### ****5. Evaluacija****

Evaluacija sustava za sintezu govora uključuje kombinaciju objektivnih metrika, koje se temelje na numeričkoj usporedbi referentnih i generiranih zvučnih zapisa,, te subjektivnih metoda, koje se temelje na ljudskoj percepciji kvalitete generiranog govora. Budući da sinteza govora uključuje višedimenzionalne aspekte poput razumljivosti, prirodnosti, intonacije i govornog identiteta, nije moguće osloniti se isključivo na jednu metriku. Stoga je provedena višestruka evaluacija, uzimajući u obzir različite objektivne pokazatelje, ali i kvalitativne dojmove dobivene slušanjem generiranih uzoraka.

Za potrebe evaluacije generirane su predikcije za 50 zvučnih zapisa testnog dijela skupa BKSpeech. Svaki predikcijski zapis stvoren je na temelju istog teksta kao i odgovarajući original, koristeći Coqui TTS naredbu:

tts --text <tekst> --model\_path <putanja\_do\_modela> --vocoder\_path <putanja\_do\_vokodera> --out\_path <output.wav> --config\_path <putanja\_do\_config.json\_modela> --vocoder\_config\_path <putanja\_do\_config.json\_vokodera>

#### **5.1 Objektivne metrike evaluacije**

Objektivne metrike evaluacije predstavljaju skup kvantitativnih mjera koje procjenjuju sličnost između generiranog i referentnog zvučnog zapisa. Njihova prednost leži u mogućnosti automatizirane i ponovljive analize, no glavno ograničenje jest osjetljivost na vremenske nesinkronizacije, koje su česte u zadacima sinteze govora.

Tijekom istraživanja ispitane su različite objektivne metrike koje se uobičajeno koriste u govornoj obradi, uključujući Mel-Cepstral Distortion (MCD), Short-Time Objective Intelligibility (STOI), f0 Pearson korelaciju i Voiced/Unvoiced error (VUV). Međutim, njihova primjena na TTS izlaze u ovom radu rezultirala je iznimno niskim ili nekonzistentnim rezultatima, čak i kada je generirani govor bio semantički ispravan i perceptivno prirodan. Na primjer, MCD metrika uspoređuje spektrogramske značajke, no jako je osjetljiva na vremensko poravnanje. Čak i mali pomaci u vremenu rezultiraju velikim penalima, što ju čini nepouzdanom za sintezu govora gdje su vremenske varijacije prirodne. Slično tome, f0 Pearson korelacija mjeri sličnost kontura temeljne frekvencije, što može biti informativno za intonaciju, ali gubi na pouzdanosti ako generirani i referentni govor nisu strogo sinkronizirani. Zbog takvih ograničenja, ove metrike često ne odražavaju stvarnu percepciju kvalitete kod slušatelja i stoga nisu uključene u završnu analizu.

U ovom su radu odabrane objektivne metrike koje su robusne na vremenske razlike i usmjerene na semantičku, akustičku i perceptivnu bliskost između generiranih i referentnih signala. Korišten je niz embedding-based metrika koje uspoređuju akustičke reprezentacije govora, tekstualne metrike razumljivosti poput WER i CER, te neuronski model MOSNet, treniran za procjenu prirodnosti govora. Sve korištene metrike bit će detaljnije opisane u nastavku.

**5.1.1 Embedding-based metrike**

U kontekstu evaluacije sinteze govora, jedno od ključnih izazova predstavlja pouzdano kvantificiranje sličnosti između generiranog i referentnog signala. Klasične akustičke metrike, koje se temelje na izravnim usporedbama valnih oblika, često podbace u ovim zadacima zbog osjetljivosti na vremenska odstupanja, promjene u tempu ili varijabilne pauze koje su u TTS sustavima prirodne i ne nužno pogrešne. Kao odgovor na ta ograničenja, razvijene su metrike koje se oslanjaju na ana*lizu tzv. govornog embeddinga,* reprezentacije koja sažima fonetske, akustičke i često i semantičke značajke govora u višedimenzionalni vektorski prostor.

#### U ovom radu korištene su embedding reprezentacije dobivene iz prettreniranog Wav2Vec2.0 modela, koji za ulazni zvuk generira niz vektora duž vremenske osi. Svaki od tih vektora, dimenzije 768, kodira lokalne akustičke značajke segmenta govora i služi kao apstraktna reprezentacija njegovog zvučnog sadržaja. S obzirom na to da se svaki zvučni zapis reprezentira kao niz ovih vektora, moguće je uspoređivati generirani i referentni signal na razini cijele rečenice, pojedinih vremenskih segmenata, pa čak i kroz agregirane reprezentacije.

#### U tu svrhu implementirano je nekoliko metrika koje koriste upravo ovakve embeddinge za usporedbu generiranih i referentnih uzoraka. Prva metrika temeljena je na srednjem vektoru cijelog govora, gdje se za svaki signal računa prosječna embedding reprezentacija te se zatim mjeri kosinusna sličnos između tih prosječnih vektora. S obzirom da se uspoređuju dva vektora koja predstavljaju dva zvučna zapisa, ova mjera zanemaruje vremensku strukturu i fokusira se na ukupnu akustičku podudarnost tih dvaju zapisa. Za sustav treniran od nule, GlowTTS + HiFi-GAN, dobivena je prosječna vrijednost od 0.978, dok je fino podešeni VITS model ostvario vrijednost od 0.973, što ukazuje na vrlo visoku globalnu sličnost između generiranog i referentnog signala u oba slučaja, s blagom prednošću GlowTTS + HiFi-GAN sustava.

#### Kako bi se evaluacija proširila i na vremensku dimenziju govora, korištene su dvije metrike koje uspoređuju sekvence embedding vektora: framewise kosinusna sličnost i embedding-DTW. Obje se oslanjaju na algoritam dinamičkog vremenskog uparivanja (Dynamic Time Warping, DTW), koji omogućuje usporedbu dvaju zvučnih zapisa čak i kad se razlikuju u duljini ili tempu.

U ovom kontekstu, sekvenca označava niz vektora koje Wav2Vec2.0 model generira iz audio signala. To su tzv. frame embedding reprezentacije, i svaka predstavlja jedan kratak vremenski segment govora. Budući da generirani i referentni zvuk često nemaju jednaku duljinu (npr. jedan traje 3 sekunde, a drugi 4 sekunde), oni će imati različit broj embedding vektora. To znači da ih nije moguće direktno usporediti po poziciji jer jedan može imati npr. 100 vektora, a drugi 120.

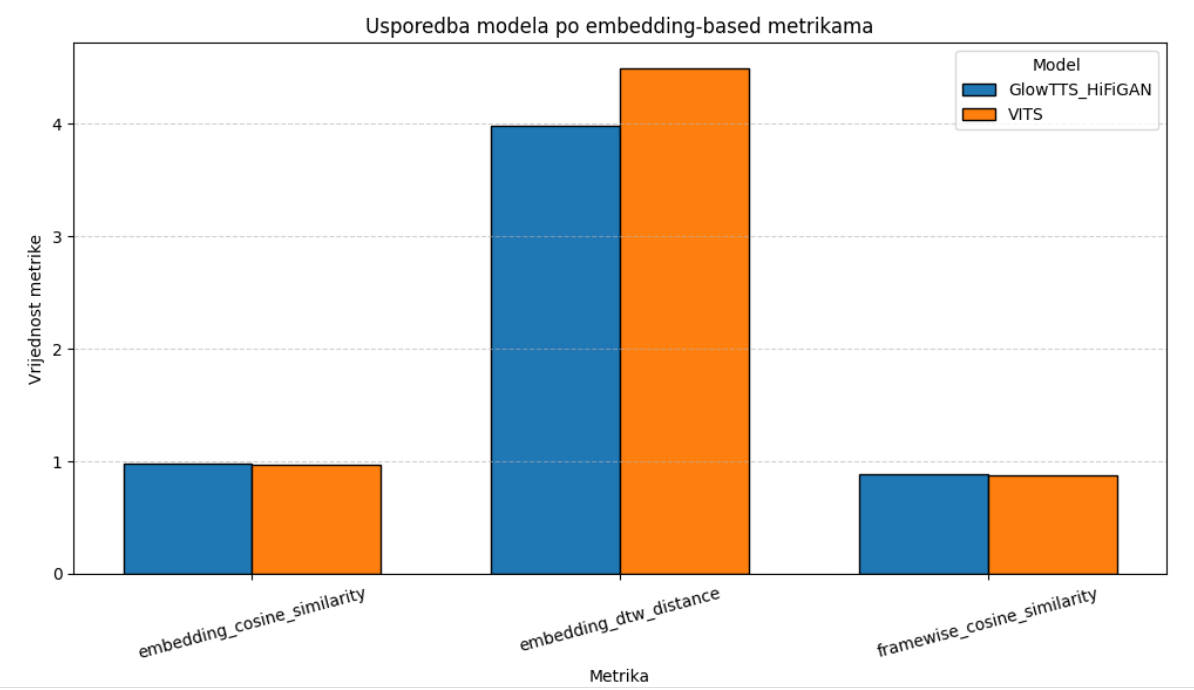
U tom slučaju DTW traži najbolji način da se vektori iz jedne sekvence sparuju s vektorima iz druge sekvence, tako da ukupna razlika među njima bude što manja. Umjesto da svaki vektor uspoređuje točno s vektorom na istoj poziciji u vremenu, DTW dozvoljava preskakanje ili ponavljanje vektora u uparivanju. Na primjer, jedan vektor iz kraće sekvence može biti uspoređen s dva uzastopna vektora iz dulje sekvence, ako to bolje odgovara stvarnom sadržaju (npr. govorene riječi izgovorene sporije). Na taj se način simulira rastezanje ili stiskanje vremena, i nalazi se najprirodnije vremensko poravnanje.

Rezultat DTW algoritma je niz parova vektora, jedan iz generiranog i jedan iz referentnog zvuka, za koje se zatim računa udaljenost te se tako čak i u slučajevima kad se dvije rečenice izgovore različitom brzinom, omogućuje usporedba njihovih stvarnih akustičkih sadržaja, a ne samo trajanja ili tempa.

Framewise kosinusna sličnost mjeri lokalnu akustičku sličnost između referentnog i generiranog zvučnog zapisa. Nakon DTW poravnanja, računa se kosinusna sličnost za svaki par uparenih vektora iz sekvenci embeddinga, a konačna vrijednost dobiva se kao njihov prosjek . Budući da kosinusna sličnost mjeri kut između vektora, ova metrika naglašava sličnost akustičkog obrasca, neovisno o apsolutnoj energiji signala. No, i dalje osjetljivo bilježi razlike u intonaciji, ritmu i pauzama jer su one sadržane u samim vektorima. Prema dobivenim rezultatima, sustav GlowTTS + HiFi-GAN je u ovoj metrikama ostvario višu prosječnu sličnost (0.887) u odnosu na VITS (0.870), što ukazuje na konzistentniju akustičku izvedbu generiranog govora.

Embedding-DTW koristi isti DTW algoritam za poravnanje, ali potom računa ukupnu euklidsku udaljenost između uparenih vektora. Za razliku od kosinusne sličnosti, ova mjera uzima u obzir i kut i amplitudu vektora, što je čini osjetljivijom na sveukupnu akustičku razliku, uključujući razlike u glasnoći, tempu i govornoj artikulaciji. Manja udaljenost označava veću sličnost između dviju sekvenci. GlowTTS + HiFi-GAN je i ovdje postigao bolji rezultat (3.98) u odnosu na VITS (4.49), što dodatno potvrđuje njegovu stabilnost i preciznost u vremenskoj strukturi govora.

Na slici 5.1 vidljivo je da sustav GlowTTS + HiFi-GAN postiže bolje rezultate u svim prikazanim embedding-based metrikama. Ima višu globalnu (embedding\_cosine\_similarity) i lokalnu (framewise\_cosine\_similarity) sličnost s referentnim signalom, te nižu vremensku udaljenost (embedding\_dtw\_distance), što ukazuje na vjernije i stabilnije generiranje govora u odnosu na VITS.



#### Slika 5.1: Usporedba modela po embedding-based metrikama (GlowTTS+HiFiGAN vs. VITS).

#### Uz ove metrike koje su primarno fokusirane na akustičku strukturu, dodatno je korišten Resemblyzer, model treniran na zadatku identifikacije govornika. On generira jedinstveni vektor svakog govora, tzv. speaker embedding, koji sažima govorne značajke za identitet govornika, poput boje glasa, visine i ritma. Kosinusna sličnost između embedding reprezentacija referentnog i generiranog govora u ovom slučaju mjeri koliko je dobro očuvan govorni identitet. Rezultati su pokazali da oba modela uspješno održavaju sličnost s izvornim govornikom, s prosječnim vrijednostima od 0.825 za GlowTTS + HiFi-GAN i 0.832 za VITS, pri čemu VITS blago nadmašuje GlowTTS + HiFi-GAN u ovom aspektu.

#### **5.1.2 WER i CER**

Uz akustičku sličnost, važan aspekt kvalitete govorne sinteze jest i **razumljivost** generiranog govora. Kako bi se ona kvantificirala, u evaluaciju su uključene i **tekstualne metrike** poput Word Error Rate (WER) i Character Error Rate (CER), koje kvantificiraju razlike između automatski transkribiranog teksta generiranog i referentnog zvučnog zapisa.

U ovom radu WER i CER izračunani su na temelju transkripcija dobivenih pomoću Whisper modela automatskog prepoznavanja govora (**Automatic Speech Recognition,** ASR) koji omogućuje visoko preciznu transkripciju i za hrvatski jezik. Za svaki generirani zvučni zapis iz testnog skupa, automatski je izvedena transkripcija korištenjem Whisper modela. Istim postupkom transkribirani su i referentni (*ground truth*) zvučni zapisi, kako bi usporedba bila konzistentna i pravedna. Prije usporedbe, svi su tekstovi prošli postupak normalizacije gdje su uklonjeni interpunkcijski znakovi, velika slova i višestruki razmaci, čime se osigurava robusnost metrike na površinske razlike koje ne utječu na sadržaj.

**Word Error Rate (WER) mjeri omjer broja pogrešaka (ubacivanja, brisanja i zamjena riječi) u odnosu na ukupan broj riječi u referentnom tekstu. Računa se prema formuli:**

gdje su - broj zamjena (substitutions), - brisanja (deletions), - umetanja (insertions), a - ukupan broj riječi u referentnom tekstu.

#### **Character Error Rate (CER) slijedi istu logiku, ali se umjesto riječi uspoređuju pojedinačni znakovi. CER je često osjetljiviji na manje izmjene, poput nedostajućih slova ili tipografskih grešaka, što ga čini korisnim u jezicima s fleksibilnim riječima i morfologijom poput hrvatskog.**

#### **Iako obje metrike mogu ponekad biti stroge – penaliziraju čak i manje razlike koje ne utječu nužno na razumijevanje – vrlo su važne u kontekstu evaluacije semantičke vjernosti, jer otkrivaju slučajeve u kojima se značenje rečenice promijenilo ili izgubilo zbog pogrešnog izgovora.**

#### **U provedenoj analizi, GlowTTS + HiFi-GAN sustav pokazao je nešto niži prosječni WER od 12.58% i CER od 3.45%, dok je VITS ostvario prosječni WER od 14.61% i CER od 4.21%. Ovi rezultati ukazuju na nešto veću semantičku točnost i precizniji izgovor kod GlowTTS + HiFi-GAN sustava. Iako razlike nisu dramatične, stabilno niže greške na razini riječi i znakova sugeriraju bolju kontrolu nad fonemskom realizacijom i segmentacijom u generiranom govoru GlowTTS + HiFi-GAN modela.**

#### **5.1.3 MOSNet**

Kako bi se procijenila **prirodnost** generiranog govora bez uključivanja ljudskih ocjenjivača, u evaluaciji je korišteni i model **MOSNet**. Riječ je o dubokoj neuronskoj mreži posebno treniranoj za predikciju ljudskih procjena, tzv. Mean Opinion Score (MOS), na temelju akustičkog signala.

MOS je uobičajena metrika u govornoj tehnologiji kojom se kvaliteta zvučnih uzoraka ocjenjuje na skali od 1 (loše) do 5 (izvrsno), a tradicionalno se temelji na **subjektivnim dojmovima ispitanika**. MOSNet model, s druge strane, koristi duboko učenje za automatsko predviđanje takvih subjektivnih ocjena, uzimajući kao ulazni podatak sam zvučni signal, bez potrebe za ljudskim evaluacijama. Model je treniran na velikim skupovima govora s pratećim MOS ocjenama, čime uči povezivati akustičke karakteristike sa subjektivnim doživljajem kvalitete.

U ovom radu MOSNet je primijenjen na svaki generirani zvučni zapis u testnom skupu, čime je dobivena prosječna predikcija prirodnosti. Dodatno, model je korišten i na referentnim (ground truth) zapisima, kako bi se procijenila njegova ocjena stvarnog, ljudskog govora. Ta referentna vrijednost služi kao pokazatelj koliko strogo model ocjenjuje čak i prirodne uzorke.

Rezultati pokazuju da MOSNet dosljedno predviđa niže ocjene za sintetički govor u odnosu na stvarni, što potvrđuje njegovu osjetljivost na artefakte i nepravilnosti u generaciji. Prosječni rezultat za GlowTTS + HiFi-GAN iznosi 2.89, dok je za VITS 2.79. Usporedbe radi, prosječan MOS za referentne zvučne zapise iznosi 3.27, što ukazuje da ni stvarni govor ne postiže maksimalne vrijednosti prema standardima MOSNet modela, ali se razlika između stvarnog i generiranog govora još uvijek jasno uočava.

### Na slici X prikazan je kutijasti dijagram (boxplot) razlike između MOSNet ocjene stvarnog (engl. ground truth, GT) i generiranog govora po modelima. Što je ta razlika manja, to model proizvodi govor s prirodnošću bližom stvarnom.

### Vidljivo je da GlowTTS + HiFiGAN ima niži medijan razlike (~0.33) i manji raspon vrijednosti, dok VITS pokazuje šire varijacije i viši medijan (~0.45), što sugerira da VITS model proizvodi govor koji je percipiran kao manje prirodan u odnosu na referentni.

### 

### Također, manji broj outliera kod GlowTTS modela ukazuje na veću stabilnost predikcije prirodnosti, dok su kod VITS-a prisutne veće oscilacije. Iako MOSNet omogućuje objektivnu procjenu prirodnosti, on ipak predstavlja samo aproksimaciju ljudske percepcije. Budući da ne može u potpunosti zamijeniti ljudsku prosudbu o izražajnosti, razumljivosti ili zvučnoj autentičnosti, u nastavku rada provodi se i subjektivna evaluacija, kako bi se dobila sveobuhvatnija slika o kvaliteti generiranog govora.

### ****5.2 Subjektivne metrike****

Za razliku od objektivnih metrika koje se temelje na algoritamskim izračunima i modelima, subjektivne metrike oslanjaju se na ljudske ocjene i dojam slušatelja te predstavljaju zlatni standard u evaluaciji sustava za sintezu govora. One mjere kako ljudi percipiraju ključne aspekte poput prirodnosti, razumljivosti, emocionalnog izraza, sličnosti s izvornim govornikom i ukupne ugodnosti slušanja. Dok objektivne metrike pružaju korisnu kvantitativnu procjenu, subjektivne evaluacije temelje se na stvarnom dojmu korisnika i obuhvaćaju čimbenike koje je teško precizno kvantificirati, ali koji su presudni za stvarni doživljaj kvalitete govora.

U istraživačkoj zajednici i benchmark platformama (npr. Papers with Code, LJSpeech leaderboard PROVJERI TO), subjektivne evaluacije, osobito **MOS (Mean Opinion Score),** redovito se koriste za rangiranje i usporedbu TTS modela. U takvim evaluacijama ljudski slušatelji ocjenjuju zvučne uzorke na temelju prirodnosti, razumljivosti i sličnosti s referentnim govornikom, često kroz testove slušanja koji uključuju ocjenjivanje ili usporedbe između različitih modela. Takav pristup omogućuje neposredan uvid u stvarnu percepciju kvalitete iz perspektive krajnjih korisnika.

### **Iako objektivne metrike, poput onih temeljenih na embedding reprezentacijama ili automatskom prepoznavanju govora, omogućuju brzu i dosljednu evaluaciju, one imaju svoja ograničenja. Ne mogu u potpunosti obuhvatiti nijanse koje ljudski slušatelj intuitivno prepoznaje, poput emocionalnog tona, prirodnog tempa ili suptilnih nepravilnosti u intonaciji koje sintetizirani govor mogu učiniti neprirodnim.**

### **Budući da je krajnji korisnik TTS sustava upravo čovjek, važno je razumjeti kako on doživljava generirani govor. Subjektivne metrike stoga ostaju nezamjenjiv dio evaluacije, osobito u kontekstu stvarnih primjena, jer kvaliteta govora ne podrazumijeva samo tehničku ispravnost, već i prirodan, uvjerljiv i razumljiv doživljaj.**

### ****5.2.1 MOS (Mean Opinion Score)****

MOS, odnosno Mean Opinion Score, je najčešće korištena subjektivna metrika u području sinteze govora. Temelji se na evaluaciji generiranih govora od strane ljudi, gdje se svakom uzorku dodjeljuje ocjena na skali od 1 do 5:

* 5 – potpuno prirodan (teško razlikovati od stvarnog govora)
* 4 – prirodan, s manjim nepravilnostima
* 3 – razumljiv, ali potpuno neprirodan
* 2 – teško razumljiv
* 1 – nerazumljiv

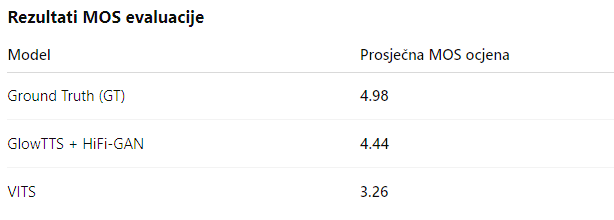
U istraživanjima i usporedbama TTS sustava, MOS ocjene se obično prikupljaju putem formalnih slušnih testova s većim brojem ispitanika. U takvom kontekstu, ispitanici ocjenjuju nasumično prikazane uzorke bez znanja o modelu iz kojeg dolaze, čime se smanjuje mogućnost pristranosti. Rezultati se zatim agregiraju i prikazuju kao srednja vrijednost ocjena po modelu, često uz standardnu devijaciju radi prikaza varijabilnosti dojma među korisnicima.

U okviru ovog rada provedena je subjektivna evaluacija pomoću *Microsoft Forms*. Test je sadržavao ukupno **30 zvučnih zapisa**, od kojih je:

* 10 pripadalo kategoriji izvornih (*ground truth*) snimki,
* 10 su bile generirane pomoću modela **GlowTTS + HiFi-GAN**,
* a 10 pomoću modela **VITS**.

Važno je naglasiti da se radi o **istih 10 govora** u sve tri kategorije, svaki originalni uzorak bio je sintetiziran korištenjem oba modela. Redoslijed snimki u formi bio je nasumično izmiješan kako bi se spriječilo prepoznavanje i usporedba modela, a svaki ispitanik ocjenjivao je snimke neovisno, fokusirajući se isključivo na **dojam prirodnosti**.

Testu je pristupilo **12 sudionika**, koji su za svaki uzorak dali ocjenu na skali od 1 do 5, prema jasnim uputama koje su definirale što pojedina ocjena predstavlja. Prikupljene ocjene zatim su agregirane po svakom modelu. Rezultati su prikazani u tablici ispod, gdje su navedene prosječne MOS vrijednosti:



Dobiveni rezultati pokazuju da model GlowTTS + HiFi-GAN generira govor koji ljudski slušatelji ocjenjuju kao vrlo prirodan, s minimalnim odstupanjima u odnosu na izvorni govor. Rezultat od 4.44 iznimno je dobar te je usporediv s rezultatima vodećih modela objavljenima na istraživačkim platformama poput Papers with Code. Nasuprot tome, VITS model je ostvario rezultat od 3.26, što je i dalje solidna ocjena razumljivosti, ali značajno niža u pogledu percepcije prirodnosti. Ova evaluacija potvrđuje da sustav GlowTTS + HiFi-GAN u ovom okruženju pokazuje veću vjerodostojnost u generiranju prirodnog govora.

Dodatno, na slici X prikazana je raspodjela MOS ocjena po svakom pojedinom zvučnom zapisu iz ispitnog skupa, što omogućuje detaljniji uvid u ponašanje modela na razini pojedinačnih primjera. Iz vizualizacije je vidljivo da je GlowTTS + HiFi-GAN u gotovo svim slučajevima ostvario rezultat blizak referentnim snimkama, dok je VITS pokazivao veće varijacije i odstupanja. Ovakav prikaz dodatno potvrđuje konzistentnost i stabilnost performansi modela GlowTTS + HiFi-GAN.

