Sinteza govora iz teksta upotrebom dubokog učenja

Džijan, Matej

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Sveučilišni studij

SINTEZA GOVORA IZ TEKSTA UPOTREBOM DUBOKOG UČENJA

Diplomski rad

Matej Džijan

Osijek, 2020.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Obrazac D1: Obrazac za imenovanje Povjerenstva za diplomski ispit

Osijek, 21.09.2020.

Odboru za završne i diplomske ispite

Imenovanje Povjerenstva za diplomski ispit

Ime i prezime studenta:	Matej Džijan
Studij, smjer:	Diplomski sveučilišni studij Računarstvo
Mat. br. studenta, godina upisa:	D-979R, 24.09.2019.
OIB studenta:	24126366791
Mentor:	Izv.prof.dr.sc. Ratko Grbić
Sumentor:	
Sumentor iz tvrtke:	
Predsjednik Povjerenstva:	Izv. prof. dr. sc. Emmanuel-Karlo Nyarko
Član Povjerenstva 1:	Izv.prof.dr.sc. Ratko Grbić
Član Povjerenstva 2:	Petra Đurović
Naslov diplomskog rada:	Sinteza govora iz teksta upotrebom dubokog učenja
Znanstvena grana rada:	**Umjetna inteligencija (zn. polje računarstvo)**
Zadatak diplomskog rada:	Sinteza govora predstavlja operaciju pretvaranja teksta u odgovarajući govor pomoću računala. U okviru diplomskog rada najprije je potrebno napraviti pregled trenutno najefikasnijih metoda za problem pretvaranja teksta u govor. Zatim je potrebno izgraditi prikladni podatkovni skup te razviti algoritam temeljen na strojnom učenju koji će pretvarati tekst na hrvatskom jeziku u govor. Nakon implementacije izgrađenog modela u odgovarajući sustav potrebno je osmisliti i provesti evaluaciju izgrađenog modela. Tema rezervirana za: Matej Džijan
Prijedlog ocjene pismenog dijela ispita (diplomskog rada):	Izvrstan (5)
Kratko obrazloženje ocjene prema Kriterijima za ocjenjivanje završnih i diplomskih radova:	Primjena znanja stečenih na fakultetu: 3 bod/boda Postignuti rezultati u odnosu na složenost zadatka: 2 bod/boda Jasnoća pismenog izražavanja: 3 bod/boda Razina samostalnosti: 2 razina
Datum prijedloga ocjene mentora:	21.09.2020.

Potpis mentora za predaju konačne verzije rada u Studentsku službu pri završetku studija:	Potpis:
	Datum:

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

IZJAVA O ORIGINALNOSTI RADA

Osijek, 28.09.2020.

Ime i prezime studenta:	Matej Džijan
Studij:	Diplomski sveučilišni studij Računarstvo
Mat. br. studenta, godina upisa:	D-979R, 24.09.2019.
Turnitin podudaranje [%]:	3

Ovom izjavom izjavljujem da je rad pod nazivom**: Sinteza govora iz teksta upotrebom dubokog učenja**

izrađen pod vodstvom mentora Izv.prof.dr.sc. Ratko Grbić

i sumentora

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis studenta:

SADRŽAJ

1. Uvod

Sinteza govora je umjetna proizvodnja ljudskog govora pomoću računala. Sustav koji se koristi za sintezu govora se zove sintetizator govora. Tekst-u-govor (engl. *Text-To-Speech – TTS*) je sustav koji pretvara tekst s normalnim jezikom u govor, dok drugi sustavi za sintezu govora pretvaraju druge oblike jezika u govor, kao što je fonetička transkripcija teksta.

Najvažnija i najraširenija upotreba sinteze govora je za pomoć ljudima s poteškoćama. Najduže se koristi za čitače zaslona za ljude s oštećenim vidom, ali ju danas često koriste i ljudi s disleksijom i s drugim poteškoćama u čitanju. Također se koriste i za pomoć ljudima s teškim poteškoćama u govoru, najčešće upotrebom namjenskih uređaja za sintezu govora. Najpoznatiji korisnik ovakvog uređaja bio je znanstvenik Stephen Hawking. Sinteza govora je također bitna kao pomoć u analizi i procjeni poremećaja govora. Sinteza govora se koristi i u industriji zabave. U video igrama i animiranim filmovima se koristi za naraciju i dijalog [1]. U posljednje vrijeme se koristi u kombinaciji s tehnologijama za prepoznavanje govora za sustave za interakciju s mobilnim uređajima, primjerice Alexa, Google Assistant i Siri.

Prvi sustavi za sintezu govora iz teksta su se pojavili sredinom dvadesetog stoljeća. Takvi sustavi su koristili metode za sintezu govora koje su se temeljile na sklapanju dijelova snimljenog govora. Poslije su metode postale sofisticiranije, pretvarale su tekst u foneme te su pokušavale reproducirati ljudski vokalni trakt i tako sintetizirati govor [2]. Razvojem područja dubokog učenja, pojavile su se i metode za sintezu govora iz teksta temeljene na dubokim neuronskim mrežama. Jedan od prvih modela koji je koristio duboke neuronske mreže je WaveNet tvrtke DeepMind i objavljen je 2016. godine [3]. Ovaj model je pokazao da je moguće sintetizirati govor iz obrađenih jezičnih značajki. Godine 2017. je objavljen char2wav model instituta Mila, koji je mogao sintetizirati govor direktno iz teksta [4]. Poslije su se pojavili i modeli Tacotron [5], VoiceLoop [6] i mnogi drugi. Pojavom modela Tacotron 2, pokazano je da je moguće sintetizirati govor vrlo nalik na ljudski direktno iz teksta [7].

Područje sinteze govora ima razne probleme zbog kojih ne zvuče potpuno prirodno, odnosno kao ljudski govor. Jedan od tih problema su homografi, riječi koje se isto pišu, ali se različito izgovaraju (npr. lùk – biljka i lúk – oružje). Za rješavanje ovog problema se koriste razne tehnike, kao što su proučavanje konteksta (susjednih riječi) ili korištenje statistike o tome koliko se često koja riječ koristi. Još jedan problem su brojevi. Primjerice, broj 92 se može pročitati kao “devedeset dva”,

“devedeset dvije” i “devet dva”. Ovaj problem se također najčešće rješava gledanjem na kontekst u kojem se koristi. Postoji problem i s rednim brojevima kod jezika u kojima se brojevi deklliniraju, kao što je hrvatski jezik. U rečenici “1945. godine su saveznici ušli u Berlin i okupirali ga i tako 1945. godina označava kraj drugog svjetskog rata” redni broj 1945. se čita na dva različita načina. Ovaj problem bi se također rješavao gledanjem na kontekst iz kojeg bi se zaključio padež broja. Prirodno se jezici pišu i s raznim skraćenicama, ponekad te skraćenice mogu imati više značenja. Prozodija i emocionalni sadržaj također mogu uvelike utjecati na to koliko prirodno zvuči određeni sintetizator govora.

Problemi nastaju i kod same evaluacije rezultata sustava za sintezu govora. Iako postoji objektivan način evaluacije ovakvih sustava, tzv. sustavi za automatsko prepoznavanje govora, ti sustavi sami po sebi nisu jednostavni za napraviti. Postoji i jednostavnija metoda za evaluaciju sustava za sintezu govora, ali se koristi samo za razumljivost sintetiziranog govora, ne i za prirodnost [8]. Češće korišten način za evaluaciju sustava za sintezu govora je ispitivanje osoba kako bi se dobilo prosječno subjektivno mišljenje, ali takva je evaluacija više primjenjiva za usporedbu, nego za apsolutnu ocjenu sustava. Primjerice, kad bi se sintetizator ocjenjivao na skali od 1 do 5, ne znači svakoj osobi ocjena 1 isto.

Postoje razni sustavi za sintezu govora za rasprostranjene jezike poput engleskog, ali za manje rasprostranjene jezike poput hrvatskog, nisu dostupna mnoga rješenja. Također ne postoje ni gotovi podatkovni skupovi koji bi se mogli koristiti za treniranje sintetizatora za hrvatski jezik, stoga se u ovom radu pokušava riješiti taj problem manjka dobrih sintetizatora za hrvatski jezik.

U okviru diplomskog rada izgrađen je model za sintezu govora iz teksta za hrvatski jezik na temelju vlastitog izrađenog skupa podataka. Zatim je provedena evaluacija predloženog rješenja korištenjem anketa. U anketama su ispitanici ocjenjivali dane zvučne isječke, uspoređivali sintetizirani govor s prirodnim te određivali pogreške u različitim sintetiziranim rečenicama.

U drugom poglavlju je pojašnjena osnovna terminologija koja se koristi u radu i napravljen je kratki pregled nekoliko postojećih rješenja za sintezu govora iz teksta. U trećem poglavlju se detaljnije obrađuje model za sintezu govora iz teksta korišten u ovom diplomskom radu. Također je u trećem poglavlju predstavljena izrada i pripremanje podatkovnog skupa za treniranje, kao i treniranje i odabir konačnog modela za sintezu govora iz teksta na hrvatskom jeziku. U četvrtom poglavlju je opisana evaluacija predloženog rješenja za sintezu govora na hrvatskom jeziku. Na kraju je dan zaključak rada

s prijedlozima za buduća poboljšanja predloženog modela za sintezu govora iz teksta na hrvatskom jeziku.

2. Postojeća rješenja za sintezu govora iz teksta

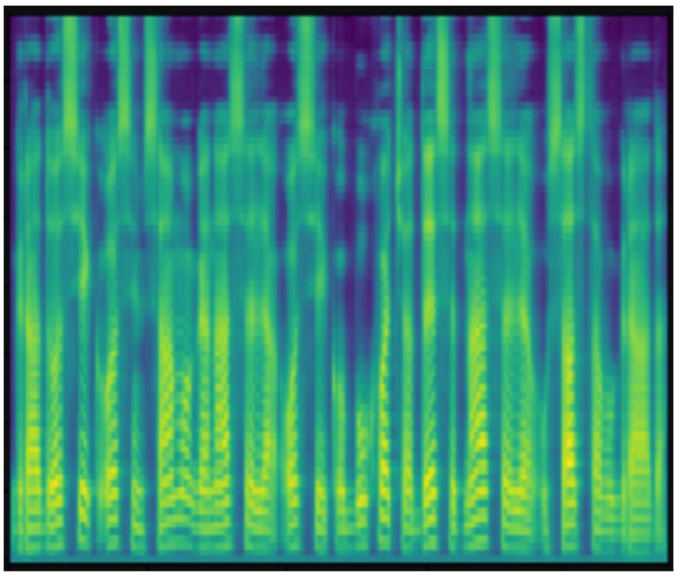
Budući da se za ovaj diplomski rad koristi strojno učenje, u ovom poglavlju je dana osnovna terminologija vezana za strojno učenje i za sintezu govora iz teksta. Nakon toga je predstavljeno nekoliko metoda za sintezu govora iz teksta.

2.1. Osnovna terminologija

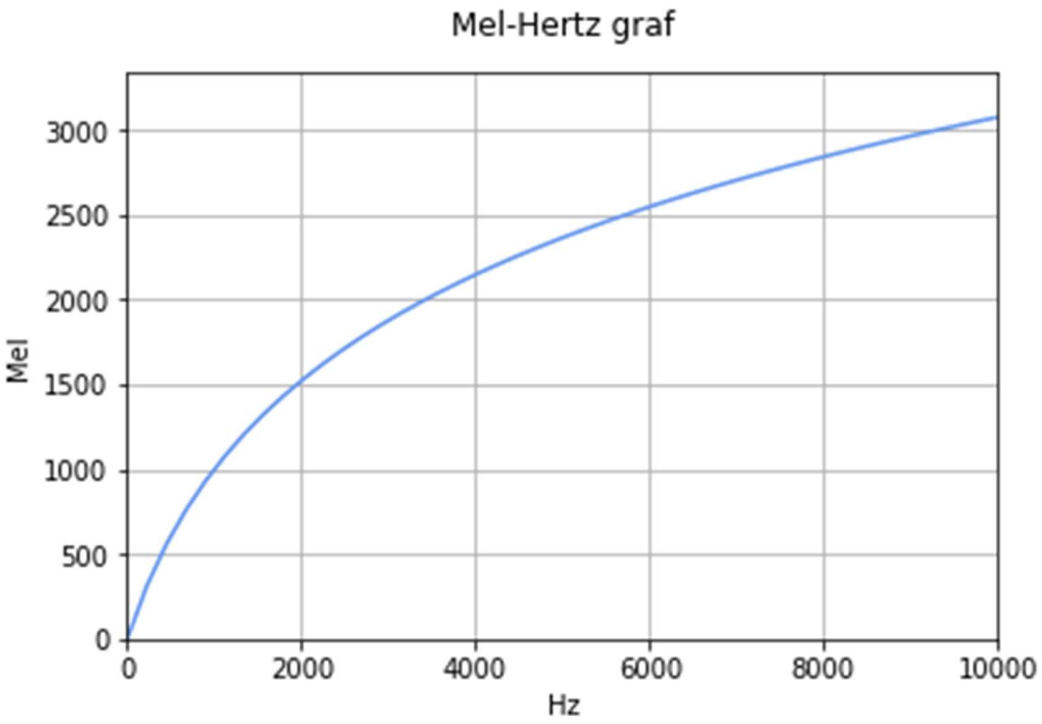
Strojno učenje je grana umjetne inteligencije koja se bavi oblikovanjem algoritama koji svoju učinkovitost poboljšavaju na temelju empirijskih podataka [9]. Postoji nadzirano, nenadzirano i podržano učenje (engl. *supervised, unsupervised i reinforcement learning*). U ovom radu se koristi nadzirano učenje. Kod nadziranog učenja se, pri treniranju modela, predaju ulazni podaci zajedno s pripadajućim izlaznim podacima, u ovom radu su to izgovorene rečenice uz pripadajući transkript.

Jedan od osnovnih modela u okviru strojnog učenja je neuronska mreža. To je slojevita struktura pri čemu se svaki sloj sastoji od niza neurona koji procesiraju informacije od ulaza ka izlazu (engl. *feedforward neural network*). Najčešće je svaki neuron nekog sloja povezan sa svakim neuronom iz prethodnog sloja i takva se mreža naziva potpuno povezana mreža (engl. *fully connected network*). Duboko učenje koristi velik broj slojeva za postupno izvlačenje značajki iz neobrađenih podataka. Jedna od najčešće korištenih dubokih mreža je konvolucijska neuronska mreža (engl. *Convolutional Neural Network* – CNN) koja osim standardnih potpuno povezanih slojeva, sadrži i konvolucijske slojeve i slojeve sažimanja [10]. Primjerice, CNN u obradi slike se može koristiti za detekciju rubova, a kombinacijom većeg broja CNN-ova mogu se realizirati kompliciraniji koncepti kao što su prepoznavanje lica. Kod sinteze govora iz teksta se osim CNN-ova koriste i razni drugi modeli neuronskih mreža.

U ovom radu se koriste *mel* spektrogrami kao posredni izlazi koji se zatim pretvaraju u valne oblike zvučnog signala. Spektrogrami su vizualna reprezentacija frekvencijskog spektra nekog signala u ovisnosti o vremenu (Slika 2.1.). *Mel* spektrogrami su spektrogrami u kojima su frekvencije pretvorene u *mel* ljestvicu. Odnos između *mel* ljestvice i Hertz ljestvice prikazan je na slici 2.2.

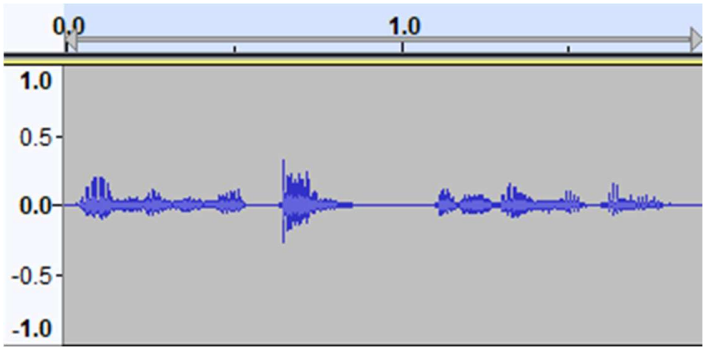


Slika 2.1. Primjer spektrograma



Slika 2.2. Graf *mel* ljestvice naspram Hertz ljestvice

Potpuni izlazi su valni oblici zvučnog signala (engl. *waveform*). Valni oblik signala je oblik grafa kao funkcije vremena, gdje je na y-osi amplituda signala (Slika 2.3.).



Slika 2.3. Primjer valnog oblika zvučnog signala

Za evaluaciju rezultata u području sinteze govora se najčešće koristi srednja ocjena slušatelja (engl. *Mean Opinion Score* – MOS). U kontroliranim uvjetima osoba sluša odgovarajuće zvučne zapise i ocjenjuje ih s obzirom na sveukupnu kvalitetu. Najčešće se koristi skala od 1 do 5, gdje je 1 najgora ocjena, a 5 najbolja. Da bi se dobio MOS, uzima se aritmetička sredina ocjena svih slušatelja. MOS se često navodi uz standardnu pogrešku aritmetičke sredine ocjena ili uz interval pouzdanosti.

2.2. Pregled metoda za sintezu govora iz teksta

Za sintezu govora iz teksta postoje klasične metode i metode zasnovane na strojnom učenju. Klasične metode uključuju spojnu metodu, sintezu baziranu na pravilima, artikulacijsku sintezu i druge. Novije metode su bazirane na strojnom učenju, odnosno na dubokom učenju. Neke od metoda za sintezu govora iz teksta su predstavljene u nastavku.

2.2.1. Spojna metoda

Spojna metoda je jedna od najstarijih metoda za sintezu govora iz teksta [2]. Postoje tri vrste spojne metode sinteze govora: sinteza izborom jedinica, sinteza difona (prijelazi sa zvuka na zvuk) i sinteza specifična za domenu. Sve tri metode su bazirane na spajanju dijelova snimljenog govora. Sinteza izborom jedinica koristi velike baze snimljenog govora, gdje se svaki snimljeni izgovor segmentira na manje dijelove (foneme, difone, slogove, morfeme, riječi, fraze, rečenice). Sinteza difona koristi minimalnu bazu podataka koja sadrži sve difone nekog jezika. Sinteza specifična za domenu spaja riječi i fraze i koristi se u sustavima u kojima je izlaz ograničen na određenu domenu (npr. rasporedi vožnji u željeznicama). Generalno, ova metoda daje sintetizirani govor koji zvuči vrlo prirodno, ali zbog razlika u prirodnim varijacijama u govoru i prirode automatiziranih tehnika za segmentaciju valnih oblika ponekad rezultira greškama u izlazima.

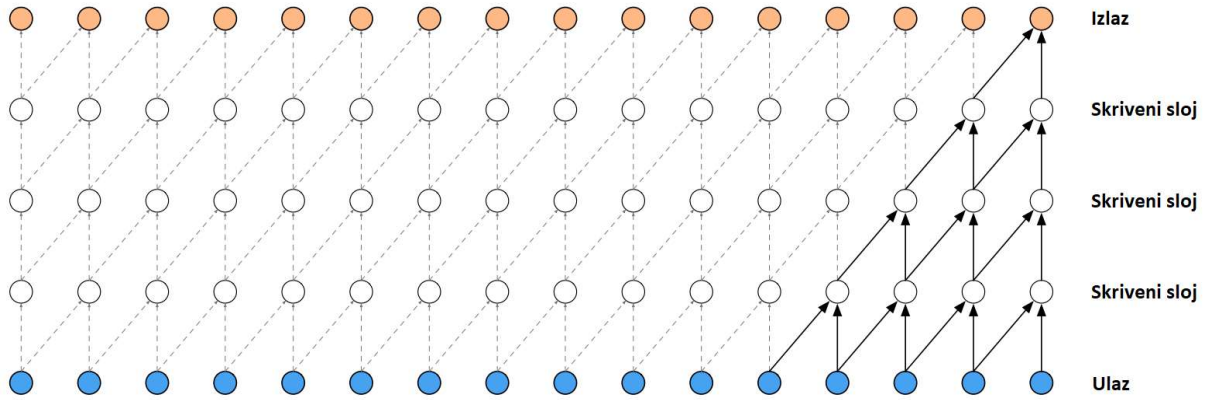
Tablica 2.1. Rezultati evaluacije za spojnu metodu iz 1998. [11]

	I	II
MOS	3.46	3.91
INTELL	3.48	3.98

U tablici 2.1. dani su rezultati evaluacije jednog od takvih sustava iz 1998. godine. Kao podatkovni skup za treniranje ovog skupa korištene su snimke profesionalne čitačice na engleskom jeziku. Evaluacija je podijeljena u dvije skupine, u prvoj kategoriji su rečenice s Harvarda i iz poslovnih vijesti (I), a u drugoj kategoriji su rečenice u obliku obavijesti (II). Svaka od rečenica je ocijenjena na skali od 1 do 5 za sveukupnu kvalitetu (MOS) i za razumljivost (INTELL – engl. *intelligibility*). Rezultati su evaluirani na 44 slušatelja. Rezultati prve kategorije s MOS-om od oko 3.5 ukazuju na to da su rezultati na granici s prihvatljivima, dok su rezultati druge kategorije prihvatljivi, ali ne izvrsni [12].

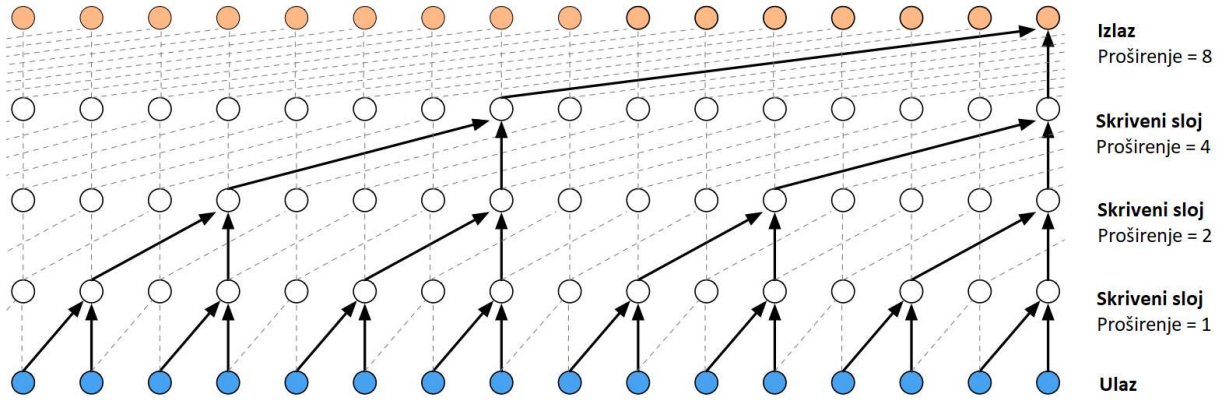
2.2.2. WaveNet

WaveNet [3,13] je neuronska mreža za generaciju sirovih zvučnih valova. Model je u potpunosti probabilistički i autoregresivan. U ovom generativnom modelu, svaki zvučni uzorak ovisi o prethodnim zvučnim uzorcima. Uvjetna vjerojatnost je modelirana kao stog konvolucijskih slojeva. Na slici 2.4. je prikazana vizualizacija stoga uzročnih konvolucijskih slojeva. Plavi kružići predstavljaju ulazne zvučne uzorke, dok narančasti predstavljaju izlazne zvučne uzorke. Bijeli kružići su skriveni konvolucijski slojevi.



Slika 2.4. Vizualizacija stoga uzročnih konvolucijskih slojeva [3]

Problem ovog pristupa je što treba velik broj slojeva da se poveća osjetljivo polje, odnosno broj prethodnih uzoraka koji utječu na sljedeći. Za rješavanje ovog problema, WaveNet koristi proširene uzročne konvolucijske slojeve (slika 2.5.). U ovom slučaju na sljedeći zvučni uzorak značajno utječe više prethodnih uzoraka. U primjeru s 3 skrivena sloja, 16 prethodnih uzoraka utječe na izlaz za razliku od 5 prethodnih uzoraka kod običnih uzročnih konvolucijskih slojeva.



Slika 2.5. Vizualizacija stoga proširenih uzročnih konvolucijskih slojeva [3]

Model je treniran na podatkovnom skupu dužine 24.6 sati za engleski jezik i na podatkovnom skupu od 34.8 sati za mandarinski jezik. Za oba su korišteni zvučni isječci profesionalnih čitačica.

U tablici 2.2. su prikazani rezultati evaluacije za WaveNet. Model je testiran na 100 rečenica koje nisu bile dio skupa za treniranje. Svaka rečenica je testirana na 8 ispitanika. Rezultati su uspoređeni s dva starija modela (parametarski i spojni modeli) i s prirodnim govorom s dvije vrste kompresije (8-bit μ-law i 16-bit linearni PCM). Za WaveNet model dobiven je MOS od 4.21 ± 0.081 za engleski jezik, što je znatno bolje od prethodnih modela za sintezu govora, ali i dalje nije na razini sa stvarnim govorom koji je dobio MOS od 4.55 ± 0.075 za engleski jezik u ovom radu. Dodatni problem ovog modela je što je proces generiranja zvuka kompjuterski vrlo zahtjevan.

Tablica 2.2. MOS i standardna pogreška aritmetičke sredine ocjena za WaveNet za engleski i mandarinski jezik

	američki engleski jezik	mandarinski kineski jezik
LSTM-RNN parametarski model	3.67 ± 0.098	3.79 ± 0.084
Spojni model baziran na HMM	3.86 ± 0.137	3.47 ± 0.108
WaveNet	4.21 ± 0.081	4.08 ± 0.085

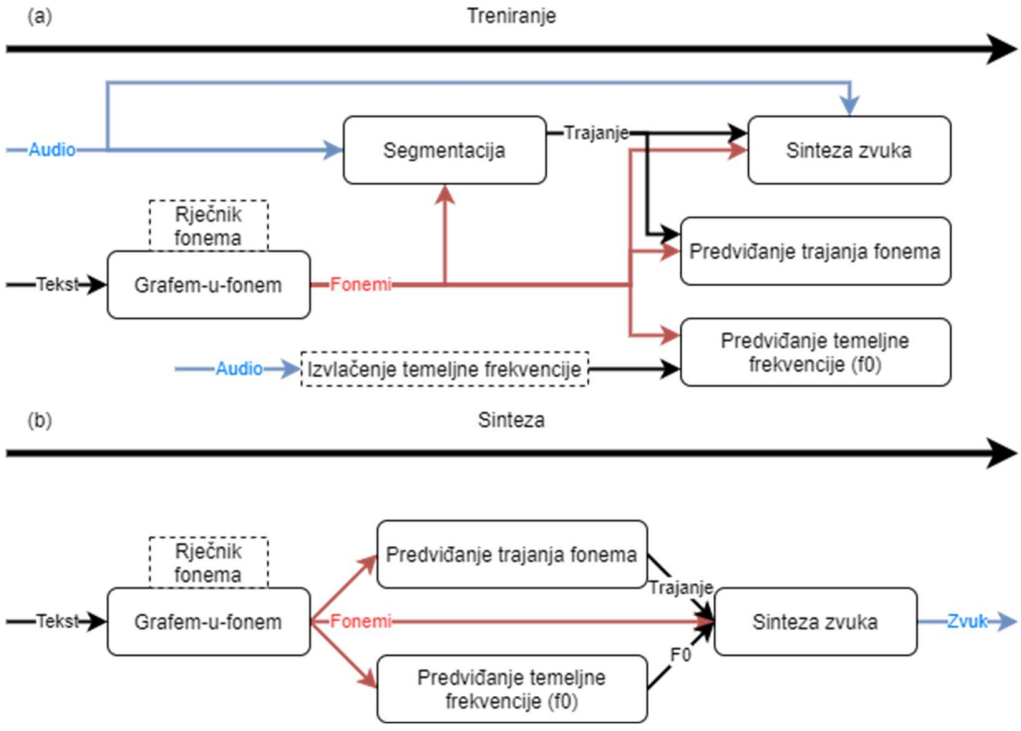
Prirodni govor (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Prirodni govor (16-bit linearni PCM)	4.55 ± 0.075	4.21 ± 0.071

2.2.3. Deep Voice 1

Deep Voice 1 [14] je sustav za pretvaranje teksta u govor razvijen korištenjem dubokih neuronskih mreža. Sastoji se od pet glavnih dijelova:

- Segmentacijski model za lociranje granica između fonema korištenjem dubokih mreža,
- Model za pretvaranje grafema u foneme,
- Model za predviđanje trajanja fonema,
- Model za predviđanje temeljne frekvencije, i
- Model za sintezu zvuka korištenjem varijante WaveNeta sa smanjenim brojem parametara.

Dijagram modela Deep Voice 1 prikazan je na slici 2.6. i prikazuje proceduru za treniranje (a) i proceduru za sintezu govora (b).



Slika 2.6. Dijagram Deep Voicea 1: (a) procedura za treniranje i (b) procedura za sintezu [14]

Model za pretvaranje grafema u foneme pretvara engleska slova u foneme. Segmentacijski model identificira gdje svaki fonem počinje i završava u zvučnom isječku. Model za predviđanje duljine fonema predviđa duljinu fonema za svaki fonem u slijedu. Model za predviđanje temeljne frekvencije predviđa je li određeni fonem izgovoren ili ne. Model za sintezu zvuka zatim kombinira izlaze iz modela za pretvaranje grafema u foneme, modela za predviđanje duljine i modela za predviđanje temeljne frekvencije. Rezultati ovog modela za engleski jezik su prikazani u tablici 2.3. Postignuti rezultati, iako zadovoljavajući, lošiji su od prethodno opisanog WaveNeta. Dodatan nedostatak ovog modela je što se svaki dio mreže trenira zasebno.

Tablica 2.3. MOS i interval pouzdanosti od 95% za Deep Voice 1 za engleski jezik [14]

	MOS
Deep Voice 1 (najbolji rezultati)	3.94 ± 0.26
Prirodni govor	4.45 ± 0.16

2.2.4. Tacotron 2

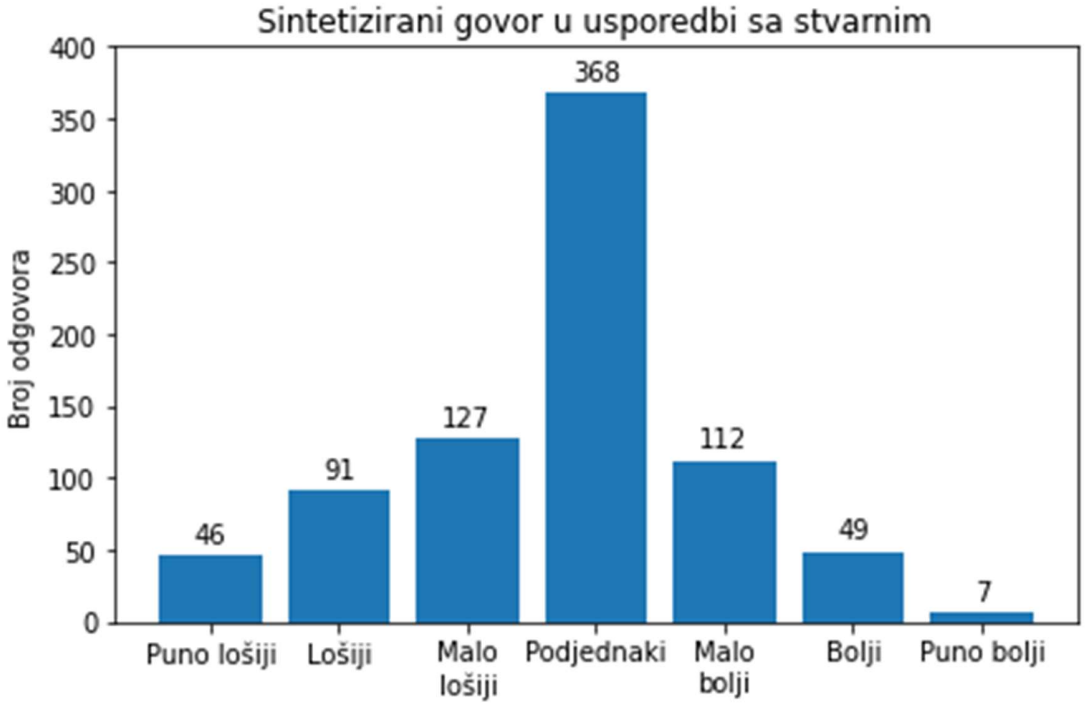
Tacotron 2 [7] je arhitektura neuronske mreže za sintezu govora iz teksta. Sustav je sastavljen od ponavljajuće slijed-u-slijed (engl. *sequence-to-sequence*) neuronske mreže za predviđanje značajki koja mapira ugradnje na razini znakova (engl. *character embedding*) u *mel-scale* spektrograme. Popraćena je izmijenjenim WaveNet modelom koji služi kao vokoder za sintezu valnih oblika u vremenskoj domeni iz tih spektrograma.

Model je treniran na podatkovnom skupu koji sadrži 24.6 sati govora profesionalne čitačice na engleskom jeziku. Rezultati su evaluirani na nekoliko načina. Prvi od njih je prikazan u tablici 2.4. gdje su prikazani MOS-ovi za Tacotron 2 i prirodni govor. Vidi se da je umjetno generirani govor postigao rezultate vrlo blizu rezultatima prirodnog govora.

Tablica 2.4. MOS i standardna pogreška aritmetičke sredine ocjena za Tacotron 2 za engleski jezik [7]

	MOS
Tacotron 2	4.526 ± 0.066
Prirodni govor	4.582 ± 0.053

Drugi od načina na koji je ovaj model testiran je usporedba generiranog zvuka s izgovorenim. Svaki par zvučnih isječaka ocjenjivači su trebali ocijeniti ocjenom od -3 (sintetizirani zvuk je puno lošiji) do 3 (sintetizirani zvuk je puno bolji). Dobiven je ukupni prosjek od -0.270 ± 0.155, što znači da su ocjenjivači imali malu preferencu za stvarni govor. Na slici 2.7. se vidi točna distribucija odgovora (800 ocjena na 100 isječaka).



Slika 2.7. Sintetizirani govor u usporedbi sa stvarnim [7]

Pri ručnoj obradi na novih 100 rečenica, utvrđeno je da nije bilo rečenica koje sadrže ponovljene riječi, šest koje sadrže krivo izgovorene riječi, jedna koja sadrži preskočenu riječ i 23 za koje je utvrđeno da sadrže neprirodnu prozodiju, kao što su naglašavanje krivih slogova ili riječi te neprirodan ton.

Zadnji način na koji je evaluiran ovaj model je bilo ocjenjivanje isječaka generiranih iz 37 naslova iz novina kako bi se testirala generalizirajuća sposobnost modela. Tacotron 2 je dobio MOS od 4.148 ± 0.124.

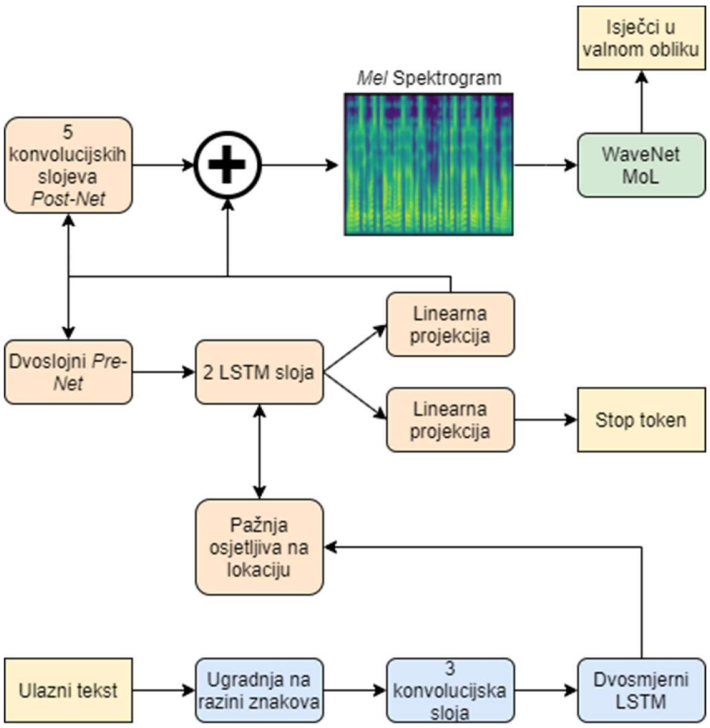
Iako ovaj model ima manjih problema s prozodijom, i dalje daje izvrsne rezultate vrlo slične pravom govoru. Velik problem ovog modela, kao i općenito svih metoda zasnovanih na dubokom učenju, je što zahtijeva velik skup podataka za treniranje.

3. Prijedlog rješenja za sintezu govora iz teksta za hrvatski jezik

Za realizaciju sinteze govora za hrvatski jezik, u ovom diplomskom radu je korišten prethodno opisani model Tacotron 2. U nastavku je detaljnije opisan ovaj model i njegovi dijelovi.

3.1. Tacotron 2

Tacotron 2 se trenutno smatra jednim od najboljih modela za sintezu govora iz teksta [15]. Nastao je na temelju starijih modela Tacotrona [5] i WaveNeta [3]. Tacotron 2 za ulaz ne koristi nikakva kompleksna lingvistička niti akustična svojstva. Generira govor koji može biti vrlo sličan ljudskom korištenjem neuronskih mreži treniranih na podacima koji sadrže samo primjere govora s pripadajućim tekstualnim transkriptom.



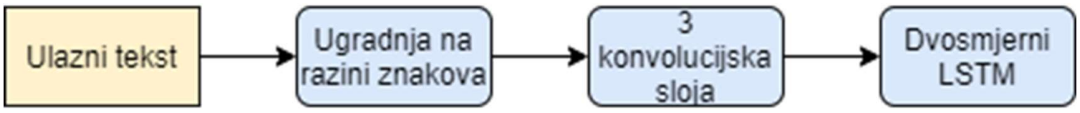
Slika 3.1. Model Tacotron 2 [7]

Na slici 3.1. su prikazani razni dijelovi modela Tacotron 2. Model je podijeljen na dva veća dijela: ponavljajuće slijed-u-slijed neuronske mreže za predviđanje značajki koja predviđa slijed *mel* spektrograma iz ulaznog niza znakova i izmijenjene verzije WaveNeta koja generira valne oblike u vremenskoj domeni prema predviđenim *mel* spektrogramima.

Mel spektrogrami za treniranje su izračunati kroz kratkotrajnu Fourierovu transformaciju (STFT – engl. *Short-Time Fourier Transformation*) koristeći okvire (engl. *frame*) veličine 50 ms, okvirne skokove (engl. *frame hop*) veličine 12.5 ms i koristeći Hannov prozor (engl. *Hann window function*). STFT veličine se transformiraju u *mel* skalu korištenjem 80-kanalne *mel* banke filtara (engl. *filterbank*) raspona 125 Hz do 7.6 kHz, popraćene logaritamskom kompresijom. Prije kompresije, izlazne veličine iz banke filtara su odrezane na minimalnu vrijednost od 0.01 kako bi se ograničio dinamički raspon u logaritamskoj domeni. U nastavku je objašnjen princip rada i svrha svakog od manjih dijelova Tacotron 2 modela.

3.1.1. Koder

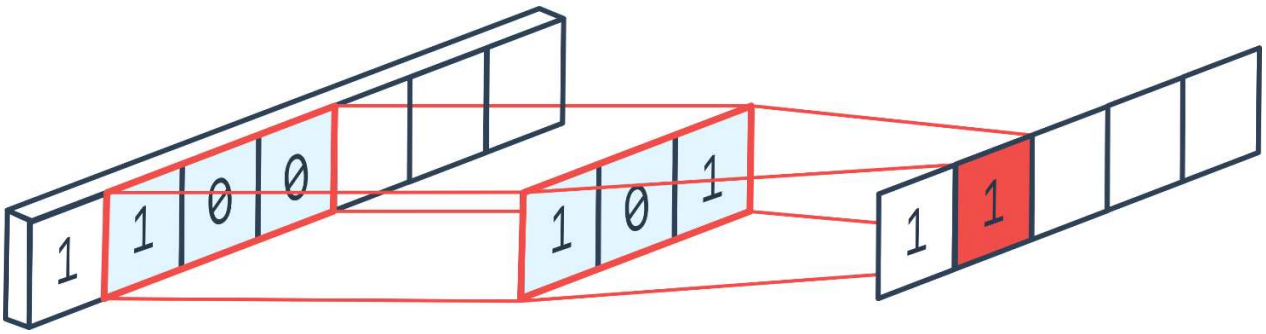
Koder modela Tacotron 2 (Slika 3.2.) pretvara dani ulazni niz znakova u značajke koje dekoder poslije koristi za predviđanje spektrograma. Ulazni znakovi se prikazuju korištenjem naučene 512-dimenzionalne ugradnje na razini znakova.



Slika 3.2. Koder Tacotrona 2

Ugradnja na razini znakova funkcionira na način da svakom znaku daje 512-dimenzionalni vektor koji predstavlja taj znak. Ovaj vektor ovisi o samom znaku i o drugim znakovima koji se pojavljuju uz taj znak.

Nakon ugradnje na razini znakova, slijede tri konvolucijska sloja od kojih se svaki sastoji od 512 filtera oblika 5×1, odnosno svaki filter obuhvaća 5 znakova. Nakon njih slijedi normalizacija serije (engl. *batch normalization*) i ReLU aktivacijska funkcija.



Slika 3.3. Simbolički prikaz 1D konvolucije [16]

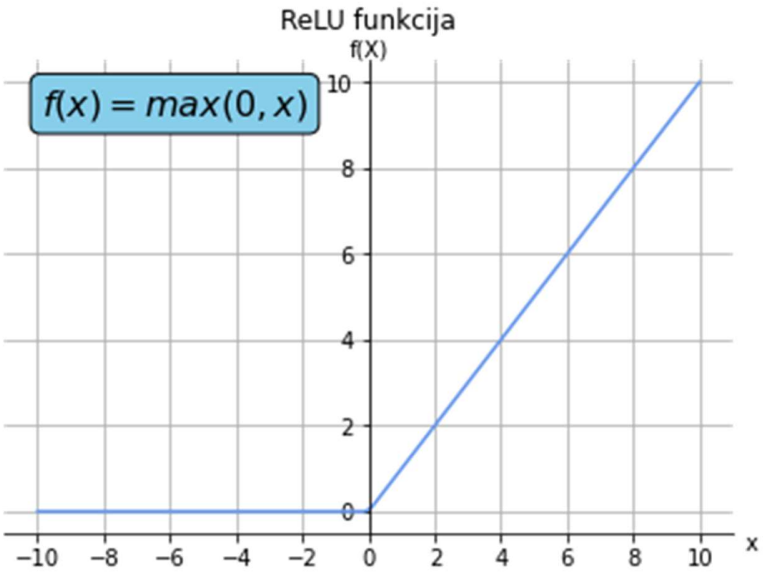
Na slici 3.3. je simbolički prikazana 1D konvolucija. Na slici je prikazana konvolucija s filterom oblika 3×1. Ovakav filter se zatim pomiče po cijelom ulaznom vektoru i ovisno o svom obliku daje izlazni vektor. Na gore navedenom primjeru se 1 iz ulaznog vektora množi s 1 iz filtera, zatim 0 iz ulaznog vektora s 0 iz filtera, zatim 0 iz ulaznog vektora s 1 iz filtera i sve se na kraju zbraja u 1 u izlaznom vektoru (1×1 + 0×0 + 0×1 = 1).

Normalizacija serije funkcionira na način da prvo normalizira izlazne podatke, u ovom slučaju iz konvolucijskih slojeva, zatim množi normalizirani izlaz nekim parametrom i na kraju tom rezultatu dodaje neki drugi parametar. Svi ovi parametri se mogu trenirati. Cilj normalizacije serije je ubrzavanje procesa treniranja i smanjivanje utjecaja rubnih vrijednosti.

ReLU aktivacijska funkcija ima sljedeći oblik:

$$f(x) = max\,(0, x) \tag{3-1}$$

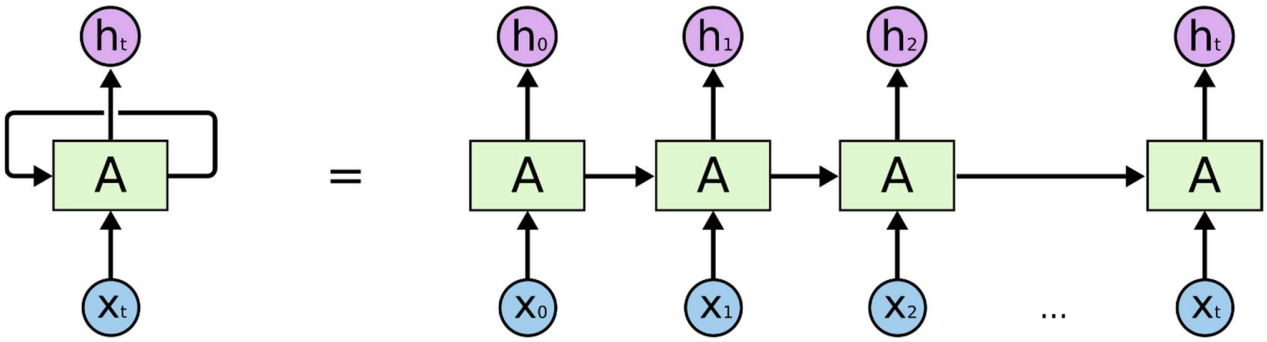
Funkcija vraća maksimalnu vrijednost između 0 i ulazne vrijednosti, odnosno, sve ulazne vrijednosti manje od 0, imat će 0 za izlaz (slika 3.4.).



Slika 3.4. ReLU aktivacijska funkcija

Izlaz iz posljednjeg konvolucijskog sloja se prosljeđuje u jedan dvosmjerni LSTM s 512 jedinica (256 u svakom smjeru) kako bi se dobile kodirane značajke ulaznog teksta. Mreže s dugotrajnom

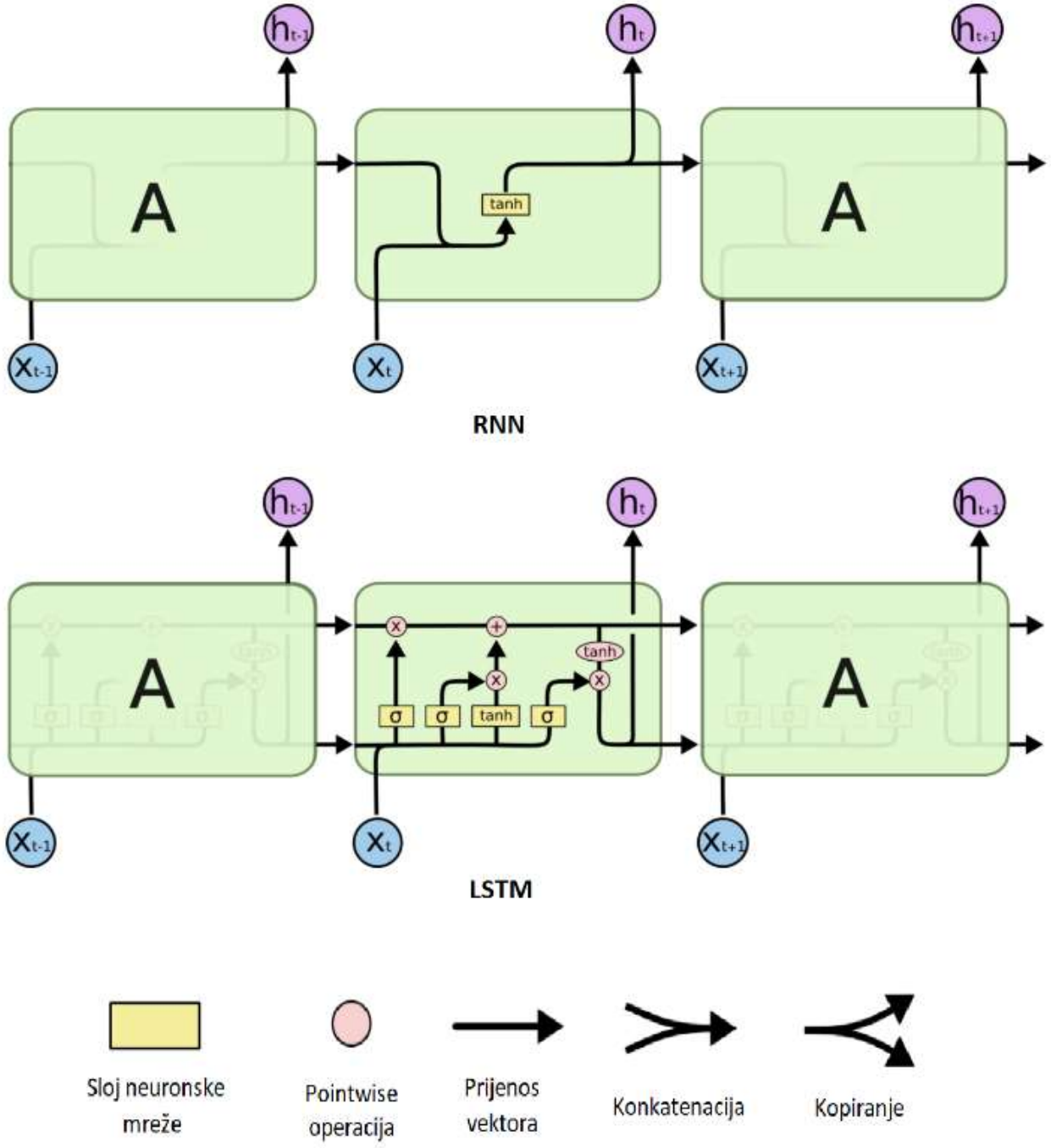
kratkotrajnom memorijom (engl. *long short term memory* – LSTM) su posebna verzija povratnih neuronskih mreža (engl. *recurrent neural network* – RNN). RNN (slika 3.5.) prima na ulazu x_t i daje izlaz h_t. Kod RNN-a postoji povratna veza, odnosno izlaz ne ovisi samo o ulazu, nego i o stanju mreže. Ovakva petlja omogućava informaciji da se proslijedi iz jednog koraka u sljedeći. RNN ima problema s tzv. *vanishing gradients* i *exploding gradients*, odnosno dugačke sekvence mogu uzrokovati da pogreška postane prevelika ili premala. Za rješavanje tog problema se koristi LSTM.



Slika 3.5. Prikaz RNN-a i "odmotanog" RNN-a [17]

LSTM-ovi su dizajnirani da riješe problem dugotrajne zavisnosti, odnosno situacije u kojima nije dovoljno uzeti u obzir samo nedavni kontekst.

RNN ima samo jednu informaciju koju prosljeđuje, trenutno stanje mreže, i ima jedan sloj neuronske mreže koji se trenira (slika 3.6.). LSTM ima 3 dijela u svom modulu: dio za zaboravljanje, dio za dodavanje informacije u stanje ćelije i dio za izlaz. LSTM umjesto samo jednog izlaza ima i stanje ćelije koje utječe na sljedeće stanje ćelije i izlaz iz LSTM-a. Dio za zaboravljanje odlučuje koliko će stanje prethodne ćelije utjecati na trenutnu ćeliju. Dio za dodavanje informacije u stanje ćelije odlučuje koliko će izlaz iz prethodne i ulaz u ovu ćeliju utjecati na trenutnu ćeliju. Posljednji dio je dio za izlaz u kojem se odlučuje koliko će izlaz prethodne ćelije i ulaz trenutne utjecati na izlaz te se to spaja sa stanjem trenutne ćelije. Izlaz i stanje ćelije se zatim prosljeđuje na sljedeću ćeliju. Svi ovi dijelovi LSTM-a se treniraju.



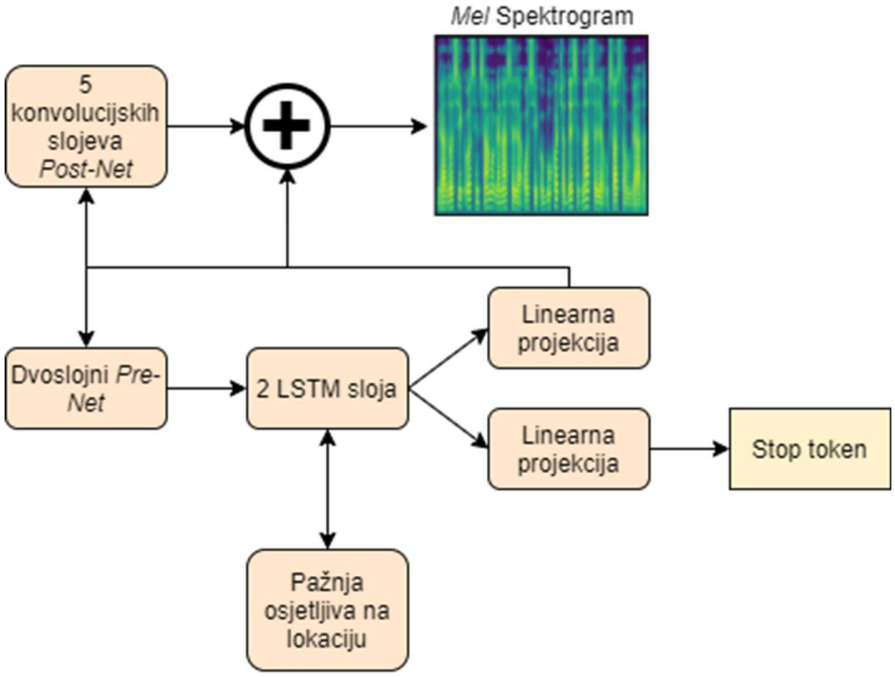
Slika 3.6. Prikaz modula RNN-a i LSTM-a [17]

Dvosmjerni LSTM je kombinacija dva jednosmjerna LSTM-a. Jedan LSTM prolazi od početka ulaznog niza prema kraju, kao inače, dok drugi LSTM prolazi od kraja istog ulaznog niza prema početku. Izlazi se zatim spajaju. Na ovaj način na izlaz ne utječe samo ono što dolazi prije nekog

znaka, nego i ono što dolazi nakon istog. Nakon ovog dvosmjernog LSTM-a se kodirane značajke prosljeđuju dekoderu.

3.1.2. Dekoder

Dekoder (Slika 3.7.) uzima kodirane značajke iz kodera i predviđa *mel* spektrograme jedan po jedan okvir, u koracima. Spektrogram se dalje predaje u WaveNet Vocoder koji pretvara *mel* spektrograme u valne oblike.

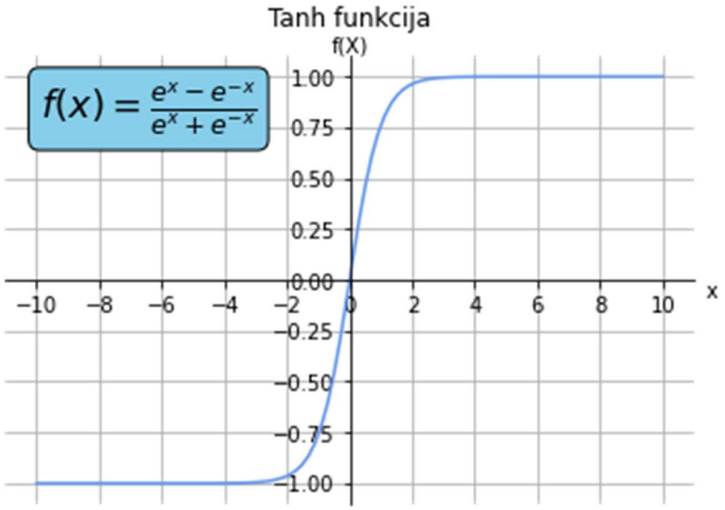


Slika 3.7. Dekoder Tacotrona 2

Mreža pažnje (engl. *attention network*) služi kako bi se pažnja usmjerila samo na određene dijelove ulaza. Mreža pažnje sumira cijeli kodirani slijed u vektor konteksta fiksne duljine za svaki korak predviđanja okvira. Pažnja osjetljiva na lokaciju (engl. *location sensitive attention*) se koristi kako bi težine pažnje iz prethodnih koraka dekodera utjecale na ulaz u trenutni korak. To potiče model da se kreće naprijed dosljedno kroz ulaz, ublažavajući potencijalne pogreške gdje se dekoder ponavlja ili zanemaruje neke slijedove.

Predviđanje iz prethodnog koraka prolazi kroz *pre-net* neuronsku mrežu koja sadrži dva potpuno spojena sloja napravljenih od 256 skrivenih ReLU jedinica. Izlaz iz ovog *pre-neta* se ujedinjuje s izlazom iz mreže pažnje i provlači se kroz stog načinjen od dva jednosmjerna LSTM sloja s 1024

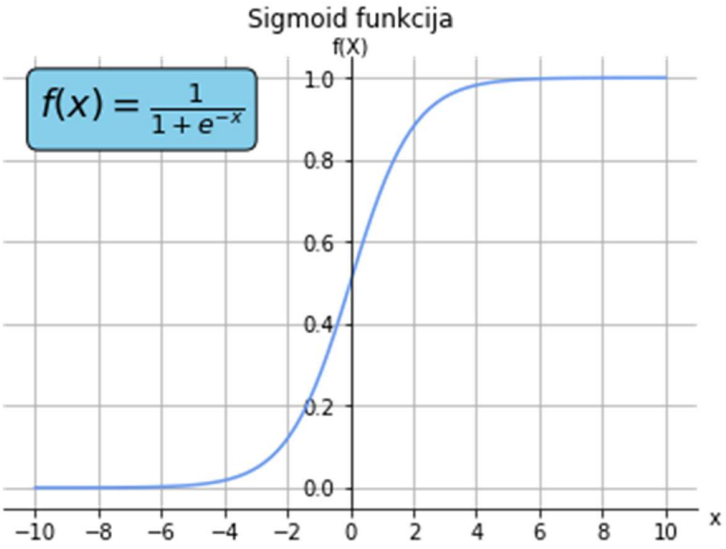
jedinice. Zatim se izlaz iz LSTM slojeva ujedinjuje s kontekstom iz mreže pažnje te se projicira kroz linearnu transformaciju. Na posljetku se predviđeni spektrogram provlači kroz 5-slojni konvolucijski *post-net* koji predviđa ostatak koji se dodaje na predviđanje iz prethodnog koraka kako bi se poboljšala sveukupna rekonstrukcija. Svaki od slojeva iz *post-neta* se sastoji od 512 filtera oblika 5×1 s normalizacijom serije, nakon koje je tangens hiperbolni (*tanh*) (slika 3.8.) aktivacijska funkcija na svim slojevima osim zadnjem.



Slika 3.8. Tanh aktivacijska funkcija

Za treniranje se minimizira suma srednjih kvadratnih pogrešaka (engl. *Mean Squared Error* – MSE) vrijednosti predviđenog okvira prije *post-neta* i poslije za pomoć konvergenciji modela.

Paralelno s predviđanjem okvira spektrograma, izlaz iz LSTM-a dekodera ujedinjuje se s kontekstom iz mreže pažnje i projicira u skalar. Taj skalar se zatim provlači kroz sigmoid aktivacijsku funkciju (slika 3.9.) s ciljem predviđanja vjerojatnosti je li izlazni slijed završio. Ovaj "stop token" se koristi za vrijeme zaključivanja kako bi model mogao dinamički odlučiti treba li zaustaviti generiranje umjesto da uvijek generira isječak jednake duljine. Točnije, generiranje prestaje na prvom okviru gdje ova vjerojatnost prelazi prag od 0.5.



Slika 3.9. Sigmoid aktivacijska funkcija

Za konvolucijske slojeve koristi se regularizacija u obliku nasumičnog izbacivanja neurona tijekom učenja mreže (engl. *dropout*) s vjerojatnošću od 0.5. Za LSTM slojeve tip regularizacije koji se koristi je *zoneout* s vjerojatnošću od 0.1. Kako bi se dobila varijacija za vrijeme zaključivanja, ispadanje s vjerojatnošću 0.5 se primjenjuje samo na slojeve iz *pre-neta*.

3.1.3. WaveNet

Zadnji dio modela Tacotron 2 je WaveNet vocoder koji pretvara izlaz iz dekodera u obliku *mel* spektrograma u zvučni valni oblik, odnosno zvučnu datoteku [7]. Način na koji obični WaveNet funkcionira ukratko je objašnjen u drugom poglavlju ovog rada. Tacotron 2 koristi modificiranu verziju WaveNet-a. Kao i u originalnoj arhitekturi, ima 30 proširenih konvolucijskih slojeva grupiranih u 3 ciklusa širenja (engl. *dilation cycles*). Kako bi radio s okvirnim skokovima veličine 12.5 ms okvira spektrograma, koriste se samo 2 sloja naduzorkovanja (engl. *upsampling*) u stogu za uvjetovanje umjesto 3 sloja. Za izračunavanje distribucija logističke smjese, izlazni stog se provlači kroz ReLU aktivaciju popraćenu linearnom projekcijom kako bi se predvidjeli parametri (srednja vrijednost, logaritamska skala, težina smjese) za svaku komponentu smjese. Gubitak se računa kao negativna log-vjerojatnost stvarnog isječka. Za izradu predloženog modela u ovom radu se WaveNet nije trenirao, nego samo koder i dekoder, a za WaveNet je korišten Nvidiin model treniran na velikom broju različitih glasova [18].

3.2. Podatkovni skup

U okviru rada napravljen je vlastiti skup podataka jer autoru nisu dostupni odgovarajući podaci za učenje modela za sintezu govora iz teksta na hrvatskom jeziku. Zbog ovakvog podatkovnog skupa, konačni model sintetizira glas nalik na autorov. Podatkovni skup potreban za treniranje modela Tacotron 2 sastoji se od zvučnih datoteka snimljenog govora te pripadajućeg tekstualnog transkripta za svaku od datoteka. Za tekstove za čitanje su korišteni članci s Wikipedije, knjige (Povratak Filipa Latinovicza, Bajke braće Grimm i Sapiens, Kratka povijest čovječanstva) i članci iz novina. U tekstovima su svi brojevi i skraćenice ispisani, primjerice skraćenica "DNK" je ispisana kao "de-en-ka". Za snimanje zvučnih datoteka korišten je mikrofon slušalica Logitech g230 i osobno računalo. Nakon snimanja zvučnih signala, istima je uklonjena buka te su izrezane nepotrebne pauze. Zvučni signali su zatim spremljeni u WAV formatu (engl. *Waveform Audio File Format*). Konačni podatkovni skup sadržava 5 sati i 39 minuta obrađenih zvučnih isječaka s pripadajućim transkriptom.

Tablica 3.1. Prikaz primjera podataka iz podatkovnog skupa

Ime datoteke	Tekstualni transkript
DZ001-0009.wav	Osijek je grad u istočnoj Hrvatskoj.
DZ001-0010.wav	Smješten je u ravnici na desnoj obali rijeke Drave između šesnaestog i dvadeset četvrtog kilometra od ušća u Dunav.
DZ001-0011.wav	Najveći je grad u Slavoniji, četvrti po veličini grad u Hrvatskoj, te je industrijsko, upravno, akademsko, sudsko i kulturno središte Osječko-baranjske županije.
DZ001-0012.wav	Osijek je grad s najviše zelenila i zelenih površina u Hrvatskoj;
DZ001-0013.wav	na području grada nalazi se sedamnaest parkova u ukupnoj površini od tristo devedeset četiri metra kvadratna.

U tablici 3.1. prikazan je isječak podataka iz podatkovnog skupa. U drugom stupcu se nalazi tekstualni transkript govora koji je snimljen, dok se u prvom stupcu nalazi ime datoteke s pripadajućim tekstom. Datoteke su napravljene po uzoru na LJSpeech podatkovni skup [19] koji je korišten za treniranje

Nvidiine implementacije Tacotron 2 modela, stoga datoteke nisu prevelike, pretežno sadrže jednu rečenicu ili dio rečenice.

3.3. Treniranje modela za sintezu govora iz teksta

Za treniranje je korišten Tacotron 2 model, točnije Nvidiina implementacija istog [20]. Treniranje je obavljeno treniranjem na predtreniranom modelu za engleski jezik kako bi se postigla brža konvergencija modela. Treniranje je također isprobano i bez korištenja predtreniranog modela, ali nisu postignuti zadovoljavajući rezultati zbog male veličine podatkovnog skupa za hrvatski jezik.

3.3.1. Predobrada podataka

Podaci su snimljeni kao stereo zvuk s frekvencijom uzorkovanja 44100 Hz, ali pošto je predtrenirani model, kao i WaveNet vocoder, bio treniran s mono zvukovima s frekvencijom uzorkovanja 22050 Hz, zvučni isječci u podatkovnom skupu su također prerađeni u taj oblik.

Skup podataka je zatim podijeljen na 3 dijela: skup za treniranje, skup za validaciju i skup za testiranje. Skup za treniranje se sastoji od 3205 datoteka s pripadajućim transkriptom i ima ukupno trajanje od oko 5 sati i 30 minuta. Ovaj skup se koristi za samo treniranje modela. Skup za validaciju se sastoji od 100 datoteka s pripadajućim transkriptom i koristi se za validaciju i poslije za odabir konačnog modela. Skup za testiranje se također sastoji od 100 datoteka s pripadajućim transkriptom. Dio testnih datoteka je poslije korišten za evaluaciju rezultata treniranog modela.

Pošto u engleskom jeziku nema nekih znakova kojih ima u hrvatskom jeziku (ć, č, š, ž), ti su znakovi dodani u skup znakova. Također, umjesto korištenja čistača (engl. *cleaners*) za engleski jezik, korišteni su osnovni čistači. Ovi čistači pretvaraju sav tekst u mala slova i uklanjaju nepotrebne razmake, npr. rečenica „Osijek je grad u istočnoj Hrvatskoj. „ bi bila pretvorena u „osijek je grad u istočnoj hrvatskoj.“, bez nepotrebnih razmaka iza točke.

3.3.2. Hiperparametri

Hiperparametri korišteni za treniranje modela su podijeljeni u nekoliko skupina. Pojedine skupine hiperparametara su u nastavku objašnjene.

Prvo skupina hiperparametra su hiperparametri za proces učenja. Oni definiraju način na koji će se učenje provoditi. Hiperparametri za proces učenja, s njihovim vrijednostima i kratkim objašnjenjima nalaze se u tablici 3.2. Ovi hiperparametri su preuzeti od Nvidiine verzije modela [20]. Broj epoha

određuje koliko puta će svaki ulazni podatak proći kroz mrežu. Broj iteracija po kontrolnoj točki samo određuje koliko iteracija učenja će biti između svakog spremanja kontrolne točke.

Tablica 3.2. Hiperparametri za proces učenja

Hiperparametar	Vrijednost	Kratko objašnjenje
epochs	800	Epoha je jedan prolazak cijelog podatkovnog skupa kroz mrežu
iters_per_checkpoint	1000	Broj iteracija prije snimanja kontrolne točke
seed	1234	Sjeme za nasumične događaje
distributed_run	False	Hoće li se trenirati na više grafičkih kartica
cudnn_enabled	True	Biblioteka za ubrzavanje korištenjem grafičke kartice za duboke neuronske mreže

Sljedeći su hiperparametri za zvuk. Oni definiraju zvučne datoteke podatkovnog skupa, kao i spektrograme koji će se generirati modelom, pa i zvučne datoteke koje će se generirati iz spektrograma korištenjem WaveNeta. Hiperparametri za zvuk, s njihovim vrijednostima i kratkim objašnjenjem se nalaze u tablici 3.3. Ove vrijednosti ovise o zvučnim isječcima koji se koriste, a pošto je treniranje nastavljeno na Nvidiin model za engleski jezik, vrijednosti ovih hiperparametara su identične onima korištenim za Nvidiin model [20].

Tablica 3.3. Hiperparametri za zvuk

Hiperparametar	Vrijednost	Kratko objašnjenje
sampling_rate	22050	Brzina uzorkovanja zvučnih signala
hop_length	256	Veličina okvirnog skoka, približno 12.5 ms za 22050 Hz sampling rate

win_length	1024	Duljina prozora, približno 50 ms za 22050 Hz sampling rate
n_mel_channels	80	Broj kanala u *mel* spektrogramu
mel_fmin	0	Minimalna frekvencija zvučnog signala na *mel* skali
mel_fmax	8000.0	Maksimalna frekvencija zvučnog signala na *mel* skali

Sljedeći su hiperparametri modela. To su hiperparametri za razne dijelove modela. U tablici 3.4. su prikazani hiperparametri modela s pripadajućim vrijednostima i kratkim objašnjenjem podijeljeni u nekoliko skupina. Ove vrijednosti hiperparametara su također preuzete od Nvidiine verzije modela [20]. Ove vrijednosti definiraju strukturu modela opisanu ranije u ovom poglavlju.

Tablica 3.4. Hiperparametri modela

Hiperparametar	**Vrijednost**	**Kratko objašnjenje**
symbols_embedding_dim	512	Dimenzija vektora uležištenja znakova
Parametri kodera		
encoder_kernel_size	5	Veličina kernela u konvolucijskim mrežama kodera
encoder_n_convolutions	3	Broj konvolucijskih slojeva u koderu
encoder_embedding_dim	512	Dimenzija vektora uležištenja kodera
Parametri dekodera		
decoder_rnn_dim	1024	Broj jedinica u LSTM-u dekodera
prenet_dim	256	Broj ReLU jedinica u *pre-netu*

max_decoder_steps	1000	Maksimalan broj koraka dekodera, služi kako se ne bi došlo u beskonačnu petlju generiranja spektrograma
gate_threshold	0.5	Prag za prestanak generiranja spektrograma
p_attention_dropout	0.1	*Dropout* za mrežu pažnje
p_decoder_dropout	0.1	*Dropout* za sve konvolucijske slojeve i *pre-net*
Parametri *post-neta*		
postnet_embedding_dim	512	Dimenzija vektora uležištenja *post-neta*
postnet_kernel_size	5	Veličina kernela konvolucije *post-neta*
postnet_n_convolutions	5	Broj konvolucijskih slojeva u *post-netu*

Posljednji su hiperparametri optimizacije. Predstavljaju hiperparametre po kojima se treniraju neuronske mreže. Hiperparametri optimizacije, s njihovim vrijednostima i kratkim objašnjenjem se nalaze u tablici 3.5. Vrijednosti ovih hiperparametara su preuzete od Nvidiinog modela [20], osim vrijednosti *batch_size* koja je smanjena sa 64 na 32 zbog korištenog sklopovlja.

Tablica 3.5. Hiperparametri optimizacije

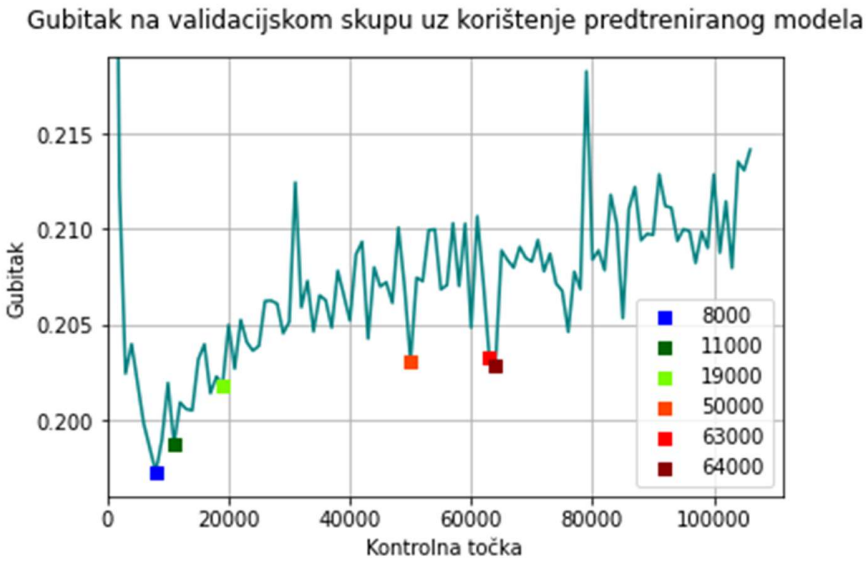
Hiperparametar	Vrijednost	Kratko objašnjenje
learning_rate	1e-3	Određuje veličinu koraka na svakoj iteraciji dok se kreće prema minimumu loss funkcije, veći learning rate ubrzava treniranje, ali može dovesti do "preletanja"

weight_decay	1e-6	Parametar koji utječe na loss funkciju, ali također smanjuje težine pa tako sprečava da ranije iteracije imaju prevelik utjecaj [22]
batch_size	32	Veličina serije za treniranje, veća serija omogućuje brže treniranje, ali je zahtjevnije za sklopovlje

3.3.3. Treniranje modela i odabir konačnog modela

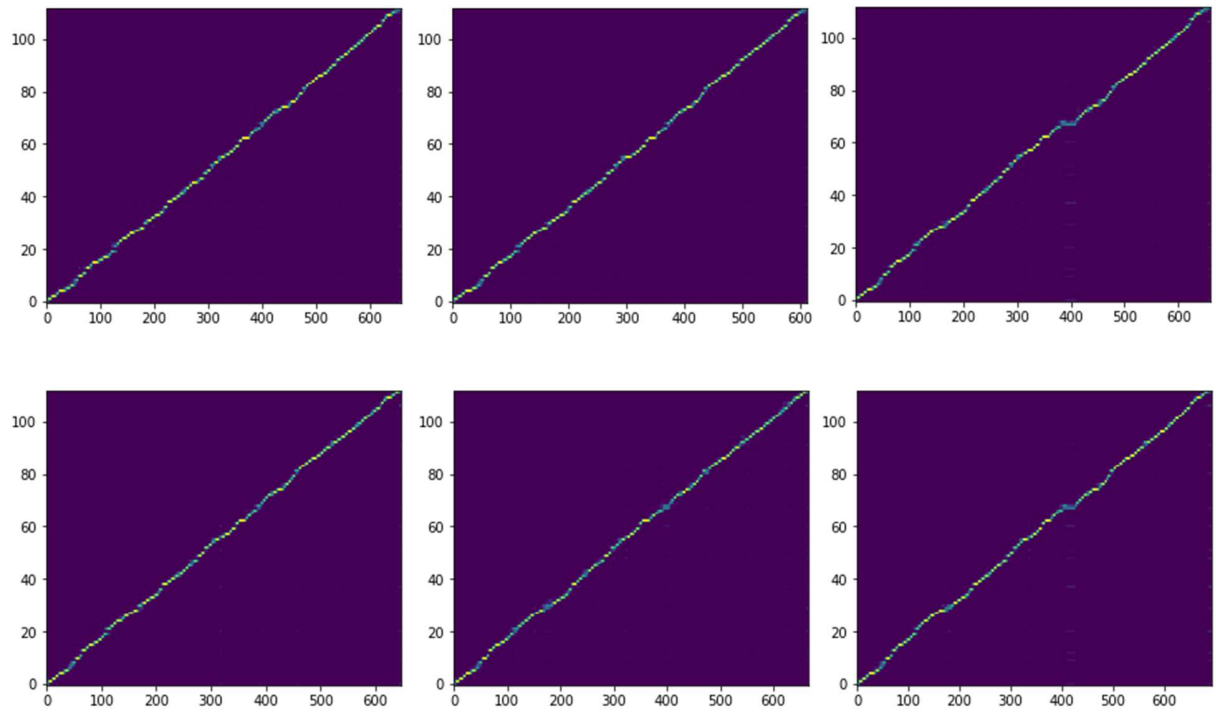
Obavljeno je treniranje bez korištenja predtreniranog modela za engleski jezik i uz korištenje. Iako je gubitak na validacijskom skupu tijekom treniranja bez korištenja predtreniranog modela (Slika 3.10.) bio sličan gubitku uz korištenje predtreniranog modela (Slika 3.11.), model nije dao razumljiv zvuk. Treniranje uz korištenje predtreniranog modela je prekinuto nakon 106000 iteracija jer je gubitak na validacijskom skupu podataka (engl. *validation loss*) rastao već velik broj iteracija (vidi sliku 3.11.) i pretpostavilo se da model počeo pretjerano usklađivati na podatke za učenje (engl. *overfitting*).

Slika 3.10. Gubitak na validacijskom skupu podataka tijekom treniranja bez korištenja predtreniranog modela



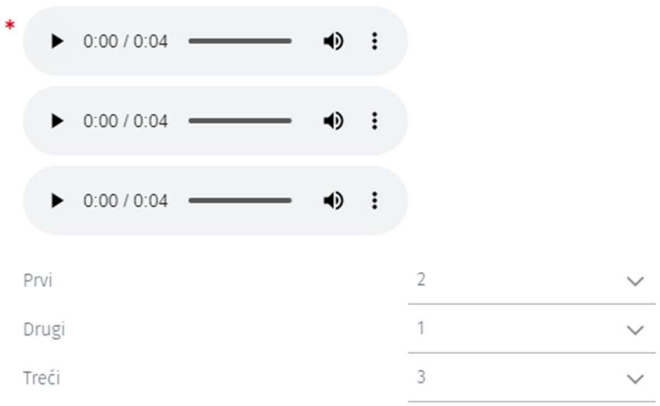
Slika 3.11. Gubitak na validacijskom skupu podataka tijekom treniranja uz korištenje predtreniranog modela

Prema gubitku na validacijskom skupu podataka, odabrano je nekoliko kontrolnih točaka za daljnji odabir konačnog modela (8000, 11000, 19000, 50000, 63000, 64000 kao što je obojenim kvadratima naznačeno na slici 3.11.). Odabir je dalje sužen po poravnanjima (engl. *alignment*) kontrolnih točaka na šest rečenica. Poravnanja su slijed težina pažnje iz dekodera [20]. Te težine predstavljaju koliko je svaki okvir predviđenog spektrograma povezan s prethodnim okvirima. Što je krivulja poravnanja bliža dijagonali, odnosno što je manje prosječno odstupanje od dijagonale, to je izlaz iz modela bolji, s manje prekida i nepotrebnih stanki. Primjer poravnanja za jednu rečenicu je prikazan na slici 3.12. U primjeru sa slike, kontrolne točke 8000, 11000, 50000 i 63000 imaju bolja poravnanja od kontrolnih točaka 19000 i 64000.



Slika 3.12. Primjer poravnanja za jednu rečenicu (kontrolne točke 8000, 11000 i 19000 u gornjem redu, 50000, 63000 i 64000 u donjem redu)

S obzirom na poravnanja, izabrane su kontrolne točke 8000, koja je imala najmanji gubitak, 50000 i 64000 za daljnji odabir. Odabrane kontrolne točke su zatim iskorištene za generiranje 10 različitih rečenica i izvršeno je subjektivno testiranje. Nakon generiranja svakog zvučnog isječka, na istom je proveden postupak uklanjanja buke (engl. *denoise*). Ovaj postupak uklanja pristranost posljednjeg dijela modela, modificiranog WaveNeta. Pristranost modela može uzrokovati propuštanje veza između značajki i izlaza (engl. underfitting). Prije provođenja ovog postupka, zvučni isječci su sadržavali kratke intervale vrlo visokog tona. Postupak uklanjanja buke je proveden i na svim sintetiziranim rečenicama korištenim za evaluaciju konačnog modela. Tri rečenice su bile iz testnog skupa, dok je sedam bilo potpuno novih. Za svaku od rečenica su ispitanicima u web anketi bili prikazani zvučni isječci za svaku od kontrolnih točaka u nasumičnom redoslijedu (Slika 3.13.). Ispitanik je tada trebao sortirati svaki od isječaka po tome koliko prirodno zvuči, gdje 1 označava da najbolje zvuči, a 3 da najlošije zvuči. Rezultati ove ankete provedene na 12 osoba su prikazani u tablici 3.6. Rezultati prikazuju prosječnu poziciju za svaku od rečenica za svaku kontrolnu točku.



Slika 3.13. Primjer pitanja iz ankete za odabir kontrolne točke

Tablica 3.6. Rezultati ankete za odabir kontrolne točke

Rezultati	**8000**	**50000**	**64000**
Prva rečenica	1.17	2	2.83
Druga rečenica	1.33	2.67	2
Treća rečenica	2.08	2.17	1.75
Četvrta rečenica	2.33	1.5	2.17
Peta rečenica	1.92	1.42	2.67
Šesta rečenica	1.75	1.33	2.92
Sedma rečenica	2.58	1.25	2.17
Osma rečenica	2.75	1.75	1.5
Deveta rečenica	2.08	1.83	2.08
Deseta rečenica	2	1.92	2.08
Ukupno	**1.999**	**1.784**	**2.217**

Iz tablice 3.6. je vidljivo da je kontrolna točka na 50000 iteracija dobila najbolje ocjene i tako je ta kontrolna točka izabrana kao konačni model. Ovaj model je detaljno evaluiran u idućem poglavlju.

4. Evaluacija predloženog rješenja za sintezu govora iz teksta

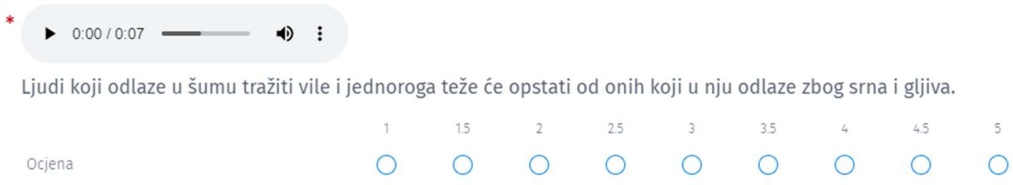
Za evaluaciju sintetizatora govora se koristi srednja ocjena slušatelja (MOS). MOS koristi skalu od 1 do 5 sa značenjima ocjena kako je prikazano u tablici 4.1. Prema [7] koristi se skala od 1 do 5 s koracima od 0.5 i takav korak je odabran i u ovom radu. Evaluacija je provedena u četiri dijela. Svaki od tih dijelova naveden je i opisan u nastavku.

Tablica 4.1. MOS skala [12]

Ocjena	Značenje
5	Izvrsno (*excellent*)
4	Dobro (*good*)
3	U redu (*fair*)
2	Loše (*poor*)
1	Vrlo loše (*bad*)

4.1. Subjektivna ocjena kvalitete izgrađenog modela na testnom skupu

Prva anketa se sastojala od 80 zvučnih isječaka, 40 sintetiziranih i 40 pročitanih koji su slušatelju predstavljene u nasumičnom redoslijedu. Transkripti 40 pročitanih isječaka korištenih u ovom testiranju su korišteni za sintetiziranje drugih 40 isječaka korištenih u ovom testiranju. Iako rečenice iz testnog skupa nisu ni na kakav način sudjelovale u treniranju modela, bile su iz istih izvora kao rečenice iz skupa za treniranje i iz validacijskog skupa. Primjer pitanja iz prve ankete prikazan je na slici 4.1.



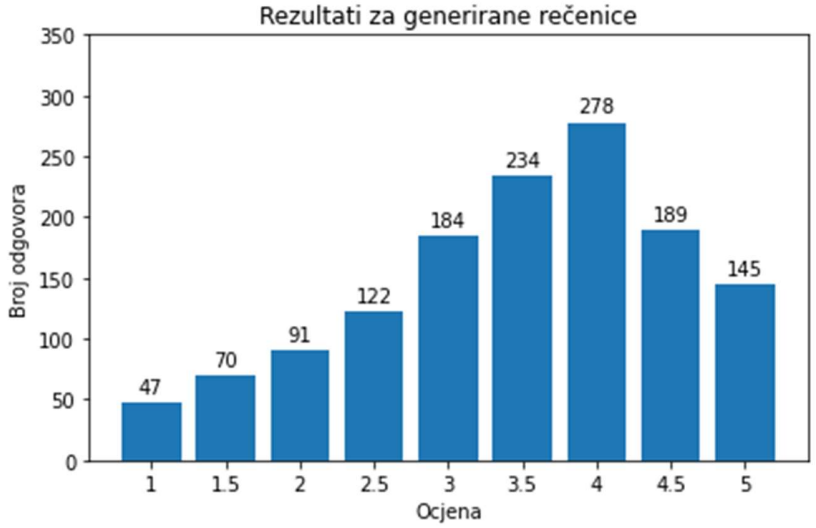
Slika 4.1. Primjer pitanja iz prve ankete

Rezultati prve ankete su prikazani u tablici 4.2. U tablici su prikazani MOS i standardna pogreška aritmetičke sredine ocjena za sintetizirane rečenice i za pročitane rečenice. Anketa je provedena na 34 ispitanika.

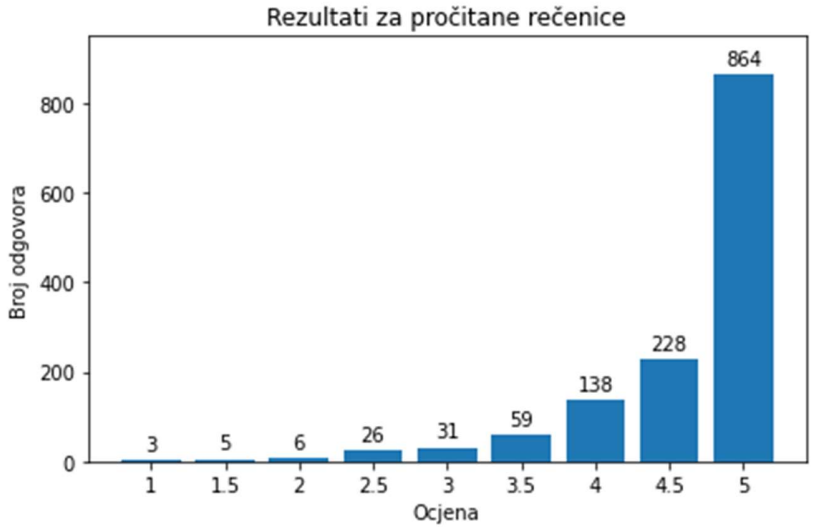
Tablica 4.2. Rezultati prve ankete

	MOS
Sintetizirane rečenice	3.454 ± 0.029
Pročitane rečenice	4.621 ± 0.018

Distribucija za ocjene na sintetiziranim rečenicama je prikazana na slici 4.2., dok je distribucija za ocjene na pročitanim rečenicama prikazana na slici 4.3.



Slika 4.2. Distribucija ocjena prve ankete za sintetizirane rečenice



Slika 4.3. Distribucija ocjena prve ankete za pročitane rečenice

Na slici 4.3. je vidljivo da je velika većina ocjena za pročitane rečenice bila 5, ali nekolicina vrlo loših ocjena upućuje na to da neki ispitanici nisu shvatili što trebaju u anketi. Usprkos tome nije utvrđeno da je itko od ispitanika glasao na obrnut način, tj. da je smatrao da je 5 najniža ocjena, a 1 najviša. Iz slike 4.2. se vidi da su ocjene za sintetizirane rečenice bile puno više raspodijeljene, ali najviše ih je oko 3.5-4, što zajedno s MOS-om od 3.454 upućuje na to da je sintetizirani govor na granici s prihvatljivim [12].

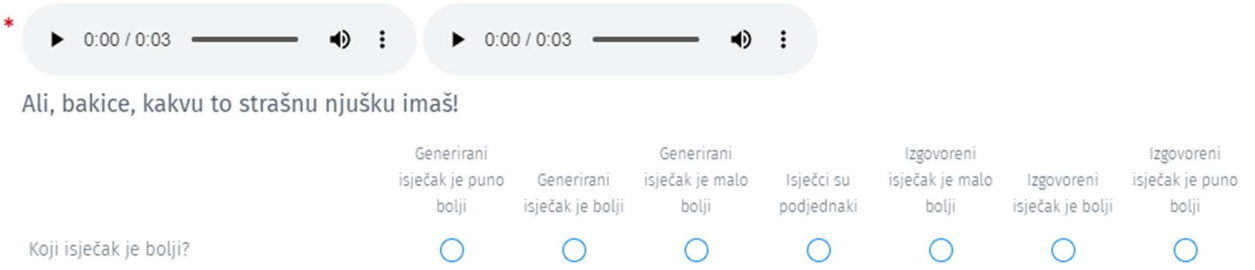
4.2. Usporedba sintetiziranih rečenica s izgovorenim

Druga provedena anketa se također sastojala od 80 isječaka, 40 sintetiziranih, 40 pročitanih. Ovaj put su sintetizirani isječci bili uspoređivani s pročitanim na skali od -3 do 3 s objašnjenjima prikazanim u tablici 4.3.

Tablica 4.3. Skala druge ankete

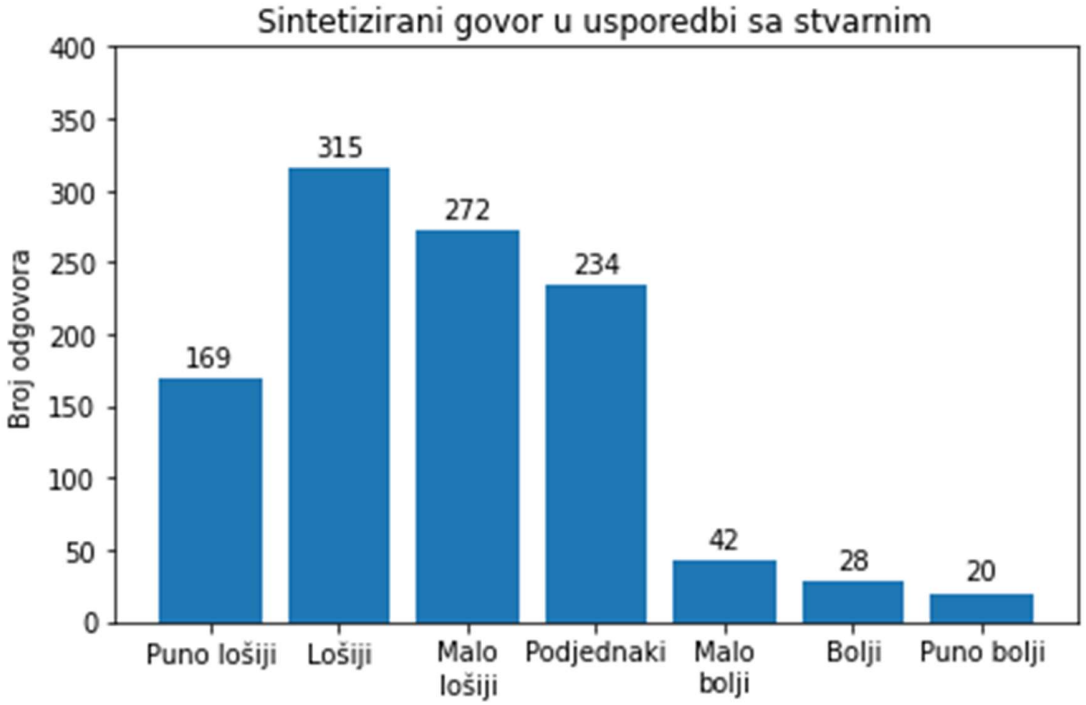
Ocjena	Značenje
3	Generirani isječak je puno bolji
2	Generirani isječak je bolji
1	Generirani isječak je malo bolji
0	Isječci su podjednaki
-1	Izgovoreni isječak je malo bolji
-2	Izgovoreni isječak je bolji
-3	Izgovoreni isječak je puno bolji

Na slici 4.4. prikazan je primjer pitanja iz druge ankete, lijevo je prikazan generirani isječak, a desno izgovoreni.



Slika 4.4. Primjer pitanja iz druge ankete

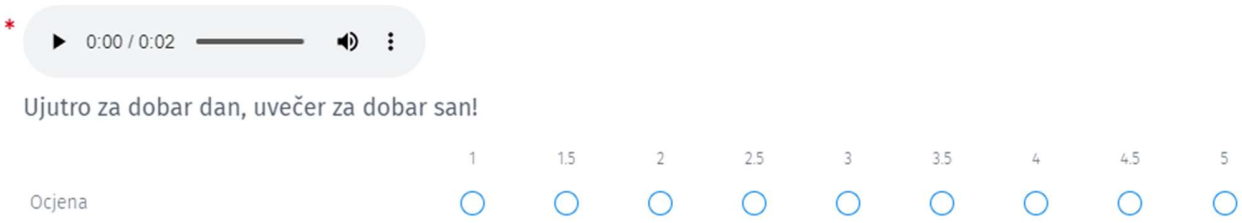
Ova anketa je provedena na 27 ispitanika s rezultatom od –1.158 ± 0.041 što ukazuje na preferenciju za izgovorenim isječcima. Postignuti rezultati su znatno lošiji od rezultata modela Tacotron 2 za engleski jezik (−0.270 ± 0.155) [7]. Distribucija rezultata prikazana je na slici 4.5. Iz grafa je vidljivo da je jako malo ispitanika za jako mali broj isječaka preferiralo sintetizirani isječak, dok je većina preferirala pročitani isječak.



Slika 4.5. Distribucija rezultata druge ankete

4.3. Subjektivna ocjena kvalitete izgrađenog modela na potpuno novim rečenicama

Treća anketa je provedena na sličan način kao prva anketa, ali ovaj put bez pročitanih rečenica i s potpuno novim rečenicama. Rečenice su naknadno izabrane i podijeljene su u 5 kategorija: vijesti iz Sportskih novosti, odredbe iz Narodnih novina, objave s Twittera i Instagrama, slogani i knjige. Svaka od kategorija je imala po 8 rečenica. Cilj ove podjele u kategorije bio je ispitati je li model bolji u sintetiziranju rečenica za neke kategorije nego za druge. Ispitanici su opet ocjenjivali svaki od isječaka na skali od 1 do 5. Primjer jednog pitanja iz treće ankete prikazan je na slici 4.6. Anketa je provedena na 30 ispitanika te su rezultati i primjeri rečenica za svaku od kategorija prikazani u tablici 4.4.

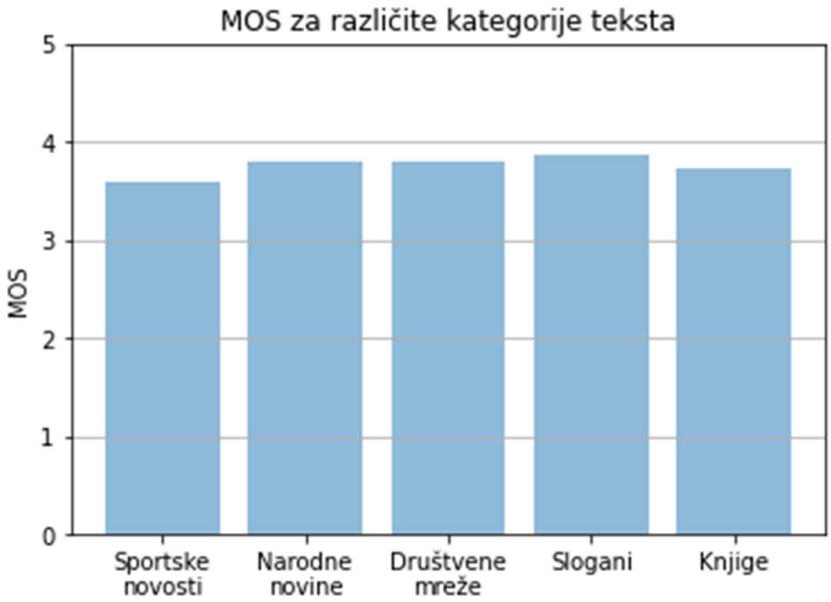


Slika 4.6. Primjer pitanja iz treće ankete

Tablica 4.4. Rezultati treće ankete

Kategorija	MOS	Primjer rečenice
Sportske novosti	3.6 ± 0.065	Real poslao veznjaka na posudbu u Milan! Samo godinu dana i to bez prava na otkup ugovora igrača.
Narodne novine	3.796 ± 0.061	Potpora se dodjeljuje za ribarska plovila koja viju zastavu Republike Hrvatske.
Društvene mreže	3.802 ± 0.066	Moram pohvaliti sestru koja me svaki dan strpljivo slika za Instagram
Slogani	3.869 ± 0.066	Ujutro za dobar dan, uvečer za dobar san!
Knjige	3.723 ± 0.066	Raskoljnjikovu se je trgao glas i nije mogao jasno da izgovara riječ

Iz tablice 4.4. i slike 4.7. mogu se vidjeti MOS-ovi za različite kategorije. Najlošije rezultate je dobila kategorija Sportske novosti, zatim kategorija Knjige, pa kategorija Narodne novine, zatim kategorija Društvene mreže, a najbolje rezultate je dobila kategorija Slogani. Kategorija Sportske novosti najviše odstupa od ostalih vjerojatno zbog velike količine imena, od kojih su većina strana imena. Kategorija Slogani ima najkraće rečenice, pa to može biti uzrok najboljim rezultatima.



Slika 4.7. Usporedba MOS-a za različite kategorije teksta

Iako je bilo za očekivati da će prva anketa imati bolje rezultate, treća anketa s potpuno novim rečenicama je rezultirala MOS-om od 3.758 ± 0.029, što je znatno bolje od MOS-a iz prve ankete (3.454 ± 0.029). Ovakvi rezultati upućuju na to da je model uspješno istreniran jer i na potpuno novim rečenicama daje zadovoljavajuće rezultate, odnosno nije došlo do pretjeranog usklađivanja na podatke za učenje.

4.4. Analiza pogrešaka u sintetiziranim rečenicama

Četvrta anketa je provedena s ciljem prepoznavanja pogrešaka u pojedinim rečenicama. Nasumično je odabrano 20 sintetiziranih rečenica, 10 iz testnog skupa i po dvije iz svake od novih kategorija. Zatim su ispitanici označili probleme koji se pojavljuju u svakoj od rečenica. U ručnoj analizi uz pomoć ankete dobiveni su rezultati prikazani u tablici 4.5.

Tablica 4.5. Analiza problema u sintetiziranim rečenicama

Kategorija problema	Broj rečenica s problemom
Nema problema	4
Ponavljanje riječi	0
Preskakanje riječi	2

Neprirodna prozodija (krivi naglasak na riječi ili slogu, neprirodan ton i slično)	16
Predugačka pauza	1
Tih zvuk	3

Iz tablice se vidi da je velika većina problema bila u prozodiji, većinom u krivom naglašavanju sloga u riječi. Nije bilo ponavljanja riječi, a preskakivanje riječi se dogodilo u dva slučaja, jednom je preskočena riječ “s”, jednom riječ “je”. Također se dogodilo u nekoliko slučajeva da je zadnjih nekoliko slova rečenice ostalo neizgovoreno. Na nekoliko mjesta je bilo jedno slovo ubačeno. Model također sintetizira glas koji zvuči robotski, tj. pomalo neprirodno.

S obzirom na rezultate ovog diplomskog rada, rezultate originalnog rada s modelom Tacotron 2 i ogromnu razliku u količini podataka za treniranje u ova dva rada (24.6 sati i 5.65 sati), može se pretpostaviti da su lošiji rezultati ovog rada prouzrokovani malom količinom podataka. Također je važno za naglasiti da je rečenice za originalni rad čitala profesionalna čitačica, dok za ovaj rad nije čitao profesionalni čitač s profesionalnom opremom za snimanje.

5. Zaključak

Trenutno ne postoje dobri sintetizatori govora iz teksta za hrvatski jezik dostupni javnosti, što je bila motivacija za ovaj diplomski rad. Za potrebe ovog rada su najprije istražena postojeća rješenja ovog problema za druge jezike. Usporedbom rješenja za druge jezike, odabran je model Tacotron 2 za potrebe ovog diplomskog rada. Tacotron 2 je model u području dubokog učenja, stoga je potrebna velika količina podataka koju je autor izradio samostalno. Iako je izrađen prilično velik skup podataka (5.65 sati), i dalje nije bila dovoljna količina podataka za samostalno treniranje modela, stoga je korišten predtrenirani model za engleski jezik koji je uvelike olakšao i ubrzao treniranje modela.

Nakon postupka treniranja modela Tacotron 2, provedena je njegova evaluacija pomoću anketa. Ispitivanjem na testnom skupu gdje su rečenice vrlo slične onima iz skupa za treniranje je postignut MOS od 3.454, dok je ispitivanjem na potpuno novim rečenicama postignut MOS od 3.758, u usporedbi s prirodnim govor za koji je postignut MOS od 4.621. Usporedbom sintetiziranih isječaka i izgovorenih, ispitanici su pokazali preferenciju za izgovorenim isječcima. Daljnjom analizom sintetiziranih rečenica je utvrđeno da je većina grešaka vezana za prozodiju, točnije greške u naglascima slogova. Također je sintetizirani glas zvučao pomalo "robotski", iako je bio vrlo sličan autorovom.

Iako sintetizator generira razumljive rečenice, ovi rezultati bi se mogli značajno poboljšati. Prvi i najjednostavniji način za poboljšanje rezultata bi bilo povećanje podatkovnog skupa za treniranje. Drugi od mogućih načina poboljšanja rezultata bilo bi korištenje predtreniranog modela za neki jezik sličniji hrvatskom, poput ruskog. Još jedan od načina poboljšanja rezultata bilo bi korištenje profesionalnog čitača ili čitačice.

Literatura

[1] E. Miller, Speech Synthesis Software for Anime Announced, AnimeNewsNetwork, 2007. godine, dostupno na: https://www.animenewsnetwork.com/news/2007-05-02/speech-synthesis-software [14.9.2020.]

[2] W. I. Hallahan, DECtalk Software: Text-to-Speech Technology and Implementation, Digital Technical Journal, 1995. godine, dostupno na: https://www.hpl.hp.com/hpjournal/dtj/vol7num4/vol7num4art1.pdf [19.9.2020.]

[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WAVENET: A GENERATIVE MODEL FOR RAW AUDIO, arxiv, 2016. godine, dostupno na: https://arxiv.org/pdf/1609.03499.pdf [19.9.2020.]

[4] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, Y. Bengio, CHAR2WAV: END-TO-END SPEECH SYNTHESIS, Mila, 2017. godine, dostupno na: https://mila.quebec/wp-content/uploads/2017/02/end-end-speech.pdf [19.9.2020.]

[5] Y. Wang, RJ 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, R. A. Saurous, TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS, arxiv, 2017. godine, dostupno na: https://arxiv.org/pdf/1703.10135.pdf [19.9.2020.]

[6] Y. Taigman, L. Wolf, A. Polyak, E. Nachmani, VOICELOOP: VOICE FITTING AND SYNTHESIS VIA A PHONOLOGICAL LOOP, arxiv, 2018. godine, dostupno na: https://arxiv.org/pdf/1707.06588.pdf [19.9.2020.]

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu, NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS, arxiv, 2018. godine, dostupno na: https://arxiv.org/pdf/1712.05884.pdf [7.9.2020.]

[8] R. Ullmann, R. Rasipuram, M. Magimai.-Doss, H. Bourlard, Objective Intelligibility Assessment of Text-to-Speech Systems Through Utterance Verification, Idiap Publications, 2015. godine,

dostupno na:
https://publications.idiap.ch/downloads/papers/2015/Ullmann_INTERSPEECH_2015.pdf
[14.9.2020.]

[9] E. Alpaydın, Introduction to Machine Learning, The MIT Press, Cambridge, Massachusetts, 2010. godine

[10] R. Prabhu, Understanding of Convolutional Neural Network (CNN) — Deep Learning, Medium, 2018. godine, dostupno na: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148 [19.9.2020.]

[11] Y. Stylianou, Concatenative Speech Synthesis using a Harmonic plus Noise Model, ISCA Archive, 1998. godine, dostupno na:
https://www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_261.pdf [7.9.2020.]

[12] P. Kamp, Mean Opinion Score (MOS), Twilio Docs, 2020. godine, dostupno na:
https://www.twilio.com/docs/glossary/what-is-mean-opinion-score-mos [19.9.2020.]

[13] A. van den Oord, S. Dieleman, WaveNet: A generative model for raw audio, deepmind, 2016. godine, dostupno na:
https://www.deepmind.com/blog/article/wavenet-generative-model-raw-audio [9.7.2020.]

[14] S. O. Arık, M. Chrzanowsk, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, M. Shoeybi, Deep Voice: Real-time Neural Text-to-Speech, arxiv, 2017. godine, dostupno na: https://arxiv.org/pdf/1702.07825.pdf [10.9.2020.]

[15] D. Mwiti, A 2019 Guide to Speech Synthesis with Deep Learning, Heartbeat (Medium), 2019. godine, dostupno na:
https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd
[13.7.2020.]

[16] 1D Convolution block, Peltarion, dostupno na: https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/1d-convolution-block [8.9.2020.]

[17] C. Olah, Understanding LSTM Networks, github, 2015. godine, dostupno na:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/ [8.9.2020.]

[18] WaveGlow, github, dostupno na: https://github.com/NVIDIA/waveglow/ [19.9.2020.]

[19] K. Io, The LJ Speech Dataset, keithito, 2017. godine, dostupno na: https://keithito.com/LJ-Speech-Dataset/ [19.9.2020.]

[20] Tacotron 2, github, dostupno na: https://github.com/NVIDIA/tacotron2 [9.9.2020.]

[21] R. Mama, Tacotron 2, github, dostupno na: https://github.com/Rayhane-mamah/Tacotron-2 [9.9.2020.]

[22] D. Vasani, This thing called Weight Decay, towards data science (Medium), 2019. godine, dostupno na: https://towardsdatascience.com/this-thing-called-weight-decay-a7cd4bcfccab [9.9.2020.]

Sažetak

Ovaj diplomski rad se bavi problematikom sinteze govora iz teksta. Cilj rada bio je realizirati model za sintezu govora iz teksta za hrvatski jezik. Za početak je dana osnovna terminologija ovog područja. Nakon toga su uspoređena neka od postojećih rješenja za problem sinteze govora iz teksta. Zadatak ovog diplomskog rada riješen je korištenjem modela Tacotron 2, koji je detaljno objašnjen. Za potrebe učenja, prvo je izrađen podatkovni skup sastavljen od pročitanih rečenica na hrvatskom jeziku s pripadajućim transkriptom. Nakon izrade podatkovnog skupa, model je prilagođen za hrvatski jezik. Model je treniran na predtreniranom modelu za engleski jezik što je uvelike ubrzalo i olakšalo treniranje. Nakon treniranja, odabran je konačni model koji je dalje evaluiran. Model je evaluiran korištenjem anketa u kojima su ispitanici ocjenjivali sintetizator na testnim rečenicama, uspoređivali ih s izgovorenim rečenicama i ocjenjivali sintetizator na rečenicama podijeljenim u kategorije. Također je provedena i analiza pogrešaka u sintetiziranim rečenicama.

Ključne riječi: duboko učenje, sinteza govora, Tacotron 2

Abstract

Text to speech using deep learning

This thesis deals with the issue of speech synthesis from text. The aim of the paper was to realize a text-to-speech model for the Croatian language. To begin with, the basic terminology of the area was given. After that, some of the existing solutions for text-to-speech were compared. The task of this thesis was solved using the Tacotron 2 model, which is explained in detail. For training purposes, a data set consisting of read sentences in Croatian with the accompanying transcript was first created. After creating the data set, the model was adapted for the Croatian language. The model was trained on a pre-trained model for the English language which greatly accelerated and facilitated the training. After training, the final checkpoint was selected which was then used for evaluation. The model was evaluated using surveys in which subjects rated the synthesizer on test sentences, compared them to spoken sentences, and rated the synthesizer on sentences divided into categories. Error analysis of the synthesized sentences was also performed.

Keywords: deep learning, speech synthesis, Tacotron 2

Životopis

Matej Džijan rođen je 17. veljače 1997. godine u Vinkovcima. Nakon završene Osnovne škole Ivana Gorana Kovačića u Vinkovcima, 2011. godine upisuje Gimnaziju Matije Antuna Reljkovića u Vinkovcima, smjer Prirodoslovno-matematički, te ju završava 2015. godine. Obrazovanje nastavlja iste godine na Elektrotehničkom fakultetu u Osijeku (današnjem Fakultetu elektrotehnike, računarstva i informacijskih tehnologija) na kojem upisuje preddiplomski studij računarstva. 2018. godine stječe naziv univ.bacc.ing. te upisuje diplomski studij računarstva, izborni blok Informacijske i podatkovne znanosti, na Fakultetu elektrotehnike, računarstva i informacijskih tehnologija. 2020. dobiva Dekanovu nagradu za uspjeh u studiranju.