли је запис краћи од тридесет секунди, као и тип извора и тематска област снимка.

Глава 4

Анализа квалитета података

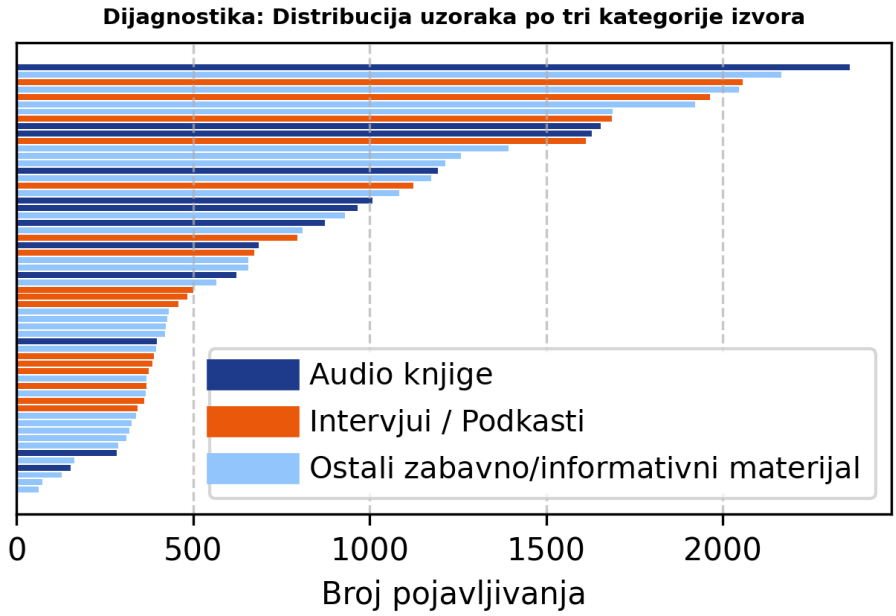
Иако је у претходној фази успешно генерисан и упарен основни скуп података, сирови извори са собом неизбежно доносе различите аномалије и неконзистентности. Због тога се поставља кључно питање у којој мери су овако генерисани подаци погодни за обучавање модела. Да би се обезбедила стабилност процеса учења, неопходно је утврдити постојање различитих типова података, као и проверити да ли су они равномерно распоређени по класама и изворима.

Оваква статистичка анализа је пресудна јер директно диктира успешност финог подешавања модела и спречава појаву пристрасности. Расподела броја узорака по категоријама извора приказана је на слици 4.1, а збирне вредности по категоријама у табели 4.1.

Табела 4.1: Број узорака по категоријама извора

Категорија извора	**Број узорака**
Аудио књиге	11.823
Интервјуи и подкасти	13.570
Остали забавни и информативни садржаји	22.398
Укупно	47.791

Ова расподела је детаљније приказана на графикону, где сваки појединачни стубић означава специфичан ресурс, попут конкретне књиге, емисије или подкаста, из којег су аудио подаци извучени. Са графика се јасно види да број узорака значајно варира међу изворима. Међутим, за крајњи циљ рада ова подела по изворима и медијским форматима није примарна. Модел за



Слика 4.1: Расподела узорака по категоријама извора

препознавање говора не групише податке на основу порекла датотеке, већ се фокусира искључиво на акустичка својства самог гласа. Због тога је следећи, кључни корак у анализи и обради корпуса пребацивање фокуса са извора на саму боју гласа и препознавање појединачних говорника.

4.1 Подела узорака по боји гласа

За анализу и класификацију аудио узорака према говорницима примењен је унапред обучени модел `speechbrain/spkrec-ecapa-voxceleb`, који је развила организација SpeechBrain [12]. Примарни циљ овог модела није препознавање и праћење семантичког садржаја, односно текста који је изговорен, већ дубока анализа акустичких својстава, као што су тоналитет и боја гласа. У зависности од задатка, модел се може користити за верификацију говорника или, као овде, за утврђивање ко је и у ком тренутку говорио унутар корпуса.

Архитектура модела заснована је на неуронској мрежи ECAPA-TDNN (енг. *emphasized channel attention, propagation and aggregation in time-delay neural network*) [13]. Модел је оптимизован да занемарује амбијентални шум и позадинску буку, док фокус усмерава искључиво на фундаменталне вокалне

атрибуте, фреквенцију и боју гласа, назалност, резонанцију вокалног тракта, као и на специфичан темпо и динамику самог говора. Као крајњи резултат, модел пресликава аудио запис било које дужине у јединствени компактни векторски простор, у виду векторске репрезентације од 192 димензије.

Унутар векторске репрезентације свака вредност описује по један акустички аспект говора, па су гласови тим сличнији што су њихови вектори ближи у вишедимензионалном простору. На основу те удаљености и емпиријски одређеног прага осетљивости узорци се деле у групе према идентитету гласа. Модел се учитава са платформе Hugging Face и извршава на графичкој картици ако је доступна.

Обрада се одвија у пакетима од осам узорака, јер би истовремено учитавање свих 47.791 записа препунило радну и графичку меморију. Сваки запис се чита са диска, вишеканални записи своде се на моно усредњавањем канала, а оштећене датотеке се прескачу да не би прекинуле поступак. Пошто записи нису исте дужине, краћи се допуњавају нулама до најдужег у пакету, уз пратећи вектор стварних дужина који моделу служи као маска, да би анализирао само стварни говор, а додате нуле занемарио.

Тако припремљен пакет пролази кроз мрежу, која за сваки запис даје вектор од 192 вредности. Ти вектори се потом пребацују са графичке картице у радну меморију и чувају за наредни корак, груписање по сличности гласа.

Након издвајања и чувања векторских репрезентација, њихово поређење обавља се алгоритмом BIRCH. Он је изабран због малих захтева над радном меморијом, због тога што не захтева да број говорника буде унапред познат и због тога што посао завршава у једном пролазу кроз податке.

Алгоритам BIRCH функционише као високо оптимизован систем архивирања који векторске репрезентације обрађује секвенцијално, једну по једну, градећи организовану структуру у облику стабла обележја (енг. *CF tree*). Уместо да оптерећује радну меморију памћењем свих појединачних вектора, алгоритам за сваку групу, односно скуп узорака са сличним гласом, рачуна и чува само основне статистичке податке. Ти подаци су број вектора N, који служи као бројач узорака, њихов збир LS, помоћу којег се рачуна просечни центар гласа, као и збир квадрата координата SS, који одређује геометријску ширину и границе те групе. Приликом обраде сваке нове векторске репрезентације, алгоритам се спушта кроз гране стабла тражећи најближи постојећи лист који садржи најсличнији глас. У том тренутку примењује се кључно правило

прага, постављеног на вредност 0,7, ако се нови снимак геометријски уклапа у границе тог листа, придружује се тој групи, а њена статистика се ажурира. Са друге стране, ако је разлика гласа већа од дозвољеног прага, алгоритам отвара потпуно нов лист на стаблу за тог новог говорника. По завршетку једног пролаза кроз све податке, алгоритам брзо пребројава и финализује преостале гране на стаблу, чиме самостално и без унапред задатог ограничења идентификује коначан број јединствених говорника у бази.

Интерно рачунање удаљености унутар алгоритма BIRCH подразумевано је ограничено на еуклидску метрику L_2 [14]. Када нова векторска репрезентација стигне у систем, алгоритам је не пореди са сваким појединачним гласом из историје, него удаљеност $D(x, C)$ између вектора x и групе C рачуна у једном кораку, искључиво из већ сачуваних сума:

$$D(x, C) = \sqrt{x^2 - \frac{2 \cdot x \cdot LS}{N} + \frac{SS}{N}}\,.$$

Ако је та удаљеност мања од задатог прага, узорак се придружује групи, а када више група задовољава услов, бира се геометријски најближа.

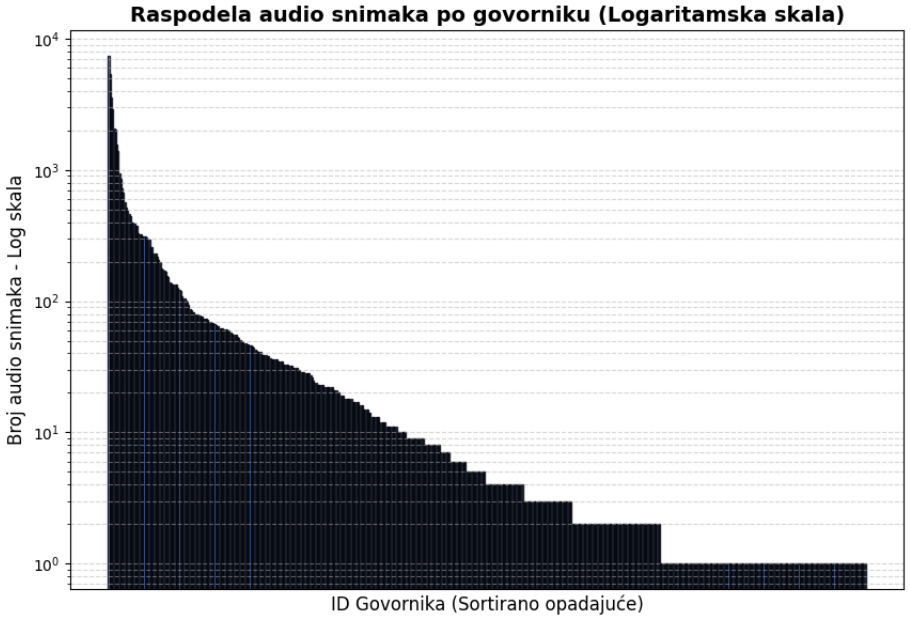
Примена еуклидске удаљености на груписање гласова носи, међутим, акустички проблем. Та метрика пореди и интензитете вектора, дакле јачину и гласноћу говора, док боју гласа носе искључиво правац и смер вектора. Мера која је овде примеренија изведена је из косинусне сличности, која мери само угао између два вектора. Њена комплементарна вредност, косинусна удаљеност, дефинише се као јединица умањена за косинусну сличност и занемарује разлике настале због гласноће микрофона или трајања снимка.

Овај проблем је решен применом нормализације L_2 над свим векторским репрезентацијама пре њиховог прослеђивања алгоритму, чиме су сви вектори сведени на јединичну дужину. У таквом нормализованом простору, квадрат еуклидске удаљености постаје директно и линеарно пропорционалан косинусној удаљености, што омогућава алгоритму да преко подразумеване и брзе метрике L_2 заправо посредно изводи угаону класификацију говорника.

4.2 Анализа група говорника

Након описа начина на који алгоритам BIRCH препознаје и групише гласове, следећи корак у анализи квалитета података јесте преглед добијених резултата. Да би се утврдило колико добро алгоритам раздваја говорнике

и каква је стварна структура скупа података, спроведена је статистичка анализа формираних група. Визуелни преглед ове расподеле, који јасно приказује уочене аномалије и структуру прикупљеног корпуса, дат је на слици 4.2.



Слика 4.2: Расподела аудио снимака по групама говорника (логаритамска скала)

На слици 4.2 приказано је како су аудио снимци распоређени по групама говорника. Пошто између најчешћих и најређих говорника постоје огромне разлике у броју снимака, график је приказан на логаритамској скали. Та скала омогућава да се јасно виде сви делови приказа, од оних где има на хиљаде снимака, па све до оних где их има свега неколико, што би на обичној линеарној скали било потпуно непрепознатљиво. Овакви подаци одговарају стварним условима, где је број узорака по говорнику веома неуједначен. На графику се јасно види неколико изузетно доминантних говорника, попут наратора аудио књига, који држе највећи део корпуса са чак 23.000 узорака, након чега крива нагло пада ка такозваном дугом репу. Тај реп чини огроман број мањих група са свега једним или два снимка. Због овакве структуре, где поједини минимални узорци заправо представљају шум или неразговетне делове, овај корпус је неопходно додатно филтрирати и средити пре него што се подаци пошаљу на обучавање и коначно коришћење.

Први корак у пречишћавању корпуса је уклањање шума и изолованих података који би могли негативно да утичу на обучавање модела. Одлучено је да се из скупа података потпуно уклоне сви говорници који имају мање од 10 узорака, јер се појављују изузетно мали број пута и не пружају довољно материјала за стабилно препознавање својстава гласа. Након примене овог филтера, из корпуса је издвојен 871 узорак који потиче од говорника са недовољним бројем снимака или представља шум, чиме је, уз изостављање оштећених и неупотребљивих датотека које нису могле да се обраде, укупан број валидних аудио снимака сведен на 46.555. Ипак, ови уклоњени подаци нису одбачени, него су сачувани као засебан скуп, у даљем тексту скуп непознатих говорника, изван и обучавања и оцењивања. Пре тога, у фази провере конзистентности текста (одељак 4.3), из њега је поређењем транскрипата уклоњен стварни шум попут музике.

Други корак пречишћавања тиче се неуједначене заступљености говорника. Поједини говорници имају диспропорционално велики број записа, што у обучавању води преприлагођавању на њихове гласове и нарушава способност уопштавања. Зато је број узорака по говорнику ограничен на највише 500, а вишак је издвојен случајним узорковањем.

Тиме је корпус подељен на два дела. Први, у даљем тексту основни корпус, обухвата 23.089 записа укупног трајања 88,62 сата и из њега се стратификацијом добијају скупови за обучавање, валидацију и тестирање. Други, у даљем тексту допунски скуп, чини издвојени вишак снимака доминантних говорника и броји 23.466 узорака у трајању од 83,18 сати. Тај скуп није одбачен него је намењен фази дестилације знања, о чему је реч у одељку 6.3.1, и пролази кроз исту проверу конзистентности описану у одељку 4.3. Иако тиме основни корпус чини приближно половину почетног материјала, смањење није проблем, јер су постојећи модели за српски језик, попут оног обученог на корпусу Јужних вести [22], развијени и валидирани на знатно мањем обиму материјала, нешто преко педесет сати.

4.3 Провера конзистентности текста

Поред акустичке анализе и груписања говорника, спроводи се и додатни ниво контроле квалитета који се фокусира искључиво на валидацију генерисаног текстуалног садржаја унутар корпуса. Главни циљ ове фазе

обраде јесте поређење сличности између референтних транскрипата, који представљају тачне ознаке, и транскрипата које је аутоматски генерисао модел `whisper-large-v3`. Оба текстуална записа складиштена су унутар пратећих датотека, референтни текст у пољу за коначни транскрипт, а излаз модела Whisper у засебном пољу. Пре самог поступка мерења, обе текстуалне ниске пролазе кроз идентичну фазу нормализације. Овај корак подразумева превођење свих знакова у мала слова, уклањање знакова интерпункције, као и претварање нумеричких записа у речи. Анализом коефицијента сличности између овако припремљених ниски омогућава се прецизно издвајање и трајно одбацивање оних узорака код којих је откривено критично одступање од референтног текста.

Овај поступак филтрирања примењује се над целокупним скупом података и уклања узорке код којих постоји критично одступање између два транскрипта. Праг је постављен тако да задржава уобичајене грешке у препознавању говора, које су чак и пожељне јер верније осликавају стварне услове, а одбацују се само сегменти чији садржај уопште не представља говор, попут чисте музике, аплауза или неразговетног шума. За такве снимке два модела независно генеришу битно различите транскрипте, њихов коефицијент сличности пада испод дефинисаног прага и они се аутоматски одбацују. Иста провера примењује се и на два раније издвојена скупа, допунски скуп и скуп непознатих говорника, чиме се управо овде спроводи раније најављено уклањање музике и позадинске буке. Притом се чист и квалитетан део допунског скупа задржава, јер ће имати кључну улогу у каснијем процесу дестилације знања.

4.3.1 Мера сличности транскрипата

За реализацију овог задатка искоришћена је класа SequenceMatcher из стандардне библиотеке difflib програмског језика Python. Овај алгоритам се заснива на Ратклиф-Обершелповом механизму, познатијем као гешталт поклапање образаца (енг. *Gestalt pattern matching*). Идеја алгоритма заснива се на рекурзивном принципу поделе проблема на потпроблеме. У првом кораку, алгоритам испитује обе ниске и идентификује најдужи заједнички подниз који се у њима појављује без прекида. Када пронађе то главно поклапање, алгоритам фиксира тај сегмент и више не посматра његове елементе. Преостале, неусклађене делове текстуалних секвенци са леве и десне

стране посматра као нове подниске. Поступак се затим рекурзивно понавља за леве, а потом и за десне преостале делове, све док постоје заједнички елементи. На овај начин, алгоритам успешно налази како примарна, тако и секундарна поклапања кроз дубину текста, занемарујући локалне аномалије или премештања речи. Квантитативна мера сличности изражава се кроз коефицијент поклапања R, који се рачуна према формули:

$$R = \frac{2 \cdot M}{L_a + L_b},$$

где L_a и L_b представљају укупне дужине ниски које се пореде, а M укупан број поклопљених знакова. Вредност M добија се сабирањем дужина свих k заједничких подсеквенци пронађених кроз све нивое рекурзије, дакле $M = \sum_{i=1}^{k} m_i$, где је m_i дужина i-те такве подсеквенце. Сваки знак урачунат у M мора бити упарен у обе ниске, чиме резултат представља чист удео идентичног текста у односу на укупну дужину посматраних узорака.

У развијеном решењу, као критична граница прихватљивости постављен је емпиријски праг од 0,40. Сваки пар код којег коефицијент сличности падне испод те вредности аутоматски се сврстава у критично одступање и уписује у контролну датотеку са комплетним приказом оба текста. Ова повратна информација омогућава прецизно издвајање и трајно уклањање неусаглашених текстуалних сегмената, чиме се финални корпус припрема за потребе стабилног надгледаног учења. Управо овим прагом регулишу се и претходно издвојени скупови, односно допунски скуп и скуп непознатих говорника. Код сегмената који садрже музику или нечист звук, два независна транскрипта се битно разликују, па њихов коефицијент сличности пада испод 0,40, чиме се музика и позадински шум поуздано откривају и уклањају управо кроз ово поређење, пре него што ти скупови уђу у фазе дестилације знања и финалног оцењивања.

Применом овог алгоритма над целокупним скупом података уклоњени су снимци чији садржај чине искључиво музичка подлога или превелик шум. Основни корпус се притом са почетних 23.089 свео на 22.390 узорака. Поступак је примењен и над два раније издвојена скупа. Скуп непознатих говорника смањен је са 871 на 808 узорака, док је допунски скуп сведен са 23.466 на 23.366 узорака.

Глава 5

Претходна обрада података

Припрема квалитетног корпуса за обучавање и оцењивање једна је од најважнијих фаза у развоју система за препознавање говора, јер директно утиче на стабилност конвергенције и способност уопштавања мреже. Текстуална провера конзистентности транскрипата већ је обављена у одељку 4.3, па се у овом поглављу описују преостала два нивоа контроле квалитета. Први обухвата статистички контролисану стратификацију корпуса, чиме се осигурава равномерна заступљеност свих аудио извора кроз подскупове података. Након тога, применом алгоритма за присилно поравнање ригорозно се проверава акустички квалитет сигнала и уклања позадинска бука, чиме подаци улазе у фазу обучавања без акустичких и текстуалних аномалија.

5.1 Стратификација

Стратификација је статистичка метода поделе скупа података на хомогене подскупове, стратуме, пре него што се изврши њихова коначна расподела у скуп за обучавање, валидациони и тест скуп. У системима за аутоматско препознавање говора проста насумична подела целокупног корпуса носи висок ризик од нарушавања репрезентативности података. Уколико се подела изврши без контроле слојева расподеле, може се десити да одређени извори података, на пример специфичне емисије, аудио књиге или подкасти, буду доминантно заступљени у скупу за обучавање, док у валидационом или тест скупу потпуно изостану.

Стратификација је кључна јер гарантује да ће сваки појединачни аудио извор задржати идентичну процентуалну заступљеност у сва три подскупа.

Тиме се обезбеђује да ниједан извор не буде виђен само током обучавања или само током оцењивања, чиме валидација постаје стабилна, а тест скуп објективан показатељ квалитета рада модела у стварним условима примене.

5.1.1 Методологија поделе и добијени скупови

Као основа узет је пречишћени основни корпус од 22.390 валидних аудио узорака. Циљ алгоритма био је подела корпуса у стандардној пропорцији 80:10:10, али уз стриктно очување расподеле унутар сваке категорије, односно сваког извора звука.

Поступак је узорке расподелио у три скупа. Скуп за обучавање броји 17.912 узорака и намењен је оптимизацији тежина модела. Валидациони скуп броји 2.239 узорака и служи за праћење конвергенције и спречавање преприлагођавања током обучавања. Тест скуп такође броји 2.239 узорака и намењен је коначном оцењивању тачности препознавања говора. Наведене вредности су величине непосредно после поделе. Коначни бројеви, добијени након провере присилним поравнањем, дати су у одељку 5.3.

Провером излаза поступка потврђено је да је алгоритам задржао прецизан математички однос унутар сваког појединачног медијског формата. Табела 5.1 приказује расподелу узорака кроз скупове за неколико карактеристичних извора са различитим нивоима заступљености у корпусу.

Табела 5.1: Расподела узорака по изворима кроз скуп за обучавање, валидациони и тест скуп

Назив аудио извора	**Укупно**	**Обучавање**	**Валидација**	**Тест**
Разговор са хакером	1.033	827	103	103
Пуца Србија	970	776	97	97
Кад су цветале тикве	820	656	82	82
Мали принц (аудио књига)	398	318	40	40
Момчадија игра мафију	53	43	5	5

Као што се из резултата може приметити, чак и код изразито заступљених извора, попут интервјуа Разговор са хакером са 1.033 узорка, и мање заступљених формата, попут забавног садржаја Момчадија игра мафију са свега 53 узорка, алгоритам је задржао строгу пропорцију од 80% у скупу за обучавање и по 10% у преостала два скупа. Овако избалансирана расподела обезбеђује да модел током обучавања равномерно види све типове

акустичких окружења, шумова и говора, док тестирање на преосталих 10% даје објективан увид у квалитет обучености.

5.1.2 Издвајање тест скупа

Након стратификације тест скуп се потпуно издваја. Он се изузима из свих даљих фаза обраде и аугментације које ће се примењивати на скуп за обучавање и валидациони скуп у наставку истраживања. Тест скуп се од тог тренутка више не мења и не користи све до коначног оцењивања у поглављу 7, како би пружио објективан и непристрасан увид у квалитет рада финалног модела.

Поред тест скупа задржан је и раније поменути скуп непознатих говорника, односно узорци иницијално препознати као шум или снимци говорника са премало појављивања. И они су изостављени из процеса обучавања, али за разлику од тест скупа пролазе кроз један ниво накнадне обраде, где су у фази провере конзистентности текста, поређењем транскрипата два независна модела, очишћени од сегмената који представљају чисту музику. Тај скуп чува се као резерва непознатих говорника. Сва мерења приказана у поглављу 7 изведена су искључиво над тест скупом, док је коришћење скупа непознатих говорника остављено за даља истраживања (одељак 8.4).

5.2 Присилно поравнање

Присилно поравнање (енг. *forced alignment*) јесте поступак у обради говорних сигнала којим се аудио запис аутоматски пресликава на одговарајући текстуални транскрипт, и то по времену. За разлику од класичног препознавања говора, где модел на основу звука генерише непознат текст, поступак присилног поравнања унутар алгоритма већ поседује тачан текст. Његов задатак је да „присили” модел да пронађе тачне временске границе, почетак и крај, за сваку изговорену реч у аудио сигналу.

Поступак омогућава аутоматску детекцију акустичких изузетака кроз препознавање сегмената у којима постоји неслагање између изговореног и записаног, чиме се из корпуса уклањају лоше исечени и нејасни снимци пре него што уђу у обучавање.

Сам механизам функционише тако што већ обучени акустички модел за српски језик анализира аудио снимак и упоређује га са датим текстом. Из тог поређења модел израчунава оцену поузданости, односно процењује колико се изговорени звук заиста поклапа са написаним речима. Да би се разумело како модел долази до ових прорачуна, у наставку је детаљно приказана његова унутрашња архитектура, као и тачан облик података које прима и враћа.

5.2.1 Одабир модела за присилно поравнање

За поравнање је свесно изабран `wav2vec2-xls-r-juznevesti-sr`, а не Whisper, што је непосредна последица разлике у њиховим основним архитектурама.

Модел заснован на архитектури wav2vec 2.0 [23] користи искључиво енкодерску архитектуру и конекционистичку временску класификацију (енг. *connectionist temporal classification*, CTC) као функцију грешке при пресликавању аудио сигнала. Он обрађује аудио кроз континуалне временске кораке, оквире, и за сваки корак директно предвиђа вероватноћу појављивања одређеног знака. Ова линеарна природа рада оквир по оквир чини га погодним за присилно поравнање, јер поступак поравнања може директно да прође кроз излазну матрицу логита и пронађе оптималну временску путању поравнања за понуђени текст. Са друге стране, Whisper је заснован на сложенијој енкодер-декодер архитектури, описаној у поглављу 2. Whisper је дизајниран да посматра шири контекст и генерише текстуалне токене један за другим, због чега нема фиксну временску корелацију између појединачног оквира и слова у логитима на начин на који то има модел заснован на архитектури wav2vec 2.0. Иако Whisper поседује механизме унакрсне пажње из којих време може приближно да се процени, модел wav2vec 2.0 фино подешен на корпусу Јужних вести пружа већу прецизност и стабилност, јер је његова матрица вероватноћа већ прилагођена фонетици српског језика.

5.2.2 Методологија валидације и рачунање оцене поравнања

Да би се измерило колико је неки аудио снимак усклађен са текстом, сирове излазне вредности модела, логити, не могу се употребити непосредно. Уместо тога, те вредности прво пропуштамо кроз логаритмовану softmax функцију,

која их претвара у логаритамске вероватноће:

$$\log\left(\text{Softmax}(x_i)\right) = x_i - \log\left(\sum_j e^{x_j}\right),$$

где су x_i логити. Прелазак на логаритамске вероватноће није само промена скале. Множење великог броја малих вероватноћа дуж путање брзо води ка бројевима које рачунар не може прецизно да представи, док се у логаритамском облику производ своди на збир, па се та опасност уклања. Формула се примењује на сваки појединачни оквир аудио записа и за свако слово из речника, а њоме се на крају добија матрица логаритамских вероватноћа.

Алгоритам анализира добијену матрицу и тражи ону путању која даје највећу вероватноћу да изговорени звуци заиста одговарају датом тексту. Међутим, пошто аудио снимци трају различито и садрже различит број речи, пуки збир свих вероватноћа није праведно мерило, јер би дужи снимци природно имали другачије укупне вредности од краћих. Да би сви узорци могли објективно да се пореде, рачуна се просечна оцена поравнања по токену S_{align}, изведена из логаритамских вероватноћа дуж путање:

$$S_{\text{align}} = \frac{\sum_{k=1}^{N} s_k}{N},$$

где s_k представља логаритамску вероватноћу додељену k-том кораку на изабраној путањи, док је N укупна дужина путање поравнања, односно број токена.

Међутим, у пракси вредности ове оцене често одступају у зависности од верзије библиотеке torchaudio и начина на који она интерно нормализује бројеве током проналажења оптималне путање кроз матрицу, оне која најбоље повезује временске оквире аудио снимка са одговарајућим словима или токенима из текста. Та путања налази се Витербијевим алгоритмом [15]. Пошто се коначни опсези ових бројева не могу унапред тачно предвидети, у поступак је уграђен механизам који у реалном времену прати минималне, максималне и просечне вредности на нивоу целог корпуса. Уместо слепог ослањања на теоријске претпоставке, границе за филтрирање одређују се емпиријски. Анализом конкретних узорака који су кроз овај поступак препознати као проблематични, акустички неисправни или оптерећени шумом, утврђују се стварни прагови унутар добијених вредности, па се на

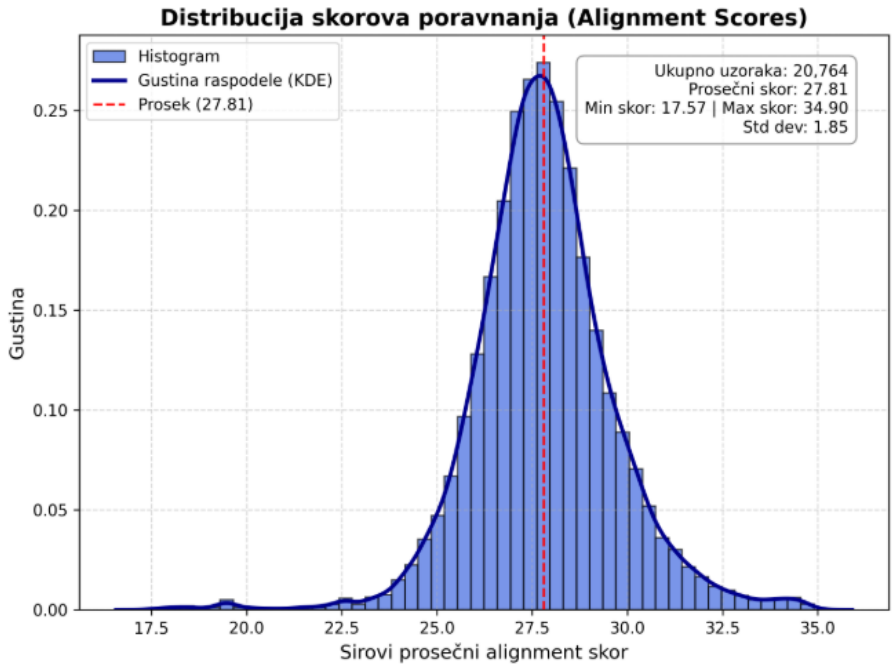
основу тих контролних тачака поставља прецизна граница за аутоматску тријажу прилагођена стварном понашању модела. У конкретној имплементацији, због интерне нормализације коју torchaudio примењује, ова оцена изражава се на позитивној скали, где веће вредности означавају боље акустичко-текстуално поклапање, док изразито ниске вредности указују на лоше или неуспело поравнање.

Ова валидација резултата поравнања примењује се над две велике, независне целине података, над скупом за обучавање и валидационим скупом добијеним стратификацијом, као и над допунским скупом уведеним у одељку 4.2. Пропуштањем оба скупа кроз алгоритам, уз ослањање на паралелно извршавање процеса, обезбеђено је да се цео корпус од 200 сати пречисти на исти начин. Узорци из обе целине који не пређу емпиријски постављени праг трајно се одбацују, тако да и основно обучавање и дестилација знања раде само са добро поравнатим и акустички исправним подацима.

5.2.3 Резултати присилног поравнања

Након успешног извршавања алгоритма за присилно поравнање над скупом за обучавање и валидационим скупом извршена је статистичка анализа добијених вредности просечне логаритамске вероватноће по токену, чија је расподела приказана на слици 5.1.

Идентичан поступак спроведен је и над допунским скупом. Резултати поравнања тог скупа показали су једнако високу чистину и стабилност. Минимална остварена вредност оцене износила је 18,08, максимална 34,65, док је просечна вредност износила 27,81. Готово идентична расподела и блискост просечних вредности између ова два независна скупа потврђују статистичку хомогеност података у погледу акустичког квалитета. Овим је обезбеђено да оба скупа прођу кроз једнако ригорозан ниво филтрирања, чиме коначни корпус улази у фазе обучавања и дестилације без уочених аномалија.



Слика 5.1: Расподела вредности оцене поравнања S_{align}

5.3 Финални скупови података након чишћења

Након стратификације, скуп за обучавање и валидациони скуп, заједно са допунским скупом, пропуштени су кроз присилно поравнање и проверу конзистентности текста описану раније. У овом кораку трајно се уклањају узорци код којих поравнање није успело, као и они код којих се транскрипти два независна модела битно разликују, на пример музика или позадински шум. Покретањем целог поступка над сачуваним подацима добијени су коначни бројеви дати у табелама 5.2 и 5.3.

Табела 5.2: Промена величине основних скупова после присилног поравнања

Скуп	**Пре овог корака**	**После**	**Уклоњено**
Скуп за обучавање	17.912	17.804	108
Валидациони скуп	2.239	2.219	20
Тест скуп	2.239	—	—

Сада имамо финалне скупове података, пречишћене и спремне за

Табела 5.3: Промена величине издвојених скупова после присилног поравнања

Скуп	**Пре овог корака**	**После**	**Уклоњено**
Скуп непознатих говорника	808	—	—
Допунски скуп	23.366	23.278	88

коришћење у процесу обучавања и оцењивања. Скуп за обучавање и валидациони скуп улазе у обучавање, тест скуп служи за сва мерења у поглављу 7, а допунски скуп за дестилацију знања. Скуп непознатих говорника остаје сачуван, али се у овом раду даље не користи.

Глава 6

Обучавање и оцењивање модела

Када су подаци очишћени од шума, изабрани и текстуално проверени, корпус је спреман за обучавање, оптимизацију и оцењивање модела. Да би се прецизно измерио ефекат различитих техника на рад система, ово поглавље је подељено на три засебна дела. Свака целина доноси другачији начин обучавања, а циљ је њихово међусобно поређење по тачности транскрипције, брзини рада и захтевима над хардверским ресурсима. Сваки од ових приступа полази од истог основног, необученог модела `whisper-tiny`.

Први део се бави стандардним финим подешавањем модела. Овај приступ је полазна тачка и најједноставнија техника, јер директно прилагођава постојећи модел својствима српског језика. Он захтева најмање графичке меморије и рачунарске снаге у односу на напредније методе.

Други део је посвећен дестилацији знања. У овој фази се знање из великог и прецизног модела учитеља преноси на мањи модел ученик. На тај начин добија се знатно лакша и бржа архитектура која ради ефикасније уз минималан пад у тачности. Резултате ове оптимизације упоредићемо директно са обичним финим подешавањем из првог дела.

Трећи део доноси итеративну дестилацију знања. То је најсложенији корак у коме се мањи модел додатно усавршава кроз неколико узастопних циклуса обучавања на специфичним деловима података.

6.1 Фино подешавање као полазна тачка

Први приступ је стандардно фино подешавање [3] модела `whisper-tiny` над припремљеним скупом за обучавање. Модел се иницијализује унапред

обученим тежинама, а затим се цела мрежа доучава на српском корпусу, при чему се као функција грешке користи унакрсна ентропија између излаза модела и референтног транскрипта. За разлику од приступа који следе, овде нема никаквог преноса знања из већег модела, једини извор надзора јесу тачне ознаке из корпуса.

Овај приступ служи као доња референтна тачка. Он показује колико се може добити самим прилагођавањем модела језику и домену, па се сваки каснији резултат мери у односу на њега.

6.1.1 Припрема података и токенизација

Полази се од коначних скупова добијених поступком описаним у поглављу 5. Сваки од њих садржи путање до аудио записа и одговарајуће транскрипте.

Приликом учитавања модела изричито се наводе језик и задатак, дакле српски језик и транскрипција. Ова два податка нису формалност. Whisper је вишејезични модел и очекује да низ токена почне посебним ознакама које му саопштавају о ком је језику и о ком задатку је реч. Навођењем језика и задатка токенизатор сам додаје префикс `<|startoftranscript|> <|sr|> <|transcribe|>`, па модел не мора да погађа језик из самог сигнала. Изостављање тих података довело би до тога да модел покушава препознавање језика, што код кратких узорака често доводи до грешке и уводи непотребан шум у обучавање.

Припрема примера своди се на токенизацију транскрипта, док се сам звук у овој фази не додирује. Разлог за то је меморија. Претварање целог корпуса у мел спектрограме унапред значило би држање неколико десетина гигабајта атрибута, при чему сваки узорак заузима исто места без обзира на стварно трајање, јер Whisper све записе допуњава до тридесет секунди. Уместо тога чува се само путања до записа, а атрибути се рачунају тек када узорак уђе у серију.

6.1.2 Формирање серије и функција грешке

Скуп података припремљен у претходном одељку садржи путање до аудио записа и токенизоване транскрипте, али не и сам звучни сигнал. Он се учитава тек када узорак уђе у серију, непосредно пре сваког корака обучавања.

Из листе узорака изабраних за једну серију прави се оно што модел заиста прима, два тензора, један са улазним атрибутима, други са тачним ознакама. Оба морају имати све редове исте дужине, а узорци у листи разликују се и по трајању снимка и по дужини транскрипта. Посао се зато дели на два дела, на обраду звука и на обраду текста.

Звучни део почиње учитавањем записа са диска и његовим свођењем на 16 kHz, након чега се од сигнала рачуна мел спектрограм. Преузорковање је обавезно зато што су мел филтери којима Whisper рачуна спектрограм дефинисани у односу на учестаност од 16 kHz [8]. Модел није научио како одређени глас звучи, него како изгледа распоређен по мел појасевима под претпоставком да секунда звука садржи тачно шеснаест хиљада одбирака. Запис снимљен на 44,1 kHz био би зато протумачен као развучен у времену, са формантима помереним ка нижим фреквенцијама, и модел би на улазу видео говор какав током обуке никада није чуо. Излаз ове обраде увек има исти облик, јер Whisper сваки запис скраћује или допуњава тишином до тридесет секунди, па се акустички део серије не мора посебно уједначавати.

Са текстом је другачије. Транскрипти остају низови различите дужине, па се морају допунити до најдужег у серији, а допуњене позиције потом обележити тако да не учествују у рачунању грешке. За то обележје користи се вредност -100, коју библиотека PyTorch подразумевано узима као ознаку позиције која се прескаче.

Оно што се на тим позицијама искључује јесте функција грешке, па је овде треба и увести. Whisper преводи секвенцу у секвенцу. Декодер производи транскрипт токен по токен, при чему сваки следећи предвиђа на основу улазних атрибута и свих претходних токена. Током обучавања претходни токени не узимају се из излаза самог модела него из тачног транскрипта, поступком присилног вођења (енг. *teacher forcing*) [1], тако да грешка на једној позицији не квари све наредне.

Задатак се тиме своди на класификацију. На свакој позицији модел над целим речником даје расподелу вероватноће, а тачан одговор је један токен из референтног транскрипта. Мера одступања јесте унакрсна ентропија [1], која узима вероватноћу додељену тачном токену и кажњава модел тим више што је та вероватноћа мања:

$$L_{\mathrm{CE}} = -\frac{1}{N} \sum_{t} \log p(y_t \mid y_{<t}, x),$$

где је x аудио запис, y_t тачан токен на позицији t, $y_{<t}$ сви токени пре њега, p вероватноћа коју је модел доделио том токену, а N број важећих позиција у низу.

Управо тај збир објашњава зашто је обележавање нужно. Без њега би N обухватило и допуњене позиције, па би модел био награђиван и за то што на њима погађа саму ознаку допуне, а тај део задатка тривијалан је за учење, јер се иста ознака понавља до краја низа. Модел би тако грешку смањивао предвиђањем допуне уместо стварног текста, и то тим лакше што су дужине у серији неуједначеније. Обележавањем се грешка рачуна искључиво над стварним токенима транскрипта и остаје упоредива међу серијама без обзира на количину допуне у њима.

Учитавање записа додатно је заштићено хватањем изузетка. Корпус чини неколико десетина хиљада датотека прикупљених из различитих извора, а обучавање траје више сати непрекидно, па би једна оштећена датотека иначе прекинула цео поступак и поништила све до тада израчунато. Уместо тога, на место таквог записа улази тишина, серија остаје потпуна, а путања таквог записа се исписује, тако да се проблематични записи могу накнадно пронаћи и уклонити из корпуса.

6.1.3 Хиперпараметри обучавања

Вредности са којима је обучавање покренуто дате су у табели 6.1. Реч је о хиперпараметрима, дакле о величинама које се задају пре обучавања и не мењају се учењем, за разлику од параметара модела, односно тежина мреже. Све су биране под истим ограничењем, а то је количина доступне графичке меморије, уз услов да доучавање остане стабилно, односно да се модел прилагоди српском језику и домену, а да притом не заборави оно што већ зна.

Мала серија и акумулација градијента иду заједно. Серија од четири примера јесте оно што стаје у меморију графичке картице, али тако мала серија даје нестабилну процену градијента. Зато се градијенти сабирају кроз четири узастопна корака, па се тежине освежавају тек након тога. Ефекат је исти као да је серија имала шеснаест примера, а меморија остаје оптерећена као да их има четири.

Стопа учења од $1 \cdot 10^{-5}$ мања је за ред величине од оне коришћене у дестилацији. Разлог је природа задатка, овде се доучава унапред обучен

Табела 6.1: Хиперпараметри финог подешавања

Хиперпараметар	**Вредност**
Величина серије по уређају	4
Акумулација градијента	4 (ефективна серија 16)
Стопа учења	$1 \cdot 10^{-5}$
Кораци загревања	100
Број епоха	5
Прецизност	мешовита (fp16)
Оцењивање	на крају сваке епохе
Чување модела	на крају сваке епохе

вишејезични модел, па превелики кораци воде брзом заборављању општег знања које модел већ поседује.

Загревање кроз првих сто корака постепено подиже стопу учења од нуле до задате вредности. На самом почетку обучавања градијенти су још нестабилни, па би пуна стопа учења одмах одвела тежине предалеко од добро подешеног полазног стања.

Мешовита прецизност значи да се већина рачуна обавља у половичној прецизности, чиме се штеди меморија и добија на брзини, док се осетљиви делови, попут сабирања градијената, и даље рачунају у пуној прецизности.

6.1.4 Оцењивање

По завршетку обучавања модел се оцењује на тест скупу, који није коришћен ни у једној фази учења. Приликом транскрипције језик и задатак наводе се изричито, из истих разлога наведених у одељку 6.1.1, како модел не би сам погађао о ком је језику реч.

Излаз модела затим се пореди са тачним транскриптом. Обе мере коришћене у раду, стопа грешке на нивоу речи и проценат подударности, заједно са нормализацијом текста која им претходи, дефинисане су на једном месту, у одељку 7.1. Оба текста се најпре нормализују и растављају на листе речи, а затим се те две листе пореде поступком описаним у одељку 4.3.1. Разлика у односу на ранију примену јесте у томе што се овде пореде читаве речи, док су се тамо поредили појединачни знакови.

Сви модели у раду оцењени су истим поступком, над истим тест скупом и уз исту нормализацију, тако да су добијене вредности међусобно упоредиве.

Резултати су приказани и упоређени у поглављу 7.

6.2 Дестилација знања

Фино подешавање описано у претходном одељку ослања се на једини извор надзора који корпус нуди, а то је тачан транскрипт. Ознака моделу каже која је реч исправна и ништа више од тога. Свака друга реч подједнако је погрешна, и она која се од тачне разликује у једном гласу и она која са изговореним нема никакве везе.

Дестилација знања, чије су основе дате у одељку 2.3, полази од тога да велики модел ту разлику зна. Поред коначног одговора, он носи и податак о томе колико је сигуран у сваку од могућих алтернатива. Мали модел који учи само из тачне ознаке види једну исправну реч, мали модел који учи и из расподеле великог модела види уз то које су речи биле блиске исправној, а које потпуно промашене. Управо та додатна информација омогућава мањем моделу да достигне тачност коју сам, из истих података, не би могао.

Мањи модел, дакле, не учи више само од података него и од већег модела који је исте податке већ савладао. У даљем тексту већи модел зове се учитељ, а мањи ученик.

6.2.1 Избор учитеља и ученика

За улогу учитеља изабран је модел `whisper-large-v3` фино подешен на српском језику [16]. Реч је о моделу представљеном у одељку 2.4 као најбоље обучено јавно доступно решење за српски у тренутку израде рада. Довољно је тачан да буде поуздан извор знања, али и превелик за употребу на скромном хардверу, што је управо разлог зашто се његово знање преноси на мањи модел уместо да се он сам користи. За улогу ученика изабран је модел `whisper-tiny`, најмања варијанта архитектуре, са око 39 милиона параметара наспрам близу милијарду и по код учитеља.

Између те две варијанте постоји техничка препрека коју је требало решити пре почетка обучавања. Дестилација почива на поређењу расподела учитеља и ученика колону по колону, при чему колона са истим редним бројем код оба модела мора значити исти токен. Тај услов овде није испуњен сам од себе.

Модели од `whisper-tiny` до `whisper-large-v2` деле исти речник од 51.865 токена, док `whisper-large-v3` има 51.866, јер уводи додатни језички токен `<|yue|>`. Тај токен није додат на крај речника него у средину, међу остале језичке ознаке, па су сви специјални токени после њега померени за једно место. Одсецање последње колоне логита, које би било довољно код претходне верзије, овде би значило да ученик учи расподеле померене за једну позицију.

Због тога је направљено пресликавање колона по имену токена, а не по положају. За сваку колону речника ученика пронађена је колона у речнику учитеља која садржи исти токен, чиме је добијен индексни низ дужине 51.865. Колона која постоји само код учитеља прескаче се, а све остале враћају се на место које ученик очекује, па су расподеле поравнате упркос разлици у речницима.

6.2.2 Функција грешке

Ученик учи из два извора истовремено, па се и грешка састоји из два члана:

$$L = \alpha \cdot L_{\mathrm{CE}} + (1 - \alpha) \cdot T^2 \cdot L_{\mathrm{KL}}. \tag{6.1}$$

Први члан, L_{CE}, јесте уобичајена унакрсна ентропија између излаза ученика и тачне ознаке, иста она уведена у одељку 6.1.2. Он обезбеђује да модел научи тачан транскрипт.

Други члан, L_{KL}, мери колико се расподела ученика разликује од расподеле учитеља. За то се користи Кулбак-Лајблерова дивергенција над омекшаним расподелама:

$$L_{\mathrm{KL}} = \mathrm{KL}(\mathrm{softmax}(z_t/T) \parallel \mathrm{softmax}(z_s/T)),$$

где су z_t и z_s логити учитеља и ученика, а T температура. Дељење логита температуром омекшава расподелу, што је T веће, разлике међу вероватноћама су мање изражене, па до изражаја долазе и алтернативе које би иначе биле занемарљиве. Управо у њима лежи знање које се преноси, јер оне говоре о односима међу погрешним одговорима, а тачан одговор ученик има и без учитеља.

Фактор T^2 није произвољан. Омекшавање смањује градијенте дестилационог члана приближно фактором $1/T^2$, па се множењем са T^2 они враћају на исту скалу као градијенти унакрсне ентропије [2]. Без њега би удео дестилације у

укупној грешци зависио од изабране температуре, што није намера, јер тај однос треба да буде одређен искључиво параметром α.

Оба члана рачунају се само на важећим позицијама. Позиције допуне, обележене на начин описан у одељку 6.1.2, искључене су и из унакрсне ентропије и из дивергенције, чиме се спречава да модел учи из празних места у низу.

Параметар α дели надзор између два извора, при $\alpha = 1$ ученик учи само из тачне ознаке, што се своди на фино подешавање, а при $\alpha = 0$ само од учитеља. Он је, заједно са температуром T, једна од кључних величина овог рада, па је утицај оба параметра детаљно анализиран у поглављу 7.

6.2.3 Начин извођења дестилације

Дестилација се изводи над истим скупом за обучавање као и фино подешавање, дакле над 17.804 узорка добијена поступком описаним у поглављу 5, тако да се два приступа разликују искључиво по извору надзора, а не по материјалу који су видели. Ради се притом над нешто мањим подскупом, јер се изостављају записи дужи од 445 токена, као и они код којих се дужина низа токена учитеља и ученика не поклапа, што је последица разлике у речницима описане у одељку 6.2.1. Рунде поновног учења, о којима је реч у одељку 6.3, узорке бирају из ширег скупа. Он настаје спајањем скупа за обучавање (17.804 узорка) са допунским скупом издвојеним у одељку 4.2 (23.278 узорака). Тако настаје заједнички скуп од 41.082 узорка, који се у наставку рада тако и назива. Сви они су прошли исто чишћење описано у поглављима 4 и 5, а тест скуп и скуп непознатих говорника у њега не улазе.

Дестилација је изведена на начин који се назива обучавањем уз присутног учитеља. Оба модела учитавају се у меморију истовремено, ученик у режиму обучавања, а учитељ у режиму закључивања, у половичној прецизности и без рачунања градијената. Свака серија тако пролази кроз оба модела, па се расподела учитеља употреби одмах, у истом кораку, и нигде се не памти.

Цена тог решења јесте то што учитељ ради непрекидно, дакле у сваком кораку обучавања и у свакој евалуацији, а добит је то што ништа не мора да се чува на диску и што ниједна вредност не мора да се одбаци. Дивергенција се овде рачуна над целим речником од 51.865 токена, без икаквог одсецања расподеле, тако да ученик види и цео реп расподеле учитеља. Управо величина

тих тензора одређује колика серија стаје у меморију, па је она држана на осам примера, уз акумулацију градијента.

Расподела учитеља добија се поступком присилног вођења, њему се истовремено предају аудио запис и тачан транскрипт, па се за сваку позицију у низу очитава расподела коју би доделио следећем токену под условом да су сви претходни токени тачни. Тако добијена расподела поравната је са тачном ознаком, што је предуслов да ученик и учитељ на свакој позицији говоре о истом месту у низу.

Пошто локални хардвер није располагао довољном графичком меморијом за оба модела истовремено, обучавање је спроведено на изнајмљеној виртуелној машини са графичком картицом на платформи Amazon Web Services.

У рундама поновног учења исти поступак не би био економичан, јер се кроз исти материјал пролази у више циклуса, па се тамо прешло на унапред издвојене расподеле учитеља, о чему је реч у одељку 6.3.1.

6.2.4 Обучавање ученика

Ученик је обучаван уз хиперпараметре дате у табели 6.2.

Табела 6.2: Хиперпараметри обучавања ученика при дестилацији

Хиперпараметар	**Вредност**
Величина серије	8
Акумулација градијента	4 (ефективна серија 32)
Стопа учења	$1 \cdot 10^{-4}$
Кораци загревања	100
Број епоха	5
Прецизност	мешовита (fp16)
Температура T	4
Однос α	0,5

Прве четири ставке одговарају онима из одељка 6.1.3, уз то што је стопа учења овде за ред величине већа, из разлога који је тамо и наведен. Последње две немају одговарајући параметар у фином подешавању, T омекшава расподеле, а α дели надзор између тачне ознаке и учитеља.

Мала серија уз четвороструку акумулацију градијента последица је величине тензора над целим речником. У меморију стаје осам примера, али

се градијенти сабирају кроз четири узастопна корака, па се тежине освежавају као да је серија имала тридесет два примера. Из скупа за обучавање притом испада неколико десетина записа, оних дужих од 445 токена и оних код којих се дужина токенизације учитеља и ученика не поклапа, па један пролаз кроз податке чини 556 корака.

6.3 Итеративно поновно учење на грешкама

Оба претходна приступа третирају све податке подједнако. Модел прође кроз цео скуп за обучавање, при чему исто пажње добијају и узорци које већ препознаје без грешке и они на којима греши. Итеративни приступ полази од претпоставке да то није најбоља расподела труда, када се једном зна где модел греши, наредни циклус обучавања може се усмерити управо на те случајеве.

Идеја је спроведена кроз затворену петљу [4]. Модел се обучава, затим се мери где греши, на основу тих мерења саставља се нов скуп за обучавање, па се модел доучава, и поступак се понавља. Свака рунда наставља од модела претходне рунде, тако да се знање гомила уместо да се сваки пут почиње изнова. Изведено је пет таквих рунди.

6.3.1 Ток једне рунде

Поступак се састоји од корака који се међусобно ослањају, па је најпре дат његов ток у целини, док појединости следе у наредним одељцима. Свака рунда одвија се у четири корака:

1. Састављање скупа за обучавање у тој рунди, по правилу описаном у одељку 6.3.3, и припрема аудио записа са аугментацијом чија јачина зависи од тога колико је пута узорак већ виђен (одељак 6.3.4).

2. Припрема расподела учитеља за изабране узорке, из логита издвојених унапред за цео заједнички скуп. Овај корак стога узима датотеке које одговарају узорцима те рунде и од њих гради попис рунде.

3. Доучавање модела претходне рунде на том скупу, у трајању од две епохе, истом хибридном функцијом грешке и истим поступком као при дестилацији (одељци 6.2.2 и 6.2.4).

4. Мерење грешке на валидационом скупу и поновно оцењивање целог заједничког скупа, чиме се добија улаз за избор узорака у наредној рунди (одељак 6.3.2).

Другим речима, рунда не уводи нов начин обучавања. Кораци два и три преузети су из дестилације у целости, једино што се мења јесте који узорци у њих улазе и у ком облику. Мења се и начин на који се долази до расподела учитеља. При дестилацији је учитељ радио уживо, у сваком кораку, што је овде неисплативо, јер се кроз исти материјал пролази у пет циклуса, па би учитељ исти посао поновио пет пута. Због тога су његове расподеле издвојене унапред, једном за цео заједнички скуп, поступком присилног вођења над референтним транскриптима, и сачуване на диск.

Чување целог вектора логита било би непрактично. Речник има 51.865 токена, па би за скуп ове величине било потребно неколико стотина гигабајта простора. Зато се за сваку позицију чува само $K = 50$ највећих вредности заједно са њиховим индексима, чиме се потребан простор смањује око петсто пута, док одбачени реп расподеле носи занемарљив део укупне масе вероватноће. Вредности се уписују у половичној прецизности, у формату safetensors, једна датотека по узорку, именована према идентификатору аудио записа. Тако је за заједнички скуп од 41.082 узорка сачуван скуп логита за 40.981 узорак, укупне величине око 940 мегабајта.

При обучавању се расподела учитеља реконструише тако што се softmax рачуна над тих педесет вредности, тако да им збир поново буде један, док се логаритамске вероватноће ученика рачунају над целим речником, па се затим очитавају на истих педесет индекса. Дивергенција се тако сабира само преко оних токена о којима учитељ уопште има шта да каже. Логити се учитавају лењо, тек када узорак уђе у серију, а при формирању серије допуњавају се на исти начин као и тачне ознаке, тако да улазни атрибути, тачне ознаке и расподеле учитеља остају међусобно поравнати. Захваљујући томе што су логити издвојени једном унапред, припрема сваке наредне рунде своди се на проналажење већ постојеће датотеке по идентификатору узорка, а попис рунде за сваки узорак повезује путању до звука, транскрипт и путању до припадајућих логита учитеља.

Прва рунда полази од модела добијеног дестилацијом, а свака наредна од модела претходне рунде. Тај детаљ је суштински, јер када би свака рунда

почињала од неподученог модела, ниједно знање се не би преносило кроз циклусе и цео поступак изгубио би смисао.

Величина скупа за обучавање у једној рунди износила је 19.100 узорака, што уз ефективну серију од шеснаест примера даје 1.194 корака по епохи, односно 2.388 корака по рунди.

6.3.2 Дијагностика грешака

Да би избор узорака био смислен, није довољно знати колика је грешка него и какве је врсте, јер различите врсте грешака траже различит одговор. Због тога се хипотеза модела и референтни транскрипт поравнавају Левенштајновим растојањем на нивоу речи [17], уз чување целог низа операција, па се свака појединачна грешка сврстава у једну од категорија приказаних у табели 6.3.

Табела 6.3: Категорије грешака препознавања

Шта се догодило	**Категорија**
Замена, фонетски блиска	акустика
Замена, фонетски далека, реч постоји у корпусу	језички модел
Замена, фонетски далека, реч не постоји у корпусу	непозната реч
Уметање уз петљу или тишину	халуцинација
Остала уметања	уметање, остало
Брисање на крају исказа	прекид
Брисање у средини исказа	тих сегмент

Разликовање фонетски блиских од далеких замена изведено је преко нормализованог Левенштајновог растојања над низом гласова, са прагом 0,30. За српски језик низ гласова добија се готово непосредно из ортографије, потребно је само спојити дирафе љ, њ и џ у по један глас. Замена код које је растојање мање или једнако прагу говори да модел није добро чуо, док замена преко прага говори да је декодер, упркос ономе што је чуо, уписао вероватнију реч.

Пошто саме називе категорија није лако повезати са оним што се у излазу модела заиста види, у табели 6.4 дат је по један стваран пример за сваку од њих, издвојен из излаза модела на валидационом скупу, истом оном над којим је рађена дијагностика из табеле 7.3. Уз сваки пример наведена је рунда из које потиче. Искази су приказани после нормализације описане у одељку 7.1

и скраћени на околину грешке, а делови у којима се хипотеза разликује од референце стоје подебљано.

Табела 6.4: Примери грешака по категоријама, из излаза модела на валидационом скупу. Звездицом је означена рунда серије са исправљеним односом чланова функције грешке (одељак 7.5), остале рунде припадају основној серији.

Категорија	**Рунда**	**Референца**	**Хипотеза модела**
акустика	5	...baš u tom **trenutku** zateže se uže ...	...baš u tom **trinutku** zateže se uže ...
језички модел	5	...odneo ruže **dao** obećanje da će ...	...odneo ruže **da** obećanje da će ...
непозната реч	5	**evropski** parlament je izglasao rezoluciju ...	**eurske** parlament je izglasao rezoluciju ...
халуцинација	1	...kako treba da da da da da da se da se ponašam ...	...kako treba da da da da da da **da da da da da da da da da** se da se **ponašašam** ...
уметање, остало	5	jedina koja ima dve jedna nacija dve države ...	jedina koja ima dve **da** jedna nacija dve države ...
прекид	2*	...podrške studentima protesta svih vrsta koji su počeli **uh**	...podrške studentima protesta svih vrsta koji su počeli
тих сегмент	2*	...samo **ne** želim da izvučeš deblji kraj buraz ...	...samo želim da izvučeš deblji kraj buraz ...

Прва три реда показују шта чини разлику међу трима врстама замена. Код акустичке грешке нормализовано растојање између *trenutku* и *trinutku* износи 0,12, дакле испод прага, па се узима да модел није добро чуо вокал у средини речи. Код језичке грешке растојање између *dao* и *da* износи 0,33, тик изнад прага, а реч коју је модел уписао не само да постоји у корпусу него је и једна од најчешћих речи језика, па је реч о томе да је декодер, упркос чутом, изабрао вероватнију реч. Код непознате речи растојање између *evropski* и *eurske* износи 0,50, а уписани низ гласова не јавља се ни у једној од 112.231 различите речи из референци корпуса, дакле декодер је саставио низ који у језику уопште не постоји.

Ред са халуцинацијом одговара оном облику грешке који је у раду више пута наведен као најштетнији, а издвојени пример показује и одакле петља често креће. Говорник се у снимку заиста замуцкује, али модел то

понављање, уместо да га верно пренесе, настави далеко преко изговореног и притом изобличи и наредну реч. У изразитијим случајевима понављање се прекида тек на граници дозвољене дужине излаза, па хипотеза садржи више речи него референца, и управо ту настају вредности стопе грешке преко сто посто. Уметање се сврстава у ову категорију само ако је хипотеза у петљи или ако у снимку има много тишине, а свако друго уметање спада у општу категорију. Такав је пети ред, где је модел у говорникову паузу после речи *dve* уметнуо везник *da*, што је уједно и једина грешка у том исказу. О прерасподели између те две категорије кроз рунде реч је у одељку 7.4.

Последња два реда деле брисања према положају. Ако низ изостављених речи допире до краја исказа, модел је прекинуо препознавање пре краја снимка, у шестом реду тако је изостала последња изговорена реч, поштапалица *uh*. Ако се брисање налази између два тачно препозната дела, реч је о сегменту који модел није чуо, најчешће због нагле промене гласноће или преклапања говорника. Пример у последњем реду показује зашто такво брисање није безазлено, изостављена је одречна речца *ne*, чиме се смисао исказа окреће у супротан, иако стопа грешке тог узорка износи свега неколико процената.

Ова подела није сама себи сврха, него одређује шта има смисла предузети. Ако преовлађују акустичке грешке, помаже јача аугментација и разноврснији услови снимања. Ако преовлађују језичке грешке, потребна је текстуална разноврсност, док додавање нових аудио записа неће помоћи. Ако преовлађују халуцинације, треба пооштрити филтер над ознакама, а не додавати податке.

6.3.3 Састав скупа за обучавање у рунди

Скуп за обучавање у свакој рунди саставља се изнова из заједничког скупа од 41.082 узорка, по правилу датом у табели 6.5, и то посебно за сваког говорника.

Табела 6.5: Правило састављања скупа за обучавање у рунди, по говорнику

Говорник	**Шта улази у скуп**
Мање од 500 узорака	улазе сви
500 и више узорака	бира се 500 узорака, 70% оних које модел најбоље зна и 30% оних на којима греши

Оних седамдесет посто добро познатих узорака није ту због учења у ужем

смислу него ради очувања већ стеченог знања. Обучавање искључиво на грешкама водило би заборављању онога што модел већ уме.

Преосталих тридесет посто не бира се просто по највећој грешци, него пропорционално по категоријама из претходног одељка. Ако код неког говорника халуцинације чине половину свих грешака, отприлике половина изабраних лоших узорака биће узорци са халуцинацијама. Унутар сваке категорије узорци се рангирају по броју грешака те врсте, а при изједначеним вредностима по сигурности модела, код халуцинација и језичких грешака предност имају самоуверене грешке, док се код осталих категорија бирају случајеви у којима је модел био најнесигурнији.

Прва рунда је изузетак, јер за њу још нема измерених грешака, па се узорци за велике говорнике бирају насумично.

6.3.4 Аугментација према броју виђања

Пошто се сваке рунде бира из истог заједничког скупа, поједини узорци природно се понављају кроз више рунди. Да их модел не би једноставно запамтио, води се евиденција о томе колико је пута сваки узорак већ виђен. Ако тај број означимо са v, јачина изобличења j рачуна се као

$$j = \min\left(1,\ \frac{v}{3}\right),$$

а њене вредности и одговарајући ефекат дати су у табели 6.6.

Табела 6.6: Јачина аугментације према броју ранијих виђања узорка

Број виђања v	**Јачина** j	**Ефекат**
0	0,00	узорак улази нетакнут
1	0,33	блага измена
2	0,67	јача измена
3 и више	1,00	пуна измена

Сви параметри аугментације извлаче се насумично, па се потом множе јачином j. При $j = 0$ све измене саме од себе испадају у нулу, брзина остаје 1,0, одјек 0 секунди, а однос сигнала и шума 40 децибела, што је практично нечујно. Узорак који се појављује први пут тако улази нетакнут, а са сваким поновним појављивањем изобличење расте.

Измене се примењују на два нивоа. На таласном облику мења се брзина репродукције, додаје се одјек синтетичким импулсним одзивом и додаје бели

шум на контролисаном односу сигнала и шума. Те измене уписују се у нову аудио датотеку. На мел атрибутима примењује се савијање фреквенцијске осе (енг. *vocal tract length perturbation*, VTLP) [19], које помера форманте тако да исти снимак звучи као други говорник, и маскирање делова спектрограма по фреквенцији и времену [18].

Измене на мел нивоу не могу се уписати у аудио датотеку, а издвајање логита учитеља и обучавање ученика два су одвојена процеса. Када би сваки од њих сам позивао генератор случајних бројева, видели би различит улаз, па би дестилација учила шум. Због тога се спецификација измена, коефицијент савијања и положаји маски, уписује у попис рунде, који оба процеса потом примењују на исти начин. Генератор случајних бројева зависи искључиво од идентификатора узорка и редног броја рунде, чиме је поступак поновљив без обзира на редослед обраде.

Глава 7

Резултати и дискусија

У претходном поглављу описани су поступци којима су добијена три модела, фино подешен, дестиловани и итеративно дорађивани. Ово поглавље доноси њихове измерене резултате и тумачење добијених бројева.

Излагање прати редослед постављених питања. Најпре се дефинишу мере и услови под којима су сва мерења спроведена, како би вредности биле међусобно упоредиве. Затим се три приступа пореде над истим тест скупом, након чега се посебно разматрају рунде итеративног поступка и промена структуре грешака кроз њих. На крају се испитује осетљивост дестилације на однос чланова функције грешке и даје се преглед потребног времена, меморије и новца, чиме се постигнута тачност доводи у везу са ценом њеног постизања.

7.1 Мере и услови мерења

Сви резултати приказани у овом поглављу добијени су над издвојеним тест скупом од 2.239 узорака, који ниједан модел није видео ни у једној фази обучавања. Валидациони скуп од 2.219 узорака коришћен је током итеративног поступка, за праћење напретка између рунди. Сви модели мерени су истим поступком, над истим скупом и уз исту нормализацију, тако да су вредности међусобно упоредиве. Коришћене су две мере.

Прва је стопа грешке на нивоу речи (енг. *word error rate*, WER) [1], стандардна мера у препознавању говора. Она се рачуна тако што се хипотеза модела и референтни транскрипт поравнају на нивоу речи, а затим се

пребројe три врсте грешака:

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

где S означава број замењених речи, D број изостављених, I број сувишно уметнутих, а N укупан број речи у референтном транскрипту. Мања вредност значи бољи модел. Пошто бројилац може премашити N, стопа грешке није ограничена на сто посто.

Поред просека по узорку наведена је и збирна вредност, добијена дељењем укупног броја грешака у целом скупу укупним бројем речи у целом скупу. Збирна мера мање је осетљива на кратке исказе, код којих једна грешка може дати непропорционално висок однос.

Друга је проценат подударности са референтним транскриптом, добијен поступком за поређење низова над листама речи, чија је формула дата у одељку 4.3.1. Већа вредност значи бољи модел. Ова мера коришћена је у ранијим фазама рада, па је задржана ради поређивости са претходно пријављеним резултатима.

Пре рачунања обе мере текст се нормализује, своди се на мала слова, уклања се интерпункција и писмо се своди на латиницу, како разлике у писму и интерпункцији не би биле бројане као грешке у препознавању. Свођење писма није формалност, јер полазни модел 0,4% излаза исписује ћирилицом, без тог корака такве реченице биле би пребројане као потпуно погрешне иако су садржајно тачне.

Уз просечну и збирну вредност наведене су и медијана стопе грешке и број узорака чија стопа грешке прелази 100%. Те две величине описују различите делове исте расподеле. Медијана показује како модел ради на реченици просечног квалитета и не помера се због неколико изразито лоших примера. Број узорака преко 100% мери учесталост потпуних промашаја, јер стопа грешке може прећи сто посто само ако модел испише више погрешних речи него што их референтни транскрипт уопште има, што је обележје петљи и халуцинација.

7.2 Поређење три приступа

Табела 7.1 приказује резултате свих испитаних приступа над истим тест скупом.

Табела 7.1: Поређење испитаних приступа на тест скупу. Осим последњег реда, сви наведени модели варијанте су модела `whisper-tiny` са око 39 милиона параметара, док учитељ `whisper-large-v3` има око 1.550 милиона. Звездицом су означене рунде изведене са исправљеним односом чланова функције грешке, $\alpha = 0{,}9$ и $T = 1$, о чему је реч у одељку 7.5; серија са основном поставком дата је у табели 7.2.

Модел	**Подуд.**	**Просечан WER**	**Збирни WER**	**Медијана WER**	**>100%**
без обучавања	29,01%	95,25%	90,21%	77,8%	171
фино подешавање	76,14%	56,80%	39,62%	23,1%	85
дестилација	74,64%	39,80%	36,25%	26,7%	39
дестилација + 1 рунда*	74,77%	38,19%	35,81%	25,9%	47
дестилација + 2 рунде*	76,03%	35,71%	33,68%	24,4%	38
учитељ `whisper-large-v3`	81,40%	30,01%	26,11%	17,6%	22

Полазни модел `whisper-tiny`, без икаквог доучавања, постиже 29,01% подударности уз просечну стопу грешке од 95,25%. Другим речима, на реченици просечног квалитета погреши готово сваку реч. То потврђује почетну претпоставку рада да најмања варијанта архитектуре Whisper, без прилагођавања, није употребљива за српски језик.

Фино подешавање подиже подударност на 76,14%, а збирну стопу грешке спушта са 90,21% на 39,62%. Тај скок долази искључиво од прилагођавања модела језику и домену, без икаквог преноса знања из већег модела.

Поређење финог подешавања и дестилације на први поглед делује противречно. По подударности је фино подешен модел бољи, 76,14% наспрам 74,64%, али по просечној стопи грешке знатно лошији, 56,80% наспрам 39,80%. Одговор дају последње две колоне. Медијана стопе грешке износи 23,1% код финог подешавања и 26,7% код дестилације, што значи да фино подешен модел на реченици просечног квалитета заиста ради нешто боље. Али број узорака чија стопа грешке прелази 100% пада са 85 на 39, дакле више него преполовљен.

Из тога следи да дестилација овде не побољшава реченицу просечног квалитета, него уклања катастрофалне излазе. Стопа грешке преко сто посто настаје када модел испише више погрешних речи него што их референца садржи, а то се дешава када упадне у петљу или почне да измишља текст. Расподела поверења учитеља очигледно спречава ученика да оде у такав излаз, док фино подешен модел, вођен искључиво тачном ознаком, нема шта

да га заустави. Тај налаз се непосредно поклапа са мерењем у одељку 7.4, где су халуцинације категорија грешака која кроз рунде најдоследније опада.

Разлика између те две мере долази од тога што оне различито третирају реп расподеле. Подударност је ограничена одоздо нулом, па најгори могући узорак доприноси исто колико и било који други потпуно погрешан. Стопа грешке нема горњу границу, па једна реченица у петљи може сама подићи просек. Управо зато су обе наведене, подударност говори колико је добијени текст у целини близу референце, а стопа грешке колико је поступак препознавања поуздан.

Ред са две рунде односи се на модел добијен поновним учењем на грешкама, уз исправно одмерене чланове функције грешке. Он спушта просечну стопу грешке на 35,71% и збирну на 33,68%, уз подударност од 76,03%, чиме по свим мерама истовремено подноси поређење са финим подешавањем. О томе је реч у одељку 7.5.

Најважније поређење је оно између дестилованог ученика и његовог учитеља. Учитељ постиже 81,40% подударности, ученик 74,64%, што значи да ученик задржава преко деведесет одсто учитељеве тачности уз око четрдесет пута мање параметара. Управо у том односу је практична вредност дестилације, заостатак од око седам процентних поена плаћен је смањењем модела за четрдесет пута.

Треба нагласити и да је вредност од 74,64%, добијена мерним поступком развијеним за овај рад, готово идентична раније пријављеној вредности од 74,5%, добијеној независним мерењем. То поклапање служи као провера исправности мерног поступка.

7.3 Резултати итеративног приступа

Итеративни поступак спроведен је у пет рунди, при чему свака рунда полази од модела претходне. Резултати на тест скупу приказани су у табели 7.2, уз вредности добијене над валидационим скупом између рунди.

Унутар серије напредак је доследан, валидациона грешка опада у свакој од пет рунди, укупно за 4,42 процентна поена, а збирна стопа грешке на тест скупу за 7,60 поена. Добитак се притом смањује из рунде у рунду, 2,29, затим 0,57, 0,72 и 0,84 поена на валидацији, што говори да поступак постепено

Табела 7.2: Резултати итеративног поступка по рундама

Рунда	Вал. WER	Тест WER	Збирни WER	Подударност
1	63,62%	72,99%	65,94%	52,87%
2	61,33%	67,87%	62,26%	54,54%
3	60,76%	60,88%	60,82%	55,04%
4	60,04%	59,05%	59,86%	54,95%
5	59,20%	59,52%	58,34%	55,38%

исцрпљује оно што се овим избором узорака може добити и оправдава заустављање на петој рунди.

Међутим, сви модели ове серије слабији су од дестилованог модела од којег је серија пошла. Прва рунда је, уместо да га побољша, подигла просечну стопу грешке са 39,80% на 72,99%. Наредне рунде тај заостатак постепено надокнађују, али га не поништавају. Другим речима, серија показује да петља ради, али да ради на штету коју је сама нанела у првом кораку. Узрок је утврђен анализом која следи у одељку 7.5, где је иста серија поновљена са измењеним односом чланова функције грешке.

7.4 Промена структуре грешака кроз рунде

Пошто избор узорака почива на класификацији грешака, значајније од укупне грешке јесте питање да ли се смањују баш оне категорије на које је поступак усмерен. Табела 7.3 приказује број грешака по категоријама на валидационом скупу, у првој и петој рунди.

Табела 7.3: Број грешака по категоријама у првој и петој рунди

Категорија	Рунда 1	Рунда 5	Промена
језички модел	12.441	11.829	−4,9%
непозната реч	7.866	6.930	−11,9%
халуцинација	7.768	4.687	−39,7%
акустика	6.851	7.016	+2,4%
тих сегмент	4.290	3.725	−13,2%
уметање, остало	2.907	3.626	+24,7%
прекид	2.048	1.863	−9,0%
укупно	44.171	39.676	−10,2%

Најизраженији помак остварен је код халуцинација, чији је број готово преполовљен. То је очекивано и представља потврду да избор делује циљано, јер су халуцинације категорија коју поступак најдоследније препознаје и најчешће бира међу тридесет посто лоших узорака.

Категорија акустике практично се није променила. И то је очекивано, јер су акустичке грешке последица квалитета самог снимка, а не тога који су узорци изабрани за обучавање. Њих не решава избор узорака него другачији услови снимања или јача аугментација.

Пораст категорије уметања за 24,7% на први поглед делује као погоршање, али је реч о прерасподели унутар исте појаве. Уметање се сврстава у халуцинацију само ако је хипотеза у петљи или ако у снимку има много тишине, како петље нестају, иста уметања прелазе у општу категорију. Укупан број уметања свих врста износи 10.675 у првој и 8.313 у петој рунди, што је пад од 22,1%.

7.5 Осетљивост на однос чланова функције грешке

Пошто ниједна рунда није надмашила полазни дестиловани модел, испитан је утицај параметара хибридне функције грешке [2], дате изразом (6.1). Са вредностима коришћеним у основној поставци, $\alpha = 0{,}5$ и $T = 4$, ефективни однос чланова износи:

$$0{,}5 \cdot L_{\mathrm{CE}} + 0{,}5 \cdot 16 \cdot L_{\mathrm{KL}} = 0{,}5 \cdot L_{\mathrm{CE}} + 8 \cdot L_{\mathrm{KL}}.$$

Дестилациони члан тако носи шеснаест пута већу тежину од члана који везује модел за референтне ознаке. Мерење над сачуваним расподелама учитеља показало је да се најизвеснији токен учитеља поклапа са референтном ознаком у свега 8,6% случајева. Комбинација та два налаза објашњава понашање серије, ученик се, под доминантним дестилационим чланом, удаљава од референтних транскрипата и приближава расподели учитеља која на нивоу појединачних токена одступа од тачне ознаке.

Да би се та претпоставка проверила, поновљена је прва рунда са истим полазним моделом, истим узорцима и истим расподелама учитеља, уз измењена само два параметра, $\alpha = 0{,}9$ и $T = 1$, чиме однос постаје $0{,}9 \cdot L_{\mathrm{CE}} + 0{,}1 \cdot L_{\mathrm{KL}}$. Резултати су дати у табели 7.4.

Табела 7.4: Утицај односа чланова функције грешке

Поставка	**Вал. WER**	**Тест WER**	**Подударност**
$\alpha = 0{,}5$, $T = 4$ (основна)	63,62%	72,99%	52,87%
$\alpha = 0{,}9$, $T = 1$ (измењена)	32,52%	38,19%	74,77%
полазни дестиловани модел	—	39,80%	74,64%

Разлика је драстична, јер на потпуно истим подацима и истом полазном моделу, промена два броја у функцији грешке мења стопу грешке на тест скупу са 72,99% на 38,19%. Уз измењену поставку рунда први пут надмашује полазни модел, за 1,61 процентни поен, чиме је потврђено да проблем није био у избору узорака него у односу чланова функције грешке.

Пошто је прва рунда са измењеном поставком дала побољшање, поступак је настављен и другом рундом, која полази од модела прве. Треба нагласити једно ограничење, избор узорака за другу рунду наслеђен је из основне серије, дакле направљен је на основу грешака модела обученог са $\alpha = 0{,}5$ и $T = 4$. Пуна итерација захтевала би поновно оцењивање целог заједничког скупа моделом прве рунде, што траје око два сата, па је овде реч о другој рунди над раније изабраним скупом, а не о потпуно самосталном циклусу. Резултати обе рунде дати су у табели 7.5.

Табела 7.5: Резултати рунди са исправљеним односом чланова функције грешке

Модел	**Подударност**	**Просечан WER**	**Збирни WER**
полазни дестиловани модел	74,64%	39,80%	36,25%
+ рунда 1 ($\alpha = 0{,}9$, $T = 1$)	74,77%	38,19%	35,81%
+ рунда 2 ($\alpha = 0{,}9$, $T = 1$)	76,03%	35,71%	33,68%

У односу на полазну тачку, две рунде доносе пад просечне стопе грешке од 4,09 процентних поена и збирне од 2,57 поена, уз пораст подударности од 1,39 поена. Важније од самих вредности јесте то што је друга рунда донела више од прве, 2,48 наспрам 1,61 процентног поена стопе грешке. Поступак се, дакле, не засићује, што је супротно понашању основне серије из одељка 7.3, где је добитак опадао из рунде у рунду. Разлика је још јаснија зато што су обе серије користиле исте узорке и исте расподеле учитеља, а разликовале се само у односу чланова функције грешке.

Валидациона стопа грешке притом је између те две рунде остала практично непромењена, 32,52% наспрам 32,73%, иако је тест скуп показао јасан напредак. Разлике те величине на скупу од 2.219 узорака налазе се у границама мерног шума, па валидациона вредност овде није довољно осетљива да прати напредак између узастопних рунди.

Тиме поступак први пут подноси поређење са финим подешавањем по свим мерама истовремено. Модел добијен са две рунде постиже 76,03% подударности наспрам 76,14% колико даје фино подешавање, дакле практично исту вредност, али уз просечну стопу грешке од 35,71% наспрам 56,80%, што је разлика од преко двадесет процентних поена. Другим речима, уз исправно одмерене чланове функције грешке дестилација праћена поновним учењем на грешкама даје модел који је по подударности изједначен са финим подешавањем, а по поузданости препознавања знатно испред њега.

Испитане су само две рунде са измењеном поставком, па остаје отворено докле би напредак ишао, као и колико би поступак добио ако би избор узорака у свакој рунди био направљен на основу грешака стварно претходног модела.

7.6 Захтеви над рачунарским ресурсима

Поређење приступа не би било потпуно без података о ресурсима које они захтевају. Ученик има око 39 милиона параметара наспрам близу милијарду и по код учитеља, што је однос од око четрдесет пута, и та се разлика непосредно види у захтевима обучавања. Дестилација, која подразумева покретање учитеља, изведена је на изнајмљеној инстанци g4dn.xlarge у региону eu-north-1, са графичком картицом Tesla T4 од 15 гигабајта и диском од 77 гигабајта, јер модел те величине није могао да стане на локални хардвер. Фино подешавање, које учитеља не захтева, изведено је локално. При дестилацији су оба модела истовремено у меморији, а највише простора заузимају тензори расподела над целим речником, због чега је серија држана на осам примера. Цео поступак је, уз мешовиту прецизност, стао у меморију наведене картице.

Цена коју дестилација плаћа јесте то што учитељ ради у сваком кораку, дакле кроз пет епоха приближно деведесет хиљада пута над скупом за обучавање, уз још пет пролаза кроз валидациони скуп. Управо зато је у рундама прећено на унапред издвојене расподеле. Оне су добијене једним

пролазом учитеља кроз цео заједнички скуп, па се припрема сваке рунде своди на очитавање већ сачуваних вредности и траје неколико секунди. Оно што у сваком циклусу заиста захтева време јесте поновно оцењивање целог заједничког скупа, неопходно за избор узорака наредне рунде, које траје око два сата, што је и разлог зашто је друга рунда у одељку 7.5 изведена над раније изабраним скупом. Обим самог обучавања износи 5.565 корака за фино подешавање, дакле пет епоха над 17.804 узорка при ефективној серији од шеснаест примера, затим 2.780 корака за дестилацију, пет епоха над истим тим скупом при ефективној серији од тридесет два примера, и 2.388 корака по итеративној рунди, две епохе над 19.100 узорака. Иако се серије разликују, укупан број виђених примера готово је изједначен, око 89.000 виђених примера у оба случаја, па разлика у резултату не потиче од различитог обима обучавања.

Захтеви над диском сведени су чувањем само педесет највећих вредности по позицији. Цео вектор логита за корпус ове величине заузео би неколико стотина гигабајта, а овако издвојене расподеле за 40.981 узорак стају у око 940 мегабајта, што је уштеда од приближно петсто пута. Без ње поступак не би био изводљив на хардверу којим се располагало.

У примени, где се модел више не обучава него само користи, однос трошкова се потпуно мења. Дестиловани ученик транскрибује око осамнаест узорака у секунди, при заузећу меморије које дозвољава рад и на скромној графичкој картици, док би учитељ за исти посао захтевао вишеструко више и меморије и времена. Укупан трошак закупа рачунарских ресурса, укључујући припрему корпуса, издвајање расподела учитеља и све фазе обучавања, био је реда величине 60 долара.

Дестилација јесте скупља од финог подешавања, али је тај додатни трошак једнократан, чим су расподеле учитеља издвојене и сачуване, учитељ више није потребан, па свака наредна рунда кошта тачно онолико колико и обично доучавање малог модела. Управо то чини итеративни поступак изводљивим и ван истраживачких услова.

Глава 8

Закључак

У овом раду испитано је колико се најмања варијанта архитектуре Whisper може приближити знатно већим системима за препознавање говора на српском језику и које технике обучавања при томе дају највећи допринос.

Рад је обухватио два подједнако значајна дела. Први је припрема корпуса, где је од око 200 сати јавно доступног аудио материјала, кроз сегментацију дугих записа, груписање по боји гласа, стратификовану поделу, присилно поравнање и проверу конзистентности транскрипата, добијен коначни заједнички скуп од 41.082 узорка. Тај део није помоћни, јер је сваки покушај обучавања над непречишћеним корпусом показао да модел брже научи шум него језик. Други део чине три приступа обучавању, фино подешавање, дестилација знања и итеративно поновно учење на грешкама, примењена над истим подацима, са истом архитектуром ученика и мерена истим поступком над истим тест скупом.

8.1 Провера хипотезе

Радна хипотеза, постављена у поглављу 1, тврди да ће дестилација знања, вођена циљаном анализом грешака, омогућити малом моделу знатно поузданије препознавање говора на српском језику у односу на исти модел обучен стандардним финим подешавањем над истом количином података. Поузданост се мери стопом грешке на нивоу речи, добитак се не очекује од побољшања реченице просечног квалитета него од уклањања потпуно промашених излаза, док подударност са референтним транскриптом остаје на нивоу који се постиже финим подешавањем.

Хипотеза је потврђена у свему што се тиче односа дестилације и финог подешавања. Стопа грешке на нивоу речи пала је са 56,80% на 39,80%. Медијана се притом није побољшала него благо погоршала, али је број узорака чија стопа грешке прелази 100% пао са 85 на 39, што значи да добитак долази одатле где је и предвиђен, из уклањања петљи и халуцинација, а не из реченице просечног квалитета. Подударност са референтним транскриптом остала је практично непромењена, 74,64% наспрам 76,14% код финог подешавања. Полазни, необучени модел `whisper-tiny` постизао је свега 29,01% подударности, што показује да без прилагођавања језику ова величина модела није употребљива.

Уз саму хипотезу, мерење је дало и налаз о односу ученика према учитељу, ученик задржава преко деведесет одсто тачности учитеља уз око четрдесет пута мање параметара, чиме заостатак од око седам процентних поена постаје разумна цена четрдесетоструког смањења модела.

Део хипотезе који се односи на вођење поступка циљаном анализом грешака потврђен је тек након исправке односа чланова функције грешке. Основна серија показала је доследан пад грешке из рунде у рунду, али ниједна рунда није надмашила модел од којег је пошла, и то не због избора узорака него због тога што је дестилациони члан вишеструко надјачавао члан који модел везује за референтне транскрипте. Са исправљеним односом, $\alpha = 0{,}9$ и $T = 1$, две рунде дају најнижу измерену стопу грешке међу свим испитаним моделима, при чему друга рунда доноси већи добитак од прве, што говори да се поступак не засићује.

8.2 Доприноси рада

Методолошки допринос представља поступак класификације грешака препознавања у седам категорија, заснован на поравнању хипотезе и референце и на нормализованом растојању над низом гласова. За српски језик низ гласова добија се готово непосредно из ортографије, што класификацију чини једноставном и поновљивом, без потребе за фонетским речником. Њена практична вредност потврђена је тиме што су категорије на које је избор узорака био усмерен заиста опале, док је категорија акустичких грешака, на коју избор не може да утиче, остала непромењена.

Практични допринос чине четири решења без којих поступак не даје

исправне резултате. Прво је пресликавање колона логита по имену токена уместо по положају, јер се речници учитеља и ученика разликују у једном уметнутом токену. Друго је чување педесет највећих вредности по позицији при издвајању расподела за рунде, чиме је потребан простор сведен са неколико стотина гигабајта на око 940 мегабајта. Треће је уписивање спецификације аугментације у попис рунде, како би издвајање расподела учитеља и обучавање ученика видели идентичан улаз. Четврто је полазак сваке рунде од модела претходне, без чега ниједно знање не прелази између циклуса.

Као засебан налаз издваја се осетљивост дестилације на однос чланова функције грешке. Фактор T^2, уведен да изједначи скалу градијената, при вишим температурама доводи до тога да дестилациони члан вишеструко надјача унакрсну ентропију. Хинтон и сарадници наводе га као техничку исправку [2], а овде је показано да у пракси може одлучити о томе да ли ће поступак побољшати или покварити модел.

Сва три модела добијена у овом раду објављена су на платформи Hugging Face, фино подешен модел[1], дестиловани модел[2] и модел добијен итеративним поновним учењем на грешкама[3]. Целокупан програмски код коришћен у изради рада, од припреме корпуса и издвајања логита учитеља до обучавања и оцењивања, доступан је у јавном спремишту[4]. Сам корпус, из разлога наведених у поглављу 3, није јавно доступан.

8.3 Ограничења

Сва мерења изведена су над једним корпусом српског језика прикупљеним из јавно доступних извора, углавном студијских и полустудијских снимака, па се закључци не могу без провере пренети на друге домене, пре свега на снимке лошијег акустичког квалитета или на спонтани говор у отежаним условима.

Ниједан узорак није проверен људском транскрипцијом. Референтне ознаке су излаз мултимодалног модела, а филтрирање се ослања на слагање тог излаза са излазом модела Whisper, па се систематске грешке заједничке обома, на пример доследно погрешно препознат термин или нагласак, овим

[1] https://huggingface.co/BogdanTomic/whisper-tiny-sr-finetuned
[2] https://huggingface.co/BogdanTomic/whisper-tiny-sr-distilled
[3] https://huggingface.co/BogdanTomic/whisper-tiny-sr-distilled-iterative
[4] https://github.com/BogBogdan/master-destilacija-whisper

поступком не могу открити. Све пријављене вредности стога мере одступање од аутоматски добијене референце, а не од људски провереног транскрипта. Треба напоменути и да рунде поновног учења узорке бирају из ширег заједничког скупа, за разлику од финог подешавања и дестилације. Свака рунда притом ради са скупом упоредиве величине, 19.100 наспрам 17.804 узорка, па разлика није у количини материјала у једном пролазу него у томе што се кроз рунде види и грађа коју остала два приступа нису имала. Допринос самог избора узорака стога се не може у потпуности раздвојити од доприноса те шире основе.

Дијагностика грешака је у сваком циклусу и мерни инструмент и основа за избор узорака, што значи да поступак не може да буде бољи од ње. Ако категоризација погреши, погрешиће и наредни скуп за обучавање, а грешка се не поништава него преноси.

На крају, обе мере рачунају се над текстом сведеним на мала слова и без интерпункције. Тако се мери оно што модел чује, али не и оно што корисник добија, јер у практичној употреби управо интерпункција и велика слова чине разлику између сировог и употребљивог транскрипта.

8.4 Правци даљих истраживања

Најнепосреднији наставак јесте понављање пуне серије рунди са исправно одмереним члановима функције грешке, уз поновно оцењивање целог заједничког скупа у свакој рунди, чиме би се проверило да ли се раст добитка примећен у првим рундама одржава кроз цео низ. Друга допуна јесте систематско испитивање равни α и T, јер су овде упоређене само две њене тачке. Трећи правац тиче се природе надзора, показано је да расподела учитеља, добијена присилним вођењем над референтним транскриптом, на нивоу појединачних токена често одступа од саме тачне ознаке, па је усклађивање учитеља и ознака пре издвајања расподела вероватно најзначајнији појединачни корак ка бољим резултатима дестилације. Четврти правац јесте употреба до сада нетакнутог скупа непознатих говорника, којим би се проверило колико добро добијени модели раде на гласовима које током обучавања нису чули. Најзад, класификација грешака могла би се проширити ослањањем на велики језички модел у улози оцењивача.

Литература

[1] Jurafsky, D., Martin, J. H. Speech and Language Processing (3. издање, радна верзија). Stanford University. https://web.stanford.edu/~jurafsky/slp3/

[2] Hinton, G., Vinyals, O., Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531. https://arxiv.org/abs/1503.02531

[3] Howard, J., Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Proceedings of ACL 2018. arXiv:1801.06146. https://arxiv.org/abs/1801.06146

[4] Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., Anandkumar, A. (2018). Born-Again Neural Networks. Proceedings of ICML 2018, PMLR 80:1607--1616. arXiv:1805.04770. https://arxiv.org/abs/1805.04770

[5] Gou, J., Yu, B., Maybank, S. J., Tao, D. (2021). Knowledge Distillation: A Survey. International Journal of Computer Vision, 129(6), 1789--1819. arXiv:2006.05525. https://arxiv.org/abs/2006.05525

[6] Graves, A., Jaitly, N. (2014). Towards End-to-End Speech Recognition with Recurrent Neural Networks. Proceedings of ICML 2014, PMLR 32(2):1764--1772. https://proceedings.mlr.press/v32/graves14.html

[7] Stevens, S. S., Volkmann, J., Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. Journal of the Acoustical Society of America, 8(3), 185--190. https://doi.org/10.1121/1.1915893

[8] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. Proceedings of ICML 2023, PMLR 202:28492--28518. arXiv:2212.04356. https://arxiv.org/abs/2212.04356

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems 30 (NeurIPS 2017). arXiv:1706.03762. https://arxiv.org/abs/1706.03762

[10] Valin, J.-M., Vos, K., Terriberry, T. B. (2012). Definition of the Opus Audio Codec. RFC 6716, Internet Engineering Task Force (IETF). https://www.rfc-editor.org/rfc/rfc6716

[11] Oliver, B. M., Pierce, J. R., Shannon, C. E. (1948). The Philosophy of PCM. Proceedings of the IRE, 36(11), 1324--1331. https://doi.org/10.1109/JRPROC.1948.231941

[12] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., и др. (2021). SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624. https://arxiv.org/abs/2106.04624

[13] Desplanques, B., Thienpondt, J., Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Proceedings of Interspeech 2020, 3830--3834. https://doi.org/10.21437/Interspeech.2020-2650

[14] Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques (3. издање). Morgan Kaufmann. (поглавље о мерама сличности и кластеровању еуклидска/L2 удаљеност)

[15] Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Transactions on Information Theory, 13(2), 260--269. https://doi.org/10.1109/TIT.1967.1054010

[16] Sagicc (2024). whisper-large-v3-sr-combined: Whisper large-v3 фино подешен на српском језику над скуповима Common Voice 13 и Google FLEURS. Hugging Face. https://huggingface.co/Sagicc/whisper-large-v3-sr-combined

[17] Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, 10(8), 707--710. https://ui.adsabs.harvard.edu/abs/1966SPhD...10..707L

[18] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Proceedings of Interspeech 2019, 2613--2617. arXiv:1904.08779. https://arxiv.org/abs/1904.08779

[19] Jaitly, N., Hinton, G. E. (2013). Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition. Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing. https://www.cs.toronto.edu/~hinton/absps/perturb.pdf

[20] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. Proceedings of LREC 2020, 4218--4222. arXiv:1912.06670. https://arxiv.org/abs/1912.06670

[21] Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., Bapna, A. (2023). FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. Proceedings of IEEE SLT 2022. arXiv:2205.12446. https://arxiv.org/abs/2205.12446

[22] Rupnik, P., Ljubešić, N. (2022). ASR training dataset for Serbian JuzneVesti-SR v1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1679

[23] Baevski, A., Zhou, H., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems 33. arXiv:2006.11477. https://arxiv.org/abs/2006.11477

Биографија аутора

Богдан Томић рођен је 21. априла 2001. године у Београду. Основне академске студије завршио је на Рачунарском факултету 19. септембра 2025. године, након чега је уписао мастер академске студије Информатике на Математичком факултету Универзитета у Београду. Током академског образовања интересовао се за области машинског учења, вештачке интелигенције, обраде природног језика и развоја софтверских система, у којима је стицао искуство кроз различите академске, истраживачке и практичне пројекте.