Универзитет у Београду
Математички факултет

Стефан Јевтић

АУТОМАТСКА ТРАНСКРИПЦИЈА ГОВОРА ЗА СРПСКИ ЈЕЗИК

мастер рад

Београд, 2026.

Ментор:

проф. др Младен Николић
Математички факултет
Универзитет у Београду

Чланови комисије:

проф. др Јелена Граовац
Математички факултет
Универзитет у Београду

проф. др Јована Ковачевић
Математички факултет
Универзитет у Београду

Датум одбране:

Наслов мастер рада: Аутоматска транскрипција говора за српски језик

Резиме: Обрада природног језика је област вештачке интелигенције која се бави интеракцијом између рачунара и природног језика. Један од кључних задатака је аутоматска транскрипција говора, која подразумева претварање говора у текст. Ова технологија има широк спектар примена, укључујући претраживање садржаја, генерисање титлова за особе са оштећењем слуха, анализу говора у правним и медицинским контекстима, као и унапређење корисничког искуства у разним апликацијама.

Предности аутоматске транскрипције говора су бројне. Она омогућава брзу и ефикасну обраду великих количина аудио садржаја, смањујући потребу за ручном транскрипцијом која је временски и ресурсно захтевна. Међутим, један од главних изазова је прецизно препознавање говора у условима позадинске буке или нејасног изговора. Такође, модели могу имати потешкоћа у обради специфичних термина, акцената или језичких нијанси, што може довести до грешака у транскрипцији.

Један од најновијих и широко коришћених модела за транскрипцију говора је Whisper који развија OpenAI. Циљ рада је проналазак оптималних вредности хиперпараметара током тренирања одређених инстанци модела, како би се постигао што ефикаснији резултат.

Кључне речи: Обрада природног језика, аутоматска транскрипција говора, OpenAI Whisper, WER, CER, семантичка анализа.

Садржај

Глава 1

Увод

Аутоматска транскрипција говора представља једну од кључних технологија у оквиру обраде природног језика (*енг. Natural Language Processing – NLP*). Са све већом доступношћу аудио и видео садржаја, потреба за прецизном и ефикасном транскрипцијом постаје све израженија. Ова технологија омогућава конверзију говора у текст, што отвара бројне могућности примене, укључујући претраживање садржаја, генерисање титлова за особе са оштећењем слуха, као и анализу говора у специфичним правним и медицинским контекстима.

Истраживање у овој области последњих година доживљава напредак без преседана, првенствено захваљујући развоју дубоког учења и примени напредних архитектура неуронских мрежа. Један од најуспешнијих савремених модела је *OpenAI Whisper*, који комбинује иновативне технике машинског учења са обимним скуповима података за обуку, постижући високу робусност у препознавању говора на различитим језицима и у сложеним акустичким условима.

Циљ овог рада је детаљно проучавање и евалуација перформанси модела *Whisper* са фокусом на српски језик. Посебна пажња посвећена је поређењу различитих верзија, конкретно *whisper-large-v2* и *whisper-large-v3*, као и оптимизацији њихових хиперпараметара, пре свега брзине учења (*енг. learning rate*) и L2 регуларизације (*енг. weight decay*). Ови параметри имају пресудну улогу у процесу конвергенције модела, те њихово прецизно прилагођавање директно одређује финалну тачност транскрипције.

Експериментални део истраживања заснован је на примени модела на три референтна скупа података који покривају различите домене говора: *classla/ParlaSpeech-SR*, који садржи транскрипте парламентарних дебата, *google/fleurs* као стандард за евалуацију на више језика, те *mozilla-foundation/common_voice_17_0*, који представља разнолик скуп волонтерских аудио записа. Емпиријска евалуација на овим корпусима открила је специфичне разлике у перформансама; док је модел *whisper-large-v3* показао супериорност у изворном режиму рада, процес финог подешавања на специфичним подацима омогућио је верзији *whisper-large-v2* остваривање значајно нижих вредности стопе грешке *WER* и *CER*. Постигнути резултати потврђују да хибридни приступ, који комбинује напредне архитектуре са пажљивом оптимизацијом хиперпараметара и статистичком провером правописа, доводи до знатно прецизније транскрипције на српском језику.

У наставку рада, у другом поглављу, биће детаљно анализирана архитектура и принципи рада Whisper модела, уз осврт на његове предности у односу на традиционалне методе. Треће поглавље је посвећено методологији евалуације, коришћеним скуповима података и метрикама за оцену перформанси. Четврто поглавље презентује резултате експеримената и анализу утицаја хиперпараметара на успешност модела. Коначно, у закључку су сублимирани главни налази истраживања и предложени потенцијални правци за даљи развој у овој научној области.

Глава 2

Модели за транскрипцију и технике тренинга

Kaldi, *Wav2Vec* и *Whisper* представљају три значајна технолошка приступа аутоматској транскрипцији говора [1], који се фундаментално разликују по архитектури, алгоритмима и методологији обраде сигнала.

Kaldi је алат отвореног кода заснован на традиционалним методама. Он се ослања на скривене Марковљеве моделе (*енг. hidden Markov models*) за моделовање временских стања система, модел мешавине Гаусових расподела (*енг. Gaussian mixture models*) за акустичку класификацију, као и на екстракцију Мел-фреквенцијских кепстралних коефицијената (*енг. mel-frequency cepstral coefficients*) за дигиталну репрезентацију аудио сигнала. Иако је познат по својој изузетној флексибилности, Kaldi захтева ригорозно ручно подешавање и висок степен експертског знања како би се постигли оптимални резултати у специфичним сценаријима.

Wav2Vec представља модернију парадигму засновану на самонадгледаном учењу (*енг. self-supervised learning*) [2]. Модел користи контрастне задатке за обуку на обимним скуповима неозначених података, што му омогућава учење ефикасних репрезентација говора без потребе за мануелном транскрипцијом сваког примера. Архитектура Wav2Vec модела комбинује конволутивне неуронске мреже (*енг. convolutional neural networks*) и трансформере за симултану обраду временских и фреквенцијских

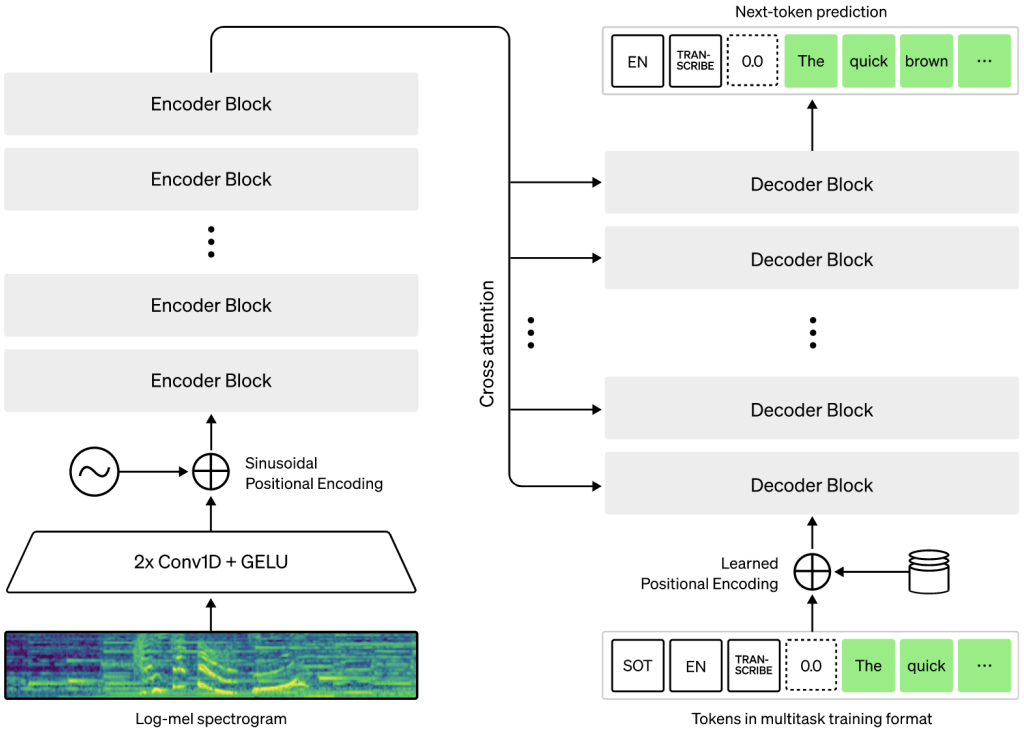
карактеристика. Иако показује изврсне резултате на малим тренинг скуповима, његова робустност може бити упитна у сложеним акустичким условима или при сусрету са терминологијом која није била заступљена током фазе обуке.

Whisper, развијен од стране компаније OpenAI, представља актуелни врхунац у домену робусне транскрипције. Налик на Wav2Vec, али уз знатно већи капацитет података, Whisper користи енкодер-декодер архитектуру засновану на трансформерима са механизмима пажње (*енг. attention mechanisms*) за ефикасно моделовање дугачких секвенци [3, 4]. Ова архитектура, обучена на екстремно разноврсној бази података која укључује бројне језике и акценте, омогућава моделу да фокусира капацитет на најрелевантније делове улазног сигнала.

Кључна предност модела Whisper лежи у његовој способности да се адаптира на различите акустичке услове и језичке структуре без додатних мануелних интервенција, што га чини погоднијим за практичну примену у односу на Kaldi и Wav2Vec. Његова прецизност и робустност позиционирају га као идеалан избор за широк спектар домена – од образовања до здравствене заштите и правних система.

2.1 Модел Whisper

Whisper користи дубоке неуронске мреже и обучен је на великој и разноврсној бази говорних записа, што му омогућава високу прецизност и робустност у препознавању различитих акцената и позадинске буке [5]. Као што је приказано на слици 2.1, овај модел користи енкодер-декодер архитектуру засновану на трансформерима за препознавање говора. Сви аудио сигнали су узорковани на 16 kHz и претворени у 80-каналну спектрограмску репрезентацију. Спектрограм је визуелни приказ који илуструје фреквенцијски садржај аудио сигнала током времена, што омогућава моделу да анализира временско-фреквентне карактеристике говора. Улазни подаци су нормализовани на опсег од -1 до 1 са математичким очекивањем приближно једнаким 0. Ова нормализација осигурава конзистентност и стабилност током тренинга модела, убажавајући проблеме са експлозијом градијената и побољшавајући конвергенцију.



Слика 2.1: Енкодер-декодер архитектура модела.

Енкодер обрађује улазне податке користећи два конволутивна слоја са GELU (*енг. Gaussian error linear unit*) активационом функцијом и синусоидним позиционим кодирањем. Синусоидно позиционо кодирање је метод који додаје информације о положају сваког елемента у низу користећи синусоидне функције са различитим фреквенцијама. Овај приступ омогућава моделу да ефикасно обрађује секвенце података и одржава редослед информација, што је кључно за прецизно препознавање говора. Трансформер блокови затим обрађују излаз из енкодера, користећи механизме пажње да би се фокусирали на најрелевантније делове улазног сигнала.

Модел користи текстуални токенизатор за енглески и прилагођава вокабулар за вишејезичне моделе како би се избегла прекомерна фрагментација. Токени представљају основне јединице података који се користе у моделима за обраду језика, као што су речи, делови речи или чак појединачни знакови. У овом контексту, токени су делови улазног текста који се претварају у нумеричке репрезентације како би их модел могао обрадити.

Декодер користи повезане улазно-излазне репрезентације токена и претходно научене информације из позиционог кодирања улазних података. Повезивање улазно-излазних репрезентација токена подразумева коришћење токена за улазни текст који модел прима и за излазни текст који модел генерише, чиме се постиже конзистентност и ефикасност у процесу декодирања. Овај приступ омогућава моделу да генерише текстуални излаз који је усклађен са улазним аудио сигналом, што је од суштинске важности за прецизну транскрипцију.

2.2 PEFT и LoRA

Адаптација великих, претходно обучених модела за различите задатке у оквиру обраде природног језика традиционално се обавља кроз фино подешавање (*енг. fine-tuning*). Овај процес подразумева ажурирање свих параметара мреже, што доводи до тога да резултујући модел задржава исти број параметара као и оригинални. Код архитектура попут *GPT-3*, која поседује 175 милијарди параметара [6], овакав приступ захтева огромне рачунарске ресурсе, чинећи га непрактичним за ширу примену у окружењима са ограниченим капацитетима.

Као одговор на ове изазове, развијене су методе ефикасне адаптације параметара (*енг. parameter-efficient fine-tuning – PEFT*). Ови приступи, међу којима се посебно издваја *LoRA* (*енг. low-rank adaptation*), омогућавају прилагођавање модела ажурирањем само малог подскупа параметара или додавањем спољних модула специфичних за задатак. Тиме се драстично смањују трошкови складиштења и израчунавања, уз одржавање високе прецизности транскрипције.

PEFT технологија почива на хипотези да се већина параметара у великом моделу не мора мењати како би се постигао одличан учинак на специфичном задатку. Уместо комплетног ажурирања, интервенише се на свега приближно 1% укупног капацитета мреже. Ово резултује значајним уштедама у времену и меморији током фазе тренинга, што ову технику чини погодном за примену чак и на мобилним уређајима или системима са ограниченом рачунарском снагом.

LoRA представља специфичну имплементацију која користи матричне декомпозиције ниског ранга како би адаптирала претходно тренирane параметре. Техника

функционише тако што „замрзава" (*енг. freezing*) постојеће параметре модела, чинећи их непроменљивим током фазе пропагације грешке уназад (*енг. backpropagation*). Уместо ажурирања оригиналних матрица параметара, тренинг се фокусира искључиво на пар матрица ниског ранга које се додају паралелно са постојећим слојевима [7]. Овим приступом се драстично смањује количина меморије потребне за складиштење градијената, јер се већина мреже третира као статична структура. По завршетку тренинга, ове помоћне матрице се комбинују са оригиналним параметрима, што значи да LoRA не уводи додатно кашњење приликом примене модела.

2.3 Скупови података

Тренинг, валидациони и тест скупови података су кључни елементи у процесу машинског учења, јер омогућавају правилно обучавање, подешавање и евалуацију модела. Тренинг скуп, који обично чини највећи део доступних података, служи за обуку модела, омогућавајући му да научи обрасце и везе унутар података. Валидациони скуп се користи за подешавање хиперпараметара и праћење перформанси модела током тренинга, како би се избегло преприлагођавање (енг. *overfitting*) и осигурала генерализација модела на нове податке. Тест скуп, који садржи податке над којима модел није обучаван, користи се за коначну евалуацију перформанси модела након завршетка тренинга, пружајући објективну меру његове успешности.

У овом раду коришћени су следећи скупови података:

1. *classla/ParlaSpeech-SR*: Овај скуп података је изграђен из транскрипата парламентарних сесија доступних у делу за српски језик ParlaMint корпуса [8]. Парламентарни транскрипти су посебно корисни за тренинг модела за препознавање говора, јер садрже формалан и структуриран језик, што доприноси робустности модела.

2. *google/fleurs*: Вишејезични скуп података, али у овом истраживању коришћени су искључиво аудио записи на српском језику. Овај скуп је погодан за тренинг модела на различитим акцентима и стиловима говора, што га чини корисним за побољшање перформанси модела у реалним сценаријима.

3. *mozilla-foundation/common_voice_17_0*: Такође вишејезични скуп података, где су за потребе овог рада узети у обзир једино аудио записи на српском језику. Овај скуп је популаран у заједници машинског учења због своје разноврсности и доступности, што га чини идеалним за тренинг модела за препознавање говора.

Процес преузимања и припреме ових скупова података обухвата неколико корака како би се осигурала њихова униформност и прикладност за тренинг модела. Први корак подразумева усклађивање скупова података тако да искључиво садрже атрибуте *audio*, који укључује звучне записе, и *transcription*, који представља транскрипцију звучног записа. Овај корак је неопходан како би се осигурала компатибилност са архитектуром Whisper модела. Други корак представља узорковање аудио сигнала из скупова података на 16 kHz, што је фреквенција коју Whisper очекује. Ово омогућава једноставну интеграцију различитих скупова података у један кохерентан скуп, осигуравајући да сви аудио записи имају исту фреквенцију узорковања.

Скуп података *classla/ParlaSpeech-SR* је преузет у облику који не садржи делове за тренинг, валидацију и тест модела, стога је било неопходно ручно поделити скуп. Овај процес је најпре подразумевао разбијање оригиналног скупа података на два дела, при чему је 80% узорака коришћено за тренинг, а 20% за привремени скуп. Привремени скуп је затим подељен на два једнака дела, тако да је једна половина коришћена за валидацију, а друга за тестирање. Овај приступ осигурава да модел буде евалуиран на независним подацима, што је кључно за процену његове способности генерализације.

Коначан скуп података је креиран обједињавањем скупова за тренинг сва три скупа података. Сличан приступ је примењен за валидациони и тест скуп, чиме је осигурана конзистентност и разноврсност података у фазама тренирања и евалуације модела. Овај кохерентан приступ омогућава моделу да научи из разноврсних извора података, побољшавајући његову робустност и прецизност у различитим акустичким и језичким условима.

2.4 Оптимизатори

Оптимизатор је алгоритам који ажурира параметре модела како би минимизовао функцију грешке, што је кључно за успешно тренирање модела у машинском учењу. У овом раду је коришћен AdamW [9], модификована верзија Adam оптимизатора (*енг. adaptive moment estimation*) која је прилагођена употреби L2 регуларизације. Adam оптимизатор је алгоритам за оптимизацију који комбинује предности адаптивне стопе учења и моментума за ефикасно ажурирање параметара неуронских мрежа. AdamW додаје L2 регуларизацију (*енг. L2 regularization*), технику која уводи казнени израз на функцију грешке како би се спречило преприлагођавање модела минимизовањем вредности параметара. Ова комбинација чини AdamW посебно ефикасним за тренинг дубоких неуронских мрежа са великим бројем параметара, као и за рад са нестационарним подацима, где су традиционални оптимизатори мање ефикасни.

Додатно, у раду је коришћен планер стопе учења (*енг. learning rate scheduler*), који динамички прилагођава стопу учења током тренинга модела. Посебно је користан у ситуацијама где је потребно фино подешавање процеса учења, јер омогућава бржу конвергенцију и бољу способност генерализације модела. Овај приступ такође помаже у избегавању осцилација и експлозије градијената током тренинга, што је чест проблем при коришћењу фиксне стопе учења. Планер стопе учења омогућава моделу да у почетним фазама тренинга користи већу стопу учења за брзо напредовање, а затим постепено смањује стопу како би се осигурала прецизна конвергенција ка оптималном решењу.

Ефикасност овог оптимизатора зависи од правилног подешавања хиперпараметара, као што су num_warmup_steps и $num\_training\_steps$. Број корака загревања (*енг.* num_warmup_steps) представља број итерација током којих се стопа учења постепено повећава од почетне до циљне вредности. Овај период загревања омогућава стабилнији почетак тренинга, спречавајући велике осцилације у градијентима које могу настати на почетку тренинга. Укупан број корака приликом тренирања (*енг.* $num\_training\_steps$) се односи на укупан број итерација током процеса тренинга модела. Овај хиперпараметар омогућава планеру да планира динамичне промене у стопи учења током целог трајања тренинга, осигуравајући оптимално завршавање процеса у предвиђеном временском оквиру.

Коришћење овог оптимизатора има значајан утицај на ефикасност и стабилност тренинга модела. Променом стопе учења, може се спречити преприлагођавање модела, као и омогућити бржу и стабилнију конвергенцију. У овом раду је коришћен *LinearLR*, који линеарно смањује стопу учења током времена. Овај приступ осигурава да модел у почетним фазама тренинга користи већу стопу учења за брзо напредовање, док се у каснијим фазама стопа учења постепено смањује како би се осигурала прецизна конвергенција. Ова комбинација AdamW оптимизатора и LinearLR планера стопе учења чини процес тренинга ефикасним, стабилним и прилагодљивим различитим сценаријима машинског учења.

2.5 Техника раног заустављања

Техника раног заустављања (*енг. early stopping*) је важан алат у процесу тренирања модела машинског учења, који има за циљ да спречи преприлагођавање модела. Преприлагођавање настаје када се модел превише прилагођава у односу на тренинг податке, учећи и њихов шум и нерелевантне детаље, што негативно утиче на његову способност генерализације на новим подацима. Овај феномен је посебно изражен када је модел превише сложен у односу на доступне податке, што доводи до тога да модел "напамет" учи тренинг скуп, губећи способност да правилно функционише на новим подацима. Уз помоћ раног заустављања решава се овај проблем праћењем перформанси модела на валидационом скупу током сваке епохе тренинга. Ако се перформансе на валидационом скупу престану побољшавати, тренирање се зауставља, спречавајући даље преприлагођавање.

Ова техника укључује постављање прага за толеранцију (*енг. patience*) који дефинише број епоха без значајног побољшања на валидационом скупу. Ако модел не покаже побољшање унутар овог прага, процес тренинга се аутоматски прекида. На тај начин, не само што се спречава преприлагођавање, већ се штеде ресурси, смањујући непотребно трошење времена и енергије на додатно тренирање које не би довело до побољшања перформанси.

Уз правилну примену, рано заустављање осигурава да модел задржи своју способност генерализације, остварујући боље резултате на тест скупу и у реалним применама. Ова техника је посебно корисна у сценаријима где је тренинг модела рачунарски захтеван или где су доступни ресурси ограничени. На пример, у случају дубоких неуронских мрежа са великим бројем параметара, тренинг може трајати сатима или чак данима, а рано заустављање може значајно смањити трошкове без губитка перформанси.

Поред тога, може се комбиновати са другим техникама регуларизације, као што је L2 регуларизација, како би се осигурало да модел не само што избегава преприлагођавање, већ и да задржи високу прецизност и стабилност. Ова комбинација техника је посебно ефикасна у задацима као што су класификација слика, препознавање говора или обрада природног језика, где су модели често изложени ризику од преприлагођавања због сложености података.

Важно је напоменути да успешна примена ове технике захтева пажљиво подешавање хиперпараметара, укључујући праг за толеранцију и величину валидационог скупа. Премали праг може довести до прераног заустављања тренинга, док превелики праг може дозволити моделу да настави да тренира без значајног побољшања. Слично, валидациони скуп мора бити довољно репрезентативан како би осигурао да перформансе на њему одражавају стварну способност модела да генерализује.

У пракси се често користи у комбинацији са техникама као што су унакрсна валидација или очување најбољег модела (енг. *model checkpointing*). Унакрсна валидација омогућава бољу процену перформанси модела, док очување најбољег модела осигурава да се сачувају параметри модела из епохе са најбољим резултатима на валидационом скупу.

2.6 Нормализација речи

Нормализација речи је поступак трансформације текста у стандардизовани облик, неопходан за постизање доследности и прецизности у задацима као што је аутоматска транскрипција говора. Ова техника елиминише варијације у формату текста (нпр. различите употребе интерпункције или величина слова) које могу ометати анализу,

претрагу или поређење транскрибованог садржаја. Посебно је важна у контексту обраде природног језика, где недоследности у тексту могу довести до грешака у алгоритмима за машинско учење или статистичкој анализи.

Алгоритам за нормализацију прослеђеног текста дизајниран је тако да као улаз прихвата текст састављен од произвољно много реченица, осигуравајући њихову потпуну унификацију кроз неколико кључних фаза. Прво, алгоритам идентификује и уклања све знакове интерпункције, попут тачака, запета и упитника, замењујући их бланко карактером. Овим поступком се обезбеђује да све речи буду раздвојене искључиво размацима, док се истовремено уклањају границе између реченица. Након тога, целокупан садржај се трансформише у мала слова, чиме се елиминишу разлике у капитализацији и олакшава даља обрада.

Затим, врши се токенизација текста на појединачне речи коришћењем бланко карактера као граничника. Алгоритам пролази кроз сваки токен и врши нумеричку конверзију – уколико је реч број, она се пребацује у текстуални облик (нпр. „23" постаје „двадесет и три"). Овај корак је од суштинске важности за транскрипцију говора, где се бројеви морају представити у пуном еквиваленту ради веродостојности података.

Након што су све појединачне речи нормализоване, алгоритам их спаја у финални низ, пажљиво контролишући размаке. У овој фази се врши транслитерација свих карактера у српску ћирилицу, заменом латиничних слова еквивалентним ћириличним знаковима. Као завршни корак, алгоритам уклања сувишне белине са почетка и краја низа, као и вишеструке унутрашње размаке. Крајњи резултат процеса је текст који се састоји од речи раздвојених размацима, што суштински представља једну реченицу без тачке на крају. Овакав формат обезбеђује максималну униформност, чинећи текст спремним за поуздану интеграцију у системе за обраду природног језика.

2.7 Фреквенција речи

Фреквенција речи представља метрику која квантификује појављивање одређене лексеме унутар текстуалног корпуса. Овај показатељ има фундаменталну примену у области обраде природног језика, а у оквиру овог рада примарно се користи за

оптимизацију модула за проверу правописа (*eng. spell checker*). Ослањањем на статистичке податке о учесталости, алгоритам може приоритетно рангирати вероватније кандидате за исправку, чиме се директно унапређује прецизност система. Овакав приступ је нарочито ефикасан код аутоматске транскрипције говора, где акустички шум често доводи до генерисања двосмислених термина.

Процес екстракције фреквенција реализован је кроз неколико фаза. Прва фаза обухвата модул за итеративно ажурирање речника: за сваку реченицу у скупу података врши се претходна нормализација текста, након чега се појединачне речи токенизују. Систем проверава постојање сваког токена у структури података – уколико је реч већ регистрована, њена вредност се инкрементира; у супротном, реч се иницијализује као нови унос са почетном вредношћу један.

Главна функција за прорачун учесталости најпре иницијализује празан речник резултата, а затим врши обједињавање података из тренинг и валидационог скупа. Оваква интеграција је неопходна како би се статистички узорак формирао на основу целокупног доступног корпуса, чиме се смањује вероватноћа појаве необухваћених речи. Сукцесивном применом претходно описане функције на оба скупа, формира се свеобухватан преглед дистрибуције лексема.

Финални излаз алгоритма је структурирани речник који садржи прецизне податке о појављивању сваке речи. Поред примарне улоге у корекцији правописа, овај речник омогућава идентификацију најчешћих термина (*eng. stopwords*), као и детекцију ретких лексема или специфичних стручних израза. Систематично пребројавање овим путем постаје основа за сваку даљу статистичку или семантичку анализу, омогућавајући бржу и рачунарски ефикаснију обраду великих количина текста у реалном времену.

2.8 Провера правописа

Провера правописа обухвата идентификацију и аутоматску корекцију лексичких неправилности, што значајно унапређује квалитет транскриптованог текста. Овим поступком се аутоматизује валидација и минимизује потреба за мануелном интервенцијом, чиме се оптимизује време потребно за финализацију садржаја.

У овом раду, механизам корекције је реализован комбинацијом структуре података Trie и Левенштајновог растојања (*енг. Levenshtein distance*). Trie, односно префиксно стабло, представља специјализовану структуру графа која омогућава ефикасно складиштење и претрагу лексикона. Сваки чвор у графу чини појединачни карактер, док путање кроз стабло формирају комплетне речи. Описана организација омогућава брзу претрагу префикса и идентификовање валидних кандидата за исправку у линеарном времену у односу на дужину речи.

На пример, уколико се у речнику налазе речи "прозор", "пројекат" и "правда", оне деле заједничке префиксне чворове унутар стабла. Претрага за префиксом „про“ ће у корену одмах елиминисати грану са речју "правда" и ефикасно усмерити алгоритам искључиво ка чворовима који генеришу легитимне кандидате "прозор" и "пројекат".

Левенштајново растојање дефинише се као мера разлике између две речи, која се израчунава као минимални број елементарних операција на нивоу карактера које укључују уметање, брисање или замену. Ове операције су неопходне за трансформацију једне речи у другу. У контексту провере правописа, наведена метрика служи као примарни критеријум за евалуацију сличности између тренутне речи која се обрађује и кандидата из речника који су њена потенцијална исправка.

На пример, ако је улазна (погрешно транскрибована) реч "прозорр", а кандидат из речника "прозор", растојање износи 1 (потребно је извршити једну операцију брисања сувишног карактера "р"). У случају трансформације речи "нос" у "мост", растојање износи 2, будући да су неопходне две операције: замена карактера "н" у "м" (чиме се добија реч "мос"), а потом уметање карактера "т" на крај речи.

Процес одређивања оптималне исправке извршава се кроз неколико фаза:

1. *Иницијализација:* Минимално растојање се дефинише као бесконачна вредност (∞), док се формира празна листа за складиштење потенцијалних фаворита.

2. *Евалуација кандидата:* За сваког кандидата генерисаног претрагом префиксног стабла, израчунава се Левенштајново растојање у односу на улазну реч.

3. *Ажурирање референтног скупа:* Уколико је израчунато растојање мање од тренутног минимума, вредност се ажурира, а листа кандидата се ресетује и попуњава датом речју. Ако је растојање идентично минималном, кандидат се додаје у постојећу листу, чиме се омогућава разматрање вишеструких једнако вероватних решења.

Уколико након исцрпне претраге листа остане празна, алгоритам задржава оригинални облик речи под претпоставком да адекватна исправка није пронађена. У супротном, врши се селекција на основу учесталости у корпусу добијеном из скупа података за обуку. За све кандидате из ужег избора анализира се њихова заступљеност у претходно формираном фреквенцијском речнику. Овај корак даје приоритет природнијим лексемама, па је коначни излаз система онај термин који има највишу вредност појављивања међу предлозима са минималним растојањем.

Глава 3

Методологија експерименталне евалуације

Експерименти у машинском учењу представљају кључни корак у процесу развоја модела, јер омогућавају одређивање оптималних конфигурација и побољшање перформанси. У овом раду, примењен је темељан и систематичан методолошки приступ који обухвата све аспекте тренинга модела, од иницијалне конфигурације до коначне евалуације резултата. Посебна пажња посвећена је варирању хиперпараметара коришћењем технике мрежне претраге (*енг. grid search*). Ова техника представља метод оптимизације хиперпараметара који систематски истражује унапред дефинисан скуп вредности хиперпараметара, тренирајући и евалуирајући модел за сваку могућу комбинацију. На тај начин се осигурава проналазак најбољих вредности хиперпараметара, чиме се побољшава тачност и поузданост модела.

Модел	Број параметара
whisper-tiny	39×10^6
whisper-base	74×10^6
whisper-small	244×10^6
whisper-medium	769×10^6
whisper-large-v2	$1,55 \times 10^9$
whisper-large-v3	$1,55 \times 10^9$

Табела 3.1: Различите верзије Whisper модела и број њихових параметара.

Whisper модел је доступан у више верзија, које се разликују по броју параметара за тренирање, што директно утиче на њихову сложеност и перформансе. Ове верзије омогућавају прилагођавање модела специфичним захтевима и ресурсима доступним за тренинг. У табели 3.1 приказане су различите верзије Whisper модела, заједно са бројем њихових параметара, што илуструје разлике у сложености и потенцијалној ефикасности.

У овом истраживању, фокус је стављен на моделе *whisper-large-v2* и *whisper-large-v3*, који су одабрани због њихове способности да постигну високе резултате кроз пажљив тренинг. Главни циљ овог поглавља представља анализа перформанси претходно поменутих модела и идентификовати оптималне вредности хиперпараметара који ће осигурати најбоље резултате.

3.1 Хиперпараметри тренинга

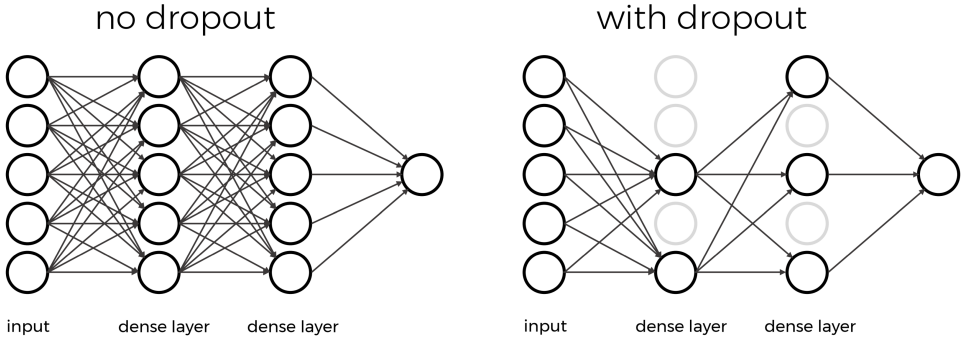
Током процеса тренинга модела, неопходно је пажљиво подесити низ хиперпараметара како би се осигурале оптималне перформансе, стабилност процеса учења и способност модела да се успешно прилагоди новим подацима. Ови параметри представљају основне управљачке механизме који дефинишу како ће се *Whisper* модел мењати и усавршавати кроз сваку итерацију излагања подацима, директно утичући на финални квалитет транскрипције.

У оквиру ове експерименталне евалуације, посебна пажња посвећена је следећим кључним конфигурационим параметрима:

- *Брзина учења*: Овај хиперпараметар се често сматра најважнијим хиперпараметром у процесу машинског учења. Он одређује величину корака којим алгоритам оптимизације ажурира унутрашње параметре модела у одговору на процењену грешку. Уколико је брзина учења постављена превисоко, модел може пребрзо конвергирати ка субнормалном решењу или постати потпуно нестабилан. С друге стране, прениска вредност може учинити процес тренинга претерано спорим. За потребе овог рада, коришћена је динамичка стратегија која балансира ове две крајности.

- *L2 регуларизација*: Како би се спречио феномен преприлагођавања (*overfitting*), где модел почиње да учи специфичне шумове и детаље из тренинг скупа уместо општих лингвистичких образаца, примењена је техника L2 регуларизације. Ова техника додаје малу казну за велике вредности параметара унутар мреже, подстичући модел да задржи једноставнију и флексибилнију структуру. Тиме се постиже боља генерализација, односно већа прецизност приликом транскрипције говора у реалним условима.

- *Искључивање неурона (енг. dropout)*: *Dropout* представља изузетно ефикасну методу регуларизације која се састоји у насумичном „искључивању" одређеног процента неурона током сваког корака тренинга, као што се може видети на слици 3.1. На овај начин, мрежа је присиљена да учи робусније карактеристике јер се не може ослонити на присуство било ког појединачног неурона. Ово значајно доприноси томе да модел задржи високу тачност чак и на подацима које никада раније није видео.

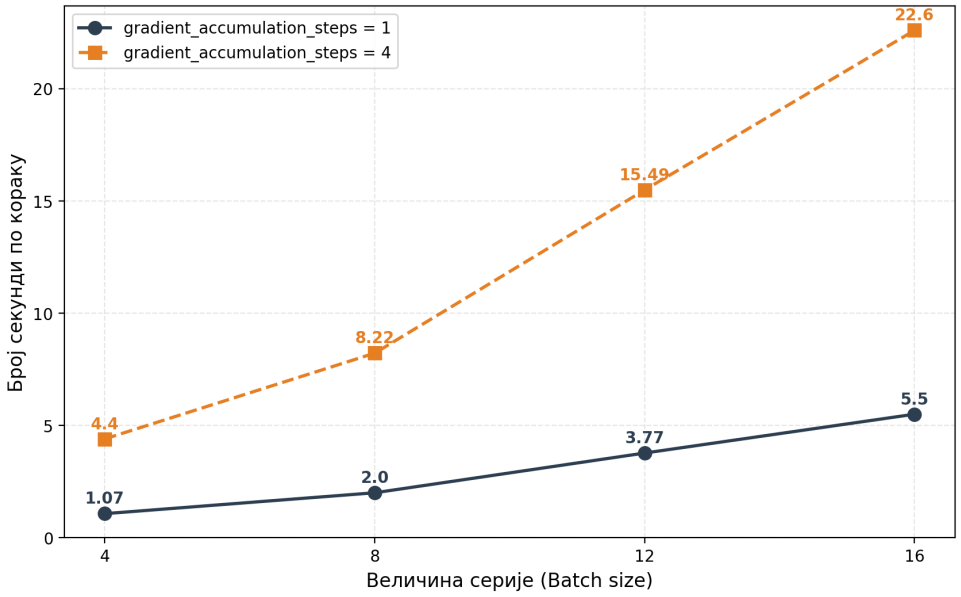


Слика 3.1: Концепт *dropout* регуларизације: оригинална мрежа у поређењу са мрежом током фазе тренинга са искљученим везама.

- *Величина серије (енг. batch size)*: Овај хиперпараметар дефинише број аудио узорака који се прослеђују кроз модел пре него што се изврши једно ажурирање његових параметара. Већа серија података омогућава прецизнију процену градијента грешке и стабилнији тренинг, али захтева значајне хардверске ресурсе у погледу видео меморије.

- *Кораци акумулације градијента (енг. gradient accumulation steps)*: Овај хиперпараметар омогућава виртуелно повећање величине серије података када су

хардверски ресурси ограничени. Уместо ажурирања вредности параметара након сваке серије, алгоритам акумулира промене кроз дефинисани број корака и тек тада врши заједничку корекцију. Тиме се постиже стабилност карактеристична за велике серије података без додатног заузећа видео меморије. На пример, као што се може видети са слике 3.2, ако је капацитет меморије 4 записа, а жељена серија 16, постављањем овог параметра на 4 модел врши четири пролаза пре ажурирања, ефективно симулирајући већи скуп података.



Слика 3.2: Анализа трајања процеса обучавања модела у односу на величину серије.

- *Максимални број корака (енг. max steps)*: Овај хиперпараметар дефинише укупно трајање процеса тренинга изражено кроз број итерација (корака), уместо кроз традиционалне епохе. Један корак подразумева обраду једне серије података и ажурирање вредности параметара модела. Постављање оптималног броја корака је кључно како би се моделу омогућило довољно времена да научи сложене говорне структуре, а да се истовремено избегне непотребно трошење рачунарских ресурса након што модел престане да показује значајна побољшања.

- *Кораци евалуације (енг. eval steps)*: Како би се процес учења држао под контролом, неопходно је периодично проверавати перформансе модела на подацима

који нису коришћени за тренинг. Хиперпараметар eval_steps одређује колико често ће се вршити та провера. Редовна евалуација омогућава праћење напретка у реалном времену и пружа увид у то да ли се модел развија у жељеном смеру, што је од суштинског значаја за препознавање тренутка када је тренинг најуспешнији.

- *Фаза загревања (енг. warmup steps)*: У почетним фазама тренинга, модел је често веома осетљив на нагле промене у вредности параметара. Коришћењем корака загревања, брзина учења се постепено повећава од нуле до циљане вредности, чиме се постиже стабилан почетак обуке и спречава рана дивергенција модела.

Пажљивом комбинацијом наведених хиперпараметара, обезбеђен је оквир који Whisper моделу омогућава да ефикасно апсорбује специфичности језика из скупа за обучавање, уз задржавање робусности неопходне за квалитетну примену у практичним задацима препознавања говора.

3.2 Евалуација модела

Методологија евалуације у оквиру овог истраживања конципирана је тако да омогући ригорозну и објективну процену перформанси система за аутоматску транскрипцију. Процес обухвата систематичну припрему скупа података за тестирање, дефинисање квантитативних метрика, спровођење експерименталних тестова и компаративну анализу добијених резултата.

Почетна фаза евалуације подразумева креирање скупа података за тестирање, који се састоји од аудио записа и припадајућих референтних транскрипата (*енг. ground truth*). Ови транскрипти представљају еталон на основу којег се врши поређење предвиђања модела са стварним текстуалним садржајем. Квантификација прецизности и ефикасности модела врши се применом две стандардизоване метрике: стопе грешке на нивоу речи (*енг. word error rate – WER*) и стопе грешке на нивоу карактера (*енг. character error rate – CER*) [10].

Ове метрике представљају индустријски стандард у области обраде природног језика, јер омогућавају егзактно мерење лексичких и морфолошких девијација. Прорачун се заснива на израчунавању минималног броја операција уређивања – супституција (S), брисања (D) и додавања карактера (I) – неопходних да би се предвиђена реч трансформисала у референтну, у односу на укупан број јединица (речи или карактера) у референтној речи (N):

$$\text{WER} = \frac{S + D + I}{N_{words}} \quad \text{и} \quad \text{CER} = \frac{S + D + I}{N_{chars}}$$

Иако се вредности метрика типично налазе у интервалу $[0, 1]$, при чему нула представља апсолутну тачност, у специфичним случајевима екстремно ниске прецизности ови показатељи могу премашити вредност 1. Оваква појава указује на критично одступање предвиђања модела од референтног улаза.

Упркос широкој примени, неопходно је констатовати инхерентна ограничења WER и CER метрика. Примарни недостатак лежи у изостанку семантичке и контекстуалне евалуације. На пример, мање правописне или граматичке варијације које суштински не ремете разумљивост поруке бивају кажњене једнако као и грешке које потпуно мењају значење реченице.

Током експерименталне фазе, генерисани транскрипти се подвргавају процесу нормализације ради постизања униформности са референтним текстом. Овај поступак често обухвата и примену алгоритма за проверу правописа, чија је методологија описана у поглављу 2.8, како би се редуковао утицај очигледних лексичких грешака.

Компаративна анализа резултата пре и након процеса корекције омогућава прецизно утврђивање доприноса провере правописа финалним перформансама система. Сви подаци се систематично документују у излазним датотекама како би се осигурала репродуцибилност истраживања. Овакав методолошки оквир пружа чврсту основу за интерпретацију резултата и идентификовање простора за даљу оптимизацију.

3.3 Тренирање модела са фиксним вредностима хиперпараметара

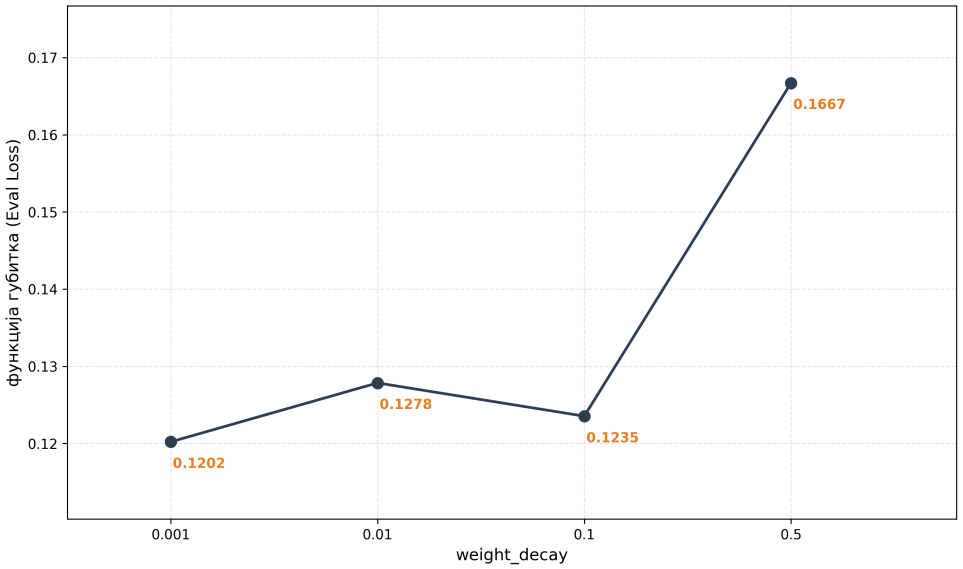
У иницијалној фази тренинга модела whisper-large-v2, коришћене су следеће вредности хиперпараметара: learning_rate = 0.001, max_steps = 4000 и eval_steps = 1000. Ови хиперпараметри су изабрани на основу препорука и стандарда утврђених у научним радовима из сродних области [11, 12]. Важно је напоменути да у иницијалном тренирању није коришћен оптимизатор, те вредност за weight_decay није била дефинисана.

Затим, током тренинга новије верзије модела whisper-large-v2, фиксиране су вредности за weight_decay на 0.001 и за learning_rate на 0.0009. Ови параметри су такође изабрани у складу са препорукама из научне литературе и сматрају се стандардним за ову област истраживања. Тренинг је спроведен до max_steps = 16000 корака, што представља износ који је значајно већи у односу на претходну верзију модела. Овај пораст је мотивисан експерименталним потребама за побољшањем перформанси модела. Евалуација на валидационом скупу података је вршена након сваких eval_steps = 500 корака.

3.4 Варирање вредности хиперпараметара

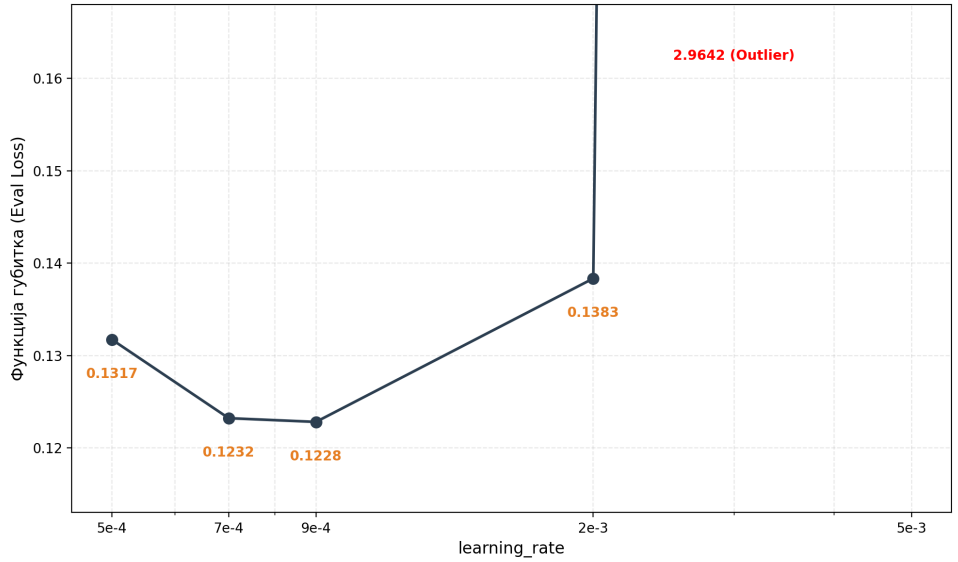
Фиксирањем вредности хиперпараметара у претходним експериментима није било могуће добити свеобухватну слику о перформансама модела за различите комбинације ових хиперпараметара. Циљ овог истраживања је био да се идентификују оптималне вредности хиперпараметара weight_decay и learning_rate, како би се омогућиле што боље перформансе модела. Овај процес укључује систематско тестирање различитих вредности ових хиперпараметара и анализу њиховог утицаја на функцију грешке на валидационом скупу података.

У првој фази експеримената са моделом whisper-large-v2, фокус је био на анализи утицаја хиперпараметра weight_decay. Изабране су четири различите вредности овог параметра 0.001, 0.01, 0.1 и 0.5. Најнижа вредност функције грешке остварена је са weight_decay = 0.001, што се може видети са слике 3.3.



Слика 3.3: Утицај промене `weight_decay` на функцију грешке.

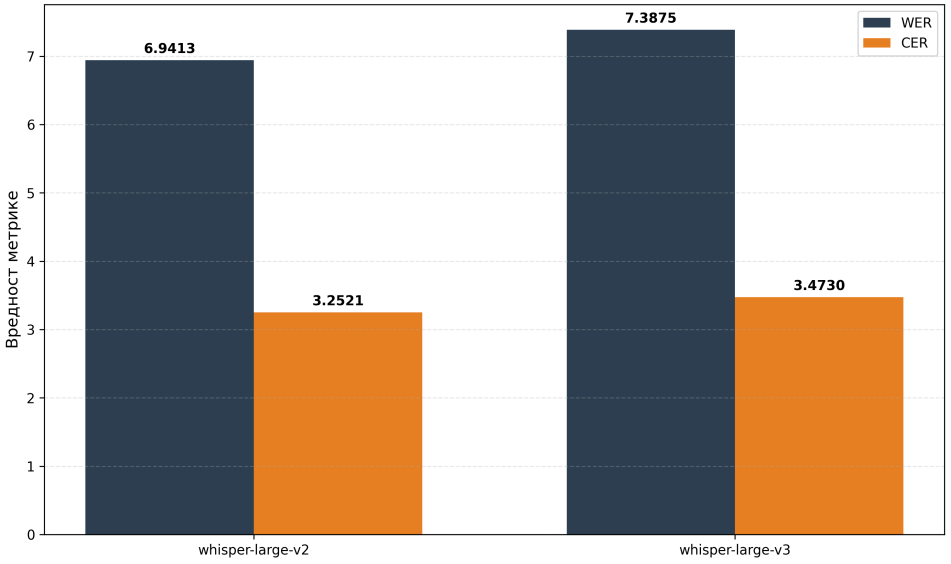
Након тога, иста оптимална вредност за weight_decay је коришћена за анализу утицаја хиперпараметра learning_rate. Тестиране су вредности 0.0005, 0.0007, 0.0009, 0.002 и 0.005. Као што је приказано на слици 3.4, најбољи резултат постигнут је са learning_rate = 0.0009, чиме се постиже најбоља стабилност и ефикасност током тренинга.



Слика 3.4: Утицај промене `learning_rate` на функцију грешке.

Након идентификације оптималних вредности хиперпараметара, процес финог подешавања спроведен је кроз серију контролисаних експеримената. За модел whisper-large-v2, тренинг је планиран до максималних 16000 корака, уз периодичну евалуацију на сваких 500 корака. Применом learning_rate од 0.0009 и weight_decay од 0.001, уз интеграцију алгоритма за исправку грешака, модел је постигао стабилну конвергенцију.

Код напредније итерације модела, whisper-large-v3, примењена је софистицираниja техника мрежне претраге за оптимизацију кључних хиперпараметара. Иако је почетни оквир такође предвиђао 16000 корака, уведена је техника раног заустављања. Овај механизам је аутоматски прекинуо процес након приближно 9000 корака, након што је утврђено да даљи тренинг не доприноси смањењу губитка, чиме је спречено преприлагођавање. Најбољи резултат за v3 верзију постигнут је са вредностима learning_rate = 0.002 и weight_decay = 0.005, а упоредни преглед кључних перформанси за оба модела приказан је на слици 3.5.



Слика 3.5: Упоредни приказ перформанси `v2` и `v3` модела.

Глава 4

Резултати и дискусија

Ово поглавље доноси детаљан преглед резултата евалуације модела whisper-large-v2 и whisper-large-v3, изведених кроз серију експерименталних тестирања. Примарни фокус истраживања постављен је на поређење перформанси у два кључна модалитета: иницијалном стању пре финог подешавања и након обуке спроведене уз дефинисање оптималних хиперпараметара. Додатно је истражен ефекат алгоритма за проверу правописа како би се прецизно квантификовао његов допринос финалној прецизности генерисане транскрипције.

Поређење стања пре и после фазе тренинга пружа егзактне податке о ефикасности процеса прилагођавања модела специфичним подацима. Истовремено, критички су сагледана унутрашња ограничења стандардних метрика као што су WER и CER, што отвара простор за дискусију о алтернативним методологијама у будућим радовима. Намера ове анализе превазилази пуко излагање бројчаних вредности; она нуди свеобухватну интерпретацију експерименталног процеса уз идентификацију предности и недостатака коришћених приступа. Такав контекстуални оквир је неопходан за правилно разумевање резултата и постављање чврстог темеља за даљи научни развој у домену препознавања говора.

4.1 Утицај броја корака и хиперпараметара током тренинга

Одабир оптималних вредности хиперпараметара представља кључни корак приликом тренирања модела у машинском учењу. Проналажење правилног баланса између броја корака у тренингу и вредности хиперпараметара, као што су learning_rate и weight_decay, од суштинске је важности за постизање максималних перформанси. Адекватне вредности осигуравају потпуну конвергенцију и стабилну генерализацију модела, што директно утиче на његову способност да ефикасно решава сложене лингвистичке задатке. Стога, пажљиво подешавање ових параметара и систематско тестирање различитих комбинација кроз експерименталне итерације представљају темељ развоја успешних система за препознавање говора.

У специфичним задацима као што је аутоматска транскрипција говора, процес обуке захтева значајан временски оквир како би модел успешно мапирао акустичке и језичке карактеристике у текстуалну форму. На овај начин се перформансе прилагођавају специфичностима циљног језика. У складу са тиме, модел whisper-large-v3 је подвргнут интензивном тренингу како би се искористио његов пуни потенцијал и осигурао висок ниво прецизности у финалној транскрипцији.

Значајан допринос стабилизацији модела whisper-large-v3 пружило је фино подешавање хиперпараметара. Иако је примењена техника мрежне претраге идентификовала њихове оптималне вредности, природа дубоких неуронских мрежа оставља простор за анализу међузависности ових параметара у различитим окружењима. Наменски одабрани learning_rate и weight_decay директно су контролисали динамику конвергенције: правилно дефинисан learning_rate спречио је осцилације у тренингу, док је weight_decay обезбедио неопходан степен регуларизације. Овакав приступ је омогућио моделу да задржи флексибилност приликом обраде аудио записа различитог квалитета, без губитка прецизности на познатим подацима.

У будућим истраживањима, фокус се може поставити на даљу свеобухватну експлоатацију хиперпараметарског простора ради постизања додатне финоће у адаптацији модела. Поред већ анализираних вредности learning_rate и weight_decay, значајан потенцијал за унапређење лежи у систематском варирању параметара као што су

batch_size и eval_steps, чиме би се додатно испитала равнотежа између стабилности градијента и учесталости валидације. Примена напредних техника, попут Бајесове оптимизације, могла би пружити још дубљи увид у сложене нелинеарне односе између ових варијабли унутар Whisper архитектуре. Континуирано експериментисање са дужином тренинг процеса и динамичким променама конфигурације омогућило би још прецизнију калибрацију модела whisper-large-v3. Ови кораци представљају природни наставак овог рада ка потпуном разумевању утицаја целокупног скупа конфигурационих параметара на финалну стабилност и тачност система.

4.2 Поређење модела без претходног тренирања

У Табели 4.1 приказани су резултати евалуације модела whisper-large-v2 и whisper-large-v3 у њиховом изворном стању. Ови подаци пружају кључан увид у иницијалне капацитете архитектура и служе као основа за разумевање ефеката касније оптимизације.

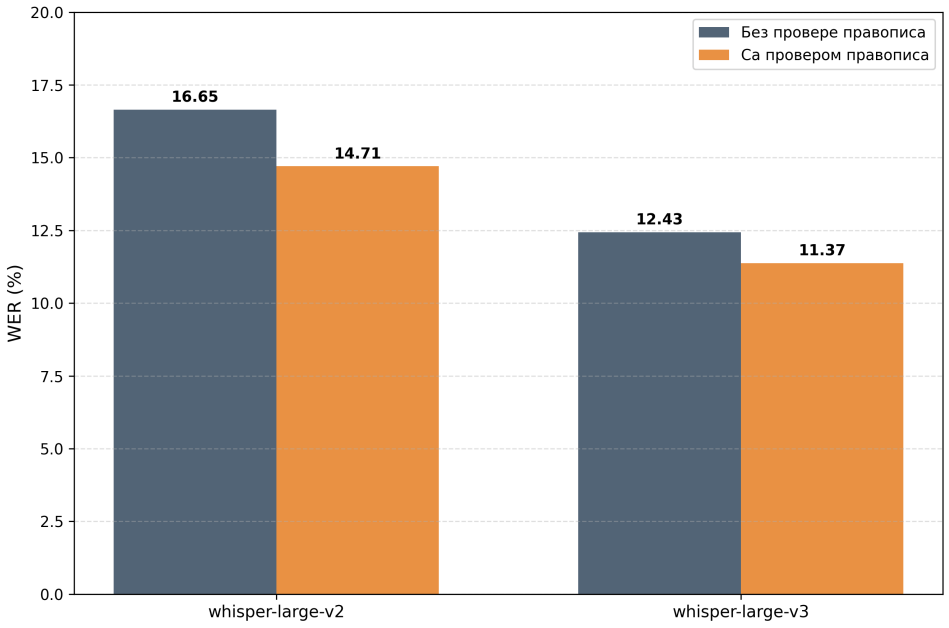
Експериментални подаци указују на то да whisper-large-v3 поседује супериорне почетне перформансе у односу на свог претходника. На пример, без примене додатних техника корекције, v3 верзија остварује WER од 12.431 и CER од 5.077, док v2 под истим условима бележи приметно више вредности грешке WER од 16.647 и CER од 6.383. Оваква разлика потврђује да унапређења у архитектури верзије v3 модела доносе бољу прецизност транскрипције чак и без накнадног прилагођавања специфичном домену.

Модел	*Провера правописа*	WER	CER
whisper-large-v2	не	16.647	6.383
whisper-large-v2	да	14.706	6.128
whisper-large-v3	не	12.431	5.077
whisper-large-v3	да	11.372	4.976

Табела 4.1: Резултати евалуације модела без претходног тренирања.

Интеграција модула за проверу правописа довела је до позитивног помака код оба модела, посебно у оквиру WER метрике. Код модела whisper-large-v2, релативно побољшање износи приближно 11.66%, док је код whisper-large-v3 забележен пад

грешке од око 8.52%. Како се на слици 4.1 може видети, иако се апсолутна разлика смањује, доследност побољшања указује на важност пост-процесирања за финални квалитет текста.



Слика 4.1: Упоредни приказ утицаја провере правописа на `v2` и `v3` моделе.

Иако ова техника ефикасно елиминише словне грешке, она има ограничен утицај на семантичке неспоразуме настале услед лошег аудио сигнала. То указује на потенцијал за будућа истраживања која би обухватила контекстуалне језичке моделе у фази корекције текста.

4.3 Резултати након тренирања модела

Резултати евалуације модела whisper-large-v2 и whisper-large-v3 показују значајне разлике у перформансама у зависности од подешавања хиперпараметара, дужине тренинга и примене провере правописа. У наставку су детаљно анализирани резултати за оба модела, уз упоређивање њихових перформанси у различитим сценаријима.

Табела 4.2 приказује резултате евалуације модела whisper-large-v2 на тест скупу кроз две фазе. У првој фази, модел је трениран до 4000 корака са фиксираном

Број корака	learning_rate	weight_decay	Провера правописа	WER	CER
4000	0.001	None	Не	18.4511	6.8129
4000	0.001	0.001	Не	12.9279	4.94
16000	0.0009	0.001	Не	7.0294	3.2322
16000	0.0009	0.001	Да	6.9413	3.2521

Табела 4.2: Резултати евалуације модела whisper-large-v2 на тест скупу.

вредношћу хиперпараметра learning_rate на 0.001 и без коришћења хиперпараметра weight_decay. Резултат евалуације се огледа у вредностима грешака WER од 18.4511 и CER од 6.8129. Након варирања вредности хиперпараметра weight_decay, уз исте вредности за број корака током тренинга и learning_rate, модел је остварио значајно боље резултате: WER од 12.9279 и CER од 4.94. Ово указује да увођење регуларизационог хиперпараметра weight_decay током тренинга омогућава бољу генерализацију модела. У другој фази је процес тренинга модела продужен на 16000 корака са оптимизованим вредностима хиперпараметара. Примена провере правописа је допринела смањењу грешака, посебно на нивоу речи.

Број корака	learning_rate	weight_decay	Провера правописа	WER	CER
16000	0.002	0.001	Не	8.423	3.7497
16000	0.002	0.001	Да	8.0861	3.7444
9000	0.002	0.005	Не	7.5813	3.4662
9000	0.002	0.005	Да	7.3875	3.473

Табела 4.3: Резултати евалуације модела whisper-large-v3 на тест скупу.

Табела 4.3 приказује резултате евалуације модела whisper-large-v3 на тест скупу. Модел је такође трениран у две фазе, са различитим комбинацијама хиперпараметара и дужином процеса тренинга. Иако је модел whisper-large-v3 новије генерације, показао је лошије перформансе у односу на whisper-large-v2. Важно је напоменути да је тренинг за whisper-large-v3 у другој фази био прекинут након 9000 корака због активације раног заустављања, што је могло утицати на коначне резултате.

Након варирања вредности хиперпараметара learning_rate и weight_decay, утврђене су њихове оптималне вредности за обе верзије модела. За модел whisper-large-v2, оптимална вредност за learning_rate износи 0.0009, док је weight_decay постављен на 0.001. Оптимална вредност хиперпараметра learning_rate за модел whisper-large-v3 је 0.002, а weight_decay је овог пута постављен на 0.005. Ове вредности су изабране

на основу резултата мрежне претраге и експеримената са различитим комбинацијама хиперпараметара.

У Табели 4.4 приказани су резултати евалуације модела на тест скупу након тренирања са оптималним вредностима хиперпараметара. Модел whisper-large-v2 остварио је WER од 7.03 и CER од 3.232 без провере правописа, док је са провером правописа WER смањен на 6.941. С друге стране, модел whisper-large-v3 остварио је WER од 7.581 и CER од 3.466 без провере правописа, а са провером правописа WER је смањен на 7.387.

Модел	Провера правописа	WER	CER
whisper-large-v2	не	7.03	3.232
whisper-large-v2	да	6.941	3.252
whisper-large-v3	не	7.581	3.466
whisper-large-v3	да	7.387	3.347

Табела 4.4: Резултати евалуације модела whisper-large-v2 и whisper-large-v3 на тест скупу за оптималне вредности хиперпараметара.

Иако је било очекивано да ће модел whisper-large-v3 остварити боље резултате након тренирања, модел whisper-large-v2 показао је боље перформансе, што може бити изненађујући резултат. Један од могућих разлога за овакав исход је чињеница да ниједан whisper-large-v3 модел током варирања хиперпараметара није трениран до максималног броја корака од 16000. Ово је могло утицати на његову способност да апсорбује пун обим тренинг података, што је резултирало субоптималним перформансама.

4.4 Ограничења метрика WER и CER

Иако су метрике WER и CER широко прихваћене и корисне за мерење тачности модела аутоматске транскрипције, оне имају одређена ограничења која могу утицати на потпуност и прецизност евалуације. Највећи недостатак ових метрика је то што оне не узимају у обзир семантички смисао или контекст речи. На пример, ако модел направи мању граматичку грешку која не утиче на разумевање смисла реченице (као што је замена једног синонима за други или непотпуна интерпункција), метрика ће

и даље казнити ту грешку као да је значајна. Ово може довести до непотпуне оцене перформанси модела, посебно у ситуацијама где је контекст кључан за правилно тумачење смисла.

Ова ограничења су посебно изражена у задацима где је семантичка тачност важнија од дословне тачности. На пример, ако модел транскрибује реченицу "Ана иде у продавницу" као "Ана иде у тржни центар", WER и CER ће ову грешку третирати као значајну, иако је смисао реченице остао исти. Ово указује на потребу за развојем метрика које би узимале у обзир семантичку сродност и контекстуалну релевантност, а не само дословно поклапање са референтним текстом.

Будући истраживачки напори требало би да буду усмерени ка развоју напреднијих система за корекцију који интегришу морфолошку прецизност са дубљом семантичком анализом, уз истовремено избегавање економски неодрживе парадигме *LLM as a judge*. Уместо тога, императив је имплементација метрика заснованих на векторским репрезентацијама речи (*енг. word embeddings*) или специјализованим моделима заснованим на трансформерима за процену контекстуалне блискости попут модела BERT. Овакав приступ би омогућио високу семантичку тачност и прецизнију евалуацију у ситуацијама где је контекст кључан за разумевање смисла, али на рачунарски знатно ефикаснији начин у односу на генерисање одговора путем великих језичких модела.

Други аспект ограничења метрика WER и CER може бити повезан са квалитетом аудио записа из скупа података. Лош квалитет звука, позадинска бука или нејасни изговори могу довести до грешака у транскрипцији које нису последица лоших перформанси модела, већ лошег улазног сигнала. Ово отвара могућност за будућа истраживања која би се фокусирала на побољшање квалитета улазних података или развој метрика које би узимале у обзир квалитет аудио записа при евалуацији.

Један од праваца би могао бити развој метрика заснованих на семантичкој сличности, које би омогућиле прецизнију оцену перформанси модела у контекстуално сложеним ситуацијама. Други правац би могао бити интеграција провере правописа са семантичком анализом, што би омогућило исправљање грешака које не утичу на смисао, али ипак побољшавају корисничко искуство.

4.5 Анализа карактеристичних грешака модела whisper-large-v2

Након спроведене евалуације модела whisper-large-v2 на тест скупу, идентификован је низ специфичних грешака које указују на ограничења акустичког модела у одређеним контекстима, као и на изазове у процесу накнадне корекције текста. Ове неправилности се могу класификовати у неколико категорија: фонетске вишезначности, халуцинације модела, варијације у записивању репетиција говора, као и грешке узроковане ограничењима коришћеног речника за српски језик. Разумевање ових појава је од суштинског значаја за даљу оптимизацију система, јер оне директно утичу на коначну вредност WER и CER метрика.

У табели 4.5 приказани су карактеристични примери одступања предвиђања модела у односу на референтне транскрипте. Ови узорци илуструју типичне сценарије у којима модел не успева да изврши тачно превођење акустичког сигнала у текст, често под утицајем језичких модела који фаворизују вероватније, али контекстуално нетачне речи.

Референтни текст	Предвиђање модела	Тип грешке
државни орган	државниорган	Неадекватно спајање речи
погрешио сам	спасибо	Халуцинација (страни језик)
ти ти ти	т т т	Неадекватна транскрипција репетиције
хајде да то	цхајде да то	Грешка у транслитерацији
онај период од	истекнао	Лексичка супституција
кућа	врућа	Погрешно предвиђена реч (акустичка сличност)
у скупштини	у скупштину	Погрешан падеж (непостојећи облик у речнику)

Табела 4.5: Примери карактеристичних грешака у транскрипцији

Детаљнијом анализом целокупног скупа података утврђено је да се већина преосталих грешака може свести на два доминантна проблема. Први проблем су акустичке халуцинације, односно ситуације у којима модел генерише потпуно погрешан низ речи или погрешну појединачну реч коју није могуће кориговати лингвистичким правилима. Други кључни изазов представљају ограничења лексичког фонда, где модел исправно препозна реч, али будући да се тај тачан морфолошки облик не налази у

речнику за проверу правописа, систем га накнадно деформише покушавајући да га усклади са познатим терминима.

Ови изазови су значајно превазиђени употребом богатијег речника српског језика који обухвата широк спектар деклинација и конјугација за сваку лексему [13]. Проширени лексички фонд послужио је као робусна основа за систем за проверу правописа, омогућавајући систему да препозна реч чак и када је модел транскрибује у мање учесталом падежном облику. На овај начин, осигурано је да систем препознаје разноврсне аутентичне облике речи, чиме је постигнута већа прецизност у финализацији транскрибованог садржаја и боља генерализација на нетипичне говорне уносе.

Глава 5

Закључак

Предмет овог мастер рада обухватио је анализу и евалуацију архитектура за аутоматско препознавање говора на српском језику са примарним фокусом на Whisper моделе компаније OpenAI. Истраживање је конципирано са циљем компаративне анализе перформанси различитих итерација модела, идентификације оптималних конфигурација хиперпараметара, те квантификације доприноса модула за проверу правописа финалном квалитету транскрипције.

Кроз ригорозан експериментални оквир, тачност система је евалуирана применом стандардизованих метрика WER и CER. Иако су ови показатељи омогућили објективно поређење са референтним истраживањима, анализом су потврђена и њихова инхерентна ограничења у домену семантичке интерпретације. Чињеница да поменуте метрике фаворизују дословну лексичку подударност у односу на контекстуални смисао указује на императив имплементације напреднијих семантичких метрика у будућим истраживачким парадигмама.

Емпиријски резултати указују на то да претренирани whisper-large-v3 остварује супериорније почетне перформансе у поређењу са v2 варијантом. Међутим, након процеса финог подешавања, модел whisper-large-v2 демонстрирао је виши степен адаптивности и боље финалне резултате. Оваква диференција наглашава критичан

значај прецизне калибрације хиперпараметара, пре свега брзине учења и регуларизације, на процес конвергенције модела. Примењена техника мрежне претраге потврдила је своју валидност, али је истовремено отворила простор за примену софистицираних метода оптимизације, попут Бајесове оптимизације или еволутивних алгоритама.

Интеграција модула за проверу правописа резултирала је значајном редукцијом грешака, посебно на нивоу речи. Ипак, граница ефикасности ове методе директно је условљена квалитетом иницијалног акустичког препознавања. Оваква каузалност сугерише да будући правци развоја треба да теже ка дубљој интеграцији семантичких и контекстуалних језичких модела унутар самог процеса корекције.

У закључку, иако анализирани модели показују импресивне капацитете, ово истраживање поставља темељ за даљи развој евалуационих оквира који би обухватили семантичку сличност и контекстуалну релевантност. Поред унапређења самог модела, будући радови требало би да обухвате и технике претпроцесирања аудио сигнала, чиме би се додатно минимизовао утицај акустичког шума на прецизност транскрипције.

Литература

[1] Andrew Seagraves. Benchmarking top open source speech recognition models: Whisper, wav2vec 2.0, and kaldi. *Deepgram*, 2022.

[2] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends, 2024. URL https://arxiv.org/abs/2301.05712.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[4] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

[8] Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, and José Aires. Multilingual comparable corpora of parliamentary debates ParlaMint 4.0, 2023. ISSN 2820-4042. URL http://hdl.handle.net/11356/1859. Slovenian language resource repository CLARIN.SI.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[10] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.03.005. URL https://www.sciencedirect.com/science/article/pii/S1877050918302187. 1st International Conference on Natural Language and Speech Processing.

[11] Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning, 2020. URL https://arxiv.org/abs/2002.11770.

[12] Yunpeng Liu, Xukui Yang, and Dan Qu. Exploration of whisper fine-tuning strategies for low-resource asr., 2024. URL https://link.springer.com/article/10.1186/s13636-024-00349-3.

[13] Јован Турањанин. spisak-srpskih-reci. https://github.com/turanjanin/spisak-srpskih-reci, 2022.