

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Diplomski studij računarstva

Diplomski rad

**Evaluacija jezičnih modela za potrebe
audiorehabilitacije**

Rijeka, kolovoz 2025.

Luka Illich
0069087894

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Diplomski studij računarstva

Diplomski rad

**Evaluacija jezičnih modela za potrebe
audiorehabilitacije**

Mentor: prof. dr. sc. Ivo Ipšić

Komentor: pred. dr. sc. Andrea Andrijašević

Rijeka, kolovoz 2025.

Luka Illich
0069087894

Rijeka, 10.03.2025.

Zavod: Zavod za računarstvo
Predmet: Računalna obrada govora i jezika

ZADATAK ZA DIPLOMSKI RAD

Pristupnik: **Luka Illich (0069087894)**
Studij: Sveučilišni diplomski studij računarstva (1400)
Modul: Programsко инженерство (1441)
Zadatak: **Evaluacija jezičnih modela za potrebe audiorehabilitacije / Evaluation of language models for the purposes of audio rehabilitation**

Opis zadatka:

Slušne vježbe koje se često provode tijekom rehabilitacije slušanja/audiorehabilitacije mogu se sastojati od specifičnog leksičkog materijala zasićenog glasovima koje osoba s oštećenjem sluha slabije čuje i razabire. Koristeći slobodno dostupne jezične modele treniranim na velikim skupovima podataka (LLM) generirajte rečenice i tekstove koji sadrže isključivo riječi obogaćene glasovima koji zadovoljavaju specificiranim fonetskim zahtjevima. Koristeći generativne modele za tvorbu govora generirajte audio zapise tekstova sa specificiranim fonetskim zahtjevima. Automatski generirane tekstove i snimke potrebno je evaluirati i provjeriti da li zadovoljavaju specificiranim fonetskim zahtjevima.

Rad mora biti napisan prema Uputama za pisanja diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Zadatak uručen pristupniku: 21.03.2025.

Mentor:
prof. dr. sc. Ivo Ipšić

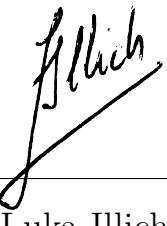
Predsjednik povjerenstva za
diplomski ispit:
prof. dr. sc. Miroslav Joler

Komentor:
pred. dr. sc. Andrea Andrijašević

Izjava o samostalnoj izradi rada

Izjavljujem da sam samostalno izradio ovaj rad.

Rijeka, kolovoz 2025.



Luka Illich

Zahvala

Zahvaljujem mentoru, prof. dr. sc. Ivi Ipšiću, te komentoru, pred. dr. sc. techn. Andrei Andrijašević, na stručnoj pomoći i korisnim savjetima pri izradi ovog diplomskog rada.

Posebnu zahvalnost dugujem svojoj obitelji na neiscrpnoj podršci, strpljenju i razumijevanju koje su mi pružali kroz cijeli studij. Njihova vjera u mene bila je temelj na kojem sam gradio svoj put. Veliko hvala i mojim prijateljima i kolegama na poticajima, razgovorima i zajedničkim trenucima koji su mi olakšali studentske dane. Na kraju, posebno zahvaljujem svojoj djevojci Ivi, koja je bila moj najveći oslonac i motivacija. Njena podrška, ohrabrenje i strpljenje dali su mi dodatnu snagu da ustrajem i završim ovaj važan korak u životu.

Sadržaj

Popis slika	x
Popis tablica	xii
1 Uvod	1
1.1 Motivacija i ciljevi rada	1
1.2 Struktura diplomskog rada	2
2 Jezični modeli za generiranje teksta i govora na hrvatskom jeziku	4
2.1 Veliki jezični modeli i njihova primjena	4
2.1.1 Povijesni razvoj i vrste jezičnih modela	5
2.1.2 Prethodna istraživanja na području audiorehabilitacije i LLM-a	6
2.2 Fonetske osobitosti hrvatskog jezika	7
2.2.1 Podjela fonema u hrvatskom jeziku	7
2.2.2 Fonemske klase korištene u radu	8
2.3 Sinteza govora	9
2.3.1 Teorijski okvir sinteze govora	9
2.3.2 Modeli sinteze u audiorehabilitaciji	10

Sadržaj

3 Zadatak i plan istraživanja	12
3.1 Definicija problema	12
3.2 Gruba specifikacija zadatka	13
3.3 Hipoteze i istraživačka pitanja	14
4 Metodologija	17
4.1 Odabir i opis jezičnih modela	17
4.1.1 Odabir modela	17
4.1.2 Model DeepSeek-R1-Distill-Qwen-32B	18
4.2 Priprema i dizajn eksperimenta	19
4.2.1 Specifikacija zadataka	19
4.2.2 Parametri izvođenja	20
4.2.3 Generiranje i obrada podataka	21
4.3 Analiza fonemske saturacije	22
4.4 Sinteza govora	23
4.5 Evaluacija rezultata	24
4.6 Računalno okruženje	25
5 Rezultati	27
5.1 Kvantitativni rezultati generiranja tekstova	27
5.1.1 Statistička analiza uspješnosti zadataka	27
5.1.2 Vrijeme izvođenja zadataka	28
5.1.3 Sažetak ključnih nalaza	30
5.1.4 Kvalitativna analiza i subjektivna procjena smisla	30
5.1.5 Analiza distribucije fonema u generiranim tekstovima	36
5.2 Rezultati statističkih testova	39
5.2.1 Utjecaj parametra MAX_NEW_TOKENS	39

Sadržaj

5.2.2	Uspješnost po fonetskim klasama i razinama saturacije	41
5.2.3	Ostale varijable	43
5.3	Rezultati sinteze govora	45
5.3.1	Opis generiranih audio zapisa	45
5.3.2	Procjena kvalitete sintetiziranog govora	47
5.3.3	Vizualizacija sintetiziranih audio zapisa	48
6	Rasprava	53
6.1	Interpretacija rezultata	53
6.1.1	Ostale varijable	54
6.1.2	Sinteza govora	56
6.2	Ograničenja provedenog istraživanja	57
6.3	Prijedlozi za buduća istraživanja i poboljšanja	57
6.3.1	Mogućnosti unapređenja TTS modela za hrvatski jezik	58
7	Zaključak	60
Bibliografija		62
Sažetak		65
A Programska okolina		67
A.1	Postupak izrade Conda/Python okruženja	67
B Dodatni materijali i repozitorij		69

Popis slika

5.1	Uspješnost generiranja po fonetskim klasama.	29
5.2	Prosječno vrijeme izvođenja po parametru MAX_NEW_TOKENS	30
5.3	Distribucija postojanja riječi na Hrvatskom jezičnom portalu.	32
5.4	Distribucija pogodnosti tekstova za sintezu govora.	35
5.5	Distribucija učestalosti fonema u generiranim riječima	37
5.6	Distribucija učestalosti fonema u generiranim rečenicama	37
5.7	Kutijasti dijagrami broja fonema po fonetskim klasama i razinama saturacije za riječi	38
5.8	Kutijasti dijagrami broja fonema po fonetskim klasama i razinama saturacije za rečenice	39
5.9	Utjecaj parametra MAX_NEW_TOKENS na završenost zadataka.	40
5.10	Uspješnost ispunjavanja fonetskih kriterija po fonetskim klasama. . .	42
5.11	Uspješnost ispunjavanja fonetskih kriterija po razinama saturacije. .	43
5.12	Distribucija vremena izvođenja po tipu zadatka i MAX_NEW_TOKENS . .	45
5.13	Distribucija generiranih audio zapisa po tipovima teksta i fonetskim klasama.	46
5.14	Uspješnost sinteze govora po fonetskim klasama.	48
5.15	Prikaz valnih oblika i spektrograma riječi - uspješni primjeri	49
5.16	Prikaz valnih oblika i spektrograma riječi - neuspješni primjeri . . .	50
5.17	Prikaz valnih oblika i spektrograma rečenica - neuspješni primjeri (1)	51

Popis slika

5.18 Prikaz valnih oblika i spektrograma rečenica - neuspješni primjeri (2) 52

Popis tablica

5.1	Statistika završenosti zadataka prema vrsti, parametrima i fonetskim karakteristikama	28
5.2	Vrijeme izvođenja završenih zadataka (u sekundama) po vrsti zadataka i parametru	29
5.3	Postojanje generiranih riječi na Hrvatskom jezičnom portalu	31
5.4	Subjektivna procjena smislenosti rečenica	31
5.5	Kategorizacija pogodnosti za sintezu govora	35
5.6	Utjecaj parametra MAX_NEW_TOKENS na završenost zadataka	40
5.7	Uspješnost ispunjavanja fonetskih kriterija po fonetskim klasama	41
5.8	Uspješnost ispunjavanja fonetskih kriterija po razinama saturacije	42
5.9	Rezultati dodatnih statističkih testova	44
5.10	Distribucija generiranih audio zapisa po vrsti teksta i fonetskim klasama	46
5.11	Rezultati subjektivne procjene kvalitete sinteze govora	47
5.12	Uspješnost sinteze govora po fonetskim klasama	48

Poglavlje 1

Uvod

1.1 Motivacija i ciljevi rada

Slušna rehabilitacija predstavlja ključni korak u procesu ponovnog usvajanja komunikacijskih vještina kod osoba s oštećenjem sluha. Kroz audiorehabilitacijske vježbe nastoji se poboljšati prepoznavanje i razumijevanje fonema — najmanjih jedinica zvuka u jeziku — koje su često posebno izazovne za ovu populaciju. U hrvatskom jeziku, kao i u drugim jezicima, određene skupine fonema mogu biti teže razlučive, osobito kod osoba koje koriste slušna pomagala ili su nedavno podvrgnute operativnom zahvatu slušnog organa. Dosadašnje metode izrade vježbi oslanjaju se na ručno odabiran leksički materijal, što postupak čini sporim, subjektivnim te često neoptimalnim u pogledu fonetske distribucije.

Razvoj velikih jezičnih modela (*Large Language Models - LLM*), treniranih na golemim korpusima tekstova, otvorio je mogućnost automatiziranog generiranja jezičnog materijala, prilagođenog unaprijed zadanim fonetskim kriterijima. Ovakvi modeli, osim brzine i prilagodljivosti, nude i potencijal za stvaranje novih, raznovrsnih vježbi koje mogu individualno odgovarati specifičnim potrebama svakog korisnika audiorehabilitacije. Osim tekstualnog materijala, napredak u području generativnih modela za sintezu govora omogućuje proizvodnju audio zapisa generiranih tekstova, čime se stvaraju cjeloviti materijali za slušne vježbe.

Glavna motivacija ovog rada proizlazi iz potrebe za objektivnim, automatiziranim

Poglavlje 1. Uvod

pristupom generiranju i evaluaciji jezičnog materijala za audiorehabilitaciju. Cilj je osigurati postupak koji će, korištenjem slobodno dostupnih alata, omogućiti izradu kvalitetnih tekstova i audio zapisa bogatih fonemima koji su od posebnog interesa za rehabilitacijski proces. Ovim radom želi se ispitati mogu li suvremeni generativni modeli ispuniti stroge fonetske zahtjeve tipične za audiorehabilitacijske vježbe, kolika je njihova objektivna uspješnost te praktična primjenjivost.

Osnovni cilj diplomskog rada je stvoriti metodologiju za automatsko generiranje tekstova i govora bogatih ciljanim fonemskim skupinama, evaluirati rezultate kroz statističke metode i subjektivnu procjenu te pružiti uvid u prednosti i nedostatke upotrebe velikih jezičnih modela i modela za sintezu govora u kontekstu audiorehabilitacije. Također, rad ima za cilj identificirati mogućnosti poboljšanja postojećih pristupa te pridonijeti dalnjem razvoju metodologije izrade rehabilitacijskih materijala na hrvatskom jeziku.

1.2 Struktura diplomskega rada

Diplomski rad sustavno i logički obuhvača sve korake istraživanja, od polazne motivacije i postavljenih ciljeva do analize i interpretacije dobivenih rezultata. Rad je podijeljen na sljedeća poglavlja.

U poglavlju 2 „Jezični modeli za generiranje teksta i govora na hrvatskom jeziku“ izlaže se teorijska podloga: prikazan je povijesni razvoj i klasifikacija jezičnih modela, sažetak ključnih istraživanja u području audiorehabilitacije, fonetske osobitosti hrvatskog jezika i klasifikacija fonema te osnovni principi sinteze govora.

U poglavlju 3 „Zadatak i plan istraživanja“ definiraju se konkretna istraživačka pitanja i hipoteze, daje gruba specifikacija zadatka te je opisan eksperimentalni plan koji obuhvača različite vrste zadataka (popisi riječi i rečenice) te razine fonemske saturacije.

Poglavlje 4 „Metodologija“ detaljno opisuje izbor i konfiguraciju jezičnih (LLM) modela i modela za sintezu govora (TTS), pripremu skupa podataka i dizajn eksperimenta, uključujući parametre izvođenja (npr. MAX_NEW_TOKENS), postupak generiranja i obrade rezultata, metodu analize fonemske saturacije, kriterije evaluacije te

Poglavlje 1. Uvod

tehničke detalje računalnog okruženja.

U poglavlju 5 „Rezultati“ prikazuju se kvantitativni ishodi generiranja teksta (uspješnost dovršenosti zadataka i vrijeme izvođenja), kvalitativna analiza uz ilustrativne primjere i subjektivnu procjenu koherentnosti, rezultati statističkih testova te ocjena kvalitete generiranih audio zapisa.

Poglavlje 6 „Rasprava“ posvećeno je interpretaciji dobivenih rezultata, analizi utjecaja parametara i različitih fonetskih klasa, raspravi o ograničenjima istraživanja i prijedlozima za unaprjedenje metodologije.

Na kraju, u poglavlju 7 „Zaključak“ sažimaju se ključni nalazi, ističe doprinos rada u području audiorehabilitacije, ocjenjuje primjenjivost razvijene metodologije i daju smjernice za buduća istraživanja.

Poglavlje 2

Jezični modeli za generiranje teksta i govora na hrvatskom jeziku

2.1 Veliki jezični modeli i njihova primjena

Jezični modeli predstavljaju srž suvremenog pristupa automatskoj obradi i generiranju prirodnog jezika. Njihov razvoj oslanja se upravo na mogućnost učenja lingvističkih uzoraka iz velikih skupova tekstualnih podataka te na sposobnost modela da generira smislen, gramatički i semantički ispravan tekst. Veliki jezični modeli, poznati kao LLM (eng. *Large Language Models*), temelje se na naprednim algoritmima dubokog učenja i sadrže milijarde parametara, što im omogućuje izuzetnu fleksibilnost i kontekstualnu razinu razumijevanja jezika.

Primjena jezičnih modela je višestruka: koriste se u sustavima za automatsku obradu upita korisnika, izradi sažetka tekstova, strojnim prijevodima, generiranju dijaloga, izradi kreativnih tekstova, edukativnim alatima, pa sve do medicinskih aplikacija. Njihovo razumijevanje jezičnih struktura omogućuje i prilagodbu na specifične uvjete, poput ciljane upotrebe fonema u tekstu, što je ključno za audiorehabilitacijske potrebe.

2.1.1 Povijesni razvoj i vrste jezičnih modela

Razvoj jezičnih modela započeo je s primjenom jednostavnih statističkih pristupa, poput n-gram modela, koji su predviđali iduću riječ u tekstu na temelju prethodnih nekoliko riječi i njihove učestalosti. Prva značajna prekretnica nastupila je početkom 2000-ih, s uvođenjem neuronskih mreža i modela temeljenih na vektorskom prikazu riječi, poznatih kao *word embeddings* (poput Word2Vec[1] i GloVe[2]). Ovi modeli omogućili su dublje semantičko razumijevanje jezika u usporedbi s ranijim pristupima.

Napredak je dodatno ubrzan pojmom dubokih neuronskih mreža, osobito s predstavljanjem transformer arhitekture (Vaswani et al., 2017)[3], koja je revolucionirala automatsku obradu jezika. Transformeri, zahvaljujući naprednoj obradi konteksta i mogućnostima paralelizacije, omogućili su razvoj velikih jezičnih modela s mili-jardama parametara, poput BERT, GPT-2, GPT-3, Llama, T5 i drugih. Moderna generacija LLM-ova stoga posjeduje sposobnost generiranja konzistentnih i tematski prilagođenih tekstova te razumijevanja složenih odnosa među riječima, kako na općoj tako i na kontekstualnoj razini.

Moderna evolucija transformera (2020.–2025.) donijela je značajne optimizacije izvorne arhitekture. Među ključnim poboljšanjima ističe se prijelaz s učenih pozicijskih kodiranja na Rotary Positional Embeddings (RoPE) [4], koji omogućuju precizniju obradu pozicijske informacije u dugim sekvencama. Multi-Head Latent Attention [5] zamijenio je standardnu multi-head pažnju zbog bolje računalne učinkovitosti, dok se SwiGLU aktivacijska funkcija [6] pokazala superiornjom od ReLU funkcije u kontekstu jezičnih modela.

Doba skaliranja (2020.–2022.) obilježeno je eksponencijalnim rastom parametara – od GPT-3 sa 175 milijardi parametara, preko PaLM-a sa 540 milijardi [7], do današnjih multimodalnih modela poput GPT-4V [8] i Google Gemini [9] koji istovremeno obrađuju tekst, slike i zvuk. Ovo razdoblje također je donijelo tehnike preciznog podešavanja učinkovitosti parametara (eng. *parameter-efficient fine-tuning*) poput LoRA-e [10], što je omogućilo široj zajednici pristup obuci velikih modela.

Najnovija generacija (2023.–2025.) uvela je modele za zaključivanje (eng. *reasoning models*) poput DeepSeek-R1 i OpenAI o1, koji koriste napredne tehnike

Poglavlje 2. Jezični modeli za generiranje teksta i govora na hrvatskom jeziku

provjere i samoprovjere. Paralelno se razvijaju i modeli za specifične domene koji postižu vrhunske rezultate u užim područjima uz manje resursa.

Danas se jezični modeli prema tehnološkoj osnovi i primjeni mogu podijeliti na:

- Klasične n-gram modele,
- Neuronske modele s vektorskim reprezentacijama riječi,
- Duboke sekvencijalne modele (RNN, LSTM),
- Modele temeljene na transformer-architekturi (npr. OpenAI GPT serija [11], BERT [12], T5 [13], Llama [14], DeepSeek-R1 [15], DeepSeek-R1-Distill-Qwen [16]),
- Precizno podešene (eng. *finetune*) modele za specifične zadatke i jezike.

Svaka od navedenih skupina donosi određene prednosti i ograničenja, no transformeri i LLM-ovi danas predstavljaju ključne alate u području generiranja i obrade prirodnog jezika.

2.1.2 Prethodna istraživanja na području audiorehabilitacije i LLM-a

Audiorehabilitacija koristi razne metode i materijale za poboljšanje sluha i prepoznavanja zvučnih signala, prvenstveno kod osoba s oštećenjem sluha. Tradicionalno, vježbe su se izrađivale ručno, uz angažman stručnjaka za fonetiku, subjekta i često ponavljane testove. Cilj je bio izraditi leksički materijal zasićen fonetskim klasama koje su problematične za konkretnu skupinu korisnika.

Jedno od istraživanja na ovom području predstavlja rad pod nazivom "Generating Speech Material for Auditory Training Exercises Using ChatGPT Chatbot" [17]. U ovom istraživanju autori su ispitali mogućnost korištenja ChatGPT modela za generiranje jezičnog materijala prilagođenog audiorehabilitaciji.

Rezultati njihovog istraživanja pokazali su da ChatGPT model može uspješno generirati izolirane riječi standardnog hrvatskog jezika za dvije ciljane fonemske klase s određenim razinama saturacije. Ovo istraživanje predstavlja važan početni korak u

Poglavlje 2. Jezični modeli za generiranje teksta i govora na hrvatskom jeziku

evaluaciji potencijala velikih jezičnih modela za stvaranje specijaliziranog materijala za slušne vježbe. Autori su potvrdili da modeli poput ChatGPT-a mogu automatizirati proces generiranja materijala prilagođenog fonetskim potrebama, značajno ubrzavajući postupak i omogućujući lakšu prilagodbu individualnim potrebama korisnika.

Usprkos izazovima — poput kontrole kvalitete i smislenosti generiranih tekstova te ograničenja u kvaliteti sintetskog govora kod hrvatskog jezika — ova početna istraživanja potvrđuju opravdanost upotrebe LLM-a u segmentu audiorehabilitacije. Zahvaljujući slobodnom pristupu velikim jezičnim modelima i modelima za sintezu govora, moguće je sustavno evaluirati kvalitativne i kvantitativne karakteristike generiranih materijala, čime se otvara nova perspektiva u razvoju individualiziranih, naprednih metoda slušnog treninga i rehabilitacije za hrvatsko govorno područje.

2.2 Fonetske osobitosti hrvatskog jezika

Fonetska struktura hrvatskog jezika obilježena je raznolikom raspodjelom glasova, odnosno fonema, koji čine temelj njegove zvukovne slike. Razumijevanje podjele i funkcije tih fonema izuzetno je važno za izradu audiorehabilitacijskih vježbi, osobito kada je cilj usmjeriti pažnju na foneme koji su kod korisnika s oštećenjem sluha teže prepoznatljivi. Kvalitetna klasifikacija fonema omogućuje precizno oblikovanje jezičnih materijala, prilagođenih specifičnim potrebama rehabilitacije.

2.2.1 Podjela fonema u hrvatskom jeziku

Hrvatski jezik svoju fonetsku sliku gradi na sustavu od trideset fonema — pojedinačnih glasova koji se u govoru razlikuju na razini izgovora, artikulacije i funkcije u jeziku. Fonemi hrvatskog jezika obuhvaćaju samoglasnike (vokale) i suglasnike (konsonante), pri čemu postoje i glasovi specifični za slavenske jezike, poput dž, nj, lj ili đ. Općenito, fonemi se dijele na nekoliko osnovnih skupina:

- Samoglasnici: /a/, /e/, /i/, /o/, /u/

Poglavlje 2. Jezični modeli za generiranje teksta i govora na hrvatskom jeziku

- Suglasmici: /b/, /c/, /č/, /ć/, /d/, /dž/, /đ/, /f/, /g/, /h/, /j/, /k/, /l/, /lj/, /m/, /n/, /nj/, /p/, /r/, /s/, /š/, /t/, /v/, /z/, /ž/.

Ovakva podjela fonema osigurava temelj za daljnje klasifikacije koje su od bitnog značaja za potrebe audiorehabilitacije, budući da omogućavaju selektivno biranje i grupiranje glasova prema fonetskim zahtjevima vježbi.

2.2.2 Fonemske klase korištene u radu

S ciljem izrade jezičnih materijala prilagođenih audiorehabilitaciji, fonemi hrvatskog jezika u ovom radu su grupirani u pet funkcionalnih klasa, prema kriteriju visine i artikulacijske specifičnosti zvuka:

- Niski (N): /m/, /n/, /nj/, /b/, /p/, /u/
- Srednjeniski (SN): /v/, /g/, /o/, /h/, /l/, /lj/
- Srednji (S): /a/, /k/, /r/, /d/, /dž/, /f/, /ž/
- Srednjevisoki (SV): /č/, /e/, /š/, /t/, /đ/, /j/
- Visoki (V): /ć/, /i/, /c/, /z/, /s/.

Ova funkcionalna podjela osmišljena je kako bi se naglasile karakteristike fonema koji su u kontekstu audiorehabilitacije posebno relevantni — bilo da se radi o glasovima koji su slabije perceptibilni kod pojedinih korisnika, ili koji su specifično zanimljivi za zadatke treniranja slušne diskriminacije — sposobnost razlikovanja sličnih fonema na temelju slušanja.

Korištenjem ovakvih fonemskih klasa može se generirati leksički materijal ciljano bogat, primjerice, visoko frekventnim ili korisniku slabije razumljivim fonemima, što je u audiorehabilitaciji presudno za učinkovit trening sluha. Modeli za generiranje teksta, kao i modeli za sintezu govora, mogu tako biti usmjereni da proizvode riječi i rečenice s traženom saturacijom određene fonemske klase, čime se omogućuje objektivna, brza i ujednačena izrada personaliziranih vježbi za korisnike s posebnim slušnim potrebama.

Ovakva klasifikacija fonema nije samo tehnička, već je i temelj za statističku

analizu uspješnosti LLM modela u zadovoljenju strogo definiranih fonetskih kriterija, što je ključno za vrijednost i znanstvenu relevantnost provedenog eksperimenta.

2.3 Sinteza govora

Sinteza govora označava tehnologiju kojom se tekstualni podaci pretvaraju u zvučni govor, čime se stvara umjetno generirani audio zapis nalik prirodnom govoru. U kontekstu audiorehabilitacije, sinteza govora zauzima osobito važno mjesto jer omogućuje brzu i fleksibilnu izradu materijala za slušne vježbe, prilagođene fonetskim potrebama korisnika s oštećenjem sluha.

2.3.1 Teorijski okvir sinteze govora

Sinteza govora temelji se na algoritmima koji iz tekstualnog zapisa stvaraju zvučne signale, oponašajući prirodnu artikulaciju ljudskog govora. Tradicionalni sustavi sinteze oslanjali su se na metode temeljene na segmentima snimljenog prirodnog govora i njihovom spajanju u nove rečenice (konkatenacija). Kasniji razvoj donio je parametarske pristupe, poput formantske sinteze ili skrivenih markovljevih modela (HMM), dok se današnja rješenja temelje na dubokom učenju, neuronskim mrežama i generativnim modelima.

Revolucija dubokog učenja, koja je započela 2016. godine, donijela je značajan napredak u kvaliteti sinteze govora. WaveNet model [18] tvrtke DeepMind prvi je koristio dilatirane konvolucijske mreže za modeliranje sirovih audio valnih oblika na razini od 16.000 do 24.000 uzoraka po sekundi. Ovaj model postigao je prosječnu ocjenu kvalitete (MOS) od 4,21 u usporedbi s 3,86 za konkatenativne sustave, što je predstavljalo značajan korak naprijed. Unatoč visokoj kvaliteti, WaveNet je bio računalno neizvodljiv za primjenu u stvarnom vremenu.

Razvoj Tacotron arhitekture [19] od strane Googlea (2017.-2018.) označio je prekretnicu u pristupu sintezi govora. Ovaj sekvencijski model uveo je mehanizam pažnje za izravnu pretvorbu znakova u mel-spektrograme. Druga verzija Tacotron-a [20] u kombinaciji s prilagođenim WaveNet vokoderom postigla je ocjenu kvalitete

Poglavlje 2. Jezični modeli za generiranje teksta i govora na hrvatskom jeziku

od 4,53, što je bilo statistički nerazlučivo od prirodnog ljudskog govora (4,58).

Era transformera u sintezi govora, koja traje od 2019. godine, donijela je modele poput FastSpeech-a i FastPitch-a [21] koji koriste transformer arhitekturu za parallelno generiranje, čime je značajno smanjena latencija obrade. Vokoderi temeljeni na generativnim suparničkim mrežama, osobito HiFi-GAN [22], omogućili su brzu sintezu visoke kvalitete. Najnoviji pristup koristi difuzijske modelle koji predstavljaju trenutnu granicu mogućnosti u generiranju prirodnog govora.

Suvremeni trendovi uključuju sintezu koja integrira emocije i prozodiju, tehnike kloniranja glasa bez prethodnog treniranja te sintezu u stvarnom vremenu za konverzacijске primjene.

Moderni sustavi sinteze, poznati kao TTS sustavi, koriste napredne arhitekture poput Tacotron-a, FastSpeech-a i WaveNet-a, sve do složenih modela treniranih na velikim skupovima podataka koji omogućuju generiranje izrazito prirodnog i eksprezivnog govora. Tipičan proces sinteze odvija se kroz nekoliko ključnih koraka:

- Pretvorba teksta u foneme i prozodijske¹ jedinice,
- Predviđanje govorne intonacije i ritma,
- Generiranje zvučnog zapisa na temelju naučenih uzoraka artikulacije².

Ključni izazovi u sintezi govora ostaju jasnoća artikulacije, prirodnost intonacije, točnost fonetskog prikaza te sposobnost prilagodbe specifičnim jezičnim zahtjevima. Posebno je važna prilagodba zadanim fonemskim klasama, što je od presudne važnosti u audiorehabilitacijskim primjenama.

2.3.2 Modeli sinteze u audiorehabilitaciji

Primjena naprednih modela za sintezu govora u audiorehabilitaciji donosi niz prednosti, kako za stručnjake, tako i za krajnje korisnike. Modeli sinteze poput speech5 ili drugih precizno podešenih neuronskih modela omogućuju generiranje govora za

¹**prozodija** - proučavanje svih značajki jezika koje utječu na stvaranje ritma i akustičkih učinaka

²**artikulacija** - položaj i funkcija govornih organa pri tvorbi i izgovoru glasova

Poglavlje 2. Jezični modeli za generiranje teksta i govora na hrvatskom jeziku

bilo koji tekst, uključujući leksički materijal s posebno odabranim fonemima važnima za rehabilitacijski proces.

Najvažnije pogodnosti sinteze govora u audiorehabilitaciji su:

- **Brza izrada individualiziranih audio materijala:** Modeli za sintezu umjetnog govora omogućuju automatizirano generiranje zvučnih snimki, čime se uklanja potreba za ručnim snimanjem i obradom.
- **Precizna fonetska kontrola:** Sinteza omogućuje stvaranje rečenica i riječi s točno zadanim raspodjelom fonema.
- **Fleksibilnost i prilagodljivost:** Moguće je izraditi vježbe za točno određene skupine korisnika, s naglaskom na njihove fonetske poteškoće.
- **Objektivnost i ponovljivost:** Generirani materijali mogu se koristiti u više navrata, u različitim kontekstima, bez varijabilnosti izvedbe.

Unatoč ovim prednostima, postoje i određena ograničenja. Kvaliteta sintetskog govora na hrvatskom jeziku često je niža nego na jezicima s većim dostupnim korpusima. Nekad su potrebne dodatne obrade ili precizno podešavanje modela za željeni jezik kako bi artikulacija, prozodija i prirodnost bile na zadovoljavajućoj razini za potrebe stvarne rehabilitacije.

Zaključno, tehnološki napredak u sintezi govora omogućuje primjenu umjetno generiranog zvuka kao učinkovite potporne metode u audiorehabilitaciji. Pravilnom integracijom LLM modela za generiranje teksta s TTS modelima za sintezu govora moguće je sustavno unaprijediti izradu, provođenje i evaluaciju vježbi za trening sluha na hrvatskom jeziku, prilagođenih specifičnim fonetskim potrebama i zahtjevima ciljane populacije.

Poglavlje 3

Zadatak i plan istraživanja

3.1 Definicija problema

Slušni poremećaji značajno utječu na sposobnost komunikacije i kvalitetu svakodnevnog života osoba koje se s njima susreću. Jedan od ključnih elemenata rehabilitacije osoba s oštećenjem sluha jest provođenje slušnih vježbi osmišljenih prema specifičnim fonetskim potrebama pojedinaca. Takve vježbe zahtijevaju korištenje jezičnog materijala bogatog fonemima koje korisnici teže raspoznavaju, što tradicionalno iziskuje velik angažman stručnjaka, ručnu pripremu tekstova i zvučnih zapisa te često neuspješno adresiranje individualnih teškoća.

Glavni izazov leži u činjenici da izrada optimalnog materijala zahtijeva preciznu fonetsku kontrolu — potrebno je izabrati, generirati te evaluirati riječi i rečenice u kojima su ciljane skupine fonema dovoljno zastupljene, a da sam tekst ostaje smislen, gramatički ispravan i primjeran svakodnevnoj upotrebi. Osim toga, manualno stvaranje i obrada govornog materijala predstavlja spor, sklon greškama te teško skalabilan proces, osobito kada je potrebno razraditi više razina vježbi za različite skupine korisnika.

Razvojem velikih jezičnih modela pojavila se mogućnost automatiziranog generiranja tekstova prema zadanim fonetskim kriterijima, dok generativni modeli za sintezu govora omogućuju izradu kvalitetnih zvučnih zapisa tih tekstova. Ipak, nepoznanica je u koliko su mjeri ti modeli sposobni pouzdano zadovoljavati stroge

Poglavlje 3. Zadatak i plan istraživanja

fonetske zahtjeve, osobito na hrvatskom jeziku, te koliko je rezultat takvog pristupa funkcionalan i kvalitetan u kontekstu rehabilitacije.

Stoga se u ovom radu postavlja problem automatizacije procesa izrade audiorehabilitacijskih vježbi korištenjem naprednih jezičnih modela i modela za sintezu govora. Potrebno je razviti i evaluirati metodologiju kojom se, uz definiran fonetski okvir, mogu automatski generirati tekstovi (riječi i rečenice) i pripadajući govorni zapisi, a zatim objektivno i subjektivno procijeniti njihovu fonetsku valjanost, smislenost te praktičnu primjenjivost.

Uz samu definiciju problema, također proizlazi i nekoliko istraživačkih pitanja:

- Mogu li suvremeni jezični modeli generirati jezični materijal s kontroliranom saturacijom fonema na razini koja zadovoljava rehabilitacijske kriterije?
- U kojoj mjeri generirani materijal održava smislenost i gramatičku ispravnost?
- Koliko su pouzdani modeli za sintezu govora u reprodukciji fonetski zahtjevnog tekstualnog materijala za hrvatski jezik?
- Koje su prednosti, ali i ograničenja, integracije takvih tehnologija u suvremenu audiorehabilitacijsku praksu?

Odgovori na navedena pitanja imaju potencijal značajno unaprijediti metodologiju audiorehabilitacije te omogućiti bržu, precizniju i individualiziranu izradu materijala za trening sluha na hrvatskom jeziku.

3.2 Gruba specifikacija zadatka

Osnovni cilj ovog diplomskog rada je razvoj i evaluacija postupka automatiziranog generiranja jezičnog materijala i govora, prilagođenog fonetskim zahtjevima audiorehabilitacije, korištenjem suvremenih alata umjetne inteligencije. Rad se temelji na integraciji velikih jezičnih modela za generiranje tekstualnih sadržaja te modela za sintezu govora u cilju izrade slušnih vježbi koje su bogate fonemima specifičnim za potrebe korisnika s oštećenjem sluha.

Specifikacija zadatka diplomskega rada obuhvača sljedeće korake:

Poglavlje 3. Zadatak i plan istraživanja

Prvo, potrebno je definirati fonetski okvir — klasifikaciju fonema hrvatskog jezika u pet funkcionalnih skupina (niski, srednjeniski, srednji, srednjevisoki, visoki fonemi) koji predstavljaju ključnu dimenziju za izradu vježbi. Na temelju ove podjele razrađuju se dva tipa jezičnih zadataka: generiranje popisa od deset riječi sa zadanim saturacijom fonema određene klase te generiranje rečenica minimalne duljine s definiranim udjelom ciljane fonemske skupine.

Zatim, računarskim putem (pomoću odabranog LLM modela) ostvaruje se generiranje tekstova prema navedenim zadacima, pri čemu se eksperimentalno varira parametar `MAX_NEW_TOKENS` i razine saturacije u traženoj fonemskoj klasi. Rezultati modela bilježe se, strukturiraju i analiziraju u smislu fonetske ispravnosti, smislenosti generiranog materijala te ispunjavanja zadane saturacije fonema. Izračunava se stvarna saturacija fonema u generiranim riječima i rečenicama te se rezultati evaluiraju i statistički obrađuju.

Za materijale koji zadovolje fonetske kriterije i smislenost, provodi se sinteza govora koristeći dostupne TTS modele prilagođene hrvatskom jeziku. Generirani audio zapisi zatim se kvalitativno procjenjuju u pogledu fonetske izgovorljivosti i razumljivosti, a prema mogućnostima i tehničkim ograničenjima modela.

Na kraju, cjelokupni postupak evaluira se kroz analizu rezultata — statističke testove i subjektivnu procjenu. Izrađuje se metodološki okvir koji može poslužiti stručnjacima za brzu, objektivnu i individualiziranu pripremu vježbi za trening sluha.

Ova gruba specifikacija zadatka pruža kronološki i logički slijed provedbe diplomskog rada, od teorijske postavke, preko računalne izvedbe i evaluacije, do razvoja preporuka za buduću primjenu automatiziranih tehnologija u audiorehabilitaciji na hrvatskom jeziku.

3.3 Hipoteze i istraživačka pitanja

Polazeći od definicije problema i grube specifikacije zadatka, u ovom istraživanju formulirane su pretpostavke o mogućnostima primjene velikih jezičnih modela i modela za sintezu govora u izradi audiorehabilitacijskih materijala na hrvatskom jeziku. Hipoteze se temelje na ranijim istraživanjima u području obrade prirodnog jezika,

Poglavlje 3. Zadatak i plan istraživanja

sinteze govora te fonetski ciljane generacije teksta, uzimajući u obzir specifične izazove koje nosi hrvatski jezik i fonološka struktura.

Temeljna hipoteza rada jest da je moguće primjenom suvremenih velikih jezičnih modela automatski generirati riječi i rečenice koje zadovoljavaju unaprijed definirane fonetske kriterije, te da se ti tekstovi mogu pretvoriti u zvučne zapise upotrebotom dostupnih modela za sintezu govora, uz razinu kvalitete prikladnu za primjenu u audiorehabilitaciji.

Iz te glavne pretpostavke proizlazi nekoliko dodatnih hipoteza: prepostavlja se da razina uspješnosti generiranih rezultata ovisi o parametrima izvođenja, prije svega o postavljenoj vrijednosti **MAX_NEW_TOKENS** i zahtijevanoj saturaciji fonema unutar ciljane klase. Nadalje, očekuje se da različite fonemske klase imaju različite razine težine za model u pogledu ispunjavanja zadanih kriterija te da je generiranje smislenih rečenica zahtjevnije od generiranja pojedinačnih riječi s istim fonetskim uvjetima. Također, postavlja se hipoteza da modeli za sintezu govora, trenirani ili prilagođeni na hrvatskom jeziku, mogu proizvesti fonetski razumljiv i prepoznatljiv govor, iako kvaliteta može varirati ovisno o složenosti i fonetskom sastavu teksta.

S ciljem provjere ovih hipoteza, u radu se formuliraju sljedeća istraživačka pitanja:

- U kojoj mjeri veliki jezični modeli mogu generirati riječi i rečenice koje zadovoljavaju strogo definiranu saturaciju fonema u zadanoj klasi?
- Kako parametri izvršavanja, poput **MAX_NEW_TOKENS** i postotka tražene fonemske saturacije, utječu na konačnu ispravnost i smislenost generiranog teksta?
- Postoje li statistički značajne razlike u uspješnosti generiranja između različitih fonemskeh klasa?
- Koja je stvarna razina saturacije fonema u generiranim rezultatima u odnosu na tražene vrijednosti, te kolika je učestalost zadovoljavanja kriterija?
- Mogu li se pomoći dostupnih modela za sintezu govora proizvesti jasni i fonetski točni slušni materijali na hrvatskom jeziku?
- Koja su ograničenja i prednosti integracije LLM i TTS tehnologija u procesu izrade audiorehabilitacijskih vježbi?

Poglavlje 3. Zadatak i plan istraživanja

Odgovori na ova pitanja trebali bi omogućiti sveobuhvatnu procjenu potencijala i praktične primjenjivosti suvremenih tehnologija umjetne inteligenecije u audiorehabilitacijskoj praksi te poslužiti kao osnova za buduća poboljšanja metodologije.

Poglavlje 4

Metodologija

4.1 Odabir i opis jezičnih modela

4.1.1 Odabir modela

U svrhu ovog istraživanja, prvi korak bio je identifikacija i odabir prikladnih velikih jezičnih modela s platforme Hugging Face[23], koja predstavlja jedno od najznačajnijih i najdostupnijih mjesto za pronalazak i testiranje modela otvorenog koda. Proces navigacije na Hugging Face portalu uključivao je filtriranje dostupnih modela prema tipu zadatka, licenci i njihovoj popularnosti. Poseban naglasak stavljen je na modele koji su kategorizirani kao "text-generation" jer je cilj bio generiranje tekstualnih sadržaja s preciznom kontrolom nad ulaznim i izlaznim podacima.

Osim tipa modela, licenca je igrala važnu ulogu u odabiru. Izabrani modeli morali su imati licencu koja omogućava slobodnu izmjenu i korištenje proizvoda, čime se osigurava legalnost istraživačkog rada i buduće primjene rezultata. Također je uzeta u obzir i veličina modela kao indikacija njegove sposobnosti, ali i dostupnost korisničkog sučelja za inicijalno testiranje modela na samoj Hugging Face platformi, što je olakšalo početnu evaluaciju bez potrebe za lokalnim računalnim resursima.

Na temelju ovih kriterija izdvojeni su modeli koji su najviše odgovarali ciljevima rada te nudili kompromis između kompleksnosti, performansi i dostupnosti: izabrani su deepseek-ai/DeepSeek-R1 i njegova derivacija DeepSeek-R1-Distill-Qwen-

Poglavlje 4. Metodologija

32B, među ostalim kandidatima.

4.1.2 Model DeepSeek-R1-Distill-Qwen-32B

Tijekom početne faze eksperimenta testirano je više jezičnih modela dostupnih putem Hugging Face platforme koristeći ugrađenu "chat" funkciju, odnosno online interaktivno sučelje za generiranje tekstova. U probiru su bili uključeni:

- deepseek-ai/DeepSeek-R1[15] (izvorni model),
- deepseek-ai/DeepSeek-R1-Distill-Qwen-32B[16] (destilirana optimizirana verzija),
- microsoft/phi-4[24],
- te razne druge destilirane varijante DeepSeek-R1 modela s manjim brojem parametara.

U ovom procesu uspoređivana je sposobnost generiranja odgovora koji zadovoljavaju zadane fonetske kriterije i specifične zahtjeve za generiranje tekstova na hrvatskom jeziku. Iako su svi navedeni modeli bili dostupni za isprobavanje kroz online chat-sučelja, detaljna evaluacija je pokazala da jedino DeepSeek-R1 i njegova destilirana verzija DeepSeek-R1-Distill-Qwen-32B pružaju dovoljno kvalitetne rezultate u kontekstu fonetske saturacije, razumijevanja zadatka i generiranja gramatički korektnih tekstova.

Nakon inicijalnog testiranja i odabira najperspektivnijih modela, započeto je s procesom lokalnog izvođenja istih na dostupnoj računalnoj konfiguraciji. Vrlo brzo se pokazalo da izvorni DeepSeek-R1 model nije moguće lokalno pokrenuti zbog tehničkih ograničenja (prevelika zahtjevnost za dostupne resurse). S druge strane, DeepSeek-R1-Distill-Qwen-32B je bio kompatibilan s dostupnim računalnim resursima te je omogućavao stabilan i kontinuiran rad. Stoga su daljnja eksperimentalna istraživanja nastavljena isključivo s destiliranom verzijom modela.

Odluka o konačnom odabiru DeepSeek-R1-Distill-Qwen-32B modela rezultat je kompromisa između tehničkih mogućnosti lokalnog izvođenja, zadovoljavajuće kvalitete rezultata i zahtjeva postavljenih istraživanjem. Takav pristup omogućio je

Poglavlje 4. Metodologija

transparentnu, ponovljivu i konzistentnu evaluaciju modela za generiranje tekstova na hrvatskom jeziku, uz osiguranje potrebne razine fonetske prilagođenosti i primjenjivosti u audiorehabilitacijskom kontekstu.

4.2 Priprema i dizajn eksperimenta

4.2.1 Specifikacija zadataka

Priprema eksperimenta započela je definiranjem preciznih zadataka za generiranje tekstualnog materijala koristeći odabrani jezični model. Za istraživanje su dizajnirane dvije osnovne vrste zadataka:

- generiranje liste od deset riječi,
- generiranje rečenice minimalne duljine od 100 znakova.

Cilj ovih zadataka bio je postići kontroliranu saturaciju fonema određene fonetske klase u generiranom tekstu, što je ključno za audiorehabilitacijske vježbe. Za svaki zadatak definirana je minimalna postotna zastupljenost fonema ciljane klase (50%, 60%, 70%, 80%, 90%). Zadaci su ponavljani za svih pet fonetskih klasa: niski, srednjeniški, srednji, srednjevisoki i visoki.

Upiti (eng. *prompts*) su standardizirani tako da je svaki zadatak imao istu kontekstualnu instrukciju (system), dok se mijenjala samo korisnička instrukcija (user) vezana uz vrstu zadatka, zahtjevanu klasu i saturaciju fonema. Ovo je omogućilo konzistentnu formulaciju zadataka i mjerjenje rezultata. Za svaku od 25 kombinacija klase i saturacije ($5 \text{ klasa} \times 5 \text{ razina saturacije}$) izvedeni su zadaci za obje vrste generiranja (rijeci i rečenice), ukupno 50 zadataka. Svaki zadatak generiran je s dvije različite vrijednosti parametra `MAX_NEW_TOKENS`, što daje ukupno 100 generiranja.

Kontekstni dio upita za obije vrste zadataka glasio je:

Ti si asistent specijaliziran za lingvistiku i fonetiku hrvatskog jezika. Tvoj zadatak je generirati riječi prema specifičnim fonetskim zahtjevima i provjeriti njihovu valjanost. Koristi svoje znanje o hrvatskom standardnom jeziku i fonemima da bi stvorio odgovarajuće

Poglavlje 4. Metodologija

riječi.

Klase fonema u hrvatskom jeziku su:

Niski (N): /m/, /n/, /nj/, /b/, /p/, /u/

Srednjeniski (SN): /v/, /g/, /o/, /h/, /l/, /lj/

Srednji (S): /a/, /k/, /r/, /d/, /dž/, /f/, /ž/

Srednjevisoki (SV): /č/, /e/, /š/, /t/, /đ/, /j/

Visoki (V): /ć/, /i/, /c/, /z/, /s/".

Upiti su se razlikovali u korisničkom dijelu. Primjer teksta za zadatak "10 riječi" je:

Napiši 10 riječi takvih da svaka riječ sadrži minimalno 50% fonema koji pripadaju klasi Visoki(V), a preostali fonemi mogu biti iz drugih klasa fonema.

te za zadatak "rečenica":

Napiši rečenicu od minimalno 100 znakova, a koja sadrži minimalno 60% fonema koji pripadaju klasi Visoki(V), a preostali fonemi mogu biti iz drugih klasa fonema.

Na ovaj način eksperimentalni dizajn omogućuje pouzdanu usporedbu sposobnosti modela da zadovolje strogo zadane fonetske kriterije, uz istu pozadinsku instrukciju i precizno kontrolirane varijacije korisničkog zahtjeva.

4.2.2 Parametri izvođenja

Ključni parametar kojim se kontrolira količina teksta koju model može generirati je MAX_NEW_TOKENS, koji određuje maksimalni broj novih tokena¹ u odgovoru modela. U okviru eksperimenta odabrane su dvije vrijednosti za ovaj parametar: 1024 i 2048 tokena. Prva vrijednost predstavljala je početnu, standardnu granicu, dok je veća

¹**token** – osnovna jedinica teksta koju model procesira; može predstavljati cijelu riječ, podriječ (slog, morfem), pojedinačni znak ili interpunkcijski simbol

Poglavlje 4. Metodologija

vrijednost uvedena kako bi se ispitalo može li povećanje broja tokena doprinijeti kvalitetnijem i potpunijem generiranju tekstova, posebice za složenije zadatke poput generiranja rečenica veće duljine.

Uz `MAX_NEW_TOKENS`, svi ostali parametri modela zadržani su na zadanim vrijednostima biblioteke `transformers` radi metodološke konzistentnosti. To uključuje *temperature*, *top-k* i *top-p* parametre, koji kontroliraju razinu varijabilnosti i kreativnosti u generiranju teksta. Budući da nisu bili eksplicitno zadani u implementaciji, korištene su njihove tvorničke postavke (npr. *top-p* = 1.0, *top-k* onemogućen), čime je zadržano zadano determinističko ponašanje modela.

Korištenje dviju različitih vrijednosti `MAX_NEW_TOKENS` omogućilo je promatranje utjecaja ovoga parametra na vrijeme izvođenja zadatka te na završnost generiranog teksta — tj. dovršava li model cijeli zadatak ili se prerano zaustavlja.

4.2.3 Generiranje i obrada podataka

Generiranje tekstova provedeno je pomoću Python skripte koja je automatizirala pokretanje odabranog modela, slanje upita te prikupljanje rezultata u zasebne tekstualne datoteke za svaki pojedini razgovor, odnosno zadatak. Svaka datoteka je sadržavala sve značajne tekstove: kontekst, korisnički upit i odgovora modela. Rezultati su zatim sustavno prikupljeni i objedinjeni u Excel tablice kako bi se omogućila detaljna kvantitativna i kvalitativna analiza.

U fazi obrade podataka izračunata je stvarna saturacija fonema željene klase u generiranim tekstovima. Ručno je također verificiran subjektivni smisao rečenica i postoje li generirane riječi na službenom jezičnom portalu, što je od velike važnosti za evaluaciju validnosti materijala. Podaci su potom kategorizirani prema ispunjenju zadane saturacije, završenosti generiranja, fonetskoj korektnosti i smislenosti, što je predstavljalo osnovu za statističku analizu.

Audio zapisi, sintetizirani za tekstove koji su zadovoljili kritične kriterije, procijenjeni su kvalitativno te su spremljeni kao dodatni materijal koji ilustrira primjenu modela za sintezu govora u kontekstu audiorehabilitacije. Cjelokupni tijek generiranja i obrade podataka osmišljen je kao reproducibilan i skalabilan, s jasno definiranim

Poglavlje 4. Metodologija

koracima koji služe kao temelj za buduća istraživanja i moguće implementacije.

4.3 Analiza fonemske saturacije

Analiza fonemske saturacije predstavlja ključni korak u evaluaciji kvalitete generiranih tekstova unutar ovog diplomskog rada. Cilj ove analize jest kvantitativno i kvalitativno ocijeniti koliko generirani tekstovi, bilo liste riječi ili rečenice, zadovoljavaju predviđene fonetske uvjete pod kojima je određena klasa fonema morala biti zastupljena u minimalno zadanoj postotnoj mjeri. Takva analiza omogućuje objektivnu provjeru uspješnosti velikih jezičnih modela u postizanju preciznih fonetskih zahtjeva koji su od posebnog značenja za audiorehabilitaciju.

Proces analize započinje fonetskom segmentacijom generiranih tekstova, gdje se svaka riječ ili rečenica rastavlja na pojedinačne foneme prema fonetskom sustavu hrvatskog jezika i definiranom klasifikacijskom okviru. Na temelju te segmentacije računa se stvarna saturacija ciljane klase fonema u odnosu na ukupni broj fonema unutar teksta. Taj se izračun uspoređuje s traženom saturacijom koja je definirana kao dio eksperimentalnog zadatka. Rezultat može biti izražen u obliku postotka stvarne prisutnosti fonemske klase te binarno, kao odgovor na pitanje zadovoljava li generirani tekst zadani kriterij (DA ili NE).

Uz kvantitativne pokazatelje, posebno se vrednuje i kvalitativna ispravnost — odnosno ima li generirani tekst smisla i odgovara li gramatičkim i semantičkim pravilima hrvatskog jezika. Za zadatke s generiranjem riječi dodatno je provjeravano postoje li generirane riječi na Hrvatskom jezičnom portalu[25], što pridonosi procjeni valjanosti jezičnog materijala. Ovaj je korak važan jer se valjanost materijala ne temelji isključivo na fonetskim kriterijima — prisutnost smislenih i gramatički pravilnih riječi značajno povećava njihovu praktičnu vrijednost u audiorehabilitacijskim vježbama.

Daljnja analiza odnosi se na kategorizaciju rezultata prema različitim varijablama: saturacija, fonemska klasa, tip zadatka (rijec ili rečenice), postavke modela (npr. `MAX_NEW_TOKENS`) te završnost generiranja. Takva segmentacija omogućuje statističku obradu i detaljniju interpretaciju dobivenih rezultata, omogućujući iden-

Poglavlje 4. Metodologija

tifikaciju parametara koji najznačajnije utječu na kvalitetu i fonetsku usklađenost generiranog jezika.

Sveukupno, analiza fonemske saturacije predstavlja osnovu za daljnje statističke testove koji potvrđuju ili opovrgavaju postavljene hipoteze o sposobnosti velikih jezičnih modela u ispunjavanju fonetskih zahtjeva. Također, uz subjektivne procjene smislenosti i postojanja riječi, ova analiza omogućuje objektivan pogled na funkcionalnu primjenu generiranih tekstova u audiorehabilitacijskim programima, čineći ovaj segment diplomskog rada centralnim dijelom evaluacije i zaključaka.

4.4 Sinteza govora

U okviru ovog diplomskog rada za sintezu govora korišten je model derek-thomas/-speecht5_finetuned_voxpopuli_hr[26], koji predstavlja napredni neuronski sustav treniran i prilagođen za generiranje prirodnog govora na hrvatskom jeziku. Ovaj model temelji se na arhitekturi transformera, poznatoj po svojoj sposobnosti "hvatanja" kontekstualnih i fonetskih odnosa u tekstualnim podacima, pružajući visoku razinu prirodnosti i jasnoće generiranog govora.

Model speecht5 posebno je dizajniran za objedinjenu obradu teksta i zvuka, omogućujući efektivnu sintezu govora iz tekstualnih ulaza, pri čemu je izvedeno precizno podešavanje na VoxPopuli datasetu, prilagođenom jezicima srednje i istočne Europe, uključujući hrvatski. Ova prilagodba modelu omogućava bolje razumijevanje i reprodukciju fonetskih značajki hrvatskog jezika, što je ključno za audiorehabilitacijske svrhe gdje je precizna artikulacija posebno važna.

Model je u radu korišten za sintezu govora na temelju generiranih tekstova koji zadovoljavaju fonetske kriterije ciljane saturacije fonema. Za tekstove koji su prošli evaluaciju smislenosti i fonetske valjanosti generiranih materijala, model je proizvodio audio zapise s jasno prepoznatljivim fonemskim obilježjima, čime je omogućena daljnja audiorehabilitacijska primjena.

Iako je model bio funkcionalan i omogućavao proizvodnju razumljivog govora, kvaliteta sintetiziranog zvuka nije u potpunosti dostizala razinu prirodnog ljudskog govora potrebnu za sveobuhvatnu audiorehabilitaciju. Ova ograničenja djelomično

Poglavlje 4. Metodologija

su posljedica manjka velikih količina hrvatskog govornog materijala dostupnog za trening, kao i tehničkih ograničenja u preciznom podešavanju modela na raspoloživim računalnim resursima.

Pokušaj dodatnog preciznog podešavanja pomoću skupa podataka disco-eth/-EuroSpeech [27] s ciljem poboljšanja kvalitete izgovora i prozodijskih karakteristika, pokazao se zahtjevan u pogledu vremena i resursa, zbog čega nije dovršen u sklopu ovoga rada. Unatoč tome, korišteni model derek-thomas/speecht5_finetuned_voxpopuli_hr predstavlja dobru osnovu za sintezu govora na hrvatskom jeziku, posebno u kontekstu automatski generiranih fonetski ciljano prilagođenih tekstova.

Zaključno, primjena modela derek-thomas/speecht5_finetuned_voxpopuli_hr u okviru diplomskog rada potvrđuje potencijal integracije suvremenih tehnologija za sintezu govora s velikim jezičnim modelima u svrhu izrade sveobuhvatnih audiorehabilitacijskih materijala, uz konstataciju da je potrebna daljnja optimizacija i istraživanje za postizanje standarda kvalitete potrebnih za kliničku primjenu.

4.5 Evaluacija rezultata

Za evaluaciju generiranih tekstualnih i audio materijala korištene su pažljivo odabранe metode koje omogućuju objektivno i sustavno vrednovanje kvalitete, fonetske ispravnosti i korisnosti rezultata. Cilj je bio procijeniti sve relevantne aspekte — od pojedinih riječi, rečenica i audio zapisa, pa do cjelokupnog procesa generiranja i sinteze.

Rezultati svakog izvođenja testova evidentirani su u dvije glavne Excel proračunske tablice. Prva tablica sadrži tehničke i izvedbene karakteristike generiranih zadataka, uključujući osnovne parametre generiranja, uspjeh zadatka i performanse modela. Druga tablica obuhvaća kvalitativne i fonetske aspekte materijala, kao što su fonemska saturacija, procjena leksičke valjanosti, smislenost rečenica i prikladnost za sintezu govora. Takva podjela omogućila je jasno razgraničenje tehničkih podataka i kvalitativnih evaluacija, čime je povećana preglednost i učinkovitost analize. Na taj način bilo je moguće precizno procijeniti u kojoj mjeri generirani tekstovi udovoljavaju postavljenim fonetskim kriterijima.

Poglavlje 4. Metodologija

Osim kvantitativnih analiza, provedena je i subjektivna kvalitativna evaluacija. Valjanost generiranih riječi provjeravana je prema hrvatskom jezičnom portalu[25], dok su smislenost i gramatička ispravnost rečenica ocjenjivane temeljem stručne prosudbe, vodeći računa o jezičnim normama i primjenjivosti u audiorehabilitacijskom kontekstu.

Procjena kvalitete audio zapisa temeljila se na subjektivnim slušnim kriterijima — jasnoći, prirodnosti, razumljivosti i fonetskoj točnosti. Takva procjena bila je ključna za identifikaciju potencijala i nedostataka primijenjenog modela za sintezu govora u stvarnim uvjetima.

Statistička evaluacija provedena je primjenom različitih testova značajnosti i korelacijskih mjera (poput χ^2 testa i Spearmanove korelacije), čime su dodatno potvrđene veze između ključnih parametara (npr. MAX_NEW_TOKENS, saturacije fonema, vrste zadatka) i izvedbe modela te su validirane postavljene istraživačke hipoteze.

Ovakva sveobuhvatna kombinacija proračunskih tablica, automatiziranih analiza, subjektivnih ocjena i statističkih procjena osigurala je preciznu i višedimenzionalnu evaluaciju generiranih materijala. Takav pristup ključan je za pouzdanu procjenu primjenjivosti rezultata u audiorehabilitacijskim procesima.

4.6 Računalno okruženje

Svi eksperimenti provedeni su na računalnom sustavu opremljenom Intel Xeon E5-2620 v4 procesorom s radnim taktom od 2.10 GHz, 128 GB radne memorije i trima grafičkim karticama NVIDIA GeForce RTX 2080 Ti. Ovakva hardverska postava omogućila je istovremeno odvijanje zahtjevnih zadataka generiranja teksta i sinteze govora te efikasnu raspodjelu računalnih resursa na više paralelnih procesa. Operacijski sustav korišten tijekom eksperimenata bio je Windows 10 Pro (verzija 1809), osiguravajući stabilno i poznato radno okruženje za izvedbu svih skripti i eksperimentalnih procesa.

Python radna okolina kreirana je primjenom Conda upravitelja paketa (verzija 4.10.3), uz instalaciju Pythona 3.11. Sve knjižnice i moduli instalirani su unutar posebnog virtualnog okruženja, čime je zajamčena izolacija od potencijalnih konfliktata

Poglavlje 4. Metodologija

s ostalim projektima i mogućnost brze reprodukcije rezultata na drugim sustavima s istom konfiguracijom.

Korišteni su sljedeći ključni Python paketi:

- **transformers** – rad s velikim jezičnim modelima, uključujući učitavanje, inferenciju² i precizno podešavanje LLM modela
- **torch, torchvision, torchaudio** – PyTorch okvir za rad s neuronским mrežama i podršku za CUDA 12.4 akceleraciju
- **datasets** – preuzimanje i obrada podataka
- **accelerate** – optimizacija izvođenja na više GPU-ova
- **bitsandbytes** – kvantizacija modela radi optimizacije korištenja memorije i bržeg izvođenja
- **sentencepiece** – tokenizacija teksta (SentencePiece)
- **soundfile** – rad s audio datotekama (SoundFile)

Postava softverskog okruženja i odgovarajući hardver omogućili su stabilno lokalno izvođenje DeepSeek-R1-Distill-Qwen-32B modela uz 4-bitnu kvantizaciju, što je značajno smanjilo memorjsko opterećenje i omogućilo bržu inferenciju teksta čak i kod zadataka koji generiraju dugo sekvencijalno izlazne podatke. Zahvaljujući dostupnosti više grafičkih kartica, proces generiranja bio je moguće dodatno parallelizirati, čime su smanjeni ukupni vremenski zahtjevi za izvođenje eksperimenata i omogućeno izvođenje višestrukih pokusa bez potrebe za kompromisima između trajanja i kvalitete rezultata.

Osim toga, radna okolina bila je u potpunosti dokumentirana kroz skriptu za automatsku instalaciju i verifikaciju svih ključnih ovisnosti, čime je osigurana ponovljivost cijelog postupka istraživanja i moguća jednostavna migracija na druge slične računalne sustave.

²**inferencija** - proces primjene prethodno istreniranog modela za generiranje ili predviđanje teksta na temelju zadanog ulaza, bez dodatnog učenja ili prilagodbe tijekom tog postupka.

Poglavlje 5

Rezultati

5.1 Kvantitativni rezultati generiranja tekstova

5.1.1 Statistička analiza uspješnosti zadataka

U ovom dijelu prikazani su kvantitativni rezultati generiranja jezičnog materijala s ciljem procjene sposobnosti modela za automatizaciju audiorehabilitacijskih vježbi. Ukupno je provedeno 100 zadataka, ravnomjerno raspoređenih između dvije vrste (generiranje liste od 10 riječi i generiranje rečenice dulje od 100 znakova), pri čemu je svaka kombinacija fonetske klase, saturacije i parametra `MAX_NEW_TOKENS` bila testirana.

Model je uspješno dovršio 36 od 100 pokušaja, što čini stopu uspješnosti od 36%. Raspodjela završenosti dodatno je prikazana u Tablici 5.1, gdje se rezultati segmentiraju po vrsti zadatka, korištenom broju maksimalnih tokena, fonetskoj klasi te traženoj razini saturacije. Kod zadatka "10 riječi" uspješnost modela iznosi 28%, a kod zadatka "rečenica" 44%. Veća vrijednost parametra `MAX_NEW_TOKENS` (2048) dovodi do boljih rezultata (46%), dok niža vrijednost (1024) bilježi tek 26% uspješno dovršenih generiranja. Veća uspješnost generiranja primijećena je kod srednje (S) klase fonema (60%), dok model ima najviše poteškoća sa klsacom visokih (V) fonema, gdje je uspješnost tek 20%.

Poglavlje 5. Rezultati

Tablica 5.1 Statistika završenosti zadataka prema vrsti, parametrima i fonetskim karakteristikama

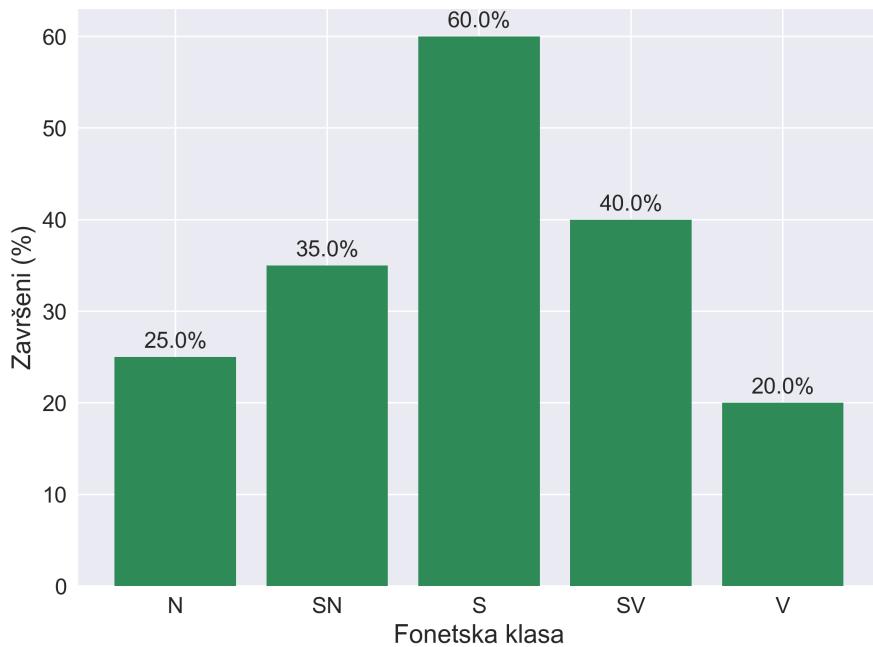
Kategorija	Završeno	Nije završeno	Ukupno
Ukupno	36 (36%)	64 (64%)	100
Vrsta zadatka			
10 riječi	14 (28%)	36 (72%)	50
Rečenica	22 (44%)	28 (56%)	50
MAX_NEW_TOKENS			
1024	13 (26%)	37 (74%)	50
2048	23 (46%)	27 (54%)	50
Fonetska klasa			
N (niski)	5 (25%)	15 (75%)	20
SN (srednjeniski)	7 (35%)	13 (65%)	20
S (srednji)	12 (60%)	8 (40%)	20
SV (srednjevisoki)	8 (40%)	12 (60%)	20
V (visoki)	4 (20%)	16 (80%)	20
Saturacija			
50%	10 (50%)	10 (50%)	20
60%	2 (10%)	18 (90%)	20
70%	6 (30%)	14 (70%)	20
80%	11 (55%)	9 (45%)	20
90%	7 (35%)	13 (65%)	20

5.1.2 Vrijeme izvođenja zadataka

Analizom podataka vremena izvođenja utvrđene su značajne razlike ovisno o postavljenim parametrima. Prosječno vrijeme generiranja s manjim brojem tokena (1024) iznosilo je 336,1 sekundu, dok je kod zadataka sa 2048 tokena generiranje trajalo 484,0 sekunde – što predstavlja povećanje od približno 44%. Standardna devijacija za veći broj tokena bila je znatno veća (174,4 sekunde), što ukazuje na varijabilnost modela s povećanim opterećenjem.

Svi navedeni rezultati mogu se izravno povezati s hipotezama iz uvodnog dijela rada: pokazuje se da povećanje broja ulaznih jedinica (tokena) utječe na uspješnost i trajanje generiranja, te da su različite fonetske klase i vrste zadataka izazovne

Poglavlje 5. Rezultati



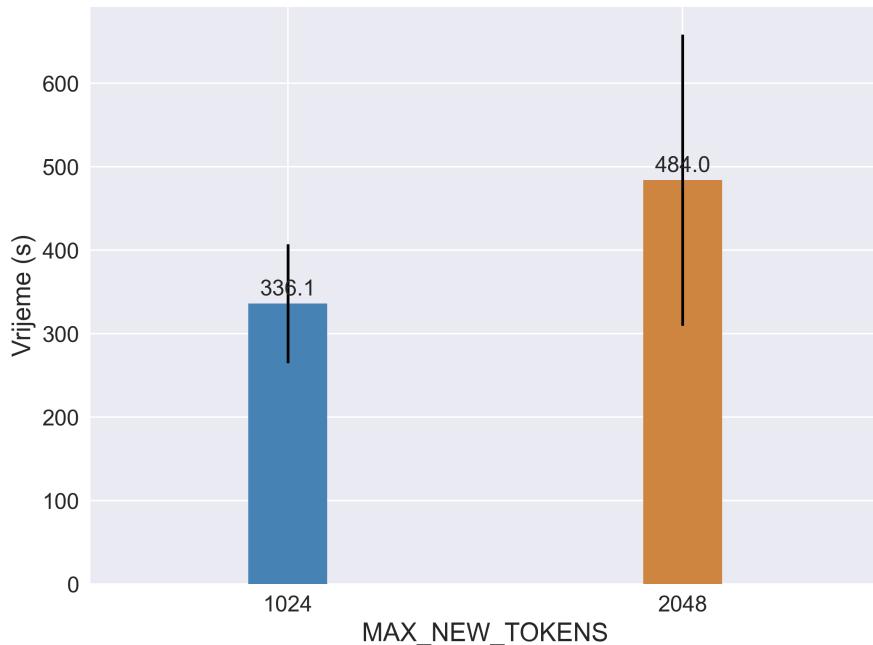
Slika 5.1 Uspješnost generiranja po fonetskim klasama.

Tablica 5.2 Vrijeme izvođenja završenih zadataka (u sekundama) po vrsti zadatka i parametru

MAX_NEW_TOKENS	Zadatak	Prosjek	Std. dev.	Broj
1024	10 riječi	385,1	22,7	4
1024	Rečenica	314,3	75,5	9
2048	10 riječi	617,5	136,1	10
2048	Rečenica	381,4	124,8	13
Ukupno 1024	Svi	336,1	71,3	13
Ukupno 2048	Svi	484,0	174,4	23

za LLM, ovisno o fonetskoj distribuciji i kompleksnosti kriterija. Ove kvantitativne analize čine temelj za kasnije interpretacije, integrirajući podatke iz tablica i grafičkih prikaza radi precizne evaluacije modela u kontekstu audiorehabilitacije.

Poglavlje 5. Rezultati



Slika 5.2 Prosječno vrijeme izvođenja po parametru MAX_NEW_TOKENS.

5.1.3 Sažetak ključnih nalaza

Model pokazuje bolju uspješnost kod kraćih zadataka (generiranje riječi) i fonema klase srednji (S), dok generiranje rečenica i fonemski zahtjevne klase visoki (V) ostaju posebno izazovne. Kvantitativna snaga analize je ograničena zadanim opsegom eksperimenta, no jasno su prikazani statistički rezultati uz napomenu da su svi nalazi interpretirani u svjetlu formuliranih hipoteza.

5.1.4 Kvalitativna analiza i subjektivna procjena smisla

Kvalitativna faza analize provedena je s ciljem procjene jezične valjanosti i praktične primjenjivosti generiranih materijala. Svaki generirani tekst podvrgnut je provjera: postojanja riječi na Hrvatskom jezičnom portalu[25], subjektivnoj evaluaciji smislenosti rečenica te pogodnosti za sintezu govora.

Poglavlje 5. Rezultati

Leksička valjanost generiranih riječi

Rezultati prikazani u Tablici 5.3 pokazuju da je tek 43,5% generiranih riječi prepoznato kao valjano u standardnom hrvatskom jeziku. Većina generiranih riječi (55,7%) su neologizmi¹ ili pseudo-rijeci koji zadovoljavaju fonetske kriterije, ali nemaju nikakvo značenje. Jedna riječ („Šetajte“) kategorizirana je kao djelomično valjana jer predstavlja gramatički ispravan oblik (imperativ drugog lica množine), iako nije eksplicitno navedena na jezičnom portalu.

Tablica 5.3 Postojanje generiranih riječi na Hrvatskom jezičnom portalu

Kategorija	Broj	Postotak
Postoji (DA)	57	43,5%
Ne postoji (NE)	73	55,7%
Djelomično valjano (NE*)	1	0,8%
Ukupno riječi	131	100,0%

Subjektivna procjena smisla rečenica

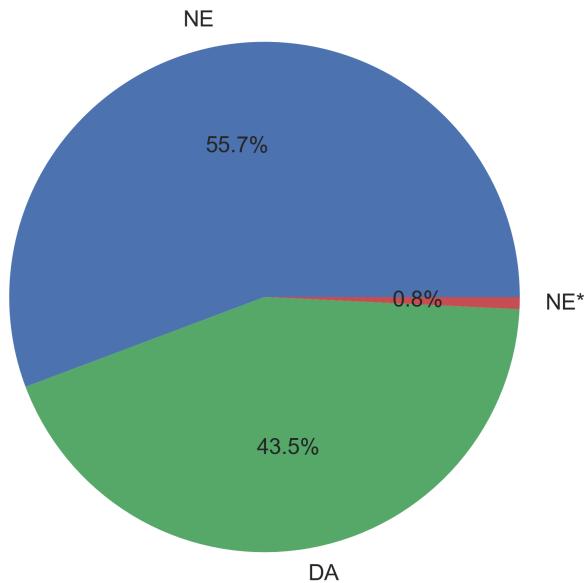
Kao što je prikazano u Tablici 5.4, od ukupno 22 generirane rečenice nijedna nije ocijenjena kao potpuno smislena na razini standardnog jezika. Dvije rečenice kategorizirane su kao djelomično smislene zbog pojave jedne neispravno generirane riječi („karakteriraju“, „studiraju“), dok je ostatak rečenice gramatički i semantički funkcionalan. Preostalih 20 rečenica nije zadovoljavalo ni gramatičke ni semantičke uvjete.

Tablica 5.4 Subjektivna procjena smislenosti rečenica

Procjena	Broj	Postotak
Nema smisla (NE)	20	90,9%
Djelomično smisлено (DA*)	2	9,1%
Potpuno smisleno (DA)	0	0,0%
Ukupno rečenica	22	100,0%

¹neologizam - ili novotvorena, novostvorena riječ ili izraz koji nije općenito prihvaćeni

Poglavlje 5. Rezultati



Slika 5.3 Distribucija postojanja riječi na Hrvatskom jezičnom portalu.

Primjeri generiranih tekstova

Da bi se bolje razumjeli praktični rezultati evaluacije, u nastavku su prikazani karakteristični primjeri generiranih tekstova koji ilustriraju različite razine uspješnosti modela u zadovoljavanju postavljenih fonetskih kriterija i jezične valjanosti.

Sljedeći primjeri predstavljaju riječi koje uspješno zadovoljavaju tražene fonetske kriterije i istovremeno postoje u standardnom hrvatskom jeziku prema Hrvatskom jezičnom portalu:

- *Mama* (klasa N, traženo 50%, postignuto 50,0%)
- *Pun* (klasa N, traženo 50%, postignuto 100,0%)
- *Voda* (klasa SN, traženo 50%, postignuto 50,0%)
- *Gol* (klasa SN, traženo 50%, postignuto 100,0%)
- *Krada* (klasa S, traženo 80%, postignuto 80,0%)

Poglavlje 5. Rezultati

Neuspješni primjeri dijele se u dvije kategorije: neologizmi koji zadovoljavaju fonetske kriterije ali ne postoje u standardnom jeziku, te postojeće riječi koje ne zadovoljavaju traženu saturaciju.

Neologizmi s ispravnom saturacijom:

- *Papuč* (klasa N, traženo 50%, postignuto 60,0%)
- *Mnog* (klasa N, traženo 50%, postignuto 50,0%)
- *Babuš* (klasa N, traženo 50%, postignuto 60,0%)

Postojeće riječi s neispravnom saturacijom:

- *Ljubav* (klasa SN, traženo 50%, postignuto 40,0%)
- *Mjesečina* (klasa N, traženo 50%, postignuto 22,2%)

Kod generiranja rečenica model se suočava s izrazito složenijim izazovom održavanja fonetskih kriterija uz semantičku koherentnost. Od ukupno 22 generirane rečenice, samo dvije zadovoljavaju traženu saturaciju fonema, a niti jedna ne postiže potpunu semantičku ispravnost.

Rečenice koje zadovoljavaju fonetske kriterije, ali nemaju smisla:

- "*Voda voda voda gol gol...*" (klasa SN, traženo 60%, postignuto 71,4%) — repetitivni niz stvarnih riječi bez semantičke vrijednosti
- "*mm, mn, mmm, mnm...*" (klasa N, traženo 90%, postignuto 97,5%) — niz glasova koji ne čine smislenu rečenicu

Djelomično smislene rečenice s gramatičkim greškama i uz nedovoljnu saturaciju tražene fonemske klase:

- "*Kad braća karakteriraju svoj karakter, često koriste razne metode analize.*" (klasa S, traženo 50%, postignuto 42,9%) — jedina greška je nepostojanje glagola "karakteriraju" dok je ostatak rečenice gramatički ispravan
- "*Na fakultetu u Zagrebu študiraju mnogi studenti iz raznih krajeva Hrvatske i drugih zemalja.*" (klasa S, traženo 70%, postignuto 29,9%) — riječ "študiraju" ne postoji u hrvatskom jeziku, trebalo bi biti "studiraju"

Poglavlje 5. Rezultati

Potpuno neuspješne rečenice (ni saturacija ni smisao):

- "*Mali rat ide kraj dubokog kašna, dok šećer i kruh padaju na stol, a šljunak i fiskalski pregled pokazivaju da je sve u redu.*" (klasa S, traženo 60%, postignuto 38,1%)
- "*Čekam šetati u času, šetnja je lijepa čast.*" (klasa SV, traženo 70%, postignuto 48,5%)

Uzroci i objašnjenja identificiranih obrazaca

Analiza pokazuje nekoliko ključnih obrazaca u funkciranju modela. Kod generiranja pojedinačnih riječi, model uspješnije balansira fonetske zahtjeve s leksičkom valjanoću kada su tražene razine saturacije umjerene (50-70%). Visoke razine saturacije (80-90%) često rezultiraju stvaranjem neologizma jer model pokušava maksimizirati udio ciljanih fonema na račun postojećih riječi u hrvatskom jeziku.

Kod rečenica, sukob između fonetskih ograničenja i semantičkih zahtjeva postaje izražena. Model često generira repetitivne nizove riječi (kao "voda gol") ili stvara nove, nepostojeće glagolske oblike ("karakteriraju", "študiraju") kako bi zadovoljio fonetske uvjete. Parametar MAX_NEW_TOKENS utječe na dovršavanje generiranja zadataka — veći broj tokena povećava stopu završenih generiranja s 26% na 46%, ali ne poboljšava kvalitetu rezultata jer duži nizovi često postaju još manje koherenti.

Ovi rezultati potvrđuju potrebu za dodatnim mehanizmima filtracije i optimizacije koji bi omogućili bolju kontrolu nad semantičkom valjanoću generiranih materijala, posebno kod složenijih tekstualnih struktura.

Kategorizacija tekstova za sintezu govora

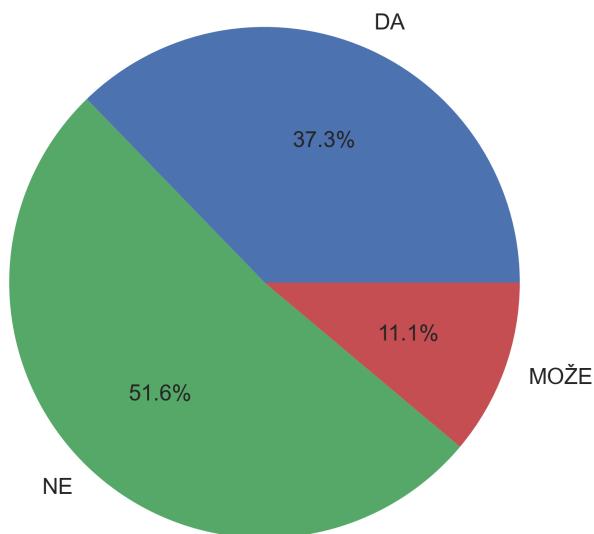
Integracijom leksičkih i semantičkih kriterija, kao i fonetske analize, tekstovi su kategorizirani prema pogodnosti za sintezu govora (Tablica 5.5). Izravno pogodni za sintezu su 37,3% materijala, dok 11,1% pripada kategoriji „uvjetno pogodno“. Tekstovi u ovoj drugoj skupini formalno ne zadovoljavaju jezičnu valjanost ili potpunu smislenost, ali sadrže elemente koji omogućuju metodološku evaluaciju funkcionalnosti

Poglavlje 5. Rezultati

TTS modela. Najveći broj, 51,6%, ne smatra se pogodnim za sintezu, prvenstveno zbog izraženih leksičkih ili semantičkih grešaka.

Tablica 5.5 Kategorizacija pogodnosti za sintezu govora

Kategorija	Broj	Postotak
Pogodno (DA)	57	37,3%
Uvjetno pogodno (MOŽE)	17	11,1%
Nije pogodno (NE)	79	51,6%
Ukupno	153	100,0%



Slika 5.4 Distribucija pogodnosti tekstova za sintezu govora.

Analiza uspješnosti po vrstama zadataka dodatno potvrđuje hipotezu iz uvoda: kod generiranja pojedinačnih riječi fonetski kriteriji zadovoljeni su u 77,1% slučajeva, dok kod rečenica udio zadovoljenih fonetskih uvjeta iznosi svega 9,1%. Ovi rezultati upućuju na važan zaključak – fonetska kontrola kod kraćih tekstualnih nizova za model je znatno lakša u odnosu na generiranje sintaktički i semantički složenih rečenica.

Poglavlje 5. Rezultati

Sažetak kvalitativnih nalaza

Kvalitativna analiza potvrdila je da automatizirana generacija tekstova za audiorehabilitaciju uspješno reproducira fonetski ciljnu saturaciju na razini izoliranih riječi, ali nailazi na ozbiljne izazove prilikom generiranja smislene rečenične strukture. Kvaliteta leksičkog materijala i subjektivna procjena smisla ostaju ključni limitirajući faktori funkcionalnosti automatiziranih pristupa u domeni audiorehabilitacije na hrvatskom jeziku.

5.1.5 Analiza distribucije fonema u generiranim tekstovima

U ovom dijelu provedena je detaljna analiza distribucije fonema u automatski generiranim riječima i rečenicama, koja omogućuje dublje razumijevanje fonetskih obrazaca koje veliki jezični model stvara pri generiranju tekstova prema zadanim kriterijima. Analiza obuhvaća učestalost pojavljivanja pojedinih fonema te distribuciju broja fonema po fonetskim klasama i razinama saturacije.

Učestalost fonema u generiranim tekstovima

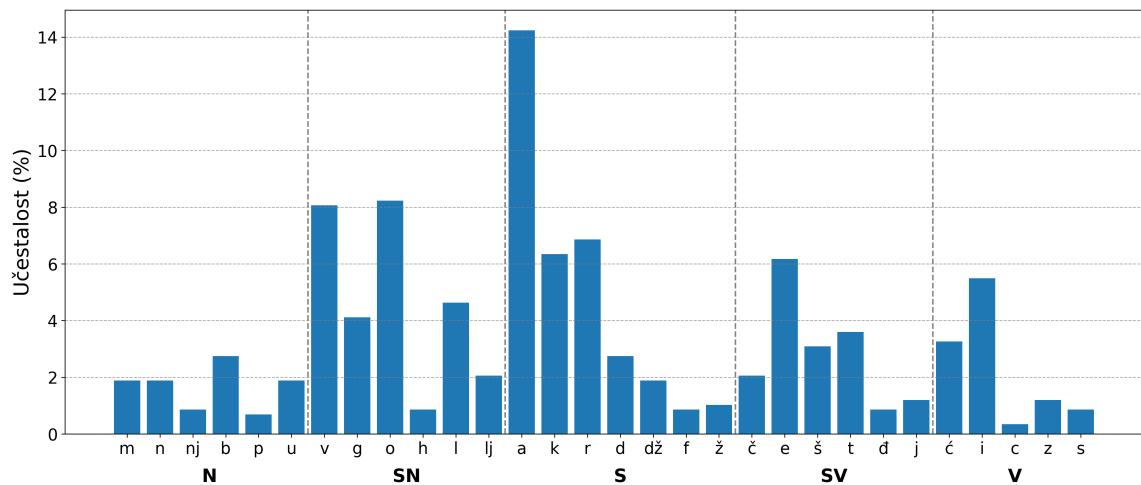
Slika 5.5 prikazuje distribuciju učestalosti svih trideset fonema hrvatskog jezika u generiranim riječima. Analiza otkriva izrazito neravnomjernu raspodjelu, gdje vokal /a/ dominira s 14,24% ukupne učestalosti, slijede ga /o/ (8,23%) i /v/ (8,06%). Ova tri najčešća fonema čine već preko 30% svih generiranih glasova. Nasuprot tome, najrjeđi fonemi poput /c/ (0,34%), /p/ (0,69%) i /f/ (0,86%) pokazuju minimalnu zastupljenost.

Kada se analizira prosječna učestalost po fonetskim klasama u riječima, klase S (srednji) i SN (srednjjeniski) postižu najviše vrijednosti (4,85% i 4,66%, respektivno), dok klase V (visoki) i N (niski) pokazuju najniže prosjeke (2,23% i 1,66%, respektivno). Ovakva raspodjela upućuje na modelovu sklonost korištenju fonema iz srednjeg spektra frekvencijskog raspona.

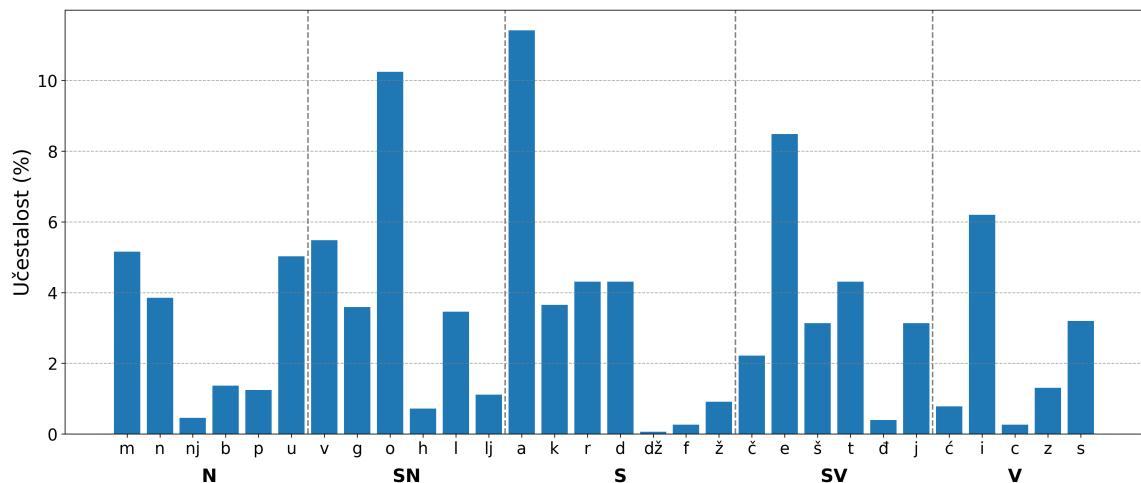
Slika 5.6 prikazuje analogne podatke za generirane rečenice. U usporedbi s riječima, rečenice pokazuju nešto uravnoteženiju distribuciju – najčešći fonem /a/ opada

Poglavlje 5. Rezultati

na 11,42%, dok se povećava zastupljenost fonema poput /m/ (5,15%) i /u/ (5,02%) koji u riječima nisu bili među prvim deset. Prosječna učestalost po klasama u rečenicama pokazuje manje ekstremne razlike: SN (4,10%), SV (3,61%), S (3,56%), N (2,85%) i V (2,35%).



Slika 5.5 Distribucija učestalosti fonema u generiranim riječima

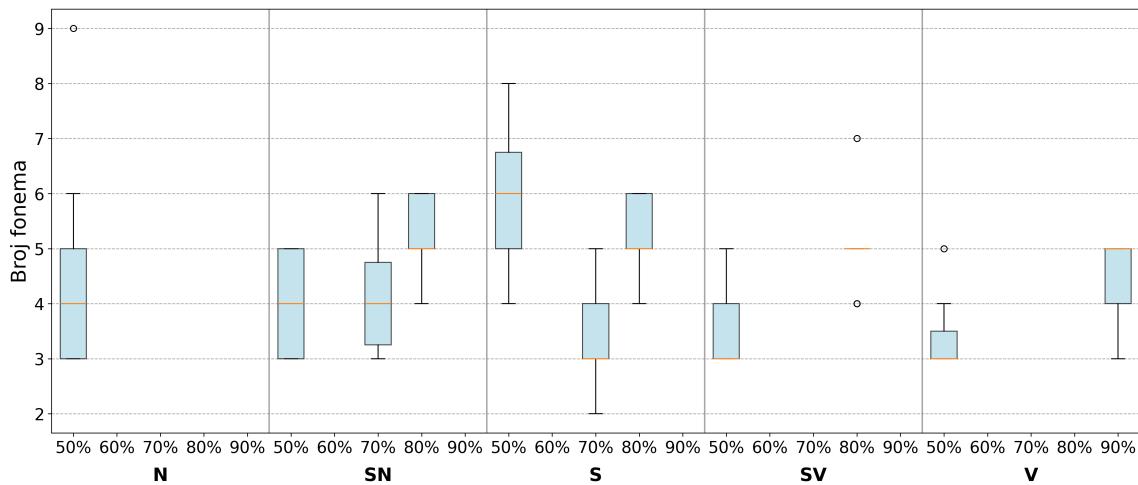


Slika 5.6 Distribucija učestalosti fonema u generiranim rečenicama

Poglavlje 5. Rezultati

Broj fonema po klasi i saturaciji

Slika 5.7 prikazuje kutijaste dijagrame broja fonema po fonetskim klasama i razinama saturacije za generirane riječi. Analiza statistika otkriva konzistentne obrasce u duljini generiranih riječi. Medijan broja fonema za većinu kombinacija kreće se između 3 i 5 fonema.

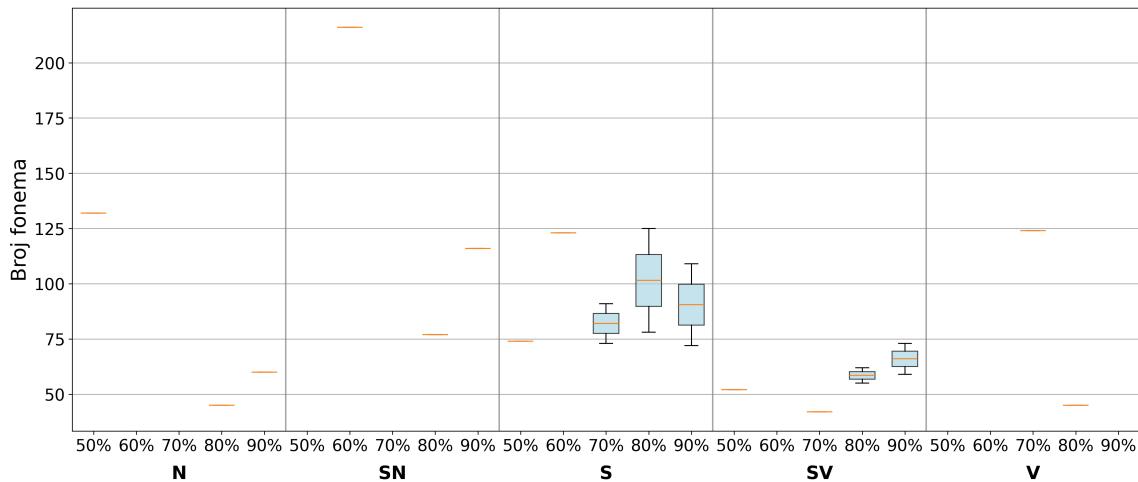


Slika 5.7 Kutijasti dijagrami broja fonema po fonetskim klasama i razinama saturacije za riječi

Slika 5.8 prikazuje distribuciju za generirane rečenice, koje pokazuju dramatično veću varijabilnost u duljini. Broj fonema po rečenici kreće se od 42 do čak 216, što odražava različite pristupe modela pri generiranju tekstova različite složenosti.

Ključni nalazi ove analize potvrđuju da model postiže bolju kontrolu nad fonetskim sastavom kod kraćih tekstualnih jedinica, dok dulje rečenice rezultiraju većom varijabilnosti i otežanim održavanjem tražene saturacije. Nelinearne veze između saturacije i duljine teksta ukazuju na složene adaptivne mehanizme koje model koristi pri pokušaju zadovoljenja fonetskih ograničenja, što ima važne implikacije za praktičnu primjenu u audiorehabilitaciji gdje su konzistentnost i predvidljivost materijala ključni faktori uspjeha.

Poglavlje 5. Rezultati



Slika 5.8 Kutijasti dijagrami broja fonema po fonetskim klasama i razinama saturacije za rečenice

5.2 Rezultati statističkih testova

U ovom dijelu prikazani su rezultati statističkih analiza koje objektivno procjenjuju utjecaj ključnih parametara, fonetskih karakteristika i tipova zadataka na uspješnost generiranja tekstova u kontekstu audiorehabilitacije. Rezultati su organizirani prema sljedećim područjima:

1. utjecaj parametra MAX_NEW_TOKENS,
2. uspješnost prema fonetskim klasama i razinama saturacije,
3. analiza drugih relevantnih varijabli.

5.2.1 Utjecaj parametra MAX_NEW_TOKENS

Rezultati analize prikazani su u Tablici 5.6. Povećanje maksimalnog broja tokena s 1024 na 2048 dovelo je do rasta uspješnosti završenih zadataka sa 26,0% na 46,0%. Ukupna uspješnost modela, kada se uzmu u obzir obje varijante, iznosi 36,0%.

Rezultati χ^2 testa ($\chi^2 = 3.516, p = 0.061$) pokazuju da utjecaj parametra MAX_NEW_TOKENS na završenost zadataka nije statistički značajan na razini tipičnog praga

Poglavlje 5. Rezultati

Tablica 5.6 Utjecaj parametra MAX_NEW_TOKENS na završenost zadataka

MAX_NEW_TOKENS	Završeno	Nije završeno	Ukupno
1024	13 (26,0%)	37 (74,0%)	50
2048	23 (46,0%)	27 (54,0%)	50
Ukupno	36 (36,0%)	64 (64,0%)	100

($p < 0.05$), premda je p-vrijednost blizu granice značajnosti. Ova činjenica ukazuje na mogućnost da bi veći uzorak mogao potvrditi statističku značajnost, kao i na praktično vidljiv pomak prema većoj uspješnosti s rastom broja generiranih tokena.

Detaljnija analiza pokazuje da veći broj tokena značajno utječe na broj generiranih riječi po zadatku ($p = 0.004$), dok se broj znakova u rečenicama ne mijenja značajno s promjenom ovog parametra ($p = 0.504$). Vizualizacija odnosa između broja novih tokena i završenosti zadatka prikazana je na slici 5.9.



Slika 5.9 Utjecaj parametra MAX_NEW_TOKENS na završenost zadataka.

Analiza vremena izvođenja pokazala je snažnu pozitivnu korelaciju s parame-

Poglavlje 5. Rezultati

trom MAX_NEW_TOKENS (Spearmanov test: $\rho = 0.583, p < 0.001$). Prosječno vrijeme generiranja zadatka s 1024 tokena iznosilo je 336,1 sekundi, dok je s 2048 tokena trajalo 484,0 sekunde. Ovi rezultati potvrđuju hipotezu da povećanje maksimalnog broja ulaznih jedinica proporcionalno povećava potrebu za računalnim resursima i produžuje trajanje generiranja.

Iz svih navedenih rezultata može se zaključiti da više vrijednosti parametra MAX_NEW_TOKENS praktično poboljšavaju dovršenost i kvantitetu generiranih riječi, uz povećanje opterećenja sustava. Statistička značajnost ovih razlika djelomično je potkrijepljena, što upućuje na potrebu za pažljivim odabirom parametara za optimalnu ravnotežu između kvalitete rezultata i tehničkih ograničenja modela.

5.2.2 Uspješnost po fonetskim klasama i razinama saturacije

Rezultati ove analize jasno pokazuju da uspješnost generiranja tekstova s traženom saturacijom fonema značajno ovisi o odabranoj fonetskoj klasi i razini saturacije. Statistički značajne razlike utvrđene su i za fonetske klase ($\chi^2 = 22.867, p < 0.001$) i za razine saturacije ($\chi^2 = 36.272, p < 0.001$), što upućuje na sustavne, a ne slučajne uzroke ovih varijacija.

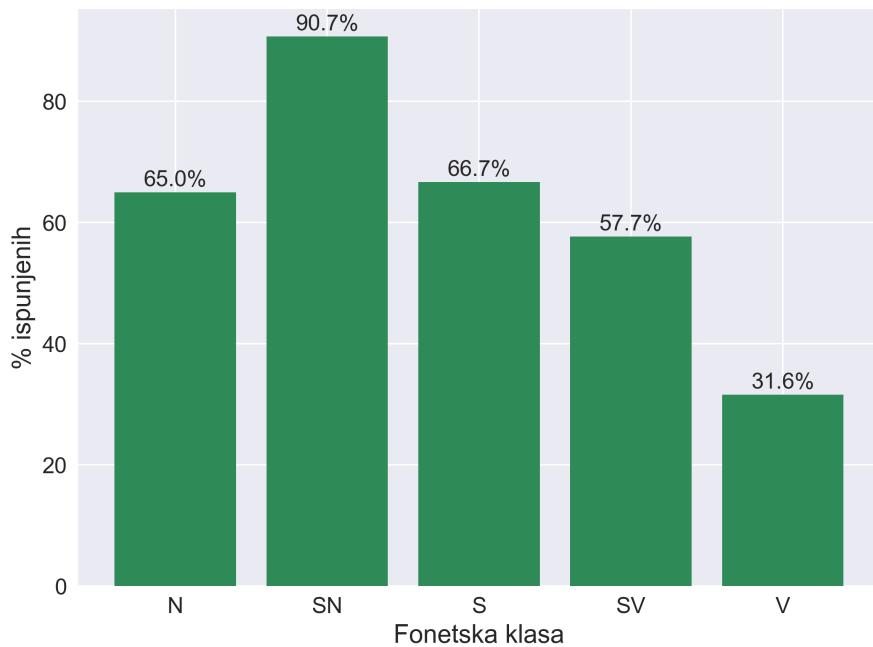
Tablica 5.7 Uspješnost ispunjavanja fonetskih kriterija po fonetskim klasama

Fonetska klasa	Uspješno	Neuspješno	Postotak uspjeha
SN (srednjjeniski)	39	4	90,7%
S (srednji)	30	15	66,7%
N (niski)	13	7	65,0%
SV (srednjevisoki)	15	11	57,7%
V (visoki)	6	13	31,6%
Ukupno	103	50	67,3%

Model je najuspješniji kod srednjjeniskih (SN) fonema (90,7%), dok najlošije rezultate postiže kod visoke (V) fonemske klase (31,6%). Ova razlika ukazuje na to da fonetske karakteristike ciljane klase direktno utječu na sposobnost modela da generira tekst koji zadovoljava stroge fonetske zahtjeve.

Daljnja analiza otkriva nelinearnu vezu između tražene razine saturacije fonema

Poglavlje 5. Rezultati



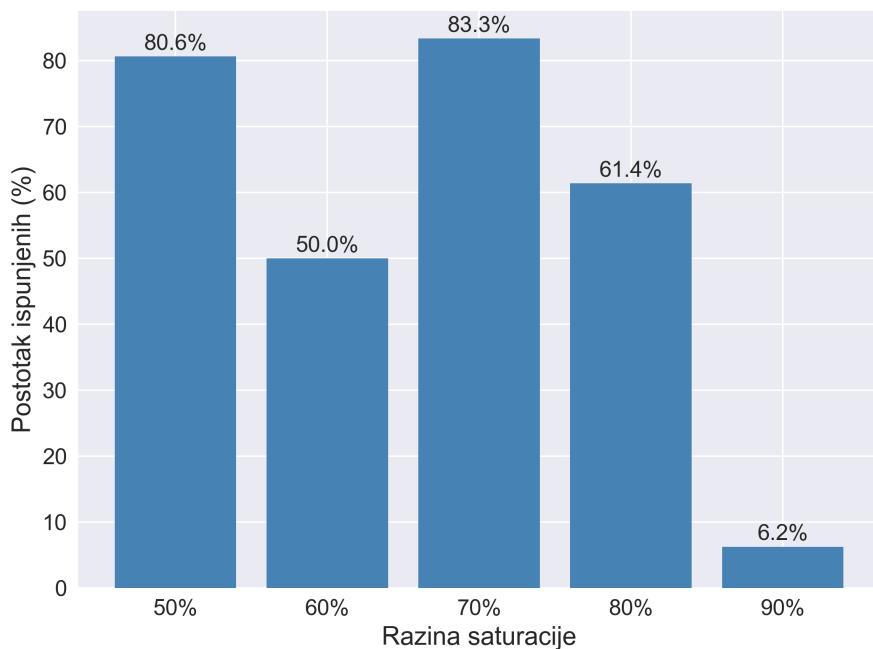
Slika 5.10 Uspješnost ispunjavanja fonetskih kriterija po fonetskim klasama.

Tablica 5.8 Uspješnost ispunjavanja fonetskih kriterija po razinama saturacije

Saturacija	Uspješno	Neuspješno	Postotak uspjeha
70%	20	4	83,3%
50%	54	13	80,6%
80%	27	17	61,4%
60%	1	1	50,0%
90%	1	15	6,2%
Ukupno	103	50	67,3%

željene klase i ukupne uspješnosti modela. Najbolji rezultati postižu se kod 50% i 70% saturacije (uspješnost 80,6% i 83,3%, respektivno), dok 60% i 80% pokazuju niže rezultate. Prema Spearmanovoj korelaciji ($\rho = -0.379, p < 0.001$), postoji značajna negativna veza između tražene saturacije i uspješnosti, što znači da zadavanje viših razina fonetske saturacije modelu predstavlja veći izazov. Kritični prag se javlja kod 90% saturacije, gdje uspješnost drastično opada – tek 6,2% tekstova zadovoljava zahtjev, što ukazuje na gubitak funkcionalnosti modela pri kombinaciji visoke saturacije

Poglavlje 5. Rezultati



Slika 5.11 Uspješnost ispunjavanja fonetskih kriterija po razinama saturacije.

i fonetskih ograničenja.

Značajna razlika utvrđena je i između tipova zadataka ($\chi^2 = 36.570, p < 0.001$). Generiranje pojedinačnih riječi postiglo je 77,1% uspješnosti, dok generiranje rečenica donosi svega 9,1% pozitivnih ishoda, potvrđujući da složenost i duljina teksta dodatno otežavaju održavanje fonetskih kriterija.

Ovi nalazi potvrđuju da fonetska klasa, razina saturacije i tip zadatka imaju snažan utjecaj na ostvarenje fonetskih kriterija, pri čemu su srednje klase i umjerene razine saturacije optimalne za generiranje tekstova primjenjivih u audiorehabilitaciji. Model se suočava s ograničenjem kod visokih zahtjeva, osobito kod složenijih zadataka poput generiranja rečenica s visokim udjelom ciljane fonemske skupine.

5.2.3 Ostale varijable

Osim osnovnih parametara, dodatne statističke analize omogućile su dublje razumijevanje međusobnih odnosa između varijabli koje utječu na kvalitetu i praktičnu

Poglavlje 5. Rezultati

vrijednost generiranih tekstova. Rezultati su sumarizirani u Tablici 5.9 i ističu nekoliko važnih uvida.

Tablica 5.9 Rezultati dodatnih statističkih testova

Statistički test	Statistika	p-vrijednost	Značajnost
χ^2 test: povezanost fonetske ispravnosti i smislenosti	0.000	1.000	Nije značajno
χ^2 test: povezanost postojanja riječi i fonetskih klasa	8.757	0.067	Nije značajno
Spearmanova korelacija: vrijeme izvođenja i saturacija	-0.100	0.323	Nije značajno
Spearmanova korelacija: vrijeme izvođenja i tip zadatka	-0.342	<0.001	Značajno

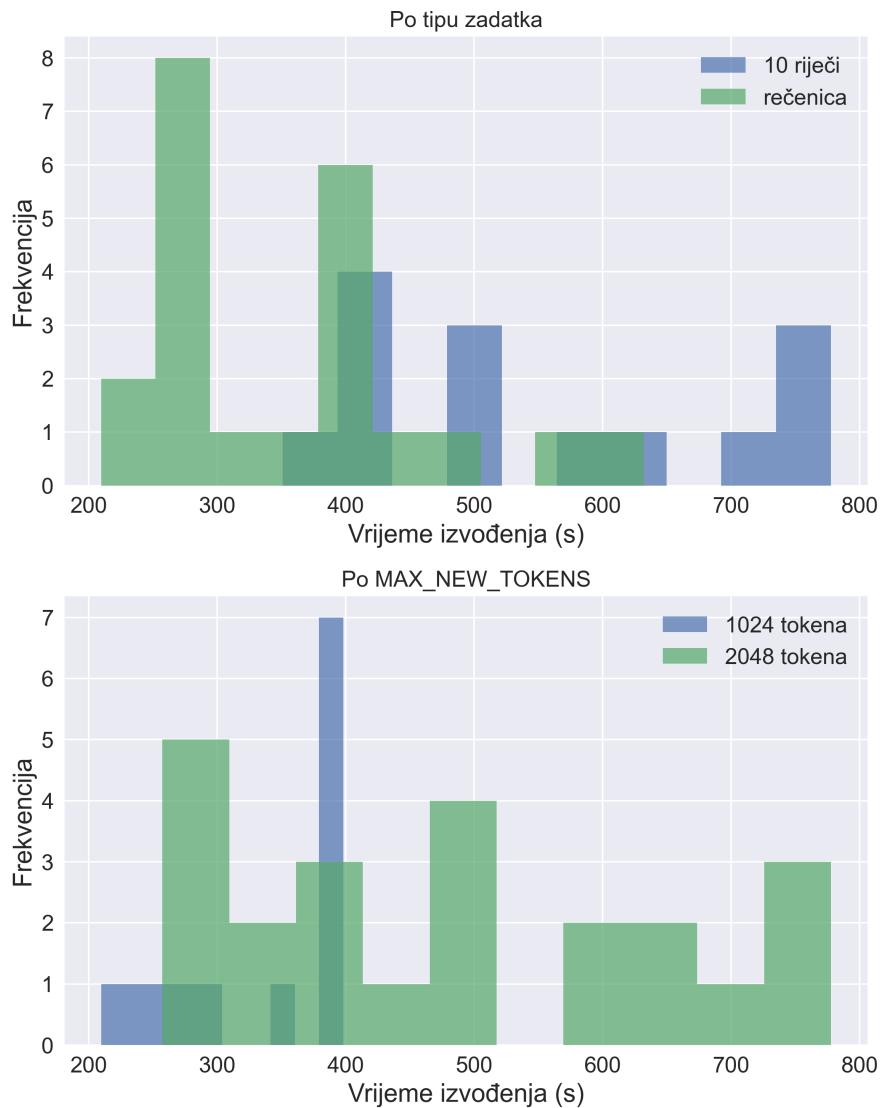
Rezultati χ^2 testa za povezanost fonetske ispravnosti i smislenosti ukazuju na to da te dvije kvalitete djeluju potpuno neovisno ($p = 1.000$), čime se opovrgava pretpostavka o njihovoj izravnoj povezanosti. Drugim riječima, model može generirati tekstove koji zadovoljavaju fonetske kriterije neovisno o smislenosti, te obrnuto.

Analiza povezanosti postojanja riječi na Hrvatskom jezičnom portalu i fonetske klase nije pokazala statističku značajnost ($p = 0.067$), iako je p-vrijednost blizu tipične granice. Takva neodređenost upućuje na mogućnost suptilnih obrazaca koje bi veći uzorak mogao detaljnije otkriti.

Spearmanova korelacija između vremena izvođenja i tražene saturacije nije značajna ($\rho = -0.100, p = 0.323$), što sugerira da složenost fonetskih zahtjeva ne utječe bitno na potrebu za računalnim resursima. Nasuprot tome, vrijeme izvođenja značajno ovisi o tipu zadatka ($\rho = -0.342, p < 0.001$): generiranje rečenica trajalo je znatno dulje od generiranja riječi, što odražava veću semantičku i sintaktičku kompleksnost dužih tekstova.

Sažeto, dodatne analize potvrđuju da su kvalitativni aspekti (fonetska korektnost i smislenost) modela neovisni, dok struktura zadatka igra ključnu ulogu u tehničkoj izvedbi i ukupnoj zahtjevnosti generiranja, što je relevantno za primjenu u audiorehabilitacijskim sustavima.

Poglavlje 5. Rezultati



Slika 5.12 Distribucija vremena izvođenja po tipu zadatka i MAX_NEW_TOKENS.

5.3 Rezultati sinteze govora

5.3.1 Opis generiranih audio zapisa

Primjenom sinteze govora generirano je ukupno 74 audio zapisa koristeći model derek-thomas/speecht5_finetuned_voxpopuli_hr[26]. Ovi zapisi obuhvaćaju dva

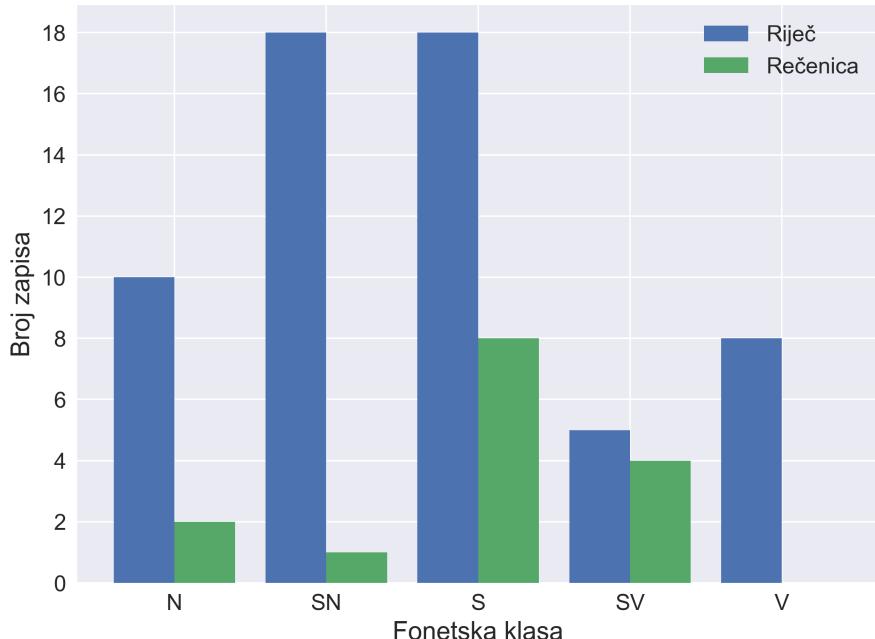
Poglavlje 5. Rezultati

tipa materijala: pojedinačne riječi (59) i rečenice (15). Prosječna duljina pojedinačne riječi bila je do 1 sekunde, dok su rečenice trajale između 4 i 11 sekundi, u skladu s duljinom generiranog teksta.

Distribucija audio zapisa po fonetskim klasama i tipu teksta prikazana je u Tablici 5.10. Najveći broj generiranih riječi pripada klasi S (20), dok su rečenice zastupljene isključivo u klasama S, SV i V.

Tablica 5.10 Distribucija generiranih audio zapisa po vrsti teksta i fonetskim klasama

Tip teksta	N	SN	S	SV	V
Riječi	12	19	20	3	5
Rečenice	0	0	6	6	3
Ukupno	12	19	26	9	8



Slika 5.13 Distribucija generiranih audio zapisa po tipovima teksta i fonetskim klasama.

Tehnički aspekti svih audio zapisa pokazuju ograničenja korištenog TTS modela. U svim snimkama zamjetan je pozadinski šum, smanjena jasnoća, kao i nekonzis-

Poglavlje 5. Rezultati

tentnosti u duljini pauza između riječi te varijacije u glasnoći koje ne odgovaraju prirodnom govoru. Ove karakteristike utječu na subjektivnu procjenu kvalitete i mogu ograničiti primjenjivost generiranih materijala u rehabilitacijskom kontekstu.

5.3.2 Procjena kvalitete sintetiziranog govora

Ocjena kvalitete sintetiziranog govora provedena je detaljnom subjektivnom procjenom svih 74 audio zapisa. Kriteriji evaluacije uključivali su prepoznatljivost izgovorenih jedinica, pravilnost naglaska te prirodnost i fonetsku točnost sukladno pravilima hrvatskog jezika.

Prema rezultatima u Tablici 5.11, svega 21,6% audio zapisa ocijenjeno je kao zadovoljavajuće kvalitete, dok je većina (78,4%) ocijenjena negativno. Analiza prema tipu teksta pokazuje izraženu razliku: dok je 27,1% audio zapisa pojedinačnih riječi ocijenjeno pozitivno, niti jedna generirana rečenica nije zadovoljila navedene kriterije.

Tablica 5.11 Rezultati subjektivne procjene kvalitete sinteze govora

Procjena	Broj zapisa	Postotak
Zadovoljavajuća kvaliteta (DA)	16	21,6%
Nezadovoljavajuća kvaliteta (NE)	58	78,4%
Ukupno	74	100,0%

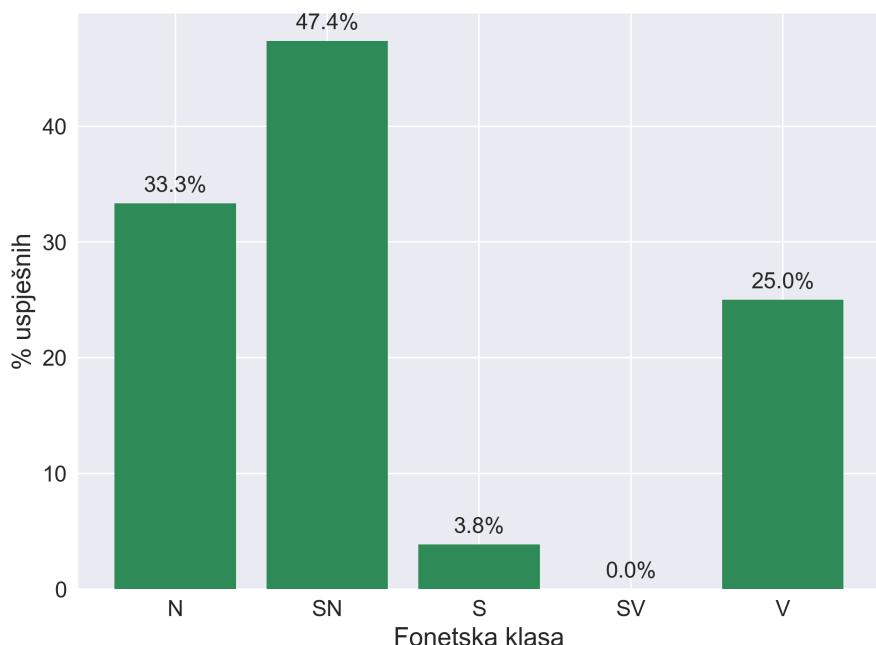
Predstavljeni rezultati po fonetskim klasama (Tablica 5.12) dodatno naglašavaju neujednačenu uspješnost modela. Najveći postotak uspješnih audio zapisa zabilježen je kod srednjjeniski fonema (SN, uspješnost 47,4%), dok su klase S i SV izrazito slabe (uspješnost 3,8% i 0,0%, respektivno).

Zaključno, kvaliteta automatski generiranih audio zapisa trenutačno je nedostatna za potrebe napredne audiorehabilitacije. Pozadinski šum, nepravilnosti u artikulaciji i izrazito loša uspješnost kod rečenica upućuju na nužnost poboljšanja i preciznog podešavanja korištenog TTS modela te obogaćivanja govornog korpusa, osobito za specifične fonetske segmente hrvatskog jezika.

Poglavlje 5. Rezultati

Tablica 5.12 Uspješnost sinteze govora po fonetskim klasama

Fonetska klasa	Zadovoljavajuće	Nezadovoljavajuće	Uspješnost
SN (srednjjeniski)	9	10	47,4%
N (niski)	4	8	33,3%
V (visoki)	2	6	25,0%
S (srednji)	1	25	3,8%
SV (srednjevisoki)	0	9	0,0%
Ukupno	16	58	21,6%

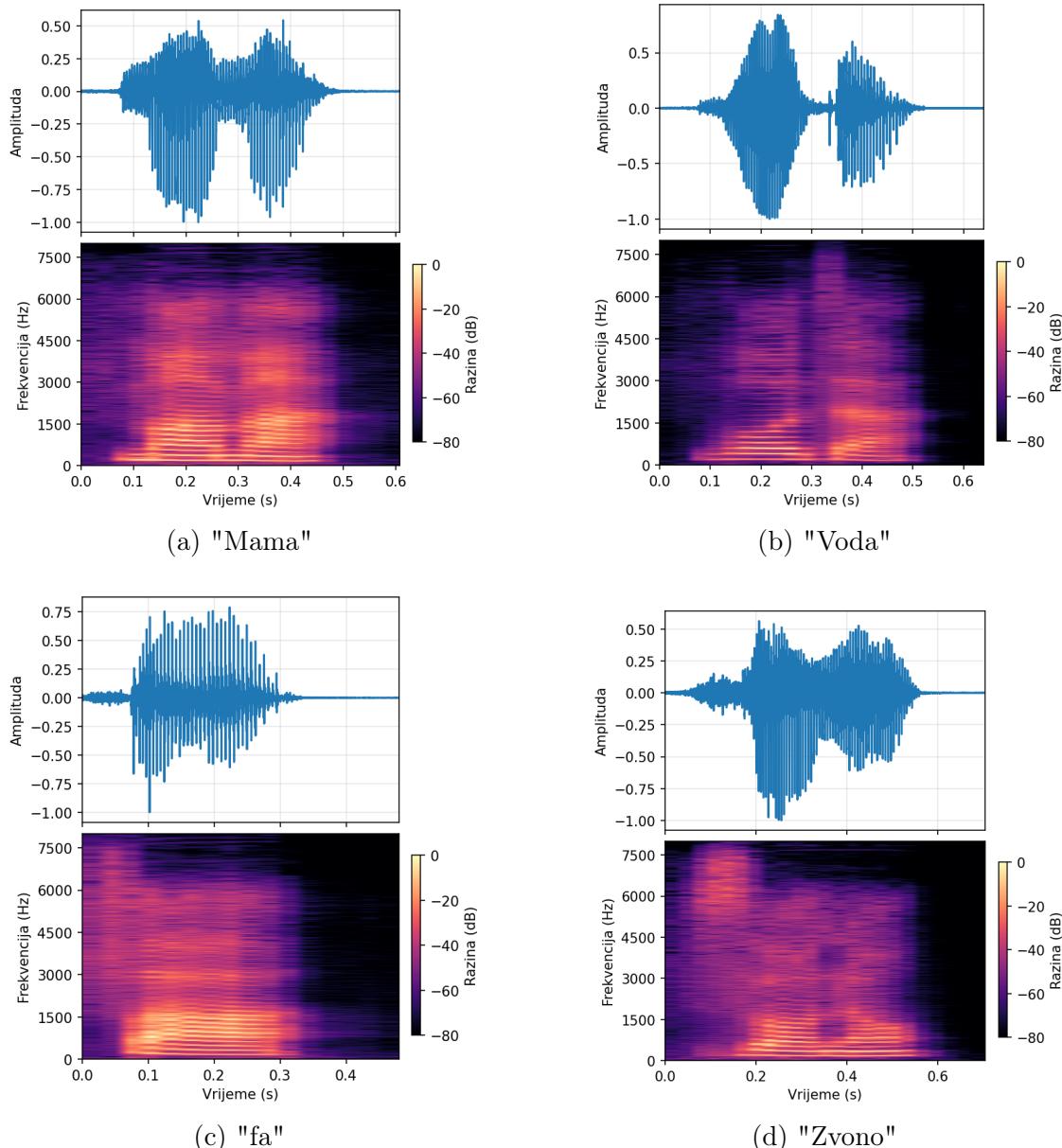


Slika 5.14 Uspješnost sinteze govora po fonetskim klasama.

5.3.3 Vizualizacija sintetiziranih audio zapisa

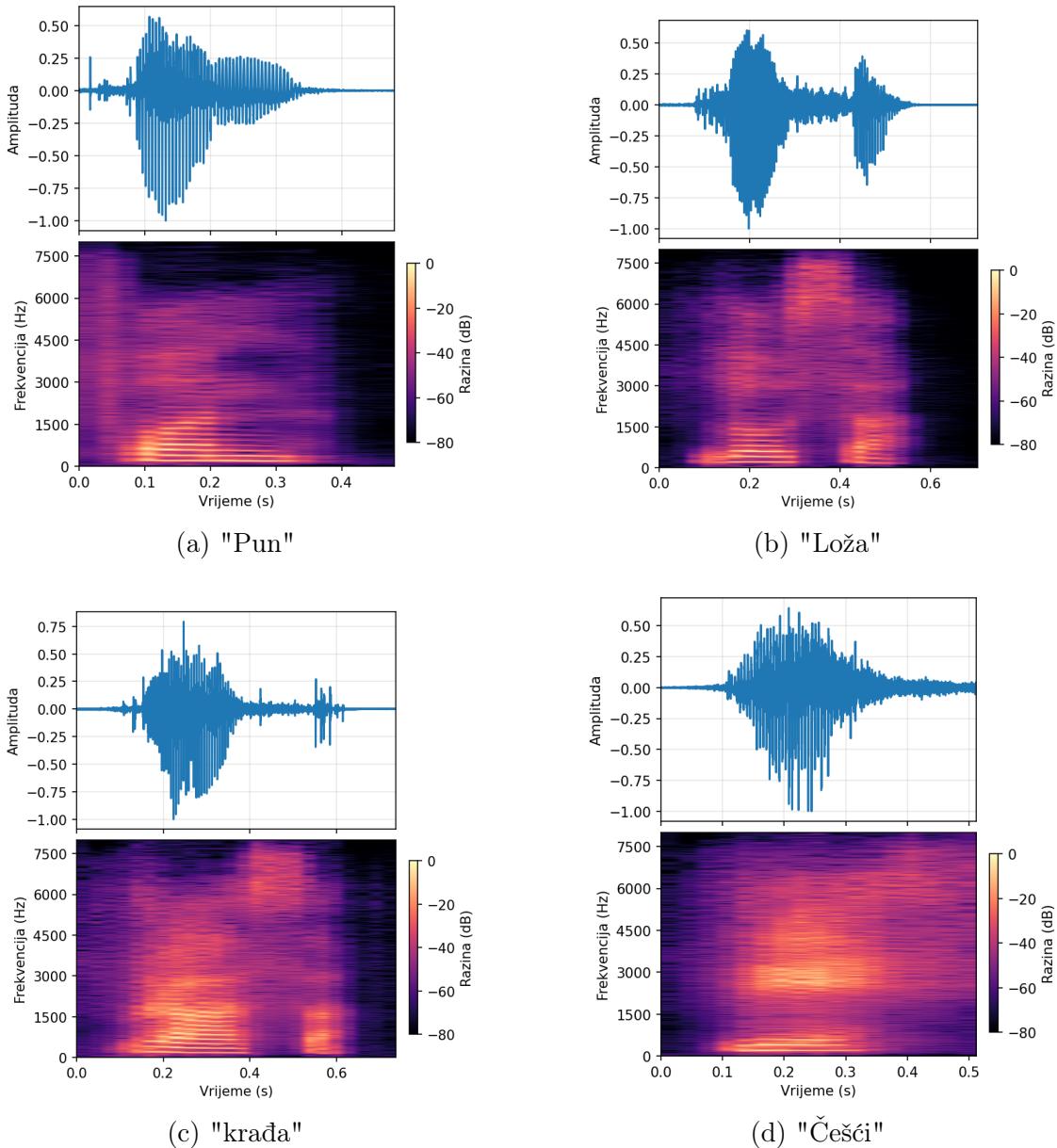
U ovom dijelu prikazani su valni oblici i spektrogrami za odabrane primjere generiranih riječi i rečenica. Svaki grafički prikaz ilustrira energetske i frekvencijske karakteristike sintetiziranog govora te omogućuje vizualnu usporedbu subjektivno zadovoljavajućih (Slika 5.15) i nezadovoljavajućih primjera (Slika 5.16, Slika 5.17, Slika 5.18).

Poglavlje 5. Rezultati



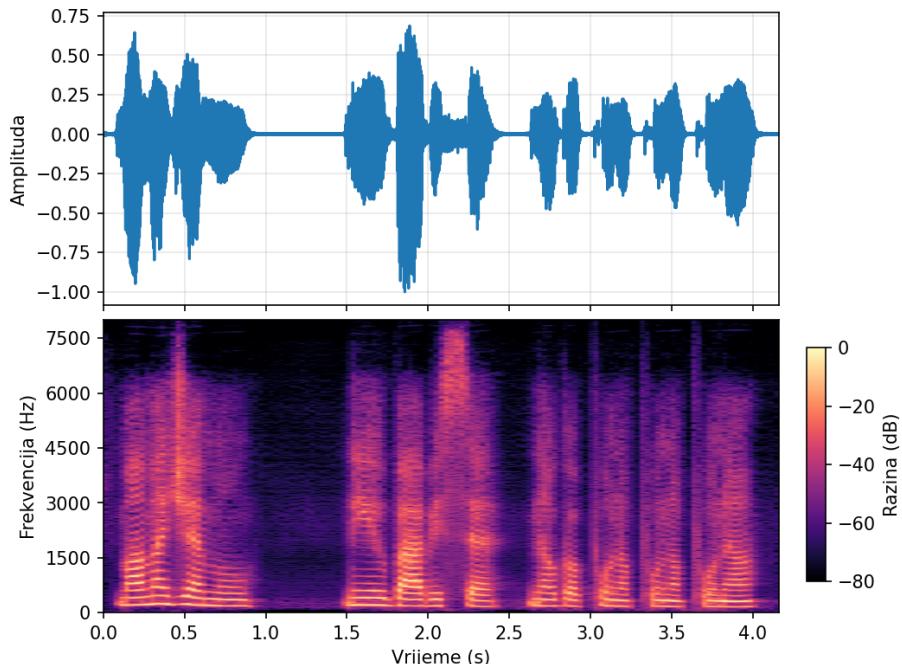
Slika 5.15 Prikaz valnih oblika i spektrograma riječi - uspješni primjeri

Poglavlje 5. Rezultati

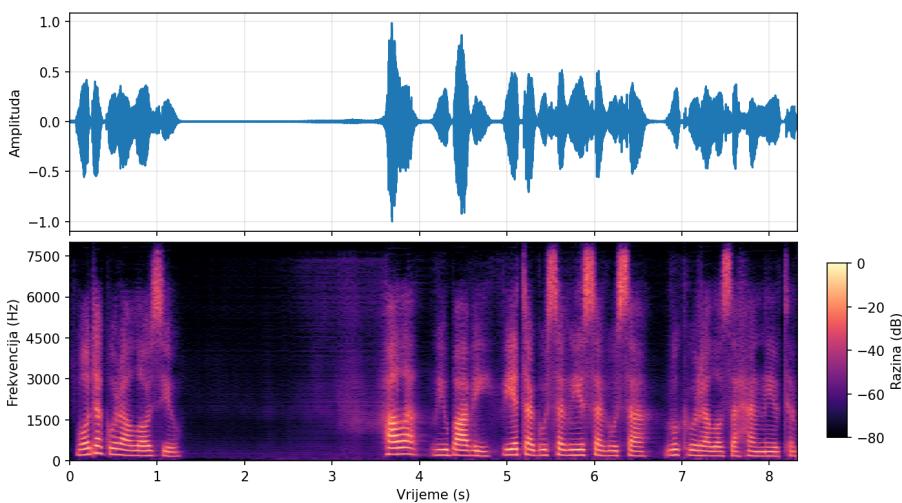


Slika 5.16 Prikaz valnih oblika i spektrograma riječi - neuspješni primjeri

Poglavlje 5. Rezultati



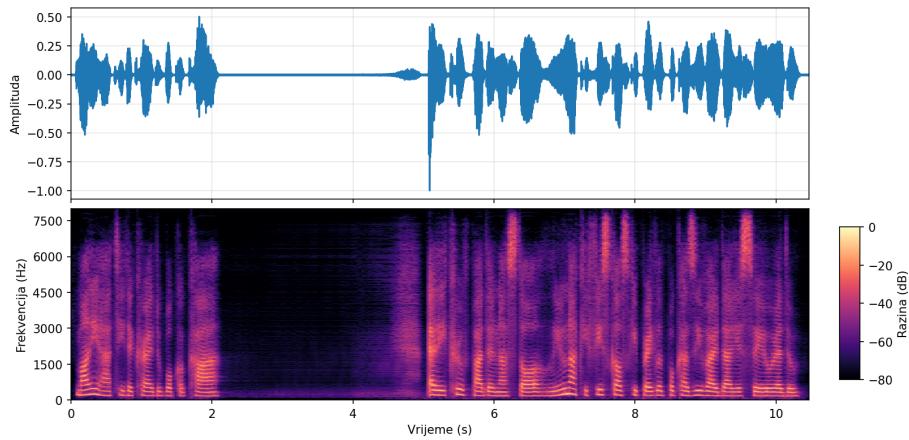
(a) "Mama je mučni u babice, mnogo puno puno puno."



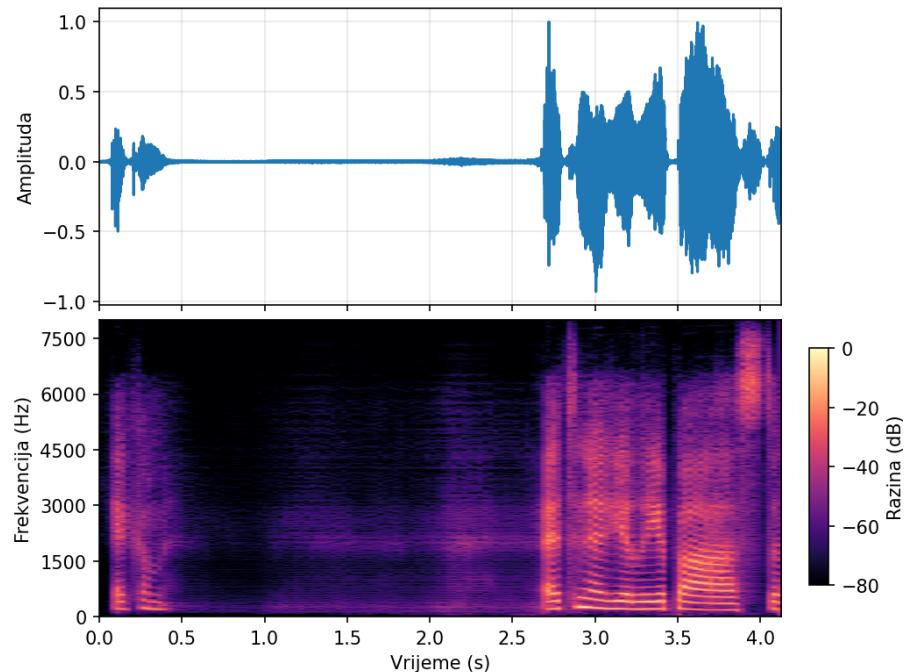
(b) "Voda gura lozju, hvala ljubavi, vjetar gusar ljeto hod, vuk gura loza hren ljut."

Slika 5.17 Prikaz valnih oblika i spektrograma rečenica - neuspješni primjeri (1)

Poglavlje 5. Rezultati



(a) "Mali rat ide kraj dubokog kašna, dok šećer i kruh padaju na stol, a šljunak i fiskalski pregled pokazivaju da je sve u redu."



(b) "Čekam šetati u času, šetnja je lijepa čast."

Slika 5.18 Prikaz valnih oblika i spektrograma rečenica - neuspješni primjeri (2)

Poglavlje 6

Rasprava

6.1 Interpretacija rezultata

Rezultati ovog istraživanja jasno pokazuju da, iako odabrani LLM model može generirati tekstove s kontroliranom saturacijom fonema, njegova sposobnost dovršavanja zadataka uvelike ovisi o parametrima generiranja i složenosti teksta. Ukupna stopa dovršenosti zadataka iznosi 36%, što potvrđuje osnovnu hipotezu o mogućnosti zadovoljavanja fonetskih kriterija, ali istovremeno ukazuje na značajna ograničenja u situacijama kada se povećaju zahtjevi – posebice kod dužih tekstova ili većeg broja tokena. Povećanje parametra `MAX_NEW_TOKENS` s 1024 na 2048 rezultiralo je porastom dovršenih zadataka s 26% na 46%, no uz prosječno vrijeme izvođenja od 336,1 sekunde do 484,0 sekunde i veću varijabilnost (povećanje standardne devijacije od 71,3 sekunde do 174,4 sekunde). Ova razmjerna veza između broja tokena i vremena obrade potvrđuje autoregresivnu prirodu generiranja teksta, gdje svaki novi token znatno povećava računski teret, što je ključno uzeti u obzir pri optimizaciji balansa između brzine i kvalitete.

Detaljna statistička analiza ističe dramatične razlike među fonetskim klasama. Srednjjeniski (SN) i srednji (S) fonemi postižu visoke stope uspješnosti (uspješnost 90,7% i 66,7%, respektivno), dok su visoki (V) fonemi znatno teži za model (uspješnost 31,6%). Ovi rezultati potvrđuju pretpostavku da akustičke i artikulacijske karakteristike fonema značajno utječu na generativnu sposobnost modela, pri čemu

Poglavlje 6. Rasprava

model pouzdanije zadovoljava fonetske zahtjeve za foneme srednjeg raspona, a slabije za foneme iz klase Visoki (V).

Analiza utjecaja tražene razine saturacije otkriva nelinearni odnos – najbolji rezultati postižu se pri 50% i 70% udjela ciljanih fonema (uspješnost 80,6% i 83,3%, respektivno), dok su razine od 60% i 80% bile manje uspješne. Ovaj neočekivani uzorak upućuje na postojanje “zona komfora” u kojima model optimalno balansira fonetske zahtjeve i leksičku raznolikost. Pojava takvih zona sugerira potrebu za dinamičkim podešavanjem razina saturacije prema specifičnoj fonemskoj klasi kako bi se maksimizirao omjer uspješnosti generiranja i kvalitete teksta.

Konačno, rezultati jasno potvrđuju da generiranje kratkih nizova (lista od 10 riječi) postiže znatno veću uspješnost fonetske kontrole (77,1%) u odnosu na generiranje rečenica (9,1%). Ovaj nalaz potvrđuje hipotezu da složenost dužeg, semantički i sintaktički povezanog teksta predstavlja znatno veći izazov za model u pogledu održavanja strogog definiranog fonetskog kriterija. Ukratko, model se pouzdano snalazi u kratkim nizovima, ali njegova sposobnost održavanja zadane saturacije dramatično opada kako raste duljina i složenost teksta.

6.1.1 Ostale varijable

Dodatne analize, prikazane u Tablici 5.9, otkrivaju važne uvide u međusobne odnose različitih aspekata performansi modela i njihove praktične implikacije za audiorehabilitaciju.

Nezavisnost fonetske ispravnosti i smislenosti

χ^2 test nije pokazao statistički značajnu povezanost između fonetske ispravnosti i smislenosti generiranih tekstova ($p = 1.000$). Ovo opovrgava pretpostavku o njihovoj uzajamnoj povezanosti i sugerira da model obrađuje fonetske i semantičke informacije kroz odvojene mehanizme. Iz praktične perspektive, pri razvoju alata za audiorehabilitaciju potrebno je zasebno optimizirati i fonetsku valjanost i semantičku koherentnost materijala.

Poglavlje 6. Rasprava

Leksička valjanost i fonetske klase

Povezanost postojanja riječi na Hrvatskom jezičnom portalu i fonetske klase nije bila statistički značajna ($p = 0.067$), iako je p-vrijednost blizu granice značajnosti. Ovaj rezultat upućuje na to da model ne favorizira pojedine fonetske klase pri odabiru postojećih riječi, ali istovremeno otvara mogućnost postojanja suptilnih obrazaca u stvaranju neologizama za specifične klase fonema.

Vrijeme izvođenja i fonetska složenost

Spearmanova korelacija između vremena izvođenja i tražene razine saturacije nije bila značajna ($\rho = -0.100, p = 0.323$). To ukazuje da povećanje fonetske složenosti materijala ne uzrokuje veće računalno opterećenje, odnosno da model fonetske uvjete provjerava učinkovito, bez dodatnog produljenja vremena generiranja.

Vrijeme izvođenja i tip zadatka

Nasuprot tome, tip zadatka pokazao je značajan utjecaj na vrijeme izvođenja ($\rho = -0.342, p < 0.001$). Generiranje rečenica traje znatno dulje nego generiranje liste riječi, što reflektira veću složenost semantičkih i sintaktičkih uvjeta kod dugih, povezanih tekstova.

Praktične implikacije

Ovi nalazi potvrđuju potrebu za dvostrukom evaluacijom u audiorehabilitacijskim alatima — zasebno za fonetsku ispravnost i semantičku koherentnost. Istovremeno, neovisnost fonetskih kriterija o računalnim troškovima omogućava primjenu zah-tjevnijih fonetskih zadataka bez ugrožavanja performansi sustava, dok je pažljivo optimiziranje tipa zadatka ključno za postizanje prihvatljivog vremena generiranja.

6.1.2 Sinteza govora

Evaluacija generiranih audio zapisa otkrila je niz ponavljanih nedostataka koji ukazuju na ograničenu prilagodbu modela hrvatskom jezičnom okruženju:

- **Neispravna prozodija i naglasak:** Primjeri jasno pokazuju neusklađenost sa standardnim prozodijskim pravilima hrvatskog jezika. Konkretno, riječ „voz“ nije pravilno naglašena s dugim /o/ (vôz), dok riječ „lova“ nije dobila pravilni naglasak na prvom slogu (lóva).
- **Utjecaj engleskog jezika:** Model pokazuje snažan utjecaj engleskog izgovora, što se manifestira u riječima poput „karakter“ (izgovoreno kao engleski *character*), „kratki“ i „dakar“, gdje se koristi tipični engleski /r/ umjesto uobičajenog hrvatskog izgovora.
- **Problemi s diakritičkim znakovima:** Fonemi označeni dijakritičkim znakovima (ć, č, ž, š, đ, dž) često se ne sintetiziraju ispravno. U nekim slučajevima, poput rečenice s riječju „studiraju“, sinteza se prekida neposredno prije izgovora glasovne jedinice „š“ te se nastavlja bez njenog izgovora.
- **Nepotpuna artikulacija:** Često se završni konsonanti riječi ne izgovaraju dovoljno jasno ili su nerazgovjetni, što je naročito uočljivo u izgovoru glasa /k/, u riječi „Vlak“.
- **Tehnički nedostaci:** Svi audio zapisi pate od pozadinskog šuma koji smanjuje jasnoću, uz neprirodne pauze i neujednačene varijacije glasnoće koje narušavaju prirodnost govora.

Ovi nalazi naglašavaju potrebu za finim podešavanjem modela za sintezu govora na skupu podataka bogatim hrvatskim govorom, s posebnim naglaskom na prozodijске i fonetske karakteristike standardnog jezika. Povećanje kvalitete govornog materijala i ciljano prilagođavanje modela ključno je za postizanje prirodnijeg, točnog i dosljednog izgovora svih relevantnih fonema u audiorehabilitacijskim materijalima.

6.2 Ograničenja provedenog istraživanja

Unatoč sveobuhvatnom pristupu i opsežnim rezultatima, ovo istraživanje ima nekoliko ključnih ograničenja koja utječu na tumačenje dobivenih rezultata i njihovu opću primjenjivost. Prije svega, obujam generiranih podataka bio je relativno ograničen—provedeno je ukupno 100 tekstualnih zadataka i 74 audio sinteze. Takav uzorak može smanjiti mogućnost detektiranja rjeđih obrazaca i ograničiti preciznost procjene varijabilnosti među fonetskim klasama i razinama saturacije.

Kvaliteta sintetiziranog govora bila je dodatno umanjena zbog nedovoljne zastupljenosti i raznolikosti hrvatskog govornog korpusa koji je korišten za precizno podešavanje TTS modela. Kao rezultat toga, u generiranim audio signalima javljali su se tehnički problemi, poput pozadinskog šuma, nepravilne prozodije, izostanaka ili pogrešnog izgovora fonema diakritičkih znakova te nerazgovjetne artikulacije završnih suglasnika.

Dizajn eksperimenta bio je ograničen na jednu (destiliranu) verziju modela DeepSeek-R1-Distill-Qwen-32B, pa rezultati ne obuhvaćaju varijacije u performansama različitih arhitektura i veličina jezičnih modela.

Na kraju, sve subjektivne ocjene smislenosti tekstova i kvalitete sintetiziranog govora proveo je jedan ocjenjivač, što povećava rizik od individualne pristranosti i smanjuje međuanalitičku pouzdanost navedenih evaluacija.

6.3 Prijedlozi za buduća istraživanja i poboljšanja

Na temelju dobivenih rezultata i identificiranih ograničenja predlaže se nekoliko smjernica za budući rad u ovom području.

Primjena više različitih jezičnih modela raznih arhitektura i veličina omogućila bi precizniju usporednu analizu performansi te olakšala identifikaciju modela optimalnih za audiorehabilitacijske zadatke.

Precizno podešavanje TTS modela na skupu podataka obogaćenim hrvatskim govornim snimkama, posebno s prozodijskim i fonetskim anotacijama, potrebno je kako bi se poboljšala kvaliteta izgovora, točnost naglaska i dosljednost artikulacije

Poglavlje 6. Rasprava

svih fonema, a posebno onih koji predstavljaju dijakritičke znakove.

Razvoj automatiziranih procedura za istovremenu kontrolu fonetske saturacije i semantičke valjanosti tijekom generiranja teksta smanjio bi potrebu za ručnom filtracijom i ubrzao pripremu relevantnih materijala.

Primjena naprednijih metoda za prikazivanje i obradu kategorijskih podataka, kao i uključivanje više nezavisnih ocjenjivača u kvalitativnim procjenama, povećali bi pouzdanost statističkih analiza i smanjili rizik od pristranosti pri ocjeni smislenosti i kvalitete sinteze.

Istraživanja koja se usmjere na dinamičko podešavanje razine fonetske saturacije unutar „zona komfora“ modela mogla bi definirati pragove pri kojima se postiže najbolji omjer između uspješnosti generiranja i prirodnosti materijala.

Provjeda ovih prijedloga omogućila bi razvoj robusnijih, skalabilnih i visoko prilagođenih rješenja za automatizirano generiranje i sintezu govora u svrhu audiorehabilitacije na hrvatskom jeziku, čime bi se dodatno unaprijedila metodologija i povećala praktična primjenjivost automatiziranih LLM i TTS modela u audioreabilitacijskim vježbama.

6.3.1 Mogućnosti unapređenja TTS modela za hrvatski jezik

Posebnu pozornost zahtjeva pitanje daljnog unapređenja modela za sintezu govora namijenjenih hrvatskom jeziku. Trenutačno korišteni model derek-thomas/-speecht5_finetuned_voxpopuli_hr treniran je na višejezičnom skupu podataka Vox-Populi, koji uključuje i hrvatski jezik. Iako takav pristup omogućuje općenitiju primjenu, on može ograničiti kvalitetu sinteze specifičnih jezičnih obilježja hrvatskog jezika.

Ciljano precizno podešavanje TTS modela isključivo na govornom materijalu na hrvatskom jeziku moglo bi znatno poboljšati kvalitetu gorovne sinteze kroz nekoliko ključnih aspekata. Prvo, omogućilo bi preciznije modeliranje fonemskih specifičnosti — točniju artikulaciju fonema kao što su /č/, /ć/, /ž/, /š/, /đ/ i /dž/, koji su jezično specifični i zahtijevaju pažljivu obradu. Drugo, unaprijedilo bi prozodijske karakteristike: hrvatski jezik ima složen sustav naglaska koji uključuje četiri različita tipa

Poglavlje 6. Rasprava

(dugoustilazni, dugouzlazni, kratkouzlazni i kratkosalazni), što je teško adekvatno obuhvatiti višejezičnim modelom.

Korištenje govornoga korpusa jednoga izvornoga govornika hrvatskoga jezika omogućilo bi dosljedniju artikulaciju, prirodniji prozodijski tok, manju varijabilnost u izgovoru te lakše prilagodbe za specifične klase fonema. S druge strane, pristup koji uključuje više govornika ponudio bi veću robusnost modela, mogućnost izbora glasa te bolju generalizaciju na različite gorone stilove.

Stoga se za optimizaciju TTS modela u audiorehabilitacijskim primjenama preporučuje sustavna usporedba sljedećih pristupa:

1. precizno podešavanje postojećeg višejezičnog modela na korpusu jednoga govornika hrvatskoga jezika;
2. treniranje novog modela na većem korpusu isključivo hrvatskoga jezika s više govornika i fonetskim oznakama;
3. hibridni pristup koji integrira prednosti navedenih metoda.

Takav istraživački postupak mogao bi znatno unaprijediti praktičnu primjenjivost automatiziranih sustava za sintezu govora u rehabilitaciji sluha na hrvatskom jeziku, čime bi se ostvario važan korak prema klinički relevantnim i jezično prilagođenim rješenjima.

Poglavlje 7

Zaključak

Ovaj rad predstavlja sustavnu analizu primjene suvremenih velikih jezičnih modela (LLM) i modela za sintezu govora (TTS) u području audiorehabilitacije na hrvatskom jeziku, s naglaskom na automatsko generiranje i evaluaciju fonetski ciljano oblikovanih tekstova i audio materijala. Rezultati jasno potvrđuju da odabrani LLM model može generirati tekstove s kontroliranom saturacijom ciljanih fonema, što je ključno za razvoj specijaliziranih slušnih vježbi koje odgovaraju fonetskim potrebama korisnika s oštećenjem sluha. Ipak, ukupna stopa dovršenosti zadatka iznosi 36%, što ukazuje na ograničene sposobnosti modela u okolnostima s povećanom složenošću i zahtjevima, osobito kod generiranja smislenih rečenica i tekstova većih dimenzija.

Analiza prema fonetskim klasama pokazala je značajne razlike u uspješnosti, pri čemu su srednjjeniski i srednji fonemi generirani s visokom preciznošću, dok su fonemi iz skupine visoki znatno teži za model. Na razini saturacije, optimalni rezultati postižu se kod umjerenih razina (50% i 70%), dok ekstremne razine, posebice 90%, značajno otežavaju generiranje kvalitetnog materijala. Uspješnost generiranja više je izražena za kratke nizove riječi nego za duže, semantički i sintaktički povezane rečenice, što ukazuje na izazove održavanja simultane fonetske kontrole i smislenosti u složenim tekstovima.

Procjena kvalitete sinteze govora pokazala je da je trenutna razina izvedbe odabranog TTS modela zadovoljavajuća u manje od četvrtine slučajeva, s primjetnim nedostacima poput pozadinskog šuma, nepravilne prozodije, pogreške u izgovoru fo-

Poglavlje 7. Zaključak

nema predstavljenim dijakritičkim znakovima te nerazgovjetne artikulacije završnih suglasnika. Ovi nalazi ukazuju na važnost ciljane nadogradnje i preciznog podešavanja modela sinteze na hrvatskom govornom korpusu, posebno s obzirom na prozodij-ske i fonetske specifičnosti jezika.

Istražene su i dodatne varijable, pri čemu nije utvrđena statistička povezanost između fonetske ispravnosti i smislenosti, što sugerira odvojenost njihovih mehanizama u modelu, te su potvrđene razlike u vremenu izvođenja ovisno o složenosti zadatka. Ova saznanja naglašavaju potrebu za dvostrukom evaluacijom u audiorehabilitacijskim sustavima, uz zasebnu optimizaciju fonetskih i semantičkih aspekata.

Unatoč važnim rezultatima, rad sadrži određena ograničenja, uključujući relativno mali uzorak generiranih podataka, ograničenja u dostupnosti i raznolikosti govornog korpusa za precizno podešavanje TTS modela, korištenje samo jedne destilirane verzije LLM modela, te subjektivnu evaluaciju provedenu jednim ocjenjivačem.

S obzirom na to, rad predlaže niz smjernica za buduća istraživanja: proširenje broja modela s različitim arhitekturama i veličinama parametara, daljnje tehnike preciznog podešavanja TTS modela na kvalitetnim hrvatskim govorima, razvoj automatiziranih procedura za simultanu kontrolu fonetske i semantičke valjanosti, uvođenje naprednih statističkih metoda za obradu kategorijskih varijabli te korištenje više nezavisnih ocjenjivača u kvalitativnoj evaluaciji. Također, istraživanja bi trebala detaljnije analizirati optimalne zone fonetske saturacije za pojedine skupine modela.

Zaključno, ovaj rad doprinosi razvoju metodologije za automatizirano generiranje i sintezu audiorehabilitacijskih materijala na hrvatskom jeziku, pružajući temelj za učinkovitiju i prilagođeniju primjenu umjetne inteligencije u području slušne rehabilitacije, ali istovremeno ukazuje na izazove koje treba premostiti za postizanje pune kvalitete i funkcionalnosti sustava.

Bibliografija

- [1] T. Mikolov *et al.* (2013) Word2vec. Google. , s Interneta, <https://code.google.com/archive/p/word2vec/> Pриступљено: 12. kolovoza 2025.
- [2] (2014) Glove: Global vectors for word representation. Stanford NLP Group. , s Interneta, <https://nlp.stanford.edu/projects/glove/> Pриступљено: 12. kolovoza 2025.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30. NeurIPS, 2017. , s Interneta, <https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- [4] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, “Roformer: Enhanced transformer with rotary position embedding,” 2023. , s Interneta, <https://arxiv.org/abs/2104.09864>
- [5] DeepSeek-AI and et al., “Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model,” 2024. , s Interneta, <https://arxiv.org/abs/2405.04434>
- [6] N. Shazeer, “Glu variants improve transformer,” 2020. , s Interneta, <https://arxiv.org/abs/2002.05202>
- [7] A. C. et al., “Palm: Scaling language modeling with pathways,” 2022. , s Interneta, <https://arxiv.org/abs/2204.02311>
- [8] “Gpt-4v(ision) system card,” 2023. , s Interneta, <https://api.semanticscholar.org/CorpusID:263218031>
- [9] G. T. et al., “Gemini: A family of highly capable multimodal models,” 2025. , s Interneta, <https://arxiv.org/abs/2312.11805>

Bibliografija

- [10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021. , s Interneta, <https://arxiv.org/abs/2106.09685>
- [11] (2025) Openai – gpt models. OpenAI. , s Interneta, <https://openai.com> Pristupljen: 12. kolovoza 2025.
- [12] (2025) Bert – bidirectional encoder representations from transformers. Google Research. , s Interneta, <https://github.com/google-research/bert> Pristupljen: 12. kolovoza 2025.
- [13] (2025) T5: Text-to-text transfer transformer. Google Research. , s Interneta, <https://github.com/google-research/text-to-text-transfer-transformer> Pristupljen: 12. kolovoza 2025.
- [14] (2025) Llama (large language model meta ai). Meta AI. , s Interneta, <https://ai.meta.com/llama> Pristupljen: 12. kolovoza 2025.
- [15] deepseek-ai. (2025) Deepseek-r1. deepseek-ai. , s Interneta, <https://huggingface.co/deepseek-ai/DeepSeek-R1> Pristupljen: 12. kolovoza 2025.
- [16] ——. (2025) Deepseek-r1-distill-qwen-32b. deepseek-ai. , s Interneta, <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B> Pristupljen: 12. kolovoza 2025.
- [17] A. Andrijašević and B. Vukelić, “Generating speech material for auditory training exercises using chatgpt chatbot,” in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, Croatian Society for Information, Communication and Electronic Technology. Opatija, Croatia: IEEE, May 2024, paper presented at MIPRO 2024.
- [18] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” 2016. , s Interneta, <https://arxiv.org/abs/1609.03499>
- [19] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” 2017. , s Interneta, <https://arxiv.org/abs/1703.10135>
- [20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” 2018. , s Interneta, <https://arxiv.org/abs/1712.05884>

Bibliografija

- [21] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech: Fast, robust and controllable text to speech,” 2019. , s Interneta, <https://arxiv.org/abs/1905.09263>
- [22] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” 2020. , s Interneta, <https://arxiv.org/abs/2010.05646>
- [23] (2025) Hugging face hub. Hugging Face Inc. , s Interneta, <https://huggingface.co> Pristupljeno: 12. kolovoza 2025.
- [24] microsoft. (2025) phi-4. Microsoft Research. , s Interneta, <https://huggingface.co/microsoft/phi-4> Pristupljeno: 12. kolovoza 2025.
- [25] (2025) Hrvatski jezični portal. Znanje, Novi Liber i Srce. , s Interneta, <https://hjp.znanje.hr/> Pristupljeno: 11. kolovoza 2025.
- [26] derek thomas. (2022) speecht5_finetuned_voxpopuli_hr. Hugging Face. , s Interneta, https://huggingface.co/derek-thomas/speecht5_finetuned_voxpopuli_hr Pristupljeno: 12. kolovoza 2025.
- [27] disco-eth. (2025) Eurospeech: A large-scale multilingual speech corpus. DISCO ETH, Hugging Face. , s Interneta, <https://huggingface.co/datasets/disco-eth/EuroSpeech> Pristupljeno: 12. kolovoza 2025.

Sažetak

Ovaj rad bavi se primjenom suvremenih velikih jezičnih modela (LLM) i modela za sintezu govora (TTS) u području audiorehabilitacije na hrvatskom jeziku. Cilj je razviti i evaluirati metodologiju automatiziranog generiranja tekstova i audio materijala koji su fonetski ciljano prilagođeni potrebama korisnika s oštećenjem sluha. Provedena su eksperimentalna istraživanja u kojima su generirani tekstovi s kontroliranom saturacijom fonema različitih klasa, uz varijacije parametra `MAX_NEW_TOKENS` i razine saturacije. Rezultati pokazuju da model uspješno generira kratke nizove s visokim udjelom ciljane fonemske klase, dok je generiranje smislenih složenijih rečenica znatno zahtjevnije. Evaluacija sinteze govora ukazuje na ograničenja trenutnih TTS modela, uključujući pozadinski šum i problematičnu artikulaciju, što naglašava potrebu za dalnjim finim podešavanjem. Analize dodatnih varijabli ukazuju na odvojenost fonetske ispravnosti i smislenosti te na važnost optimizacije zadataka radi učinkovitog generiranja. Zaključno, rad pruža temelj za razvoj skalabilnih, automatiziranih audiorehabilitacijskih rješenja prilagođenih hrvatskom jeziku, istovremeno naglašavajući izazove i smjernice za buduća istraživanja.

Ključne riječi — audiorehabilitacija, veliki jezični modeli, LLM, generiranje teksta, fonetska saturacija, sinteza govora, TTS, hrvatski jezik, fonetska evaluacija, automatizacija, umjetna inteligencija

Abstract

This thesis explores the application of state-of-the-art large language models (LLMs) and text-to-speech (TTS) synthesis models in the field of auditory rehabilitation for the Croatian language. The objective is to develop and evaluate a methodology for the automated generation of text and audio materials that are phonetically tailored to meet the needs of individuals with hearing impairments. Experimental studies were conducted in which texts with controlled phoneme saturation across different phoneme classes were generated, varying parameters such as `MAX_NEW_TOKENS` and saturation levels. Results demonstrate that the model successfully generates short sequences with a high proportion of the targeted phoneme class, while producing

Sažetak - Abstract

meaningful and complex sentences remains significantly more challenging. The evaluation of speech synthesis reveals limitations of current TTS models, including background noise and articulation issues, highlighting the need for further fine-tuning. Analyses of additional variables indicate a dissociation between phonetic accuracy and semantic coherence, emphasizing the importance of task optimization for effective generation. In conclusion, this work lays the foundation for developing scalable, automated auditory rehabilitation solutions adapted to the Croatian language, while underscoring challenges and directions for future research.

***Keywords* — auditory rehabilitation, large language models, LLM, text generation, phoneme saturation, speech synthesis, TTS, Croatian language, phonetic evaluation, automation, artificial intelligence**

Dodatak A

Programska okolina

A.1 Postupak izrade Conda/Python okruženja

Za reproducibilnost istraživanja, potrebno je uspostaviti odgovarajuće Python okruženje s potrebnim bibliotekama. Sljedeći koraci omogućuju kreiranje funkcionalnog okruženja za rad s velikim jezičnim modelima i TTS sustavima:

1. Otvoriti Anaconda Prompt
2. Pozicionirati se u željeni direktorij
3. Stvoriti novo Conda okruženje

```
conda create --prefix [putanja] python=3.11 -y
```

4. Aktivacija okruženja

```
conda activate [putanja]
```

5. Instalacija PyTorch biblioteka s CUDA podrškom

```
pip install torch torchvision torchaudio  
--index-url https://download.pytorch.org/whl/cu124
```

Dodatak A. Programska okolina

6. Instalacija dodatnih potrebnih modula

```
pip install transformers datasets accelerate  
sentencepiece bitsandbytes soundfile
```

7. Testiranje dostupnosti CUDA-e

```
python cuda-test.py
```

8. Deaktiviranje okruženja po završetku rada

```
conda deactivate
```

Dodatak B

Dodatni materijali i repozitorij

Svi skripte, ulazne i izlazne datoteke te rezultati korišteni u ovom diplomskom radu dostupni su u javnom GitHub repozitoriju:

<https://github.com/lIlich/diplomski-rad-illich>

Struktura repozitorija uključuje sljedeće direktorije i datoteke:

- **izlazi/text-generation/**
Svi tekstualni izlazi generirani velikim jezičnim modelom.
- **izlazi/text-to-speech/**
Audio zapisi generirani modelom sinteze govora.
- **rezultati/**
Rezultati evaluacije u Excel i CSV formatima.
- **skripte/**
Skripte za izvođenje eksperimenata na LLM i TTS modelima te pomoćne skripte.
- **ulazi/**
Ulagani podaci i zadaci za LLM model.

Link na repozitorij omogućuje potpunu reprodukciju, daljnju analizu i nadogradnju istraživanja.