УНИВЕРЗИТЕТ У НОВОМ САДУ

ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА

ЕКСПРЕСИВНИ ВИШЕЈЕЗИЧНИ СИНТЕТИЗАТОР ГОВОРА

ДОКТОРСКА ДИСЕРТАЦИЈА

Ментор:
Проф. др Милан Сечујски

Кандидат:
Тијана Носек

Нови Сад, 2023. године

УНИВЕРЗИТЕТ У НОВОМ САДУ **ОБРАЗАЦ – 5а**
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА[1]

Врста рада:	Докторска дисертација
Име и презиме аутора:	Тијана Носек
Ментор (титула, име, презиме, звање, институција)	др Милан Сечујски, редовни професор, Факултет техничких наука, Нови Сад
Наслов рада:	Експресивни вишејезични синтетизатор говора
Језик публикације (писмо):	Српски (латиница)
Физички опис рада:	Унети број: Страница____93_____ Поглавља____6_____ Референци___108_____ Табела____5_____ Слика____40_____ Графикона__0_____ Прилога___0_____
Научна област:	Електротехничко и рачунарско инжењерство
Ужа научна област (научна дисциплина):	Телекомуникације и обрада сигнала
Кључне речи / предметна одредница:	синтеза говора, експресивни говор, неуронске мреже, вокодер
Резиме на језику рада:	Циљ истраживања ове докторске дисертације је да испита могућност синтетизовања говора гласом говорника на језику који он никада није говорио. Креирани су вишејезични модели, како за језике чији је говорни материјал анотиран на исти начин, тако и за оне чији је говорни материјал анотиран различитим конвенцијама, што укључује и српски језик. По квалитету синтетизованог говора неки модели чак превазилазе стандардне моделе обучене на говорном материјалу на једном језику. Поред архитектуре за вишејезичне моделе, предложен је и начин адаптације таквог модела на новог говорника. Таква адаптација омогућује брзу и једноставну продукцију нових гласова задржавајући могућност синтезе на свим језицима подржаним моделом, без обзира на оригинални језик новог говорника.
Датум прихватања теме од стране надлежног већа:	
Датум одбране: (Попуњава одговарајућа служба)	
Чланови комисије: (титула, име, презиме, звање, институција)	Председник: др Татјана Грбић Члан: др Никола Ђурић Члан: др Татјана Лончар-Турукало Члан: др Јелена Николић Члан: др Никша Јаковљевић Члан, ментор: Милан Сечујски
Напомена:	

[1] Аутор докторске дисертације потписао је и приложио следеће Обрасце:
5б – Изјава о ауторству;
5в – Изјава о истоветности штампане и електронске верзије и о личним подацима;
5г – Изјава о коришћењу.
Ове Изјаве се чувају на факултету у штампаном и електронском облику и не кориче се са тезом.

UNIVERSITY OF NOVI SAD
FACULTY OF TECHNICAL SCIENCES

KEY WORD DOCUMENTATION[2]

Document type:	Doctoral dissertation
Author:	Tijana Nosek
Supervisor (title, first name, last name, position, institution)	Milan Sečujski, PhD, full professor, Faculty of Technical Sciences, Novi Sad
Thesis title:	Expressive Multilingual Speech Synthesizer
Language of text (script):	Serbian language (latin)
Physical description:	Number of: Pages_____93_____ Chapters____6_____ References__108_____ Tables_____5_____ Illustrations__40_____ Graphs____0_____ Appendices__0_____
Scientific field:	Electrical and computer engineering
Scientific subfield (scientific discipline):	Telecommunications and signal processing
Subject, Key words:	text-to-speech synthesis, expressive speech, neural networks, vocoder
Abstract in English language:	The aim of this thesis is to investigate the possibility of synthesizing speech in the voice of a speaker in a language which he had never spoken. Multilanguage models are created, both for the languages whose databases are annotated using the same conventions, and for the languages whose databases are annotated using different conventions, which includes the Serbian language. Regarding quality of synthesized speech, some models even surpass the quality of synthesis produced by standard monolanguage models. Beside architecture for multilanguage models, a method for adaptation of such models to the data of a new speaker is proposed. The proposed method of adaptation enables fast and simple production of new voices, while preserving the possibility to synthesize speech in any language supported by the model, regardless of the target speaker's original language.
Accepted on Scientific Board on:	
Defended: (Filled by the faculty service)	
Thesis Defend Board: (title, first name, last name, position, institution)	President: Tatjana Grbić, PhD Member: Nikola Đurić, PhD Member: Tatjana Lončar-Turukalo, PhD Member: Jelena Nikolić, PhD Member: Nikša Jakovljević, PhD Member, Mentor: Milan Sečujski, PhD
Note:	

[2] The author of doctoral dissertation has signed the following Statements:
5б – Statement on the authority,
5в – Statement that the printed and e-version of doctoral dissertation are identical and about personal data,
5г – Statement on copyright licenses.
The paper and e-versions of Statements are held at he faculty and are not included into the printed thesis.

Sažetak

Verbalnu komunikaciju između čoveka i mašine omogućavaju govorne tehnologije. Jedna od njih je sinteza govora na osnovu teksta, koja omogućava da se mašina obrati čoveku na njemu razumljiv način, govorom. Ova tehnologija pronalazi svoju primenu u virtuelnim asistentima, igricama, aplikacijama za pomoć osobama sa invaliditetom i dr.

U disertaciji je razmatran pristup koji se zasniva na dekompoziciji sintetizatora govora zasnovanog na veštačkim neuralnim mrežama na tri modula, a to su: modul za jezičku obrada teksta, modul za modelovanje akustičkih parametara pomoću veštačkih neuralnih mreža i modul za generisanje govornog signala. Modelovanje akustičkih parametara podrazumeva obuku modela neuronske mreže koja na osnovu lingvističkih specifikacija pristiglih od modula za jezičku obradu teksta predviđa vrednosti akustičkih parametara neophodnih za generisanje govornog signala. Ovaj modul je najčešće nezavisan od jezika i njegove varijacije u pogledu arhitekture i algoritama obuke omogućuju promene govornika, stila govora, jezika i dr.

Ovakvi sistemi omogućili su razvoj algoritama za dobijanje različitih sintetizovanih glasova sa malom količinom govornog materijala za obuku, što je dodatno povećalo mogućnosti primene sintetizovanog govora. Povećanje ekspresivnosti sintetizovanog govora, te promene stila i emocija u sintetizovanom govoru, omogućili su da sintetizovani govor zvuči prirodnije, te i prihvatljivije ljudima za upotrebu. Široka upotreba sinteze govora dovela je i do interesantne teme višejezične sinteze govora.

U ovom radu predložen je upravo model koji omogućava ekspresivnu višejezičnu sintezu govora. Model se obučava na govornoj bazi koja obuhvata mnoštvo govornika od kojih svaki govori jednim jezikom. Nakon obuke, standardni višejezični model omogućio bi sintezu govora glasom govornika isključivo na jeziku kojim taj govornik govori u bazi za obuku, dok predloženi model omogućava sintezu glasom govornika na svim jezicima koji postoje u bazi za obuku. Dakle, govor se može sintetizovati glasom govornika na jeziku kojim on nikada i nije govorio, zadržavajući izuzetno visok kvalitet u pogledu prirodnosti i razumljivosti. Ovaj model primenjen je na nekoliko različitih jezika, kako na one čije su baze anotirane po istim konvencijama za prozodijsku anotaciju, tako i na one anotirane različitim konvencijama. Prozodijska anotacija baza podrazumeva obeležavanje tipova akcenata, granica između intonativnih celina i sl. Primera radi, široko raprostranjen

skup konvencija za prozodijsku anotaciju je ToBI (eng. Tones and Break Indices), koji je originalno razvijen za američki engleski, a naknadno prilagođen i mnogim drugim jezicima. Međutim, za neke jezike postoje drugi sistemi anotacije, koji u većoj meri odgovaraju akcenatskom sistemu samog jezika, pa tako npr. i za srpski postoji sistem za anotaciju koji je već duže vreme u upotrebi, te bi se govorne baze morale anotirati ispočetka ako bi model zahtevao da se anotacija uskladi. To je razlog zašto je model predložen u ovoj disertaciji osmišljen tako da može da podrži i jezike čije su baze prozodijski anotirane različitim konvencijama. Model ne zahteva velike količine govornog materijala po jeziku, kao ni po govorniku, a ima sposobnost da iskoristi znanja iz jezika sa više materijala, zahvaljujući formiranju različitih *embedding* slojeva. Takođe, model je moguće adaptirati pomoću male količine govornog materijala novog govornika, čime se brzo mogu kreirati novi glasovi koji zadržavaju sposobnost sinteze govora na svim jezicima podržanim modelom, bez obzira na originalni jezik ciljnog govornika.

Model je evaluiran na osnovu objektivnih mera, ali je izvršena i subjektivna evaluacija kroz niz različitih eksperimenata kako bi se došlo do optimalne arhitekture, kao i načina adaptacije modela. Visok kvalitet sintetizovanog govora potvrđen je visokim ocenama od strane slušalaca.

Abstract

Verbal communication between humans and computers is made possible by speech technologies. One of them is speech synthesis from text, which enables computers to communicate with humans in a natural way, using speech. This technology is applied in virtual assistants, computer games, applications for people with disabilites, etc.

In this thesis, a neural network based approach for speech synthesis consisting of three modules is analysed. There is a module for linguistic text processing, a module for modelling acoustic parameters using neural networks and a module for generating the speech signal. The modeling of acoustic parameters is done via trained neural networks which should be able to predict acoustic features necessary for producing speech waveforms by using linguistic features gathered from the module for text processing. The module dealing with acoustic parameters is usually language independent, and its variations with respect to architecture and training algorithms allow us to change the speaker, style and language of the produced synthesized speech.

Such systems led to the development of algorithms for gathering different synthetic voices by using small speech databases for training, which broadened the spectrum of possible applications of synthetic speech. The introduction of expressiveness in synthetic speech and changes of styles and emotions, contributed to the naturalness of synthetic speech, which made it more convenient for use by humans. Wide usage of synthetic speech opens up the interesting possibility of multilanguage speech synthesis.

A model which is able to produce expressive multilingual speech is proposed in this thesis. The model is trained on a multispeaker database, where each speaker speaks only one language. After the training, a standard multilanguage model is able to produce speech in the voice of the target speaker only in the language that the speaker actually speaks in the database, but the proposed model is able to produce speech in the voice of the target speaker in any language that exists in the database. Therefore, the proposed model enables synthesis in multiple languages with voice characteristics even of speakers who do not speak the intended language, while preserving high quality of speech synthesis in terms of intelligibility and naturalness. As the model does not depend on annotation conventions, it

can be applied equally to languages with prosodic features annotated using the same conventions, e.g. ToBI, as well as to languages with prosodic features annotated using different conventions. Prosodic annotation of databases includes the labelling of accents, phrase breaks, etc. ToBI (Tone and Break Indices) is a widely used set of conventions for prosodic annotation, originally developed for American English, but subsequently adapted to a number of other languages. However, for some languages other, more suitable systems of conventions exist. For instance, for Serbian there exists a prosodic annotation scheme which has been used for a long time, which implies that the requirement for a single annotation scheme for all languages would also require complete re-annotation of all Serbian speech databases. For that reason, the proposed model is designed so as to be able to handle languages with speech databases annotated using different conventions. The model does not require the existence of massive speech databases for each language nor for each speaker, and it is able to transfer knowledge acquired from languages for which there is more training data available to those with less, owing to multiple embedding layers in its architecture. What is more, the model can even be adapted to new speakers using small databases, which enables fast production of new voices while preserving the possibility of synthesis in all languages supported by the model, regardless of the target speaker's original language.

The proposed model was evaluated by objective measures and subjective evaluation was conducted within a number of experiments in order to find the optimal architecture and method of model adaptation. High scores across listening tests confirm the high quality of synthesized speech.

Zahvalnica

Zahvaljujem se svom najužem timu, mentoru prof. dr Milanu Sečujskom i kolegama Siniši Suziću i Darku Pekaru, sa kojima sam uživala radeći na svim svojim naučnim radovima, na koje sam uvek mogla da se oslonim i od kojih sam mnogo naučila.

Takođe, želim da se zahvalim preduzeću „AlfaNum" iz Novog Sada na ustupljenim govornim bazama i računarskim resursima, kao i svim članovima komisije koji su svojim sugestijama doprineli kvalitetu ove disertacije. Veliko hvala dugujem i prijateljima koji su morali da prođu kroz mnoštvo testova slušanja.

Ogromnu zahvalnost dugujem svojoj porodici, koja mi je uvek pružala bezrezervnu podršku i oslonac, a posebno hvala Igoru, koji je morao da odsluša i razmotri svaki segment svakog eksperimenta iako se time ne bavi. Posebnu motivaciju da se ova disertacija privede kraju dala mi je moja Ema.

Sadržaj

Spisak slika

Spisak tabela

Spisak skraćenica

TTS - engl. *Text-to-Speech Synthesis*

ASR - engl. *Automatic Speech Recognition*

SPSS - engl. *Statistical Parametric Speech Synthesis*

HMM - engl. *Hidden Markov Models*

DNN - engl. *Deep Neural Networks*

NLP - engl. *Natural Language Processing*

MOS - engl. *Mean Opinion Score*

MUSHRA - engl. *Multiple Stimuli Hidden Reference and Anchor*

ToBI - engl. *Tone and Break Indices*

V/UV - engl. *Voiced/Unvoiced*

MGC - engl. *Mel Frequency Generalized Cepstral Coefficients*

GAN - engl. *Generative Adversarial Networks*

RNN - engl. *Recurrent Neural Networks*

MSE - engl. *Mean Square Error*

LSTM - engl. *Long Short-Term Memory*

1. Uvod

Najprirodniji način komunikacije među ljudima jeste govorna komunikacija. Kako su internet i brza dostupnost informacija danas od ključne važnosti, a telefoni, tableti i kompjuteri su svuda oko nas i uvek pri ruci, ne čudi potreba da govorom komuniciramo i sa mašinama. Govorna komunikacija čovek-mašina ostvaruje se pomoću dve govorne tehnologije – sinteze govora na osnovu teksta (engl. *Text-to-Speech Synthesis* – TTS), koja omogućava da se mašina „obrati" čoveku, i automatskog prepoznavanja govora (engl. *Automatic Speech Recognition* – ASR), koje omogućava da mašina „razume" čoveka [Sečujski, 2011]. Stvaranje veštačkog govornog signala, odnosno sinteza govora, oduvek je bilo interesantno ljudima, a prvi pokušaji da se konstruiše mašina koja će sintetizovati ljudski govor datiraju još iz XII veka. Međutim, sve do 1930-ih, istraživači nisu uspevali da naprave razumljiv sintetizator, pokušavajući da razviju model koji će oponašati vokalni trakt čoveka, pa i jezik i usne. Prvi model ljudskog vokalnog trakta, koji je bio u mogućnosti da proizvodi različite vokale, realizovan je 1779. Potom su na sličan način modelovani i jezik i usne, pa su mogli da se produkuju ne samo vokali nego i konsonanti. Na osnovu tog modela je 1837. konstruisana mehanička „mašina koja govori". Tek 1930. u preduzeću *Bell Labaratories* razvijen je vokoder, elektronski uređaj za analizu i sintezu govora koji je mogao da produkuje razumljiv govor. Skoro 10 godina kasnije ovaj uređaj je unapređen i od tada je naučna zajednica bila sve više zainteresovana za sintezu govornog signala. Osnovna ideja do koje se zapravo došlo bila je da se formira model izvor-filtar, jer se glotis može smatrati izvorom zvuka, dok se vokalni trakt ponaša kao filtar. Od tada su razvijani analogni elektronski uređaji koji su mogli da oponašaju ljudski govor, mada su ti prvi sintetizatori zvučali veoma robotizovano i nisu bili dovoljno razumljivi [Pantazis, 2007].

Sinteza govora je izuzetno atraktivna tema za istraživanje u oblastima obrade prirodnog govora i jezika kao i u oblasti veštačke inteligencije, sa širokom primenom u industriji. Razvoj TTS sistema zahteva znanja o jeziku i načinu na koji ljudi produkuju govor, te uključuju discipline kao što su lingvistika, akustika, digitalna obrada signala i mašinsko učenje. TTS pronalazi svoju primenu u raznim aplikacijama poput virtuelnih asistenata u bankama, pametnim kućama, pozivnim centrima, aplikacijama za zabavu kao što su igrice i audio knjige,

aplikacijama za pomoć pri čitanju starima ili slepima ili za govornu podršku osobama sa urođenim ili stečenim gubitkom govora, itd. Dve glavne karakteristike kvaliteta sintetizovanog govora su *prirodnost* i *razumljivost*. Neretko prirodnost govora doprinosi razumljivosti govora i boljem prenosu informacije [Sečujski, 2011]. Mnogo je otvorenih polja za unapređenje TTS sistema iniciranih upravo pomenutim primenama. Odavno je sintetizovani govor izuzetno razumljiv, ali nedovoljno prirodan, te zato i dalje čovek više voli kada mu se pri pozivu banke javi ljudsko biće, a ne mašina. Prijatnije je slušati audio knjigu koju čita profesionalni glumac, a ne mašina. Odsustvo ekspresivnosti u sintetizovanom govoru je glavni razlog odbojnosti prema komunikaciji sa mašinama. Način na koji je nešto izgovoreno takođe prenosi informacije, a ne samo sadržaj izgovorenog. Stoga je jedan od aktuelnih pravaca istraživanja upravo sinteza *ekspresivnog govora*.

Iako ne možemo reći da u svakodnevnoj komunikaciji ljudi naglašavaju svoje emocije, ipak njihov govor sadrži ekspresivnost u vidu npr. neke vrste pozitivnog stava govornika pri saopštavanju dobrih vesti ili negativnog pri saopštavanju loših vesti. Gotovo je nemoguće u redovnoj komunikaciji izgovoriti bilo šta, a da ne bude obojeno emocijom ili stavom prema govorniku ili izgovorenom. Često se naglašavaju određene reči sa ciljem isticanja neke informacije, ili se jasno čuje upitan ton pri postavljanju pitanja. Takođe u glasu čoveka možemo da čujemo kada je nesiguran, zbunjen ili kada ima zapovedni ton. Sve to utiče na prijatnost razgovora i doprinosi smanjenju mogućnosti nesporazuma. Pri takvom govoru, menjanju stilova govora, menjaju se osnovna frekvencija, dužina trajanja pojedinih segmenata i glasnoća, ali i neke druge stvari. Na primer, čovek može kroz osmeh saopštiti dobru vest i to će uticati na efektivnu dužinu njegovog vokalnog trakta, biće manja nego što bi bio u slučaju pućenja usta (kada se možda duri). Smanjenje efektivne dužine vokalnog trakta uticaće na smanjenje energije na nižim učestanostima, i samim tim će osoba zvučati pozitivnije. [Hamza, 2004].

Kroz istoriju su se smenjivali različiti pristupi, odnosno metode, za sintezu govora. U početku je korišćena *artikulatorna sinteza*, gde je govor produkovan simuliranjem ljudskog artikulatornog sistema, odnosno organa kao što su glotis i vokalni trakt [Coker, 1976, Shadle, 2001]. Ovakav pristup brzo je odbačen zbog teškog modelovanja i skupljanja podataka za simulaciju. Pristup koji je zaživeo i mogao zapravo da produkuje prve zadovoljavajuće

rezultate za to vreme, bio je *formantna sinteza* [Seeviour, 1976, Klatt, 1980]. Ova metoda zasniva se na skupu određenih pravila koja kontrolišu pojednostavljen izvor-filtar model. Pravila su razvijena od strane lingvista s ciljem imitiranja formantne strukture i ostalih spektralnih karakteristika ljudskog govora. Ovakav pristup se i danas smatra korisnim za neke primene (npr. *embedded* sistemi) gde je ključno da govor bude razumljiv. Međutim govor ne zvuči prirodno već robotizovano. Potom je usledila *konkatenativna metoda* [Moulines, 1990, Hunt 1996], koja je dugo vremena bila dominantna i danas pronalazi svoju primenu. Većina najboljih postojećih sistema za sintezu govora na bilo kom jeziku, pa i na srpskom, radi koristeći upravo metodu konkatenacije, koja podrazumeva selekciju i spajanje snimljenih talasnih oblika govornih segmenata. Glavna mana ovakvog pristupa je potreba za ogromnom bazom govornog signala kako bi se pokrile sve moguće kombinacije govornih segmenata da bi sintetizovan govor zvučao što prirodnije. Druga velika mana je nemogućnost promene boje glasa (stila, govornika) bez snimanja nove govorne baze. S ciljem prevazilaženja navedenih problema predložen je *statistički parametarski pristup* (engl. *statistical parametric speech synthesis* - SPSS) [Yoshimura, 2002, Tokuda, 2000, Tokuda 2013, Zen 2009] gde je ideja da se na osnovu govorne baze formiraju modeli kontekstno zavisnih fonema, na osnovu kojih će se kasnije sintetizovati govor. Dakle, umesto direktnog generisanja talasnog oblika govornog signala, generišu se akustički parametri na osnovu kojih se može produkovati talasni oblik govornog signala. Glavna prednost parametarskog pristupa jeste fleksibilnost – ne zahteva postojanje velikih, memorijski zahtevnih govornih baza u fazi sinteze, i fleksibilan je u pogledu promene karakteristika govornika i stila govora, što postaje veoma značajno širenjem primene govornih tehnologija. Kod ovog pristupa uglavnom se razlikuju tri modula: modul za jezičku obradu teksta, modul za pretvaranje lingvističkih obeležja u akustička i modul za pretvaranje akustičkih obeležja u talasni oblik govornog signala. Modul za jezičku obradu teksta ima zadatak da iz teksta u kome nisu obeleženi akcenti, intonacija, trajanja pojedinih fonema, izdvoji takve informacije na osnovu rečnika, obučenih lingvističkih modela i dr. Modul za pretvaranje lingvističkih u akustička obeležja ima zadatak da na osnovu informacija dobijenih od prethodnog modula, a pomoću obučenog modela, produkuje informacije poput trenutnih vrednosti osnovne frekvencije i spektralnih obeležja, odnosno akustičkih obeležja koja su neophodna narednom modulu kako bi mogao da rekonstruiše talasni oblik signala. Sve do 2010-ih za pretvaranje lingvističkih u akustička obeležja dominantno se koristio model

zasnovan na skrivenim Markovljevim modelima (engl. *Hidden Markov Model* – HMM). Potom je ovakav model zamenjen dubokom neuralnom mrežom (engl. *Deep Neural Network* – DNN) [Zen, 2013, Qian, 2014, Fan, 2014]. Kasnije je razvijena ideja da se DNN iskoristi za direktno generisanje akustičkih obeležja iz niza fonema, a ne lingvističkih obeležja [Wang, 2016], pa i direktno generisanje talasnog oblika govornog signala iz lingvističkih obeležja [van den Oord, 2016]. Svi ovi pristupi koji koriste DNN umesto nekog od modula TTS, nazvani su *neuralnom sintezom* [Tan, 2021]. Kasnije su razvijeni i tzv. *end-to-end* sistemi koji zamenjuju sva tri modula određenim sistemom neuralnih mreža, odnosno direktno predviđaju odbirke govornog signala na osnovu niza grafema [Ping, 2018, Ren, 2020, Donahue, 2020].

1.1. Predmet i ciljevi istraživanja

Ekspresivni višejezični sintetizator koji će biti razmatran u ovoj disertaciji zasnovan je na SPSS pristupu sa DNN modelom za predviđanje akustičkih obeležja na osnovu lingvističkih.

Danas, kada je moguće relativno brzo i sa relativno malo govornog materijala dobiti novi sintetizovani glas, otvorile su se razne mogućnosti. Recimo, svaka banka može da ima svog virtuelnog asistenta koji će govoriti glasom njihovog zaštitnog lica (npr. poznatog glumca/fudbalera), a pritom to zaštitno lice neće morati da provede sate u studiju snimajući govorne baze. Ili na primer, može da se produkuje mnoštvo knjiga glasom nekog glumca/profesora ili čak pisca. Razna istraživanja idu i u pravcu unapređenja kvaliteta govornih baza [Kuo, 2018, Valentini-Botinhao, 2016], pa se i sa bazama koje nisu snimljene u profesionalnom studiju već sadrže javna obraćanja ili su snimljene mobilnim telefonom, mogu formirati kvalitetni sintetizovani glasovi. Ovo dalje može da omogući i produkciju recimo nekog teksta glasom istorijske ličnosti za koju postoji svega nekoliko minuta snimljenog govora relativno lošeg kvaliteta. Međutim, ako se zadržimo na primeru knjiga, jasno je da je neophodno da govornik menja stil govora dok izgovara različite delove knjige, da se povremeno u glasu oseti izuzetno izražena emocija, i sl. Stoga je razvoj ekspresivnog TTS [Schröder, 2009], posebno sa idejom podešavanja nivoa ekspresivnosti i idejom transplantacije stila, dakle preuzimanja stila od govornika A, i produkcija glasa govornika B u tom stilu, iako u originalnoj bazi nema snimaka govornika B u datom stilu [Suzić, 2019], od

izuzetnog značaja. I kada postoji sintetizovan glas koji korisnik želi, i koji može da menja stil govora, zvuči dovoljno prirodno, javlja se potreba da može da menja i jezik kojim govori. S obzirom da mnoge kompanije posluju širom sveta, interesantna je ideja da korisnici mogu da dobiju informacije na jeziku kojim i sami govore, a glasom zaštitnog lica kompanije bez obzira na jezik. Još jedna veoma interesantna primena mogla bi biti za poslovne međunarodne sastanke, gde bi svaki sagovornik govorio na svom maternjem jeziku, a njegov govor bi se automatski prevodio i potom sintetizovao njegovim glasom na jeziku drugog sagovornika.

Postoje različiti pristupi za generisanje sintetizovanog govora sa različitim glasovima, stilovima ili na različitim jezicima. Metode bi se mogle podeliti na one koje se zasnivaju na direktnoj modifikaciji već generisanih signala ili akustičkih obeležja [Kaneko, 2017, Kameoka, 2018], one koje se zasnivaju na adaptaciji celog modela neuronske mreže [Delić, 2018] ili nekog njegovog dela [Fan, 2015] i metode koje se zasnivaju na istovremenom modelovanju većeg broja govornika/stilova/jezika [Hojo, 2016, Suzić, 2018, Zhang, 2019, Sečujski, 2020].

U ovom radu, akcenat je na poslednjem navedenom pristupu koji modeluje više govornika, stilova i jezika istovremeno, a koji zahvaljujući specifičnoj strukturi može da radi i sa veoma ograničenim bazama u pogledu kombinacije govornik-stil-jezik. Ovakav model omogućava produkciju kombinacije govornik-jezik koja nije viđena pri obuci modela, kao i podešavanje nivoa ekspresivnosti, iako nije u mogućnosti da produkuje govor određenim glasom i u određenom govornom stilu ako ta kombinacija govornik-stil nije bila zastupljena u podacima za obuku.

Cilj ovog istraživanja je da ispita mogućnost sintetizovanja govora glasom ciljnog govornika na jeziku na kome on nikada nije govorio. Osim toga, biće ispitane i dodatne mogućnosti koje model pruža, poput podešavanja nivoa ekspresije u govoru. Hipoteza koja će biti testirana jeste da se korišćenjem govornika A i B, i jezika X i Y, i to kombinacija A-X i B-Y u obuci, može generisati sinteza A-Y i B-X. Kvalitet sintetizovanog govora ogleda se u njegovoj razumljivosti i prirodnosti. Pored navedenog, biće testirana sličnost glasa kojim je govor sintetizovan sa glasom originalnog govornika kao i nivo emocije koju slušalac percipira u sintetizovanom govoru.

Problem generisanja kombinacije govornik-jezik koja nije viđena pri obuci modela ogleda se u tome da mnoge fonetsko-prozodijske kombinacije nisu nikada viđene za ciljnog govornika i samim tim ideja je u oslanjanju na sposobnost DNN da generalizuje [Giles, 1987], učenju od drugih govornika iz baze [Fan 2015, Hojo, 2016] i primeni tzv. *embedding* slojeva mreže [Hojo, 2018, Wu, 2015].

1.2. Organizacija disertacije

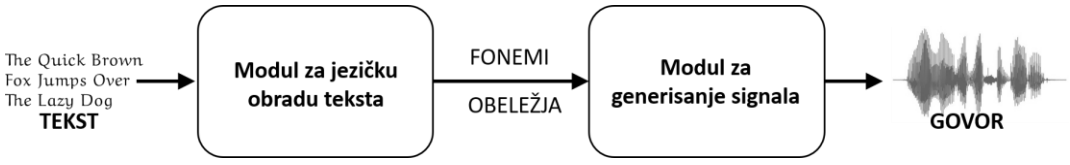
Disertacija se sastoji iz šest poglavlja. Nakon uvoda u kojem su predstavljeni predmet i ciljevi istraživanja sledi poglavlje u kojem su detaljnije opisani tipični moduli sistema za sintezu govora na osnovu teksta. Treće poglavlje uključuje teorijske osnove o DNN i detaljno opisuje modul za generisanje akustičkih obeležja pomoću DNN. Četvrto poglavlje posvećeno je pregledu metoda iz literature za uvođenje više govornika, stilova i jezika u sistem za sintezu govora. U petom poglavlju predstavljen je glavni rezultat istraživanja u okviru rada na ovoj disertaciji - ekspresivni višejezični sintetizator govora. Dati su detalji implementacije i prikazani postignuti rezultati pri različitim eksperimentima. U šestom poglavlju dati su osnovni zaključci i ideje o pravcima daljeg istraživanja.

2. Sistem za sintezu govora na osnovu teksta

Sistem za sintezu govora na osnovu teksta kao ulaz dobija tekst, a kao izlaz daje talasni oblik govornog signala. Sistem sadrži dva modula koji obavljaju manje ili više razgraničene zadatke (slika 2.1). Prvi je zadužen za obradu prirodnog jezika (engl. *Natural Language Processing* – NLP), a drugi za generisanje signala. Prvi vrši morfosintaktičku analizu ulaznog teksta kao i njegovu fonetizaciju, da bi zatim generisao prozodijska obeležja, te je uglavnom u velikoj meri jezički zavisan, a drugi pretvara ligvistička obeležja u govorni signal, i najčešće nije jezički zavisan.

Već je pomenuto da se kod parametarskih sintetizatora modul za generisanje signala rastavlja na dva modula, model za generisanje akustičkih obeležja i vokoder. Postoje i sistemi koji direktno iz teksta predviđaju odbirke govornog signala [Ren, 2020, Wang, 2017], ali i kod njih, iako značajno pojednostavljen, postoji neki oblik NLP modula. U pitanju je najčešće bar neka vrsta predobrade, odnosno normalizacije teksta. Tu se podrazumeva npr. konverzija brojeva i skraćenica u ortografske reči. Ovakvi pristupi značajno smanjuju cenu manuelne anotacije baze za obuku, ali činjenica da sistem mora samostalno da nauči poravnanje između teksta i audio frejmova dovodi do grešaka poput lošeg izgovora, ponovljenih i/ili preskočenih reči u sintezi i sl., a takođe veliki problem predstavlja i veliko kašnjenje pri sintezi [Wang, 2017, Ma, 2020]. U ovom radu akcenat će biti na parametarskom sistemu čiji se modul za pretvaranje lingvističkih obeležja u akustička zasniva na DNN, dok se za generisanje signala može upotrebiti deterministički ili neuralni vokoder.

Evaluacija TTS sistema [Suzić, 2019] vrlo je nezgodan zadatak jer ne postoje objektivne mere procene kvaliteta koje su u dovoljnoj meri usklađene sa subjektivnim utiskom. Možemo, dakle, vršiti evaluaciju korišćenjem objektivnih i subjektivnih mera. Objektivne mere predstavljaju mere odstupanja generisanih parametara od parametara izdvojenih iz prirodnog



Slika 2.1 Osnovna blok šema TTS sistema

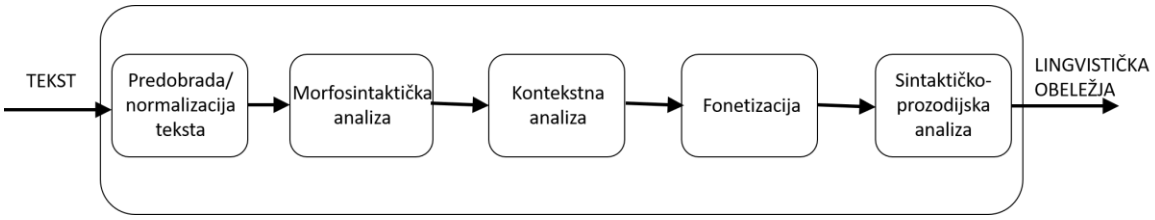
govora. Ove mere su najkorisnije u toku razvoja TTS sistema jer mogu da usmere onoga ko razvija sistem kako da podesi odgovarajuće hiperparametre sistema i da ukaže na postojanje problema kada su odstupanja velika. Međutim, ključna evaluacija je kako čovek percipira da li sintetizovani govor zvuči dobro ili ne, a objektivne mere često nisu u skladu sa ocenom slušalaca. Npr. moguće je da dodje do značajnijeg smanjenja neke objektivne mere, a da slušaoci to i ne osete, ili da se objektivne mere poboljšaju, ali da se subjektivni utisak ne poboljša. Stoga je najrelevantnija evaluacija TTS sistema upravo subjektivno ocenjivanje od strane ljudi. Za ove potrebe koriste se MOS (engl. *Mean Opinion Score*) i MUSHRA (engl. *Multiple Stimuli Hidden Reference and Anchor*) testovi. MOS testovi podrazumevaju da slušaoci neku karakteristiku govora (najčešće kvalitet, odnosno prirodnost i razumljivost) ocene ocenom od 1 (loše) do 5 (odlično). Krajnji rezultat predstavljaće prosečna ocena svih učesnika. Prirodan govor uglavnom ima ocenu preko 4,5. MUSHRA testovi korišćeni su inicijalno za poređenje kvaliteta različitih kodera. Ideja je da postoji originalna rečenica, jasno izdvojena, i nekoliko preostalih rečenica čiji kvalitet treba oceniti oceniti ocenom 1 do 100 u poređenju sa izdvojenom, originalnom rečenicom. Među rečenicama koje se ocenjuju uvek je data i originalna, kojoj bi slušalac trebao dati ocenu 100 (jer je identična izdvojenoj rečenici za koju se zna da je original), kao i jedna rečenica koja je značajno lošija od preostalih, kako bi se jasno postavila i donja granica kvaliteta. Ovakvi testovi su zgodni za direktno poređenje različitih sintetizatora. Iako izuzetno korisni, subjektivni testovi su vrlo nezgodni za sprovođenje jer su često vremenski zahtevni i pronalaženje slušalaca nije jednostavan zadatak. Još jedna velika mana jeste nemogućnost objektivnog direktnog poređenja rezultata različitih istraživanja, kako zbog subjektivne evaluacije (različite grupe ljudi, različiti uslovi slušanja i dr.), tako i zbog obuke modela na različitim bazama.

2.1 Jezička obrada teksta

Modul za jezičku obradu teksta (slika 2.2) neophodan je u fazi sinteze govornog signala, kada je dostupan samo tekst, bez ikakvih naznaka o trajanju fonema, intonaciji i sl. Proces u kom će mašina na osnovu teksta doći do navedenih informacija mora biti automatizovan, i u izuzetno velikoj meri je zavisan od jezika. Može se zasnivati na ekspertskim ili automatizovanim sistemima. Danas se sve češće pribegava automatskim sistemima za analizu

teksta i predikciju prozodije, mada često ekspertski sistemi i dalje daju bolje rezultate. U skladu sa korišćenim sistemom, u fazi obuke TTS sistema moraju postojati adekvatno fonetski, a neretko i prozodijski anotirane govorne, ali i tekstualne baze. Sve ono što sistem za predobradu teksta u fazi sinteze treba samostalno da predvidi, u fazi obuke treba da postoji naznačeno u bazi.

Pri sintezi na osnovu teksta, prvi korak je predobrada ili normalizacija dobijenog teksta u smislu identifikovanja znakova interpunkcije, brojeva, skraćenica i dr. Tekst se deli u rečenice, brojevi i skraćenice se zamenjuju odgovarajućim rečima. Ovaj deo ranije je bio uvek zasnovan na ekspertskom sistemu, odnosno nizu pravila [Sproat, 2001], a danas se neretko zamenjuje neuralnom mrežom koja se obučava na parovima reči zapisanih u obliku broja ili skraćenice, i njihovim proširenjima u pun ortografski oblik [Sproat, 2016, Zhang, 2019], ili kombinaciji ekspertskog sistema i neuralne mreže [Zhang, 2020]. Potom se vrši morfosintaktička analiza, koja podrazumeva određivanje koja je vrsta reči u pitanju, koje su vrednosti morfoloških kategorija (padež, broj, …) i eventualno koja je sintaktička uloga pojedinih reči [Sečujski, 2011]. Ovi koraci su potrebni da bi se ispravno izvršila fonetska transkripcija teksta, a potom i prozodijska. U zavisnosti od jezika, mogu biti potrebni i modeli koji rastavljaju reč na morfeme i koriste morfosintaktička pravila za njihovo povezivanje. Druga mogućnost je postojanje morfoloških rečnika koji sadrže reči datog jezika u svojim osnovnim oblicima ili u svim mogućim oblicima. Ukoliko u rečniku postoje samo osnovni oblici, moraju biti utvrđena pravila tvorbe reči kako bi se moglo utvrditi od koje reči iz rečnika je reč nastala. Postoje i različite procedure u slučaju nepronalaženja reči u rečniku, a koje se zasnivaju na formiranju analogija na osnovu standardnih prefiksa i sufiksa. Uz morfološku analizu, vrši se i analiza konteksta posmatranjem okruženja u kojem se reč nalazi s ciljem smanjenja broja mogućnosti za njenu vrstu, oblik i funkciju u rečenici. Fonetizacija, odnosno prevođenje grafema u foneme, podrazumeva dodelu niza fonema svakoj reči, što je npr. u srpskom jeziku jednostavno jer



Slika 2.2 Detaljna blok šema NLP modula

jednom slovu odgovara jedan fonem (ako se izuzmu eventualne reči stranog porekla pisane u originalnoj ortografiji), ali u mnogim jezicima to nije slučaj [Zhang, 2002, Xu, 2004]. Korišćenjem morfološke i kontekstne analize dolazi se do odgovarajućih fonema i akcentuacije – načina izgovora svake od reči [Sečujski, 2011]. Konačno, kako bi pročitani tekst zvučao razumljivo i prirodno, potrebno je utvrditi i vrednosti prozodijskih obeležja. Prozodija obuhvata sve lingvističke elemente kojih u govoru ima, a u tekstu ne. Na fizičkom nivou podrazumeva kretanje osnovne učestanosti, trajanje pojedinih segmenata i promenu energije signala, a na simboličkom su to akcenti i sl. Sintaktičko-prozodijska analiza ima zadatak da ispita preostale mogućnosti za vrstu, oblik i funkciju reči u rečenici i na taj način utvrdi unutrašnju hijerarhisku strukturu rečenice. Ona podrazumeva identifikaciju rečeničnih fraza, utvrđivanje rečenične intonacije, rečeničnog naglaska, ritma i pauze. Interpunkcija ima važnu ulogu u otkrivanju prozodije.

Obuka samog NLP modula neće biti razmatrana u okviru ove disertacije. Za potrebe obuke akustičkih modela koriste se govorne baze koje se sastoje od audio snimaka i odgovarajućih tekstualnih transkripcija. Tekstualnu transkripciju neophodno je fonetski i prozodijski anotirati. Od toga koliko je tačno i detaljno izvršena anotacija zavisiće i kvalitet sintetizovanog govora. Neophodno je imati barem fonetsku anotaciju. Fonetska anotacija je potpuno automatizovan proces i vrši se na osnovu audio snimaka i unapred obučenih lingvističkih modela za automatsko prepoznavanje govora.

Kada je u pitanju prozodijska anotacija, često se koristi sistem konvencija zasnovan na ToBI modelu (engl. *Tone and Break Indices*) [Beckman, 2005] koji se zasniva na indeksiranju tonova i granica između određenih intonacionih celina. Ovaj model razvijen je prvenstveno za američki engleski, ali su se pojavile i njegove verzije za niz drugih jezika, uključujući i srpskohrvatski [Gođevac, 2005]. ToBI opisuje tonske događaje (tonske akcente, frazne akcente i granične tonove) i intonacione celine. Tonski akcenti se javljaju kao kombinacije visokih i niskih tonova i vezuju se za naglašeni vokal u reči, a zavise najčešće od sintakse, a ređe i od semantike ili specifične namere govornika. Tonski akcenti dele se na visoke i niske, u zavisnosti da li se naglašeni vokal vezuje za visoki ili niski ton. Javljaju se u raznim kombinacijama poput L+H* i L*+H, gde je sa H označen visoki ton, sa L niski, a sa * leksički naglašen slog. Tonski akcenat nose samo reči koje govornik smatra važnim u govornoj

situaciji. Granice između intonacionih celina obeležavaju se oznakama nivoa 0 do 4. Rečenice se dele na intonacione fraze, a svaka od njih završava se oznakom nivoa 4, dok se unutar tih fraza mogu uočiti kraće fraze i razdvojiti oznakom nivoa 3. Diskontinuitet između reči koji ne utiče na promenu osnovne učestanosti označava se oznakom nivoa 2, a kada diskontinuiteta između reči nema, koristi se oznaka nivoa 1. Krajevima fraza nivoa 3 se pridružuju frazni akcenti (L-, !H- i H-, gde je ! oznaka za spuštanje nivoa u odnosu na H). Svaka intonaciona fraza nivoa 4 karakteriše se graničnim tonom L% ili H%, a kako svaka intonaciona fraza nivoa 4 sadrži bar jednu intonacionu frazu nivoa 3, postoji 6 kombinacija fraznog akcenta i graničnog tona kojima se intonaciona fraza nivoa 4 može završiti. Postoji još oznaka koje obuhvata standardan ToBI (recimo oznake za spontani govor), a neretko autori uočavaju nedostatke standardnog ToBI-ja i sami ga proširuju novim oznakama (recimo oznakama za upravni govor, umetnute fraze i isticanje određene reči) [Sečujski, 2018]. Iako je ToBI čest izbor konvencija za anotaciju jer je čitljiv od strane računara, proširiv na druge jezike i fenomene, anotacija traje i po nekoliko stotina duže od govora koji se anotira, a često se uočava i neslaganje transkripcije različitih labelatora. Postoje i drugačiji sistemi za prozodijsku anotaciju, često kreirani po specifičnosti jezika. Tako se na primer, za srpski jezik standardno identifikuju 4 vrste akcenta (kratkosilazni, dugosilazni, kratkouzlazni i dugouzlazni). Granice intonativnih fraza se takođe mogu drugačije definisati, kao i tonovi i koristiti podskup, nadskup ili potpuno drugačiji skup oznaka u odnosu na ToBI [Sečujski, 2011].

Na osnovu anotacije vrši se izdvajanje obeležja koja se prosleđuju u modul za akustičko modelovanje. Obeležja predstavljaju odgovore na unapred definisana pitanja koja se odnose na određen fonetski segment. Pitanja mogu biti leksička, pa se već na osnovu tekstualne transkripcije baze mogu dobiti odgovori, ali mogu se odnositi i na morfološke ili prozodijske karakteristike odgovarajuće lingvističke jedinice, pa se odgovor dobija na osnovu morfološke ili prozodijske anotacije. Pitanja su binarna, a neki od primera su:

- Da li je posmatrani fonem vokal (nazal, frikativ…);
- Da li je prethodni fonem vokal (tišina, isti taj fonem, …);
- Da li fonemu odgovara tonski akcenat H* (L*, L+H*, …);
- Da li posle fonema sledi granica nivoa 3 (1, 2, 4).

2.2 Modelovanje akustičkih obeležja

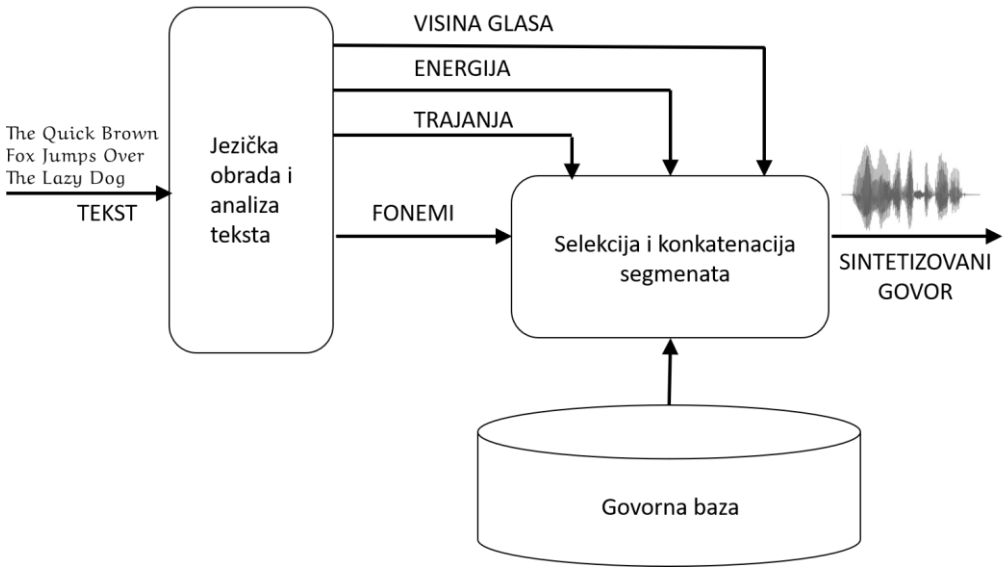
Kada računar raspolaže nizom fonema praćenih informacijom o prozodijskim obeležjima, tada je potrebno povezati apstraktne simboličke komande sa stvarnim talasnim oblicima opisanim numeričkim vrednostima. Zadatak modula za generisanje signala je da na osnovu informacija dobijenih od NLP modula generiše govorni signal. Od njega direktno zavisi razumljivost generisanog govornog signala. Realizacija ovog modula značajno se razlikuje kod različitih vrsta sintetizatora.

U nastavku će biti ukratko objašnjen način generisanja govornih signala kod dva dominantna pristupa sintezi govora – konkatenativnom i parametarskom.

2.2.1 Generisanje govornog signala kod konkatenativnog pristupa

Generisanje govornog signala kod konkatenativnih sintetizatora (slika 2.3) svodi se na problem traženja odgovarajućih segmenata u govornoj bazi koji će potom biti spajani i u nekoj meri izmenjeni. Poželjno je povezivati što veće govorne celine da bi bilo što manje čujnih prelaza između segmenata. Na taj način se delimično mogu očuvati i prozodijska svojstva govorne celine. Međutim, to bi zahtevalo preobimne govorne baze. Najbolji rezultati su se pokazali kod sistema koji ne ograničavaju trajanje segmenta, te se može uzeti i čitava rečenica iz baze ukoliko ona odgovara svim željenim prozodijskim kriterijumima. Svaki fonem u fonetskoj transkripciji sadrži vektor obeležja, npr: osnovna frekvencija, energija, trajanje, a implicitno i fonetski kontekst prethodnih i narednih fonema.

U zavisnosti od raspoloživih segmenata u bazi i željenih koje treba generisati, svakom fonetskom segmentu dodeljuje se cena korišćenja koja se određuje na osnovu razlika u njihovim obeležjima. Takođe se dodeljuje cena spoja svakom paru segmenata koji bi se mogao spojiti, a koja se određuje na osnovu razlika na granicama dva segmenta. Konačno, biraju se segmenti sa ukupnom najnižom cenom jer je kod tako odabranog niza segmenata najbolji kompromis između potrebe za njihovim prozodijskim modifikacijama i toga koliko su pogodni za međusobno povezivanje. Kao mera razlike spektralne obvojnice može se koristiti kepstralno



Slika 2.3 Blok šema konkatenativnog sintetizatora

rastojanje. Prozodijske modifikacije ne unose jednaku degradaciju u sve foneme i subfonemske jedinice pa se dodeljuju i težinski koeficijenti za različite foneme pri računanju krajnje cene.

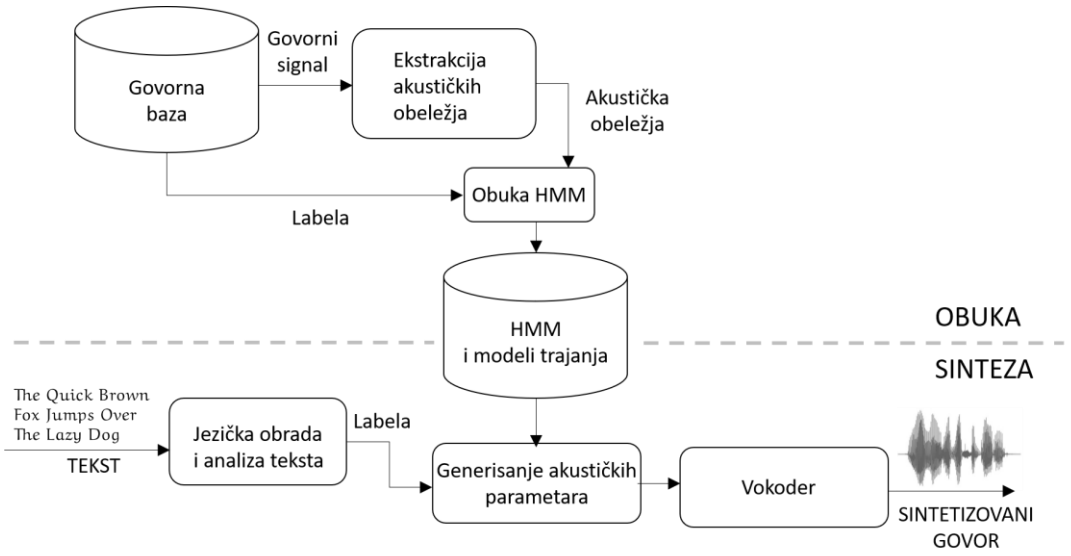
Proces pripreme segmenata za korišćenje je dugotrajan i podrazumeva ručno (ili poluautomatsko – automatsko uz proveru od strane ljudi) anotiranje čitave govorne baze, ali kada je poznato šta treba sintetizovati, sama sinteza mora da bude veoma brza. Kada su odabrani segmenti koji će biti spajani, ostaje da se modifikuju prozodijska obeležja. Željena trajanja segmenata i krive osnovne frekvencije nikada se neće u potpunosti poklopiti sa raspoloživim segmentima u bazi. Zato su neophodne tehnike modifikacije prozodijskih obeležja i povezivanja segmenata, ali takve da unesu što manju degradaciju u govorni signal. Tehnika koja se koristi je *povezivanje frekvencijski sinhronih segmenata u vremenskom domenu sa preklapanjem i sabiranjem* (engl. *Time Domain Pitch Synchronous Overlap-and-Add* – TD-PSOLA). Kod sinteze govora u vremenskom domenu postoji jedino problem nemogućnosti modifikovanja oblika spektralne obvojnice na granicama između segmenata. Taj problem prevazilazi se velikom govornom bazom u kojoj je najčešće moguće naći segmente dovoljno bliske sa akustičkog stanovišta. Sprovođenje modifikacije osnovne frekvencije direktno na talasnim oblicima podrazumeva da se od pseudoperiodičnog signala jedne periode dobije pseudoperiodičan signal druge periode, a da se ne naruši izgled spektralne obvojnice. Ako bi se radilo jednostavno sabijanje ili širenje signala u vremenskom domenu,

spektralna obvojnica bi se promenila. Postupak se zasniva na ekstrakciji frejmova govornih segmenata iz baze koja mora biti usklađena sa osnovnom frekvencijom govornog signala u trenutku analize i spajanjem tih frejmova tehnikom preklapanja i sabiranja, tako da je novo rastojanje između frejmova određeno zahtevanom osnovnom frekvencijom govornog signala datom u trenutku sinteze. S obzirom da se trajanja ciljanog i raspoloživog segmenta razlikuju, neki od frejmova polaznog signala moraju se više puta preslikati u frejmove ciljanog signala, a neki neće biti uopšte iskorišćeni. Da bi neželjeni čujni artefakti bili što manje izraženi, osnovna frekvencija i trajanja mogu se menjati, ali vrlo malo [Dutoit, 1997]. Stoga, glavni nedostatak ove metode je da se kvalitet glasa (npr. boja) ne može menjati [Schröder, 2009], te sintetizovani govor uvek zvuči kao govor iz baze. Takođe, za bilo kakvu modifikaciju u pogledu govornika ili govornog stila, morale bi se snimati i dugotrajno pripremati nove velike govorne baze.

2.2.2 Generisanje govornog signala kod parametarskog pristupa

Odavno postoje tehnike kodovanja govora i zahvaljujući njima omogućen je jednostavan i brz prenos govornog signala kroz telefonske kanale (prenose se parametri umesto brojnih odbiraka signala). Na sličan način može se razmišljati i o sintezi govora, ali umesto slanja parametrizovanog govornog signala ideja je da se ti parametri čuvaju, te da se na osnovu njih sintetizuje govor [King, 2010] kada to bude potrebno. Prva faza kod ovih sintetizatora jeste faza obuke, kada se modeli prilagođavaju govornoj bazi, odnosno, uče iz podataka u bazi koji predstavljaju obeležja izdvojena vokoderom. Tako obučeni modeli se čuvaju, umesto cele govorne baze, što je bio slučaj kod konkatenativnih sintetizatora, te se u drugoj fazi koriste za generisanje akustičkih obeležja. Na osnovu generisanih akustičkih obeležja, korišćenjem vokodera, dobija se talasni oblik govornog signala. Takvi modeli nazivaju se statističkim, a jedan od najpopularnijih pristupa bazirao se na skrivenim Markovljevim modelima (HMM), ali su, ne postižući dovoljno dobre rezultate, prevaziđeni modelima koji se baziraju na DNN.

Na Slici 2.4 prikazan je blok dijagram sistema za sintezu govora na bazi HMM, koji su predstavili tvorci HTS alata, HTS radna grupa [Yoshimura, 1999, Tokuda, 2002]. U fazi obuke se izvlače statički mel-kepstralni koeficijenti i sekvenca osnovne frekvencije, pomoću kojih se potom izračunavaju i dinamička obeležja, delta i delta-delta mel kepstralni koeficijenti (prvi i



Slika 2.4 Blok šema parametarskog sintetizatora baziranog na HMM

drugi izvod statičkih), kao i delta i delta-delta osnovne frekvencije. Spektralni i parametri osnovne frekvencije se kombinuju u jedinstvene opservacije za svaki frejm. Inicijalno, obučava se skup monofona, a zatim se ti modeli kloniraju i formiraju se modeli kontekstno zavisnih fonema za svaku kombinaciju kontekstualnih faktora koja postoji u skupu za obuku. Modeli se potom reestimiraju preko *embedded* verzije Baum-Welch algoritma, odnosno, na nivou cele rečenice, i vrši se hijerarhijska klasterizacija svih stanja na istim pozicijama u HMM-ovima. Izlazna stanja koja su završila u istom klasteru se predstavljaju kao jedna raspodela sa prosečnim parametrima radi smanjenja složenosti sistema. Modeli se ponovo reestimiraju Baum-Welch algoritmom i potom se vrši poravnavanje podataka za obuku sa dobijenim konačnim modelima preko Viterbijevog algoritma. Tako se dobijaju raspodele verovatnoća trajanja pojedinih stanja HMM-ova, a trajanje svakog stanja je modelovano jednom Gausovom raspodelom.

Tok sinteze je sledeći – sekvenca fonema koja se želi sintetizovati preslikava se u niz obučenih HMM modela koji se povezuju dajući rečenični HMM. Za modele koji ne postoje u podacima za obuku koriste se najpribližniji obučeni modeli koji se određuju pomoću stabla odluke korišćenog za klasterizaciju. Potom se iz rečeničnih HMM-ova generiše sekvenca govornih parametara (parametri pobude – osnovna frekvencija i trajanje, i mel-kepstralni

koeficijenti koji opisuju izlazni spektar), a iz njih se generiše govorni signal korišćenjem vokodera.

Postoji niz problema parametarske metode zasnovane na HMM. Količina dostupnog materijala za trening najčešće je nedovoljna za dobru estimaciju svih kontekstno zavisnih skrivenih Markovljevih modela. To rezultuje činjenicom da retko postoji dovoljan broj potrebnih kontekstnih kombinacija za tačnu obuku svakog od HMM modela. Da bi se prevazišao pomenuti problem, uvedena su stabla odluke bazirana na kontekstnom klasterovanju. Slični kontekstno zavisni HMM-ovi su grupisani u klastere u okviru kojih oni koriste istu raspodelu parametara, međutim, to ne odgovara u potpunosti većini HMM-ova u okviru klastera. Da bi se stablima odluke predstavili kompleksniji slučajevi kontekstne zavisnosti, stabla moraju biti veoma velika. Osim toga, za obuku svakog od stabala, koristi se samo deo celokupnog skupa za obuku. Sve to dovodi do natprilagođenja i smanjuje kvalitet sintetizovanog govora [Zen, 2013].

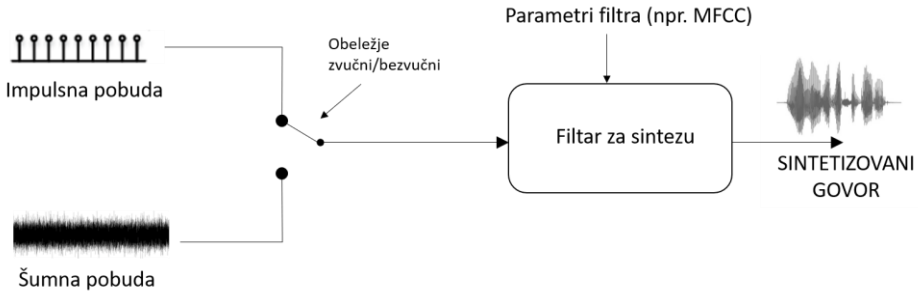
Sa razvojem hardvera, neuralne mreže su postale veoma popularne. Inspirisane su ljudskim načinom produkcije govora, za koji se smatra da ima hijerarhijsku stukturu za transformisanje informacija sa lingvističkog nivoa u talasni oblik govornog signala, kao što je to slučaj u neuralnim mrežama. Smatra se da neuralne mreže mogu postići bolje rezultate u odnosu na HMM metodu jer komplikovane kontekstne zavisnosti mogu predstaviti na kompaktniji način. Osim toga, omogućuju bolju generalizaciju modela jer se obuka vrši na celokupnom trening skupu kao i lakši rad sa visokodimenzionalnim obeležjima kao ulaznim podacima. Međutim, stabla odluke daju pravila laka za interpretaciju, dok je težine koje se koriste za različite čvorove neuralne mreže praktično nemoguće interpretirati i razumeti [Zen, 2013]. U pitanju je parametarski pristup, samo se model za generisanje akustičkih obeležja na osnovu lingvističkih zasniva na DNN umesto na HMM algoritmu.

Više o primeni DNN za modelovanje akustičkih obeležja biće dato u poglavlju 3. Danas se DNN koriste kako za zadatke modula za generisanje govornog signala, tako i za zadatke NLP modula, težeći potpuno automatizovanim *end-to-end* sistemima.

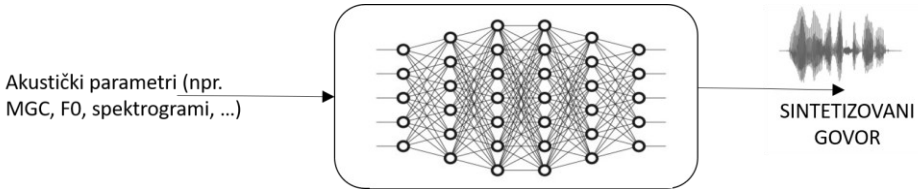
2.3 Generisanje govornog signala pomoću vokodera

Vokoder predstavlja neophodan krajnji modul kod parametarske sinteze govora. Pod pojmom vokoder podrazumeva se skup postupaka koji omogućavaju parametrizaciju govornog signala i rekonstrukciju govora na osnovu datih parametara. Ideja primene vokodera u TTS sistemu jeste da izdvoji odgovarajuća akustička obeležja iz govornih signala iz baze podataka za obuku modela, odnosno da na osnovu akustičkih parametara koje model u fazi sinteze predvidi generiše odbirke govornog signala.

Postoje dve velike grupe vokodera, deterministički i neuralni, čije su uopštene šeme prikazane na slikama 2.5 i 2.6, respektivno. Godinama su u upotrebi deterministički vokoderi koji se zasnivaju na raznim idejama, pa tako postoje fazni vokoderi [Flanagan, 1966], kanalni vokoderi [Gold, 1967], vokoderi koji estimiraju spektralnu obvojnicu [Paul, 1981], itd. Primeri naprednijih determinističkih vokodera koji se često koriste u parametarskim sintetizatorima su STRAIGHT [Kawahara, 2006] i WORLD [Morise, 2016]. Svi ovi vokoderi zasnivani su na modelu izvor-filtar. Ideja je u filtriranju šumne ili periodične impulsne pobude, u skladu sa parametrima izvora – osnovne frekvencije F0 i V/UV obeležja (engl. *Voiced/Unvoiced*) koje daje informaciju da li se generiše zvučni ili bezvučni deo govornog signala. Filtar zavisi od



Slika 2.5 Blok šema determinističkog vokodera



Slika 2.6 Blok šema neuralnog vokodera

generisanih akustičkih parametara, imitirajući karakteristike vokalnog trakta. Ideja potiče upravo od prirodnog načina produkcije govornog signala gde pobudni signal generišu glasne žice, a vokalni trakt vrši filtriranje tog signala. Vokoderi mogu da rade sa različitim akustičkim obeležjima, npr. WORLD vokoder koristi osnovnu učestanost, spektralnu obvojnicu i koeficijente aperiodičnosti. Kako bi se smanjio broj parametara, parametri spektralne obvojnice se pretvaraju u mel-frekvencijske generalizovane kepstralne koeficijente (engl. *Mel Frequency Generalized Cepstral Coefficients* – MGC). Sam vokoder podržava različite algoritme za izdvajanje osnovne učestanosti. Kao pobudu, ovaj vokoder koristi kombinaciju impulsne i šumne pobude kombinovane u frekvencijskom domenu, gde su frekvencijskim opsezima pridružene težine na osnovu koeficijenata aperiodičnosti [Ai, 2018]. Međutim, deterministički vokoderi imaju nekoliko značajnih mana. Prvo, izvor-filtar model ne uzima u obzir nelinearne efekte u prktičnoj produkciji govora. Drugo, predstavljanje vokalnog trakta ma kojim niskodimenzionim spektralnim obeležjima (npr. MGC) neminovno dovodi do gubitka informacija, odnosno gubitka spektralnih detalja i informacije o fazi. Ovi vokoderi produkuju karakteristično zujanje zbog istaknutih harmonika na višim frekvencijama, što je upravo posledica izvor-filtar modela. U boljim vokoderima ovaj problem je značajno smanjen uticanjem na korišćeni signal pobude. Razvijeni su i tzv. sinusoidalni vokoderi koji drugačije modeluju i šumnu pobudu [Hu, 2013].

Iako deterministički vokoderi mogu da postignu visok kvalitet sintetizovanog govora, smatra se da predstavljaju najslabiju kariku u TTS sistemu i da su odgovorni za uspešno razlikovanje prirodnog govora od sintetizovanog. Razvojem algoritama mašinskog učenja razvili su se vokoderi koji uče iz podataka (engl. *data-driven vocoders*) ili neuralni vokoderi [Tan, 2021]. Ova vrsta vokodera predstavlja neuronske mreže čiji su ulaz akustička obeležja (poput onih koje koristi i WORLD vokoder ili npr. mel-spektrogrami, koji su i najčešći izbor), a izlaz su odbirci govornog signala. Neuralni vokoderi mogu se podeliti u nekoliko velikih grupa. Prva predstavlja vokodere zasnovane na autoregresivnim modelima, koji koriste probabilističke modele za predviđanje odbiraka signala na osnovu prethodnih odbiraka. Iako ova vrsta vokodera postiže kvalitet sintetizovanog govora takav da ga je gotovo nemoguće razlikovati od prirodnog, sinteza je veoma spora u poređenju sa drugim pristupima. Neki od poznatijih vokodera iz ove grupe su WaveNet [van den Oord, 2016] i WaveRNN [Kalchbrenner, 2018]. Druga grupa zasniva se na generativnim suprotstavljenim mrežama

(engl. *Generative Adversarial Networks* – GAN). Ovi vokoderi prevazišli su prethodnu grupu i po kvalitetu sinteze i po brzini, zasnivajući se na osnovnoj ideji GAN modela gde generator proizvodi talasni oblik sintetizovanog govora, a diskriminator mu poboljšava kvalitet poredeći ga sa signalom prirodnog govora. Primeri vokodera su MelGAN [Kumar, 2019] i paralelni WaveGAN [Yamamoto, 2020]. Treća grupa predstavlja difuzione modele, koji su probabilistički generativni modeli zasnovani na dva glavna procesa, difuziji i oporavku. Difuzija je proces poput Markovljevog lanca u kom se Gausov šum postepeno dodaje na originalni signal dok se ovaj potpuno ne uništi. S druge strane, u procesu oporavka, procedura je obrnuta, signal se popravlja postepenim uklanjanjem šuma. Ovi vokoderi postižu izuzetan kvalitet, ali sinteza je spora. Primeri vokodera iz ove grupe su WaveGrad [Chen, 2020] i DiffWave [Kong, 2020]. Postoje i razni drugi neuralni vokoderi, a njihova glavna zajednička osobina je računarski i vremenski izuzetno zahtevna obuka.

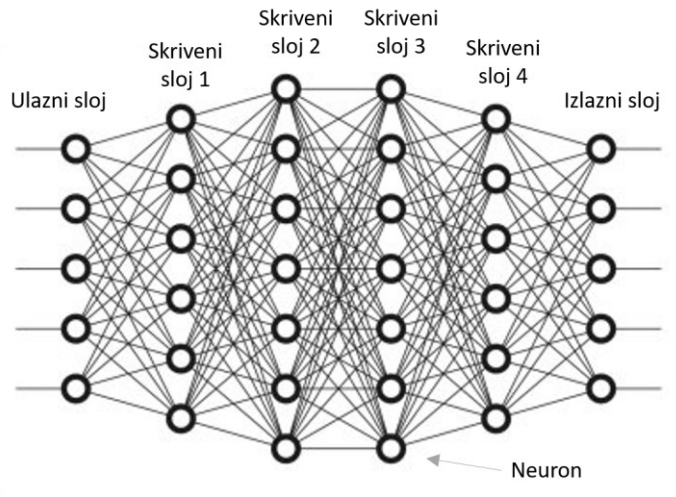
3. Primena neuralnih mreža u TTS

Kao što je ranije već pomenuto, DNN se danas u TTS koriste za različite zadatke. Mogu da se obuče za predobradu teksta, za morfološku i prozodijsku analizu, za predviđanje akustičkih parametara na osnovu lingvističkih obeležja ili na osnovu niza fonema, kao i za predviđanje odbiraka govornog signala na osnovu akustičkih ili na osnovu lingvističkih obeležja ili pak na osnovu niza fonema. Potpuni *end-to-end* sistemi imaju prednosti kao što su smanjenje potrebe za anotacijama i razvojem obeležja, izbegavanje propagacije greške među modulima u kaskadnoj vezi, kao i smanjenje cene obuke, razvoja i implementacije. Međutim, razlika između teksta i talasnog oblika govornog signala predstavlja veliki problem, jer govor sadrži ogromnu količinu redundanse u odnosu na tekst a nosi i druge lingvističke, paralingvističke i ekstralingvističke informacije. Govorni signal od 5 s sadrži oko 20 reči, odnosno može se predstaviti kao sekvenca od stotinak fonema, ali dužina talasnog oblika govornog signala je oko 80.000 (što je broj odbiraka pri frekvenciji odabiranja od 16kHz, a na višim frekvencijama razlika je još izraženija) [Tan, 2021]. Stoga se i dalje pribegava podeli TTS sistema na module, te korišćenju DNN za određene module sistema – u disertaciji će konkretno biti detaljnije razmotreno korišćenje DNN za predviđanje vrednosti akustičkih obeležja na osnovu lingvističkih.

3.1 Teorijske osnove DNN

Veštačke neuralne mreže predstavljaju jedan od algoritama mašinskog učenja. Sastoje se od velikog broja procesora (neurona) koji su organizovani u slojeve (slika 3.1). Prvi sloj neuralne mreže naziva se ulazni sloj i prima sirove podatke (ulazna obeležja), dok svaki naredni sloj kao ulaze prima izlaze iz prethodnog sloja. Poslednji, izlazni sloj, daje izlaz sistema (izlazna obeležja). Slojevi između ulaznog i izlaznog se nazivaju skriveni slojevi. Svaki neuron je povezan sa mnoštvom neurona iz prethodnog sloja, kao i mnoštvom neurona iz narednog sloja.

Još 90-ih godina neuralne mreže su korišćene za sintezu govora. Postale su popularne zbog svoje mogućnosti da predstave proizvoljnu funkciju ukoliko imaju dovoljno jedinica u



Slika 3.1 Veštačka duboka neuralna mreža sa četiri skrivena sloja

skrivenom sloju. Iako se znalo da neuralne mreže sa više skrivenih slojeva, nazvane duboke neuralne mreže – DNN, mogu predstaviti razne funkcije na mnogo efikasniji način od neuralnih mreža sa jednim skrivenim slojem (plitke neuralne mreže), takve mreže nisu korišćene zbog nedostatka računarskih resursa. Međutim, skorašnji nagli razvoj kako hardvera tako i softvera omogućio je obuku i DNN [Zen, 2013].

Većinom kada se razvijaju računarski programi, računaru se daju određena pravila i izuzeci i on na osnovu toga radi. Međutim, za kompleksnije probleme, kao što su, između ostalog, prepoznavanje i sinteza govora, vrlo je teško odrediti sva pravila, navesti sve izuzetke. Stoga su za razvoj ovih tehnologija algoritmi mašinskog učenja od izuzetnog značaja, jer se model obučava na osnovu podataka. Na osnovu tih podataka za obuku neuralna mreža treba da sazna određene pravilnosti i veze na osnovu kojih će funkcionisati.

3.1.1 Veštački neuron

Prvi veštački neuron nazvan je perceptron (slika 3.2). Ulaz u perceptron predstavlja nekoliko binarnih podataka, a izlaz je jedan binarni podatak, 0 ili 1. Težine opisuju doprinos svakog od ulaza za računanje izlaza. Izlaz zavisi od toga da li je suma proizvoda ulaza i njima odgovarajućih težina iznad ili ispod unapred određenog praga i dat je sa

$$y = \begin{cases} 0, & \sum_{j=0}^{m} \omega_j x_j \leq p \\ 1, & \sum_{j=0}^{m} \omega_j x_j > p \end{cases}, \tag{3.1}$$

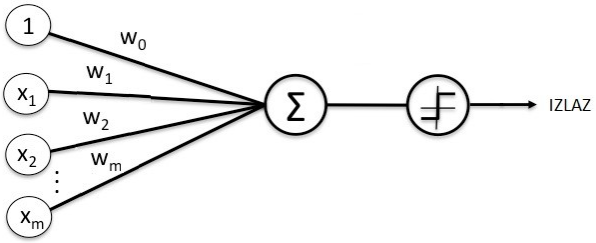
gde važi da su x_j ulazi, ω_j njima odgovarajuće težine, p je prag, y je izlaz perceptrona.

Prethodno je opisan način funkcionisanja jednog perceptrona, a neuralna mreža mogla bi se predstaviti sa nekoliko slojeva od po nekoliko perceptrona. Drugi sloj perceptrona donosi odluku na osnovu odluka iz prvog sloja perceptrona. Tu već postoji kompleksnija struktura donošenja odluka koja nije sasvim intuitivno objašnjiva. Ali može se pretpostaviti da veći broj slojeva sa velikim brojem neurona može predstavljati sistem koji na apstraktan način može donositi kompleksne odluke i donekle se može poistovetiti sa sistemom neurona koji postoje u ljudskom mozgu. Umesto praga, uvodi se termin pomeraj ili pristrasnost (engl. *bias*).

$$y = \begin{cases} 0, & \sum_{j} \omega_j x_j + b \leq 0 \\ 1, & \sum_{j} \omega_j x_j + b > 0 \end{cases} \tag{3.2}$$

U izrazu 3.2 sa b je označena pristrasnost i važi da je $b = -p$. Može se pokazati da korišćenjem mreža perceptrona može da se predstavi bilo koja logička funkcija [Nielsen, 2015].

Danas, perceptron gotovo da je u potpunosti zamenjen neuronom koji sadrži nelinearnu aktivacionu funkciju (primeri dati na Slici 3.3). Zamisao je da malim promenama težina utičemo malo na rezultat, te da takvim ponašenjem samoj neuralnoj mreži prepustimo da nauči kako da dođe do vrednosti težina i pristrasnosti. Upravo to omogućeno je uvođenjem



Slika 3.2 Šema veštačkog neurona

sigmoidalnog neurona. Za razliku od perceptrona, ulazi u neuron, kao i njegov izlaz, predstavljaju realne brojeve između 0 i 1. Izlaz neurona više ne predstavlja običnu sumu nego je relacija između ulaza i izlaza predstavljena sigmoidalnom funkcijom definisanom sa

$$y = \sigma(\boldsymbol{\omega}\mathbf{x} + b) = \frac{1}{1 + e^{-\sum_j \omega_j x_j - b}}, \tag{3.3}$$

gde je x_j ulaz, ω_j odgovarajuća težina, a b pristrasnost. U praksi se često umesto sigmoidalne funkcije koristi tangens hiperbolični ili ispravljačka linearna funkcija (engl. *Rectified Linear Unit* – ReLU) jer u zavisnosti od primene, mogu dati bolje rezultate. Izlaz neurona sa funkcijom tangens hiperbolični (tanh neuron) čiji je ulaz **x**, vektor težina **ω** i pristrasnost b dat je sa

$$y = \tanh(\boldsymbol{\omega}\mathbf{x} + b) = \frac{e^{\boldsymbol{\omega}\mathbf{x}+b} - e^{-\boldsymbol{\omega}\mathbf{x}-b}}{e^{\boldsymbol{\omega}\mathbf{x}+b} + e^{-\boldsymbol{\omega}\mathbf{x}-b}}. \tag{3.4}$$

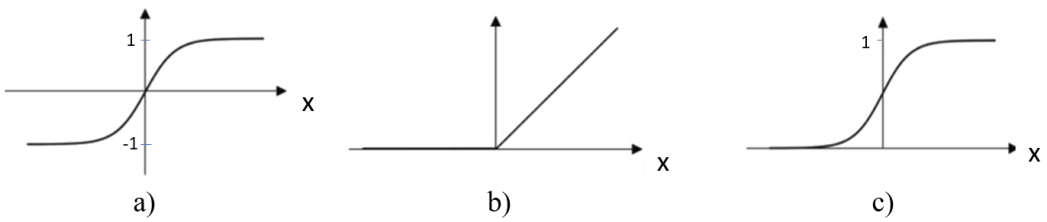
Veza između funkcije tangens hiperbolični i sigmoidalne data je izrazom

$$\sigma(z) = \frac{1 + \tanh\left(\frac{z}{2}\right)}{2}. \tag{3.5}$$

Izlaz neurona sa ispravljačkom linearnom funkcijom (ReLU neuron) čiji je ulaz **x**, vektor težina **ω** i pristrasnost b dat je sa

$$y = \mathrm{ReLU}(\boldsymbol{\omega}\mathbf{x} + b) = \max(0, \boldsymbol{\omega}\boldsymbol{x} + b). \tag{3.6}$$

Razlika između neurona sa različitim aktivacionim funkcijama ogleda se u opsegu mogućih vrednosti izlaza – kod sigmoidalnog to je 0 do 1, kod tanh neurona −1 do 1, a kod ReLU od 0 do beskonačno (slika 3.3), što znači da se normalizacija izlaza, a često i ulaza, mora vršiti na drugačije načine [Nielsen, 2015].



Slika 3.3 Aktivacione funkcije: a) tanh, b) ReLU, c) sigmoid

3.1.2 Veštačke neuralne mreže

Svaka neuralna mreža ima ulazni, izlazni i najmanje jedan skriveni sloj. Svaki sloj sačinjen je od određenog broja neurona (slika 3.1). Postoje različite arhitekture neuralnih mreža. Mreža kod koje informacija ne prolazi više puta istim mestom, nazivaju se nerekurzivne mreže (engl. *feedforward*). Kod ovih mreža jedan ulaz utiče na aktivacije svih neurona u preostalim slojevima. Mreže kod kojih postoje određene petlje, nazivaju se *rekurzivne* neuralne mreže (engl. *recurrent neural networks* – RNN). Kod takvih mreža, ponašanje neurona nije određeno samo aktivacijama u prethodnim slojevima, nego i aktivacijama u prethodnim vremenskim trenucima. Na ulaz neurona čak utiču i njegovi izlazi iz prethodnih trenutaka [Nielsen, 2015].

U procesu obuke neuralne mreže mora postojati neka procena koliko dobar rezultat neuralna mreža daje. Mera tog kvaliteta naziva se funkcija cene (engl. *cost function*) i u slučaju regresionih problema najčešće predstavlja sumu kvadrata razlike (engl. *Mean Square Error* – MSE) između prave i dobijene vrednosti (izraz 3.7), dok je u slučaju klasifikacionih problema najčešći izbor međuentropijska funkcija (izraz 3.8):

$$J_{mse} = \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{k_L}(\hat{y}_m{}^{(i)} - y_m{}^{(i)})^2, \tag{3.7}$$

$$J_{ce} = \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{k_L}\left[y_m{}^{(i)}\ln(\hat{y}_m{}^{(i)}) + \left(1 - y_m{}^{(i)}\right)\ln(1 - \hat{y}_m{}^{(i)})\right]. \tag{3.8}$$

U izrazima 3.7 i 3.8, $y_m{}^{(i)}$ predstavlja željenu vrednost izlaza m-tog neurona izlaznog sloja mreže za i-ti uzorak, $\hat{y}_m{}^{(i)}$ je izlazna vrednost m-tog neurona izlaznog sloja mreže za i-ti uzorak koju je mreža predvidela, N je ukupan broj uzoraka, a k_L je broj neurona u izlaznom (L-tom) sloju mreže.

Cilj je da se vrednost funkcije cene minimizuje. Dakle, potrebno je doći do vrednosti težina i pristrasnosti takvih da minimizuju funkciju cene. Da je u pitanju neka jednostavna funkcija, zavisna od jedne ili nekoliko promenljivih, mogli bi se naći minimumi funkcije traženjem izvoda po svakoj promenljivoj. Međutim, funkcija cene je nelinearna funkcija najčešće izuzetno velikog broja promenljivih, te se minimizacija funkcije cene vrši korišćenjem iterativnog postupka. Najčešće je u pitanju postupak zasnovan na algoritmu gradijentnog silaska (engl. *gradient descent*). Ovaj algoritam započinje od proizvoljnih vrednosti

promenljivih koje utiču na funkciju cene. Te promenljive se potom u iterativnom procesu ažuriraju u skladu sa pravilom datim u izrazu

$$\boldsymbol{\omega}^{r}{}_{j(\text{novo})} = \boldsymbol{\omega}^{r}{}_{j(\text{staro})} - \alpha \frac{\partial J}{\partial \boldsymbol{\omega}^{r}{}_{j}}, \tag{3.9}$$

gde je J opšti oblik funkcije cene, $\boldsymbol{\omega}^{r}{}_{j}$ je težinski vektor j-tog neurona u r-tom sloju mreže (skup promenljivih od kojih zavisi funkcija cene), a α je brzina učenja.

Dakle, u svakom koraku se računa gradijent funkcije i u malim koracima menja vrednost promenljivih smanjujući vrednost funkcije cene. Efikasan algoritam za računanje gradijenata je algoritam propagacije unazad (engl. *backpropagation algorithm*). Ažuriranje se ponavlja do konvergencije algoritma.

Proces obuke započinje inicijalizacijom parametara mreže (težine i pristrasnosti). Inicijalizacija se može raditi na različite načine, a najjednostavniji je da se dodele slučajne male vrednosti. Ulazni podaci ulaze u mrežu i izračunavanjem izlaza pojedinih neurona, sa trenutnim vrednostima parametara mreže, propagiraju od ulaznog ka izlaznom sloju (engl. *forward propagation*). Na osnovu dobijenih i željenih (datih u bazi za obuku) izlaza mreže, izračunava se vrednost funkcije cene. Potom se korišćenjem algoritma propagacije unazad izračunavaju parcijalni izvodi funkcije cene za svaku težinu i svaku pristrasnost. Računanje greške vrši se od poslednjeg ka prvom sloju, odakle i potiče naziv ovog algoritma.

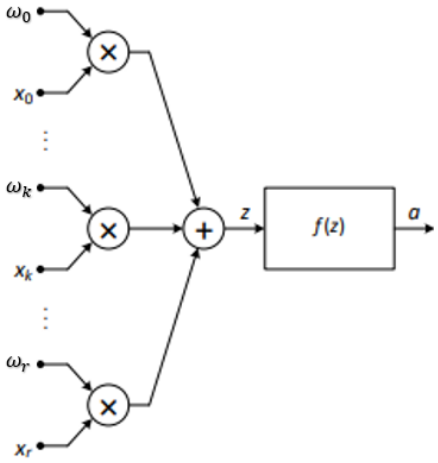
3.1.3 Algoritam propagacije unazad

Kako bi algoritam propagacije unazad bio u potpunosti jasan, prvo ćemo razmotriti uticaj težinskih faktora na izlaz. Ako posmatramo jedan neuron u izlaznom sloju i imamo samo jedan uzorak (slika 3.4), neka je $\boldsymbol{x}$ ulaz, $\boldsymbol{\omega}$ vektor težinskih koeficijenata, a je dobijeni, a y željeni izlaz neurona, dok je $f(z) = f(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x})$ aktivaciona funkcija (npr. sigmoid). Ako je funkcija cene MSE, odnosno, $J = \frac{1}{2}(a - y)^2$, onda se zavisnost funkcije cene od svakog pojedinačnog težinskog koeficijenta može naći na sledeći način:

$$\frac{\partial J}{\partial \omega_k} = \frac{\partial J}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial \omega_k} \tag{3.10}$$

$$\frac{\partial J}{\partial \omega_k} = (a - y) \cdot f'(z) \frac{\partial}{\partial \omega_k} \sum_{i=0}^{r} \omega_i x_i \tag{3.11}$$

$$\frac{\partial J}{\partial \omega_k} = (a - y) \cdot f'(z) \cdot x_k \,. \tag{3.12}$$



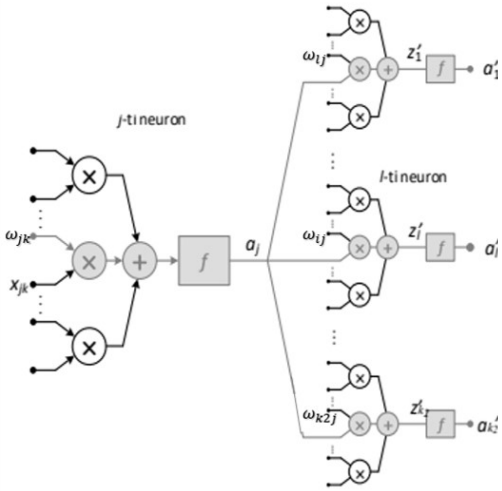
Slika 3.4 Uticaj težinskih faktora na izlaz u slučaju jednog izlaznog neurona

U slučaju kada posmatramo j-ti neuron u skrivenom sloju, težinski faktor nekog neurona utiče na funkciju cene preko svakog neurona s kojim je izlaz tog neurona povezan u narednom sloju (slika 3.5). Tada osetljivost funkcije cene na nivou neurona j računamo na sledeći način:

$$\frac{\partial J}{\partial a_j} = \sum_{l=1}^{k_2} \frac{\partial J}{\partial z'_l} \frac{\partial z'_l}{\partial a_j} \tag{3.13}$$

$$\frac{\partial J}{\partial \omega_{jk}} = \frac{\partial J}{\partial a_j} \cdot \frac{\partial a_j}{\partial \omega_{jk}} = \frac{\partial J}{\partial a_j} \cdot f'(z) \cdot x_k \tag{3.14}$$

$$\frac{\partial J}{\partial \omega_{jk}} = \delta_j x_k \tag{3.15}$$



Slika 3.5 Uticaj težinskih faktora na izlaz u slučaju j-tog neurona u skrivenom sloju

$$\delta_j = \frac{\partial J}{\partial a_j} \frac{\partial a_j}{\partial z_j} = \begin{cases} (a - y) \cdot f'(z_j), & \text{ako je neuron } j \text{ u izlaznom sloju} \\ \left[\sum_{l=1}^{k_2} \delta_l \omega_{lj}\right] \cdot f'(z_j), & \text{ako je neuron } j \text{ u skrivenom sloju.} \end{cases} \tag{3.16}$$

U nastavku je prikazan algoritam propagacije unazad po koracima, uzimajući u obzir uticaje težinskih faktora na izlaz razmotrene kroz prethodne primere.

1. ***Inicijalizacija*:** Inicijalizovati sve težine malim slučajnim vrednostima.
2. ***Propagacija unapred*:** Za svaki uzorak $\boldsymbol{x}(i)$ naći sve:

$$\left.\begin{aligned} z_j{}^{(r)}(i) &= \boldsymbol{\omega}_j{}^{(r)T}\boldsymbol{a}^{(r-1)} \\ a_j{}^{(r)}(i) &= f\left(z_j{}^{(r)}(i)\right) \end{aligned}\right\} r = 1,2,\dots,L; j = 1,2,\dots,k_r$$

   i izračunati funkciju cene za trenutne vrednosti težina

$$J = \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{k_L} e_m{}^2(i) = \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{k_L}\left(a_m{}^{(L)}(i) - y_m(i)\right)^2.$$

3. ***Propagacija unazad:*** Za svaki uzorak $\boldsymbol{x}(i)$ i svaki neuron u poslednjem sloju izračunati:

$$\delta_j{}^{(L)}(i) = \left(a_j{}^{(L)}(i) - y_j(i)\right)\cdot f'\left(z_j{}^{(L)}(i)\right), j = 1,2,\dots,k_L,$$

   a zatim, za svaki neuron u preostalim slojevima izračunati:

$$\delta_j{}^{(r-1)}(i) = \sum_{k=1}^{k_r}\delta_k{}^{(r)}(i)\,\omega_{kj}{}^{(r)}\cdot f'\left(z_j{}^{(r-1)}(i)\right), r = L, L-1,\dots,2; j = 1,2,\dots,k_r.$$

4. ***Ažuriranje težina:*** Korigovati težine na osnovu izraza

$$\boldsymbol{\omega}_j{}^{(r)}(\text{novo}) = \boldsymbol{\omega}_j{}^{(r)}(\text{staro}) - \alpha\sum_{i=1}^{N}\delta_j{}^{(r)}\boldsymbol{a}^{(r-1)}(i), r = 1,2,\dots,L; j = 1,2,\dots,k_r$$

   i ponavljati poslednja tri koraka do ispunjenja izlaznog kriterijuma.

U prikazanom algoritmu korišćene su sledeće oznake: $\boldsymbol{x}(i)$ – jedan trening uzorak; $y_m(i)$ – željeni izlaz m-tog neurona za i-ti uzorak; L – broj slojeva neuralne mreže; N– broj uzoraka za obuku; k_r – broj neurona u r-tom sloju; funkcija cene J definisana je kao MSE; $z_j{}^{(r)}$ – skalarni proizvod ulaza u j-ti neuron iz r-tog sloja i odgovarajućih težina; $a_j{}^{(r)}$ – vrednost aktivacione funkcije f j-og neurona iz r-tog sloja; $\boldsymbol{\omega}_j{}^{(r)}$ – težine između neurona iz $(r-1)$-og sloja i j-og neurona iz r-tog sloja; $\delta_j{}^{(r)}$ – osetljivost funkcije cene na nivou j-og neurona iz r-tog sloja; α – brzina učenja.

Kada se utvrde vrednosti svih izvoda, ažuriraju se parametri mreže korišćenjem neke od optimizacionih metoda u pravcu minimalnog gradijenta (pravilo ažuriranja dato u algoritmu

propagacije unazad kao korak 4, koriguje se se u skladu sa odabranom optimizacionom metodom). Veličina koraka (promena parametara) u pravcu opadanja gradijenta određena je brzinom učenja. Brzina učenja ne sme biti premala da ne bi preterano usporila proces konvergencije ka minimumu, ali ni prevelika da ga ne bi ugrozila preskakanjem globalnog minimuma funkcije cene. Različiti optimizacioni algoritmi za ažuriranje parametera mreže imaju za cilj da smanje oscilacije vrednosti funkcije cene tokom učenja, te na taj način ubrzaju proces obuke. Neki od najčešće korišćenih algoritama su stohastički algoritam opadajućeg gradijenta sa momentom (engl. *stochastic gradient descent with momentum*) i adaptivna estimacija momenta (engl. *adaptive moment estimation* – Adam). Utvrđeno je da različiti slojevi mreže uče različitim brzinama. Kada viši slojevi uče normalnom brzinom, često se dešava da se niži slojevi tokom treninga zaglave, odnosno proces učenja se toliko uspori da gotovo i da prestane. Razlog ove pojave jeste algoritam opadajućeg gradijenta. Takođe se može desiti i obrnuta situacija, da se zaglave viši slojevi. Ovaj problem naziva se problemom nestajućeg gradijenta (engl. *vanishing gradient*). Iako postoje određene alternative kojima bi se ovaj problem mogao izbeći, one nisu naročito popularne jer mogu dovesti do problema eksplodirajućeg gradijenta (engl. *exploding gradient*). Ispostavlja se da je gradijent u dubokim neuralnim mrežama nestabilan, i ovaj problem je potrebno prevazići ukoliko je to moguće. Verovatnoća pojave nestajućeg gradijenta može se umanjiti korišćenjem neke druge funkcije aktivacije umesto sigmoidalne, na primer, tangensa hiperboličnog ili ReLU. Veliki problem sa gradijentom pri obuci RNN može se prevazići na primer korišćenjem rekurzivnih neuralnih mreža sa dugom kratkoročnom memorijom (engl. *Long Short-Term Memory Recurrent Neural Network – LSTM RNN)*.

3.1.4 Problemi obuke DNN

Jedan od čestih problema kada su DNN u pitanju jeste preobučavanje mreže (engl. *overfitting*). Kako bi se izbeglo da se mreža preobuči, koristi se validacioni skup. Dakle, podaci koji su na raspolaganju se nikada ne koriste u celosti za obuku mreže, već se dele na tri skupa: skup za obuku, test skup i validacioni skup. Skup za obuku se koristi u procesu obuke mreže za poređenje predviđenih i stvarnih vrednosti izlaza, test skup služi za krajnju evaluaciju koliko je mreža dobro obučena, dok se validacioni skup koristi u okviru obuke, u svrhu prevencije

preobučavanja. Dakle, neprestano se vrši evaluacija mreže, ali na validacionom skupu i to nakon svake iteracije ili epohe. Jedna epoha predstavlja jedan prolazak svih podataka za obuku kroz mrežu. Kada se greška na validacionom skupu ustali, smatra se da je mreža obučena i proces obuke se prekida. Takav pristup se naziva rano zaustavljanje. Ipak, potrebno je odrediti parametre koji će označiti da se mreža ustalila, odnosno u koliko uzastopnih iteracija greška (ili funkcija cene) treba da bude u određenom opsegu, kao i koliki je taj opseg. Ove parametre je pogotovo teško odrediti s obzirom da se dešava da mreža nekoliko iteracija stagnira da bi kasnije nastavila da se obučava [Nielsen, 2015]. Druga mogućnost je da se unapred odredi broj iteracija, mada je teško unapred znati koliko iteracija je dovoljno. Međutim, sa stanovišta vremena neophodnog za obuku, ovakav pristup je praktičniji. Ažuriranje parametara mreže može se vršiti nakon svake epohe, ali je u praksi zbog radne memorije to najčešće neizvodljivo, ili nakon svakog uzorka, što je u praksi vrlo nepraktično. Stoga se ažuriranje parametara mreže vrši najčešće nakon prolaska jednog podskupa uzoraka zadate veličine (engl. *batch*) kroz mrežu. Parametri kao što su broj epoha, veličina *batch*-a, broj neurona u skrivenom sloju, broj slojeva i ostale veličine koje su unapred određene za mrežu, odnosno ona ne može u procesu obuke samostalno da ih koriguje, nazivaju se hiperparametri. Validacioni skup često je pogodan upravo za određivanje pogodnih vrednosti hiperparametara. Tu spadaju i brzina učenja, regularizacioni parametar i dr. Kada postoji veliki broj uzoraka za obuku, teže će doći do preobučavanja. Međutim, često prikupljanje uzoraka za obuku kao i njihovo obeležavanje predstavljaju najskuplji deo procesa.

Za sprečavanje pojave preobučavanja mreže često se koristi neka od metoda regularizacije. Najčešće korišćene metode su L1, L2 i *dropout* [Nielsen, 2015]. L2 regularizacija sprečava pojavu veoma velikih vrednosti parametara koje ukazuju na pojavu natprilagođenja. Ovaj tip regularizacije podrazumeva da se na osnovni oblik ma koje funkcije cene, C_0, doda suma kvadrata svih težina u mreži skalirana sa $\frac{\lambda}{2N}$, gde je $\lambda > 0$ regularizacioni parametar, a N je broj uzoraka za obuku (3.17)

$$C = C_0 + \frac{\lambda}{2n}\sum_{\omega}\omega^2 . \tag{3.17}$$

Kod L1 regularizacije na osnovni oblik funkcije cene, C_0, dodaje se suma apsolutnih vrednosti svih težina u mreži skalirana sa $\frac{\lambda}{N}$, gde je $\lambda > 0$ regularizacioni parametar, a N je broj uzoraka za obuku (3.18).

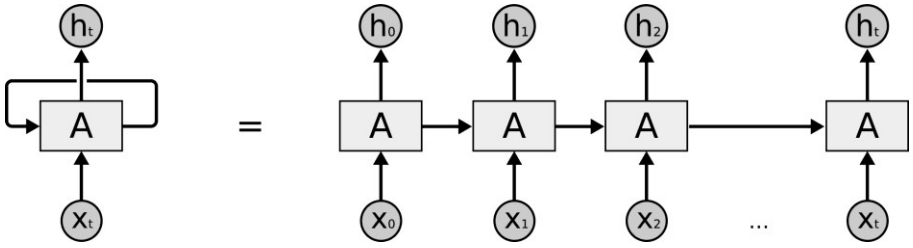
$$C = C_0 + \frac{\lambda}{N}\sum_{\omega}|\omega|. \tag{3.18}$$

Metoda *dropout* se razlikuje u većoj meri od L1 i L2 metode utoliko što ne modifikuje oblik funkcije cene, već samu mrežu. Ideja je da se u svakoj od epoha izlaz pojedinih neurona sa određenom verovatnoćom zanemaruje izjednačavanjem sa nulom. Tako se praktično smanjuje međusobna zavisnost neurona, odnosno treniraju se različite manje mreže da bi se došlo do krajnjih vrednosti pristrasnosti i težina u mreži.

Dodatno ubrzanje obuke mreže može se postići normalizacijom ulaza – normalizacijom vrednosti na određen opseg ili standardizacijom (svođenjem srednje vrednosti na 0 i varijanse na 1). Smanjenje zavisnosti između slojeva može se postići tzv. *batch* normalizacijom, što takođe dovodi do ubrzanja konvergencije. Ovim postupkom vrši se normalizacija izlaza svakog sloja mreže.

3.1.5 Rekurentne veštačke neuralne mreže

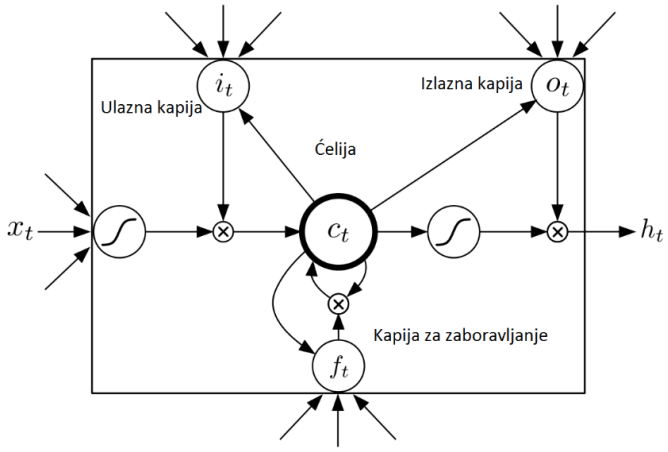
Postoje problemi, a TTS je jedan od njih, u kojima su uzorci koje dajemo kao ulaze DNN međusobno zavisni. Svaki uzorak zapravo zavisi od niza prethodnih uzoraka, odnosno predstavlja sekvencu gde postoji jasan redosled i upravo ta međusobna vremenska povezanost uzoraka značajno može da doprinese uspešnosti predviđanja od strane DNN. Međutim, da bi se uzela u obzir zavisnost od uzoraka u prethodnim trenucima, mora se iskoristiti arhitektura DNN koja sadrži povratne sprege (rekurzije), a takva je RNN. Može se reći da je RNN odmotana (engl. unrolled) kroz vremenske trenutke (korake sekvence), sa identičnim parametrima u svakom od tih trenutaka. Direktne veze se primenjuju sinhrono propagirajući izlaze neurona jednog sloja u naredni sloj u istom vremenskom trenutku, a povratne veze su dinamičke i prosleđuju informacije kroz susedne vremenske trenutke. Kao što se može otkriti posmatranjem šeme odmotane RNN sa Slike 3.6, o RNN se može razmišljati kao o redovnoj



Slika 3.6 Šema odmotane RNN

nerekurzivnoj (engl. *feedforward*) mreži, gde su parametri svakog od slojeva (direktnih i povratnih) deljeni kroz vremenske trenutke.

Međutim, ispostavlja se da je modele RNN veoma teško obučavati zbog problema nestajućeg gradijenta, koji je u slučaju RNN čak i izraženiji u odnosu na nerekurzivne mreže jer problem ne samo da propagira kroz slojeve nego i kroz vreme. Ako mreža radi dugo, gradijent postaje izuzetno nestabilan i onemogućava obuku [Nielsen, 2015]. Problem eksplodirajućeg gradijenta jednostavno je rešen odsecanjem (zadavanjem maksimalne dozvoljene vrednosti gradijenta pri obuci), dok je za problem nestajućeg gradijenta rešenje pronađeno u korišćenju LSTM neurona. Ideja je uvedena još 1997. [Hochreiter, 1997] s ciljem upravo prevazilaženja problema nestajućeg gradijenta, a ove mreže postale su dominantne za sekvencijalno učenje od 2011. Kod LSTM obični rekurentni čvorovi su zamenjeni memorijskim ćelijama. Svaka memorijska ćelija sadrži interno stanje, tj. čvor sa konekcijom sa samim sobom sa fiksnim težinskim faktorom 1. To omogućava da gradijent prođe kroz ćeliju kroz mnogo vremenskih trenutaka, a da ne dodje do njegovog nestajanja ili eksplozije. Svaka memorijska ćelija ima i nekoliko multiplikativnih jedinica nazvanih kapije (engl. *gates*), koje se koriste za regulisanje protoka informacija u memorijsku ćeliju i iz nje (slika 3.7). Ulazna kapija može da dozvoli ili da zabrani da ulaz modifikuje stanje memorijske ćelije (engl. *input gate*), dok izlazna kapija (engl. *output gate*) može da zabrani ili dozvoli da stanje memorijske ćelije utiče na neuron. Kapija za zaboravljanje (engl. *forget gate*) daje instrukciju memorijskoj ćeliji da zapamti ili da zaboravi svoje prethodno stanje, odnosno, postavi ga na 0 [Wu, 2016]. Ove jedinice imaju svoje težine na osnovu kojih funkcionišu i koje uče tokom obuke mreže, kao što sama mreža uči svoje težine.



Slika 3.7 Šema LSTM neurona

Podaci koji se prosleđuju LSTM kapijama su ulaz u neuron iz trenutnog vremenskog trenutka i skriveno stanje iz prethodnog vremenskog trenutka. Tri potpuno povezana sloja sa sigmoidalnim aktivacionim funkcijama računaju vrednosti ulaznih i izlaznih kapija, kao i kapija za zaboravljanje. Pored navedenog, neophodan je i ulazni čvor, najčešće sa *tanh* aktivacijom. Ulazna kapija određuje koliko od vrednosti ulaznog čvora treba dodati na interno stanje memorijske ćelije. Kapija za zaboravljanje određuje da li da zadrži trenutnu vrednost memorije ili da je anulira, dok izlazna kapija određuje da li memorijska ćelija treba da izvrši uticaj na izlaz u datom vremenskom trenutku ili ne. Postoje različite varijante LSTM neurona koji kombinuju neke od pomenutih kapija ili dodaju određene konekcije među pomenutim delovima LSTM. Jedna od poznatijih varijanti jeste GRU, *Gated Recurrent Unit*, što je često korišćena varijanta jednostavnija od standardne LSTM, uvedena 2014 [Cho, 2022].

Još jedan važan aspekt koji treba razjasniti jeste kako algoritam propagacije unazad radi u slučaju sekvencijalnih modela. Ideja je da se mreža odmota, odnosno da se RNN posmatra kao nerekurzivna mreža sa parametrima koji se ponavljaju u svakom vremenskom trenutku, te da se kroz takav graf radi standardna propagacija unapred, lančano pravilo, te propagacija gradijenata kroz razmotanu mrežu. Gradijenti se moraju sumirati kroz sva mesta gde se pojavljuju u razmotanom grafu. Komplikacije u suštini nastaju jer sekvence mogu biti veoma duge, te se ulazi ili u problem računanja usled ograničenog memorijskog prostora ili u problem optimizacije zbog numeričke nestabilnosti. Ovaj algoritam nazvan je algoritmom propagacije unazad kroz vreme (engl. *backpropagation through time* – BPTT) [Werbos, 1990].

Ako bismo posmatrali pojednostavljen model gde je $\boldsymbol{h}_t$ skriveno stanje, $\boldsymbol{x}_t$ ulaz, a $\boldsymbol{o}_t$ izlaz u trenutku t, imali bismo kao rezultat skrivenog i izlaznog sloja sledeće:

$$\boldsymbol{h}_t = f(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{\omega}_h), \tag{3.19}$$

$$\boldsymbol{o}_t = g(\boldsymbol{h}_t, \boldsymbol{\omega}_0), \tag{3.20}$$

gde su $\boldsymbol{\omega}_h$ i $\boldsymbol{\omega}_0$ težine, dok su f i g aktivacione funkcije skrivenog i izlaznog sloja, respektivno. Stoga imamo vrednosti $\{\dots, (\boldsymbol{x}_{t-1}, \boldsymbol{h}_{t-1}, \boldsymbol{o}_{t-1}), (\boldsymbol{x}_t, \boldsymbol{h}_t, \boldsymbol{o}_t), \dots\}$ koje su međusobno zavisne. Propagacija unapred je jednostavna, potrebno je računati trojke $(\boldsymbol{x}_t, \boldsymbol{h}_t, \boldsymbol{o}_t)$, dakle za jedan po jedan vremenski trenutak. Razlika željene i dobijene izlazne vrednosti, $\boldsymbol{y}_t$ i $\boldsymbol{o}_t$ se evaluira funkcijom cene L kroz T vremenskih trenutaka

$$L(\boldsymbol{x}_1, \dots, \boldsymbol{x}_T, \boldsymbol{y}_1, \dots, \boldsymbol{y}_T, \boldsymbol{\omega}_h, \boldsymbol{\omega}_0) = \frac{1}{T}\sum_{t=1}^{T} l(\boldsymbol{y}_t, \boldsymbol{o}_t). \tag{3.21}$$

Kod propagacije unazad, korišćenjem lančanog pravila, dobija se

$$\frac{\partial L}{\partial \boldsymbol{\omega}_h} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial l(\boldsymbol{y}_t, \boldsymbol{o}_t)}{\partial \boldsymbol{\omega}_h} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial l(\boldsymbol{y}_t, \boldsymbol{o}_t)}{\partial \boldsymbol{o}_t} \frac{\partial g(\boldsymbol{h}_t, \boldsymbol{\omega}_0)}{\partial \boldsymbol{h}_t} \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{\omega}_h}. \tag{3.22}$$

U dobijenom proizvodu problematičan član je $\partial \boldsymbol{h}_t / \partial \boldsymbol{\omega}_h$ jer $\boldsymbol{h}_t$ zavisi i od $\boldsymbol{h}_{t-1}$ i od $\boldsymbol{\omega}_h$, a i $\boldsymbol{h}_{t-1}$ zavisi od $\boldsymbol{\omega}_h$ pa se dobija

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{\omega}_h} = \frac{\partial f(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{\omega}_h)}{\partial \boldsymbol{\omega}_h} + \frac{\partial f(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{\omega}_h)}{\partial \boldsymbol{h}_{t-1}} \frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{\omega}_h}. \tag{3.23}$$

Može se pokazati da je

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{\omega}_h} = \frac{\partial f(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{\omega}_h)}{\partial \boldsymbol{\omega}_h} + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} \frac{\partial f(\boldsymbol{x}_j, \boldsymbol{h}_{j-1}, \boldsymbol{\omega}_h)}{\partial \boldsymbol{h}_{j-1}} \right) \frac{\partial f(\boldsymbol{x}_i, \boldsymbol{h}_{i-1}, \boldsymbol{\omega}_h)}{\partial \boldsymbol{\omega}_h}. \tag{3.24}$$

Jedna ideja je da se izračuna cela suma, međutim, ovaj proces bio bi vrlo spor i gradijenti bi mogli postati preveliki jer bi male promene ulaza mogle dovesti do ekstremno velikih promena izlaza, a to nije ponašanje sistema koje ciljamo. Stoga je strategija da se suma odseče, odnosno da se posmatra samo fiksan broj prethodnih vremenskih trenutaka τ. Ovakav pristup zapravo

daje aproksimaciju pravog gradijenta, prostim zaustavljanjem na $\partial \boldsymbol{h}_{t-\tau}/\partial \boldsymbol{\omega}_h$. U praksi se ovakva aproksimacija smatra dovoljno dobrom i postupak se naziva odsečenom propagacijom unazad kroz vreme (engl. *truncated backpropagation through time*) [Jaeger, 2002]. Posledica korišćenja ovakvog pristupa jeste da se model koncentriše na uticaj bliskih vremenskih trenutaka više nego na uticaj udaljenih vremenskih trenutaka, što je zapravo i poželjno jer dovodi do jednostavnijih i stabilnijih modela.

3.2 Osnovni model TTS na bazi DNN

Iako DNN mogu da budu korišćene i za jezičku obradu teksta i za modelovanje akustičkih obeležja, pa i za generisanje odbiraka govornog signala, ovde će biti akcenat na realizaciji modula koji pretvara lingvistička obeležja u akustička korišćenjem neuralne mreže, dok druga dva modula neće biti razmatrana.

Osnovni model ovakvog TTS sistema sačinjen je od dve neuronske mreže [Qian, 2014, Delić, 2017]. Jedna mreža predstavlja model za predviđanje trajanja fonema, a druga model za predviđanje akustičkih obeležja (slika 3.8). Ovakav model, kao i svaki model zasnovan na algoritmima mašinskog učenja, ima fazu obuke i fazu primene, u ovom slučaju, sinteze. U fazi obuke model za predviđanje trajanja i model za predviđanje akustičkih obeležja mogu se posmatrati i trenirati nezavisno jedan od drugog.

Model za predviđanje trajanja predstavlja neuronsku mrežu proizvoljne arhitekture (najčešće nerekurzivna mreža sa nekoliko skrivenih slojeva, rekurzivna mreža ili hibrid sa nekoliko nerekurzivnih i 1-2 LSTM sloja). Veličinu ulaznog sloja određuje broj ulaznih obeležja, što zavisi od modula za obradu teksta (NLP modul). U pitanju su uglavnom binarna obeležja koja predstavljaju odgovore na pitanja koja se tiču identiteta fonema, konteksta, morfoloških i/ili prozodijskih obeležja. Veličinu izlaznog sloja definiše broj izlaznih obeležja, a to je trajanje fonema ili češće trajanja HMM stanja na koje je fonem izdeljen (uglavnom 3 do 5) izraženih u frejmovima. Ulaz i izlaz mreže poravnati su po fonemu. Kako je u pitanju regresiona mreža, u izlaznom sloju se preporučuje korišćenje linearne aktivacione funkcije, a funkcija cene je srednja kvadratna greška. Kako bi se pripremili podaci za obuku ovakvog modela, pored jezičke obrade teksta, neophodno je izvršiti i poravnanje audio snimaka i

fonetske transkripcije i utvrditi trajanja pojedinih fonema, odnosno njegovih HMM stanja. Postoje različiti algoritmi za automatsko poravnanje, odnosno za automatsko određivanje granica između pojedinih fonema i stanja unutar fonema, a najjednostavniji je prinudno poravnanje (engl. *forced alignment*), koje potiče iz automatskog prepoznavanja govora. Ovaj algoritam zasniva se na obuci HMM modela na podacima koje treba poravnati, gde se potom postojeća transkripcija i odgovarajući audio signal pomoću Viterbijevog algoritma poravnaju međusobno, a tranzicije između obučenih HMM modela se uzimaju kao granice fonema [Li, 2016]. Ipak, za male govorne baze ovakav način poravnanja ne daje dovoljno dobre rezultate, pa su razvijena različita unapređenja ovog algoritma, od kojih je jedno predstavljeno u [Suzić, 2017]. Pored navedenog, ulazna i izlazna obeležja je neophodno normalizovati i potrebno je koristiti regularizacione metode kako bi obuka bila efikasnija. Ulazna obeležja normalizuju se na opseg (0,1), izraz 3.25, što ovde nije neophodno jer su ulazi binarni, dok se izlazna obeležja standardizuju, odnosno srednja vrednost svodi se na 0, a varijansa na 1 (3.26):

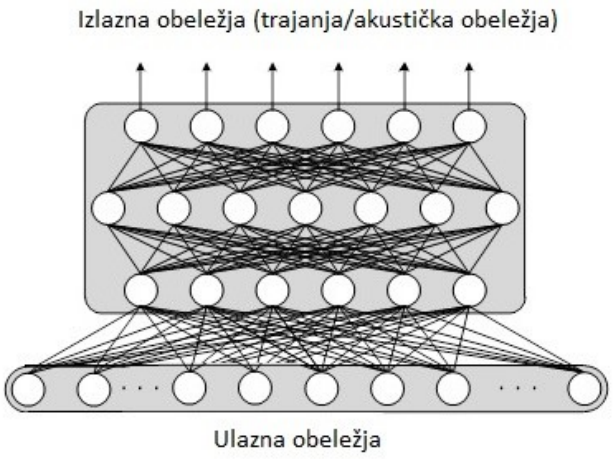
$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{3.25}$$

$$\hat{x} = \frac{x - \mu}{\sigma}. \tag{3.26}$$

Mreža za predviđanje akustičkih obeležja može biti iste ili drugačije arhitekture u odnosu na mrežu za predviđanje trajanja. Ulazi i izlazi ove mreže poravnati su na nivou frejma. Shodno tome, ulazi u mrežu isti su kao u slučaju mreže za predviđanje trajanja fonema, ali je dodato i nekoliko obeležja koja bliže određuju dati frejm [Wu, 2016]. To su obeležja koja određuju poziciju frejma u okviru HMM stanja, redni broj stanja u okviru fonema, kao i trajanje stanja, a samim tim i fonema, koje je predvidela prva mreža. Stoga, ovde je normalizacija ulaza na opseg (0,1) neophodna. Veličina izlaznog sloja određena je brojem akustičkih obeležja koja se predviđaju, i ona mogu biti različita za različite TTS sisteme. S obzirom da je i ova mreža regresiona, u izlaznom sloju se preporučuje linearna aktivaciona funkcija, a funkcija cene je srednja kvadratna greška. Za pripremu podataka za obuku modela neophodno je pomoću vokodera izdvojiti željena akustička obeležja iz originalnih audio snimaka na nivou frejma (ili recimo, u nekim sistemima, pripremiti spektrograme), što uglavnom zavisi od vokodera koji će biti korišćen za generisanje odbiraka govornog signala na osnovu datih obeležja. Izlazna obeležja se standardizuju. Takođe je preporučljivo korišćenje regularizacionih metoda. Kako

su ulazi i izlazi poravnati na nivou frejma, ova mreža ima mnogo više uzoraka za obuku u odnosu na mrežu za predviđanje trajanja, te je obuka višestruko duža.

U fazi sinteze, pomenute dve neuronske mreže vezuju se kaskadno, odnosno, na osnovu lingvističkih obeležja predviđaju se trajanja stanja fonema, a potom se te informacije koriste za izračunavanje dopunskih obeležja za ulaz u mrežu za predviđanje akustičkih parametara, na osnovu kojih se potom generiše govorni signal korišćenjem vokodera.



Slika 3.8 TTS model na bazi dubokih neuronskih mreža – primer modela NN za predviđanje trajanja fonema ili akustičkih obeležja.

4. Proširenje osnovnog TTS modela na bazi DNN

4.1 Proširenje TTS modela za više govornika/stilova

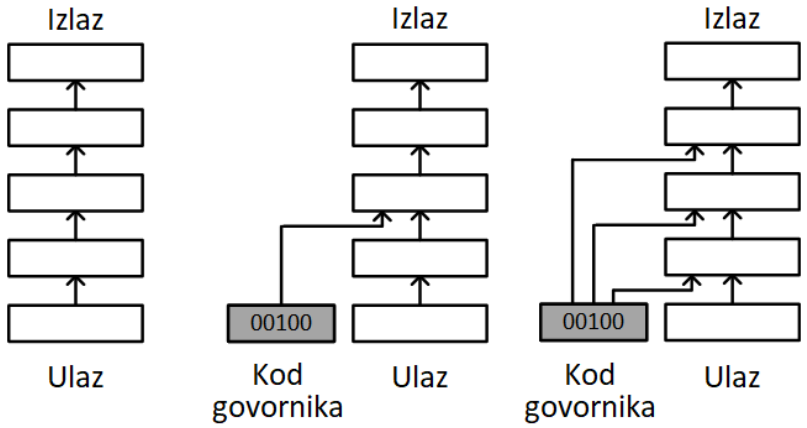
Standardni TTS modeli obučavaju se na velikim bazama govornog signala jednog govornika koji je snimljen u studiju (u povoljnom akustičkom ambijentu) ujednačenim tempom i stilom govora. Razlog za to je činjenica da bi upotreba više stilova ili više govornika dovela do uprosečavanja, te bi izlazni akustički parametri uz pomoć vokodera proizveli usrednjen glas odnosno stil. Stoga, ako se formira model koji koristi više različitih govornika ili stilova govora, ta informacija se mora na neki način naglasiti pri obuci modela. Jedan način je da se ova informacija prosleđuje kao dodatni ulaz u mrežu [Hojo, 2016, Yang, 2016]. Drugi način je da se različitim govornicima/stilovima dodele zasebni izlazni slojevi [Fan, 2015, Suzić, 2018]. Postoji mogućnost i adaptacije modela na novog govornika/stil [Delić, 2018] ili direktne konverzije audio signala [Kaneko, 2017, Kameoka, 2018]. U nastavku će detaljnije biti objašnjen svaki od pomenutih pristupa.

4.1.1 Informacija o govorniku/stilu kao dodatni ulaz

U [Hojo, 2016] predstavljena je ideja da se različiti govornici koduju jedinstvenim *one-hot* vektorima i da se takav kod prosleđuje u jedan ili više skrivenih slojeva (slika 4.1). Pomenuti kod za i-tog govornika u sistemu sa N govornika predstavlja vektor $S^i = \left[{s^i}_1, {s^i}_2, \dots, {s^i}_N\right]$, pri čemu elementi vektora S^i zadovoljavaju jednakost

$${s^i}_k = \begin{cases} 1, k = i \\ 0, k \neq i \end{cases}. \tag{4.1}$$

Na taj način, s jedne strane, mreža za svaki uzorak ima informaciju kom govorniku pripadaju ciljna akustička obeležja i shodno tome uči da za različite kodove proizvodi takva akustička obeležja kojima će se uz pomoć vokodera produkovati glas željenog govornika. S druge strane, s obzirom da različiti govornici dele sve slojeve mreže, moguće je obučiti mrežu i sa malo materijala jednog ciljnog govornika jer će se koristiti i znanja od nekog drugog govornika za kog je dostupno više materijala. Mana ovakvog pristupa je što se svi govornici smatraju



Slika 4.1 Formiranje TTS sa više govornika primenom kodova govornika

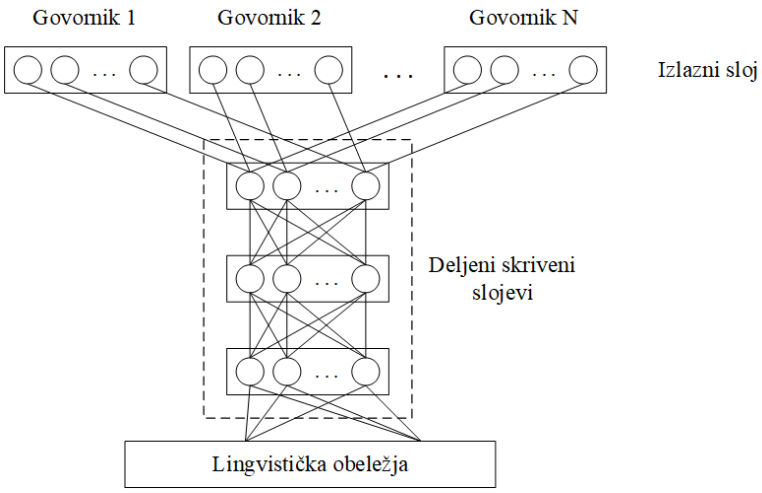
međusobno jednako udaljenim (različitim). Jedna mogućnost za prevazilaženje ovakvog problema jeste formiranje tzv. *embedding* sloja, odnosno smanjenje dimenzionalnosti prostora koji čine *one-hot* vektori [Hojo, 2018]. Na ovaj način mreži je prepušteno da samostalno tokom obuke utvrdi sličnosti i razlike između govornika te im u tom novom prostoru smanjene dimenzionalnosti dodeli odgovarajuće tačke tako da su međusobno sličniji govornici predstavljeni međusobno bližim tačkama. Kod govornika se takođe može prosleđivati u proizvoljan skriveni sloj. U [Luong, 2017] predstavljene su još neke ideje za kreiranje kodova govornika. Na primer, određene sličnosti se mogu unapred naglasiti mreži na taj način što će se uz *one-hot* kod za govornika prosleđivati i informacija o polu i starosti govornika. U [Yang, 2016] uz informaciju o polu govornika, umesto *one-hot* vektora prosleđuju se tzv. *i-vektori*. Ovi vektori dobijaju se kreiranjem univerzalnog modela govornika na osnovu većeg broja govornika, a potom se na osnovu tog modela određuje vektor jedinstven za govornika, koji se određenim metodama projektuje u prostor niže dimenzije. Sve pomenute metode zahtevaju da se za novog govornika mreža obučava ispočetka, ali se i mogu jednostavno prilagoditi da imaju mogućnost adaptacije samo dela parametara [Luong, 2017], što dovodi do lakše i brže produkcije novog govornika.

Kada je u pitanju formiranje TTS modela sa više stilova, pristupi su isti, samo se šalje kod koji opisuje stil – *one-hot* ili drugačije definisan. Ovakav pristup primenjen je na parametarski TTS baziran na DNN i na HMM u [Delić, 2018], gde su kodovi stilova definisani kao *one-hot* vektori. Osnovna razlika u TTS sa više stilova i više govornika jeste što je uvek dostupna

veoma mala količina govora u ciljnom stilu, a značajno više u neutralnom stilu, dok kod TTS sa više govornika zastupljenost pojedinih govornika u bazi može biti ujednačenija. Međutim, akustička obeležja istog govornika u različitim stilovima međusobno su mnogo sličnija nego akustička obeležja različitih govornika. Npr. kada bi se hipotetički u jednoj sintetizovanoj rečenici tokom produkcije menjao stil govora od reči do reči, a boja glasa bila ista, slušalac ne bi nužno to ni primetio ili bi mu prosto bilo prihvatljivo, dok bi u slučaju promene boje glasa od reči do reči, bilo evidentno da je u pitanju neka greška u sintezi. U radu [An, 2017] kodovi stila definisani su u trodimenzionalnom prostoru, gde je neutralni stil predstavljen kao koordinatni početak (0,0,0), a preostala tri korišćena stila predstavljaju *one-hot* vektore. Kodovi stila se tokom obuke prosleđuju ulaznom i svim skrivenim slojevima u neizmenjenom obliku, dok je pri sintezi dozvoljeno njihovo menjanje, odnosno prosleđivanje tačke u prostoru izmedju tačke koja označava određen stil i tačke koja označava neutralni stil, čime se omogućava kontrola intenziteta stila u sintetizovanom govoru. Pored *one-hot* vektora stila, u radu [Lorenzo-Trueba, 2018] istražena je mogućnost reprezentacije vektora stila kao rezultata subjektivne analize govorne baze korišćene za obuku, te vektori stila ne predstavljaju diskretne vrednosti nego kontinualne, a zasnovane na utisku o intenzitetu emocije izražene u uzorku za obuku. Pokazalo se da takav pristup doprinosi jasnoći stila u sintetizovanom govoru i takođe omogućava kontrolu jačine emocije pri sintezi.

4.1.2 Zasebni delovi mreže za svakog govornika/stil

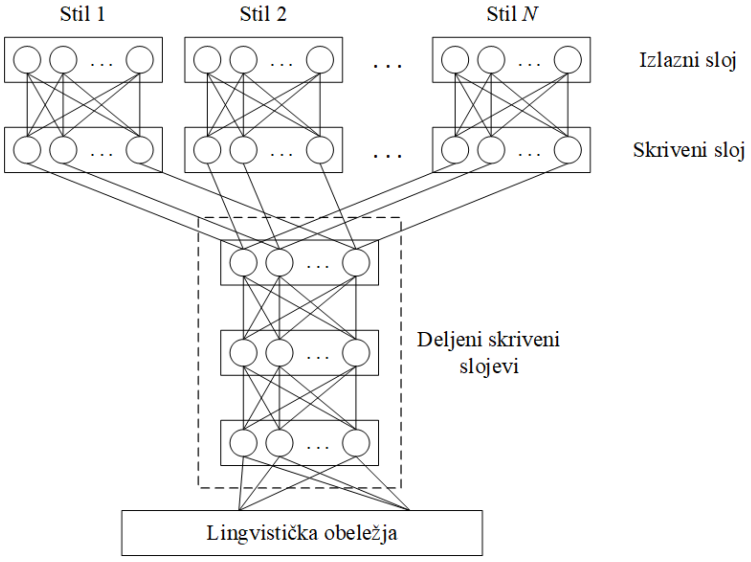
Druga ideja obuke modela sa više govornika zasnovana je na razdvajanju izlaznog sloja po govorniku [Fan, 2015]. Ovakav pristup polazi od ideje da se prvih nekoliko slojeva mreže tiče prevashodno lingvistike/jezika, a to je zajedničko za sve govornike, dok se krajnji slojevi tiču akustike, odnosno specifični su za svakog govornika. U ovakvom pristupu prvih nekoliko (ili čak svi) skriveni slojevi su deljeni (slika 4.2), odnosno celokupan materijal iz baze prolazi kroz taj deo mreže i služi za obuku njenih parametara. S druge strane, postoji onoliko izlaznih slojeva (ili izlaznih delova mreže po nekoliko slojeva) koliko ima različitih govornika, te kroz svaki od njih prolazi samo deo baze koji potiče od jedinstvenog govornika. Ovakav pristup zahteva veću količinu materijala po govorniku, kako bi se izlazni sloj (ili izlazni deo mreže) specifičan za govornika mogao obučiti. Ovaj problem može se prevazići inicijalnom obukom



Slika 4.2 Arhitektura TTS modela sa zasebnim izlaznim slojem za svakog govornika

svakog izlaznog sloja (izlaznog dela mreže) na bazi nekog većeg govornika, te bi se u procesu obuke modela sa više govornika svaki od izlaznih slojeva praktično samo adaptirao na novog govornika. Ovakav pristup je zgodan i u slučaju adaptacije na novog ciljnog govornika, gde se pritom zajednički slojevi ne bi adaptirali jer se već smatra da su nezavisni od govornika.

Slična ideja primenjena je i za TTS sa više stilova (slika 4.3), ali je, s obzirom na malu količinu materijala po stilu (izuzev neutralnog stila), inicijalna obuka svakog od izlaznih slojeva modela korišćenjem baze neutralnog govora, predstavlja neophodan korak i nije



Slika 4.3 Arhitektura TTS modela sa zasebnim izlaznim delovima mreže za svaki stil

dovoljno imati samo izlazni sloj razdvojen za svaki stil [Suzić, 2018]. Dakle, nakon obuke svakog od izlaznih slojeva na bazi neutralnog govora, radi se adaptacija svakog od izlaznih slojeva na jedan od govornih stilova.

4.1.3 Adaptacija TTS modela

Treća ideja sastoji se u prostoj adaptaciji celokupnog osnovnog TTS modela na novog govornika [Delić, 2018]. Ova ideja, iako jednostavna, daje dobre rezultate, a zasnovana je na ideji da je model obučen na govornika A bliži ciljnom modelu govornika B od modela inicijalizovanog na slučaj, te je doobuka modela govornika A na govornika B brža i zahteva manje materijala za govornika B nego standardna procedura obuke od modela inicijalizovanog na slučaj. Ovakav pristup još je poznat i kao *transfer learning*, gde se koriste znanja dobijena obukom na drugoj bazi. Može se vršiti doobuka celokupne mreže ili samo nekog njenog dela, uglavnom krajnjih slojeva.

I ova ideja može se iskoristiti i za obuku TTS modela sa više stilova [Suzić, 2018]. Postoje i modeli sa više govornika i više različitih stilova istovremeno, a koji se mogu formirati kao kombinacija nekih pomenutih ideja. Jedan takav model predstavljen je u [Sečujski, 2020], gde se svaka kombinacija govornik-stil smatra novim govornikom i koduje *one-hot* vektorom, ali se potom vrši preslikavanje u prostor niže dimenzionalnosti pomoću *embedding* sloja, u očekivanju da će na taj način mreža samostalno utvrditi veću sličnost između materijala istog govornika u različitim stilovima, nego različitih govornika.

4.1.4 Transformacija govornog signala/akustičkih parametara

Pomenute ideje zasnivaju se na direktnim intervencijama na samom TTS modelu, odnosno adaptacijama njegovih parametara. Postoje, međutim, i pristupi u kojima se produkovani sintetizovani govor jednog govornika transformiše u sintetizovani govor nekog drugog govornika na nivou akustičkih obeležja ili na nivou talasnog oblika signala. Transformacija na nivou akustičkih obeležja podrazumeva postojanje paralelnog korpusa dva govornika, te obuku TTS modela jednog govornika i obuku modela koji će prediktovana akustička obeležja transformisati u akustička obeležja ciljnog govornika [Wu, 2015]. Postoje modeli, uglavnom

zasnovani na GAN, koji imaju za cilj transformaciju govornog signala jednog govornika na drugog govornika bez upotrebe paralelnog korpusa [Kaneko, 2017, Kameoka, 2018]. Ovakve metode trebalo bi da prevaziđu probleme uprosečavanja koji su uobičajeni kod standardnih statističkih modela, kao i da omoguće konverziju na osnovu izuzetno male količine materijala ciljnog govornika.

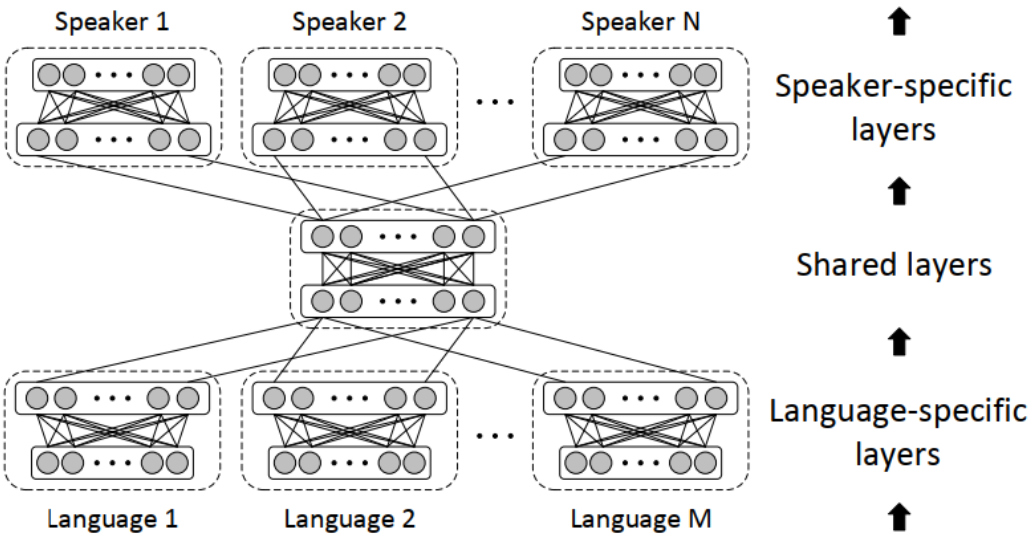
Jedna od ideja je i preslikavanje same prozodije iz referentne rečenice na sintetizovanu rečenicu, čime bi se u sintetizovanoj rečenici dobio ciljni stil govora [Kim, 2021], ali je velika mana ovakvog pristupa upravo potreba za snimanjem/pronalaženjem adekvatne referentne rečenice. U radu [Kim, 2021] rađena je sinteza sa čak 327 kodova za stil, definisanih opisno, prirodnim jezikom (npr. „u žurbi“, „pospan“, i sl.). U pitanju je model koji na ulazu dobija grafeme i kôd stila, a kao izlaz vraća spektrograme. Definisanjem stilova na prirodnom jeziku omogućena je jednostavna kontrola stila u sintetizovanom govoru, a model omogućava sintezu čak i neviđenih stilova korišćenjem lingvističkog *embedding*a, dobijenog augmentacijom postojećih kodova i sinonima. Pri sintezi je moguće koristiti ili referentni govor ili kod stila.

4.2 Proširenje TTS modela na više jezika

Problem korišćenja više govornika ili više stilova svodi se na problem koji se tiče isključivo akustičkog nivoa. Kada je u pitanju problem TTS modela koji može da vrši sintezu na više jezika, potrebno je osvrnuti se i na lingvistički/jezički nivo. Naime, ulazi u TTS sistem predstavljaju lingvistička i prozodijska obeležja, a dva jezika obično ne koriste ni isti skup fonema.

4.2.1 Proširenja DNN SPSS modela na više jezika

Jedan od najjednostavnijih pristupa jeste potpuno razdvajanje ulaza za različite jezike [Fan, 2016], slično razdvajanju izlaza za različite govornike. Ovakav pristup podrazumeva razdvajanje mreže na tri funkcionalna dela: slojevi specifični za jezik, deljeni slojevi i slojevi specifični za govornika. Mreža sa ovakvom modularnom strukturom omogućava sintezu govora sa karakteristikama glasa bilo kog govornika iz skupa za obuku na bilo kom od jezika iz skupa za obuku. Dakle, iako je za neki glas u obuci dostupan samo govor na jednom jeziku,

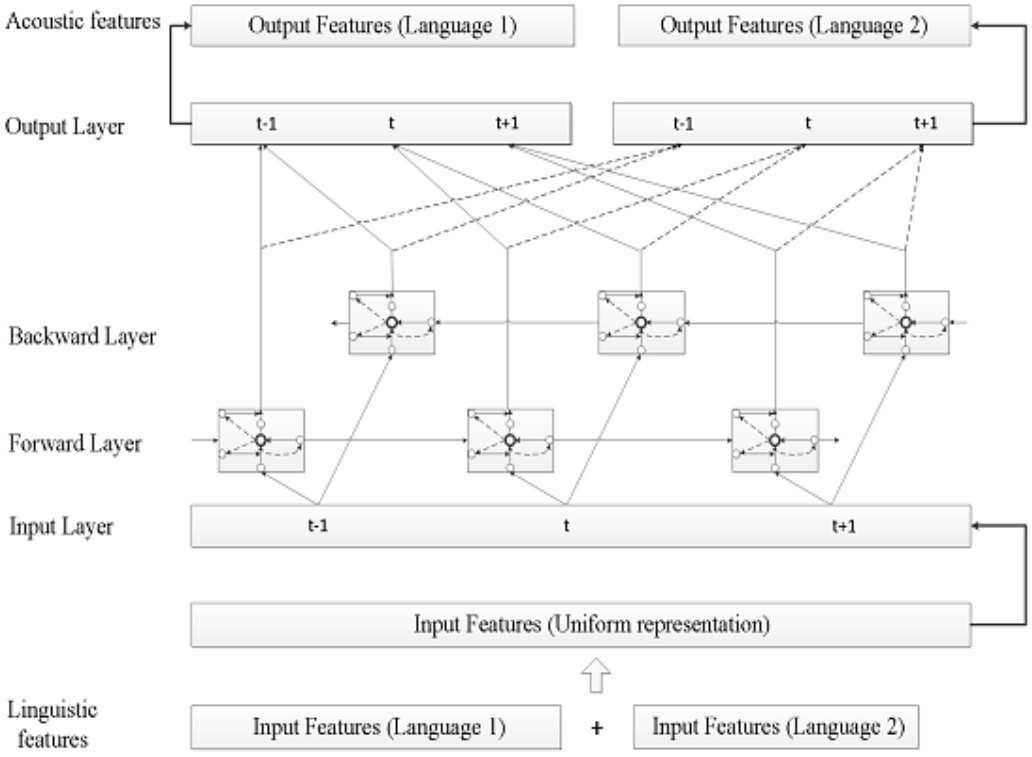


Slika 4.4 Arhitektura višejezičnog TTS modela predloženog u [Fan, 2016]

moguće je sintetizovati rečenice sa karakteristikama tog glasa na ma kom od jezika iz skupa za obuku. Šematski prikaz predloženog modela dat je na Slici 4.4. Deljeni slojevi nezavisni su od govornika i jezika i predstavljaju ključan deo koji omogućava sintezu kombinacije govornik-jezik koja nije viđena u skupu za obuku. U [Fan, 2016] predloženi model obučavan je na bazama 3 govornika, gde svaki govori i engleski i kineski, po oko 45 minuta. Obučavana je nerekurzivna mreža sa 3 skrivena sloja, a evaluacija je izvršena objektivnim i subjektivnim testovima. Utvrđeno je da su dobijeni rezultati gotovo isti kao u slučaju modela sa jednim jezikom (subjektivna ocena na MOS skali je u proseku oko 3,0, u poređenju sa prirodnim govorom ocenjenim u proseku oko 4,0). U drugom eksperimentu, za jednog od govornika izostavljen je deo baze na engleskom jeziku, a potom je sinteza izvršena upravo za tu kombinaciju govornik-jezik koja nije viđena u obuci modela. Dobijena je nešto niža ocena za prirodnost u poređenju sa rezultatima za istog govornika u slučaju sinteze modelom obučavanim na jednom jeziku (2,44 prema 2,69), i primetno niža u pogledu sličnosti glasa sa originalnim govornikom (2,13 prema 2,71). Ovakav pristup zahteva dovoljne količine materijala za svaki od jezika kako bi se svaki ulazni sloj (ili ulazni deo mreže) obučio, ali i onemogućava mreži da utvrdi sličnosti između pojedinih fonema ili prozodijskih pojava u različitim jezicima, te da iskoristi znanja o jednom jeziku za sintezu na drugom. Osim navedenog, u radu svi eksperimenti sadrže bar jednog bilingvalnog govornika, a bilingvalni

materijal je teško prikupiti, zbog čega bi bilo neophodno ispitati uspešnost pristupa u slučaju nepostojanja takvih baza.

U radu [Yu, 2016] se polazi od ideje da su ulaz i izlaz sistema, odnsono ulaz i izlaz bidirekcione LSTM mreže koja predstavlja akustički model TTS sistema, obavezno zavisni od jezika, dok skriveni slojevi mogu da budu jezički nezavisni ako se ispravno obuče. Predlaže se model koji omogućava pomenutu nezavisnost skrivenih slojeva i specifičan pristup obuke koji će omogućiti najbolje iskorišćene podataka iz svih jezika. Arhitektura predloženog modela prikazana je na Slici 4.5. Ideja se zasniva na razdvajanju izlaznog sloja za svaki od jezika, dok se ulaz ostavlja kao zajednički. Međutim, različiti jezici mogu imati različit broj lingvističkih

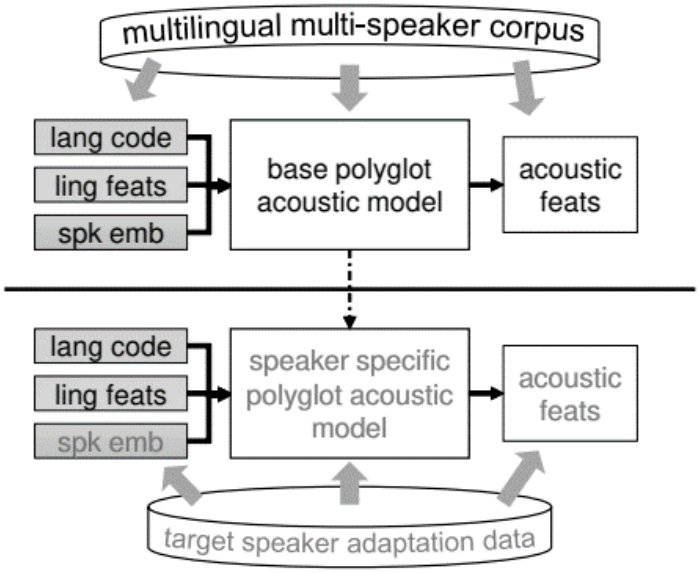


Slika 4.5 Arhitektura višejezičnog TTS predstavljena u [Yu, 2016]

obeležja (npr. broj fonema, prozodijske oznake, i dr.), što je prevaziđeno prostom konkatenacijom ulaza za svaki od jezika. Jednostavno, ako prva polovina ulaza pripada jednom jeziku, a druga polovina drugom, kada se uzorak odnosi na prvi jezik, druga polovina ulaznih obeležja predstavljaće nule i obrnuto. Skriveni slojevi su deljeni među jezicima, te se

pretpostavlja da će transformacijama formirati neku internu reprezentaciju ulaza predstavljenih na opisani način. Izlazni slojevi će koristiti tu internu reprezentaciju, što znači da će se znanja o jednom jeziku koristiti i za sintezu na drugom. Pri obuci modela korišćene su fonetski i prozodijski anotirane baze na engleskom, mandarinskom i kantonskom jeziku, sa 550 rečenica za engleski i po 2000 za druga dva jezika. Za sintezu je korišćen STRAIGHT vokoder. Rezultati su pokazali da je sinteza na osnovu višejezičnih modela višeg kvaliteta u odnosu na sintezu na osnovu jednojezičnih modela, ali nije bilo pokušaja da se sintetizuje govor u kombinacijama govornik-jezik koje nisu prethodno viđene tokom obuke.

U radu [Himawan, 2020] je predstavljen parametarski TTS model zasnovan na DNN. Osnovni višejezični TTS kao akustički model koristi jednostavnu nerekurentnu mrežu. Sistem sadrži modul koji pretvara tekst na različitim jezicima u lingvistička obeležja, model trajanja za svaki jezik posebno, akustički model za sve jezike zajedno, pri čemu koristi WORLD vokoder. Među lingvističkim obeležjima nalazi se i kod jezika u obliku *one-hot* vektora. Zadržana su samo malobrojna obeležja specifična za jezik, dok je većina lingvističkih obeležja ista za sve jezike. Ulaz u akustički model pored pomenutih lingvističkih obeležja sadrži i kod jezika i *embedding* govornika. Ideja za adaptaciju akustičkog modela prikazana je na Slici 4.6. Osnovni višejezični model obučen je na bazi sa nekoliko govornika i četiri jezika: engleski,



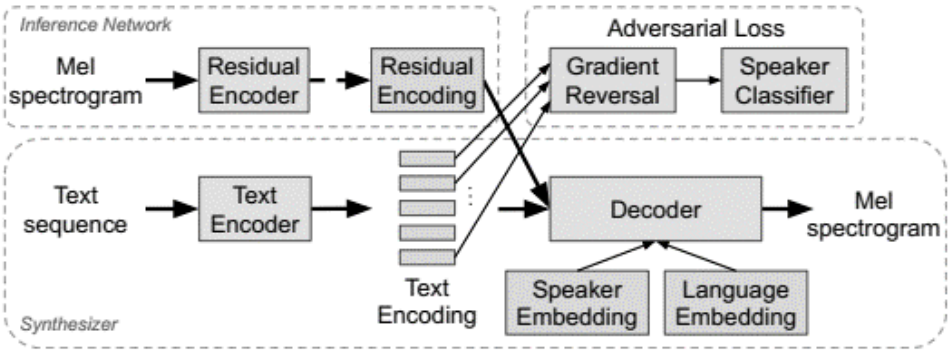
Slika 4.6 Ilustracija ideje za adaptaciju akustičkog modela iz [Himawan, 2020]

korejski, japanski i mandarinski. Količina materijala po jeziku varirala je između 37 i 87 sati, odnosno 28 i 210 govornika (ukupno 230 h, odnosno 342 govornika). Većina govornika u bazi govori jedan od 4 jezika, dok nekolicina govori 2 od 4 jezika. Dva su pristupa adaptaciji modela. Prvi podrazumeva obuku na bazi ciljnog govornika na njegovom maternjem jeziku, iako je cilj da se kasnije vrši sinteza njegovim glasom na nekom drugom jeziku. Drugi pristup podrazumeva da se adaptacija radi na bazi koja sadrži govor ciljnog govornika na maternjem jeziku i nekog drugog govornika na ciljnom jeziku. Ideja je u korišćenju regularizacije modela kako bi se sprečilo njegovo preobučavanje na maternji jezik ciljnog govornika. Za adaptaciju je korišćeno 26 do 45 minuta govora ciljnog govornika na maternjem jeziku, a za drugi pristup je na to dodato i 132 minuta drugog govornika na ciljnom jeziku. Pokušano je i sa adaptacijom pomoću bilingvalnih govornika, kada je količina materijala bila 12 do 24 minuta po govorniku po jeziku. Za evaluaciju su korišćene sinteze sa trajanjima iz originalnih rečenica, a učestvovalo je 20 slušalaca, čije su ocene predstavljene skalom od 0 do 1, za sličnost glasova (original i sinteza), kvalitet, razumljivost i prirodnost za 4 para jezika. Za sva četiri ocenjivana aspekta najbolji rezultati su postignuti u slučaju adaptacije na bilingvalni materijal ciljnog govornika, dok su najlošiji postignuti kada se koristi samo materijal ciljnog govornika na maternjem jeziku. Razumljivost je iznad 0,8 za sve modele i sve parove jezika. U pogledu prirodnosti, samo rezultati modela doobučenog bilingvalnim materijalom prelaze 0,6, a sličnost sintetizovanog glasa sa originalnim je za sve modele i parove ispod 0,75.

4.2.2 Proširenje *end-to-end* modela na više jezika

U radu [Zhang, 2019] predstavljen je višejezični model zasnovan na proširenju Tacotron [Wang, 2017] sistema za sintezu govora. Ovaj sistem omogućava sintezu u kombinaciji govornik-jezik koja nije viđena u skupu za obuku bez korišćenja ijednog bilingvalnog govornika. Iako Tacotron podrazumeva predviđanje akustičkih obeležja (spektrograma) na osnovu grafema (čistog teksta), u [Zhang, 2019] je ulaz proširen dodatnim informacijama o identitetu jezika i govornika, ali i o prozodiji i pozadinskon šumu. Predviđeni spektrogrami šalju se u neuralni vokoder, nakon čega se dobija govorni signal. Šematski prikaz sistema prikazan je na Slici 4.7. Ispitivano je nekoliko mogućnosti: kada su ulaz grafemi, kada su ulaz UTF-8 kodovani bajtovi i kada su ulaz fonemi. Kada su ulaz grafemi, skup grafema jednog

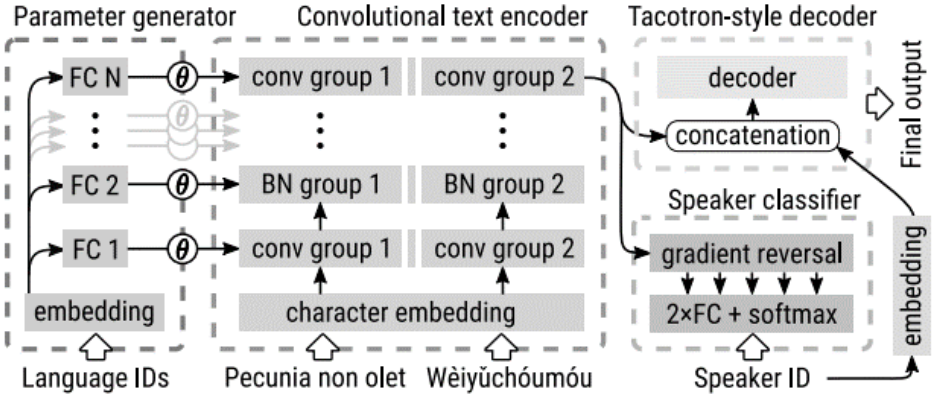
jezika konkatenira se sa skupom grafema drugog jezika što onemogućava deljenje znanja među jezicima, a povećanje broja jezika dovodi do brzog porasta broja ulaza. UTF-8 kodovanje teksta koristi 256 mogućih vrednosti, a preslkavanje grafema u UTF-8 kodove je jezički zavisno. Ovakav pristup omogućava deljene reprezentacije između jezika. Konačno, fonemi značajno olakšavaju zadatak TTS sistema jer nema potrebe za učenjem pravila izgovora za svaki jezik, ali se zahteva sistem za fonetizaciju teksta, odnosno, određivanje njegove fonetske transkripcije. Osim fonema, uključene su i informacije o tonu za kineski, odnosno primarnom i sekundarnom naglasku za španski i engleski. Pokazalo se da takav sistem, sa uključivanjem informacije o fonemima i pomenutih dodatnih informacija, daje najbolje rezultate. Osim navedenog, dodat je i sistem za kodovanje dodatnih faktora poput prozodije i pozadinskog šuma, kao i sistem za određivanje *embeddinga* jezika kao i *embeddinga* govornika, kako TTS ne bi povezao govornika i jezik kao jedinu moguću kombinaciju koja postoji na osnovu skupa za obuku. Predloženi sistem ne samo da je veoma kompleksan, nego zahteva i velike količine podataka za obuku modela. Konkretno, u [Zhang, 2019] obuka modela rađena je sa više govornika od kojih svaki govori jedan jezik – kineski, španski ili engleski. Nisu svi govornici imali isti akcenat, odnosno nisu pripadali istom regionalnom dijalektu. Eksperiment sa najmanje materijala sadržao je po jednog govornika iz svakog od tri jezika, ukupno 129 h govora, dok su neki eksperimenti rađeni sa više govornika po jeziku, ukupno 550 h govora. U subjektivnim testovima učestvovalo je svega 6 slušalaca, a dobijeni rezultati su sledeći. Kvalitet sintetizovanog govora za model sa jednim i model sa više jezika ocenjen je vrlo slično, a u oba slučaja se pokazuje najbolje kada su ulazi fonemi (prosečna ocena preko 4,0).



Slika 4.7 Arhitektura višejezičnog TTS modela predloženog u [Zhang, 2019]

Prirodnost sintetizovanog govora ocenjena je u proseku oko 4,3 za originalni par govornik-jezik, a oko 4,2 za kombinacije govornik-jezik koje nisu viđene u skupu za obuku. U pogledu sličnosti sa originalnim govornikom, u kombinaciji jezik-govornik viđenoj u obuci, ocena je u proseku oko 4,0, dok je za neviđene kombinacije prosečna ocena ispod 3,0.

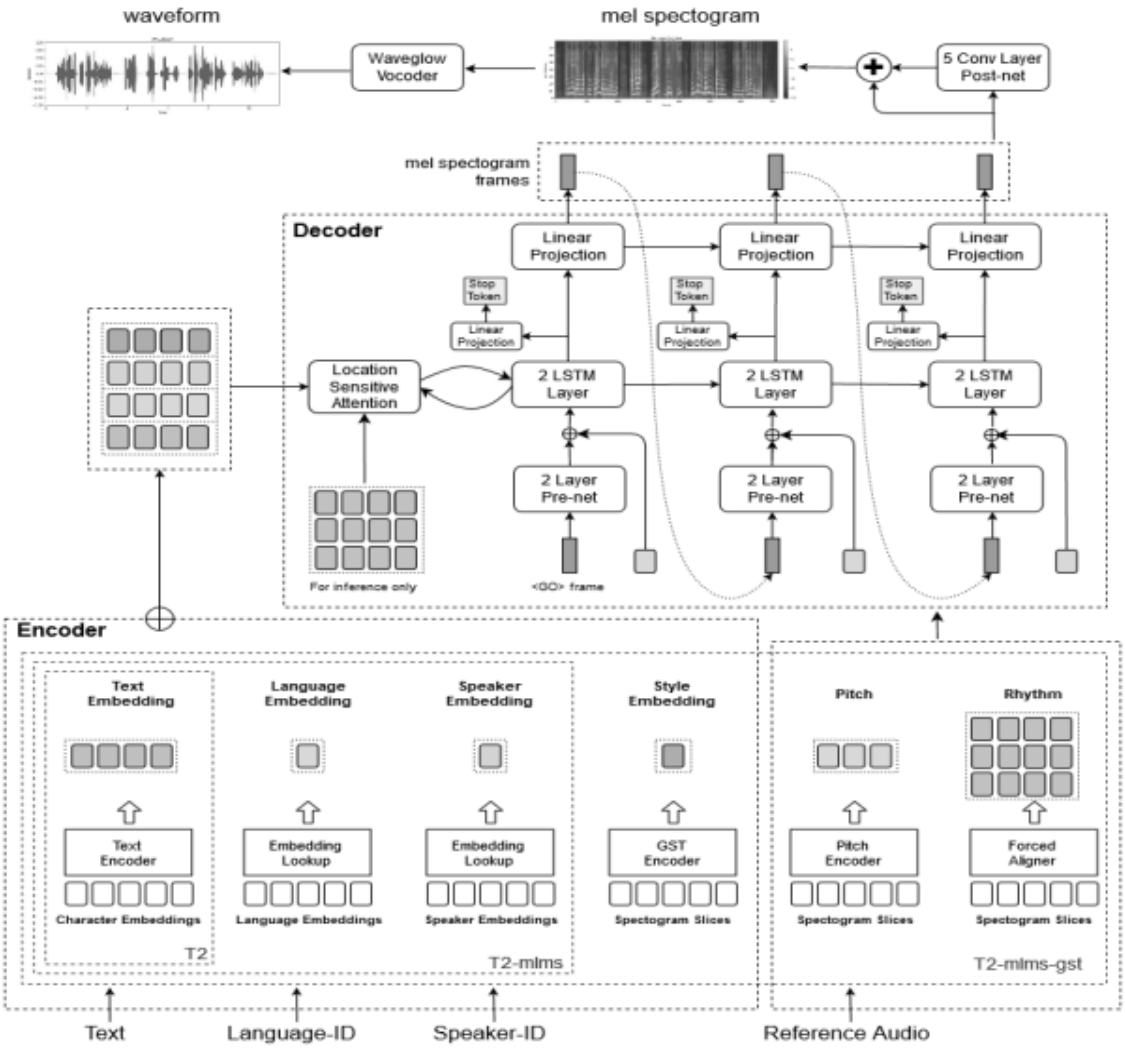
Još jedan model zasnovan na *Tacotron2 end-to-end* modelu predstavljen je u radu [Nekvinada, 2020]. Ovaj model radi sa malim količinama podataka i omogućava kvalitetnu sintezu čak i pri promeni jezika tokom jedne rečenice (engl. *code switching*). Ovaj model sadrži konvolucioni koder teksta na ulazu koji se obučava na zasebnom generatoru zasnovanom na DNN. Takođe sadrži klasifikator govornika koji omogućava odstranjivanje informacija specifičnih za govornika iz kodera. Blok šema prikazana je na slici 4.8. Kao neuralni vokoder koristi WaveRNN. Za eksperimente su korišćene baze na 10 različitih jezika, po jedan govornik, više stilova, a količina materijala po jeziku varirala je od 3,5 h do 20,9 h. U eksperimentu nije pokušavana sinteza u kombinacijama govornik-jezik koja nije viđena u obuci, a kvalitet sinteze potvrđen je samo performansama ASR sistema, odnosno u smislu razumljivosti. Za eksperiment u kom je sinteza rađena i za kombinacije govornik-jezik koje ne postoje pri obuci, dodata je govorna baza sa više govornika, ali samo 5 jezika (1-5 h po jeziku). Ovaj eksperiment testiran je korišćenjem rečenica koje sadrže reči na različitim jezicima, dakle neophodna je promena jezika u sintezi same rečenice. Ovakve rečenice ocenjene su subjektivnim MOS i MUSHRA testovima. Prirodnost je ocenjena u proseku sa 3,5, dok je



Slika 4.8 Arhitekture višejezičnog TTS modela predloženog u [Nekvinada, 2020]

tačnost izgovora ocenjena u proseku sa 3,7, a glavna uočena greška jeste preskakanje reči pri sintezi u oko 20% analiziranih rečenica.

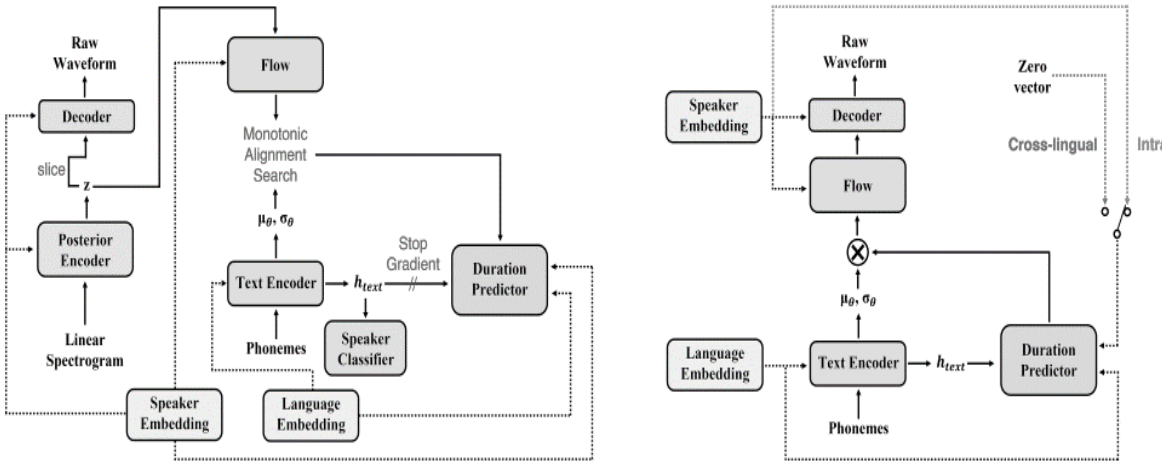
Još jedan primer prilagođavanja *Tacotron2* modela za dobijanje višejezičnog TTS predstavljen je u [Azizah, 2020]. Arhitektura modela prikazana je na Slici 4.9. Predložen je pristup hijerarhijskog *transfer learning*-a baziranog na DNN koji omogućava da novi model preuzme znanja iz prethodnog modela obučenog na velikoj bazi jednog jezika. Tako predobučeni parametri modela za poravnanje se preuzimaju za TTS model koji je adaptiran na drugi jezik za koji je dostupno manje materijala za obuku, kao i za kompleksniji model koji je adaptiran na višejezičnu bazu sa više govornika. Sličnim metodama se formira višejezični



Slika 4.9 Arhitektura višejezičnog TTS modela predloženog u [Azizah, 2020]

višegovornički model koji omogućava i transfer stila govora korišćenjem referentnog audio signala. Ulaz u celokupan sistem su grafemi, a kao vokoder se koristi WaveGlow. Za obuku je korišćeno nekoliko različitih baza, tri ciljna jezika sa manje materijala (1,7 h za prvi, 2,3 h za drugi i 7 h za treći) i engleski sa mnogo materijala (preko 24 h), a količina materijala po govorniku u bazama varira od 7 do 43 minuta. Rezultati su evaluirani od strane 9 do 20 slušalaca korišćenjem MOS skale za ocenu kvaliteta i semantički nepredvidive rečenice za ocenu razumljivosti sintetizovanog govora. Postignuta rezultati za ciljne jezike u pogledu razumljivosti su preko 97,5%, a kvalitet je ocenjen u proseku preko 4,2, dok je prirodan govor ocenjen u proseku sa 4,3. Smatra se da su dobri rezultati postignuti zahvaljujući korišćenju jezika sa sličnim lingvističkim aspektima.

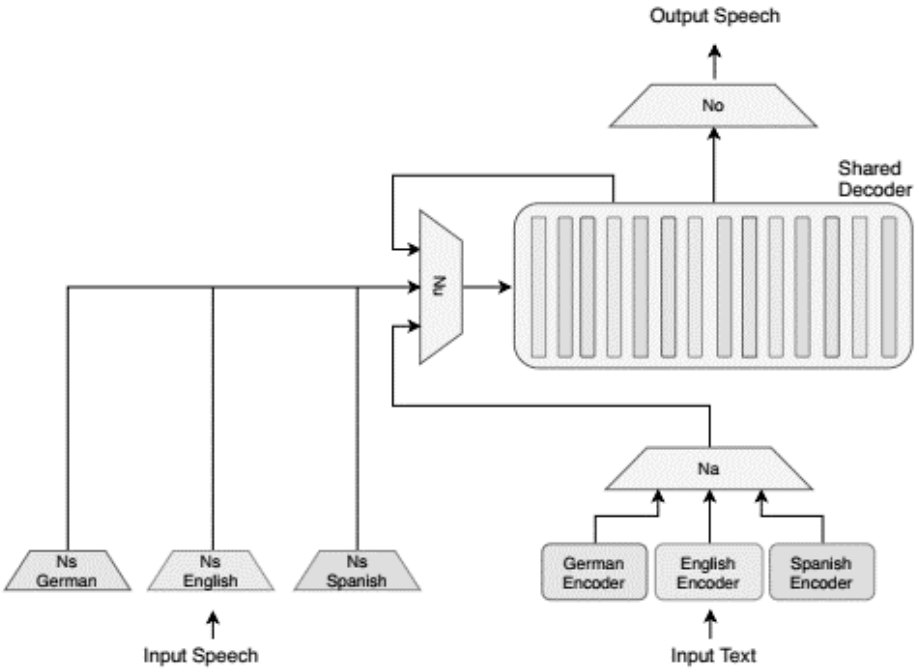
U radu [Cho, 2022] takođe je *end-to-end* model proširen radi dobijanja višejezičnog TTS. Međutim, ovde je kao osnova odabran ne-autoregresivni TTS model VITS [Kim, 2021]. Uveden je regularizacioni deo u funkciju cene zavisan od govornika, a osim toga i ideja postavljanja nultog vektora za *embedding* govornika u slučaju predviđanja trajanja. Obe novine dovele su do stabilizacije sinteze neviđenih kombinacija govornik-stil. Ovaj model na ulazu dobija sekvencu fonema, kao i *embedding* govornika i jezika, a generiše direktno talasni oblik govornog signala na izlazu. Blok šema modela data je na Slici 4.10. U eksperimentu su korišćena četiri jezika, više govornika: engleski, korejski, japanski i mandarinski, a po jeziku je količina materijala varirala od 27 h do 72 h. Rezultati su upoređeni sa javno dostupnim TTS



Slika 4.10 Arhitekture višejezičnog TTS modela predloženog u [Cho, 2022] za obuku (levo) i sintezu (desno)

višejezičnim modelom [Nekvinada, 2020]. Za subjektivnu evaluaciju korišćeni su samo primeri na engleskom zbog nemogućnosti pronalaženja evaluatora kojima je neki od preostalih korišćenih jezika maternji. Evaluaciju je vršilo svega 5 evaluatora, poredeći sintetizovan sa prirodnim govorom u smislu kvaliteta i sličnosti glasa, te ocenjujući traženo na MOS skali. Kvalitet sintetizovanog govora za postojeće kombinacije govornik-jezik ocenjen je sa 3,95, dok je prirodni govor ocenjen sa 3,99, dok je sinteza govora u kombinacijama govornik-jezik koje nisu viđene pri obuci ocenjena sa 3,82. Sličnost govornika u originalu i sintezi u slučaju viđene kombinacije govornik-jezik iznosila je 3,48, dok je prosek za neviđene kombinacije 3,34.

U radu [Nachmani, 2019] predstavljen je još jedan pristup za obuku višejezičnog TTS modela, a zasniva se na proširenju modela VoiceLoop [Taigman, 2017], koji predstavlja biološki motivisanu arhitekturu TTS sistema za više govornika koja kombinuje nekoliko neuronskih mreža (slika 4.11), ali se napominje da su predstavljena unapređenja nezavisna od vrste TTS sistema. Sistem sadrži nekoliko komponenti koje su deljene među različitim jezicima, kao i nekoliko komponenti specifičnih za svaki od jezika. Te specifične komponente



Slika 4.11 Arhitektura višejezičnog TTS modela predloženog u [Nachmani, 2019]

podrazumevaju formiranje *embedding* prostora za jezik koji će sekvencu fonema kodovati u vektorski prostor nezavisan od jezika. Zatim, mreža koja će formirati *embedding* prostor govornika, i funkcija cilja koja će identitet govornika očuvati nezavisno od jezika. Eksperimenti u [Nachmani, 2019] rađeni su sa bazama na tri jezika, engleski, španski i nemački, ukupno preko 70 h govora, a evaluacija je rađena subjektivnim testovima sa po bar 10 slušalaca. Za sintezu je korišćen WORLD vokoder, a slušaoci su ocenili sintezu vokoderom na osnovu originalnih obeležja sa 4,19, u poređenju sa 4,46 koliko je prosečna ocena originalnih snimaka. Prirodnost sinteze ocenjena je u proseku nešto preko 3,0 i za kombinacije govornik-jezik koje su viđene u obuci i za one koje nisu, dok je sličnost sinteze sa originalnim govornikom ocenjena u proseku 3,3 i za viđene i za neviđene kombinacije govornik-jezik.

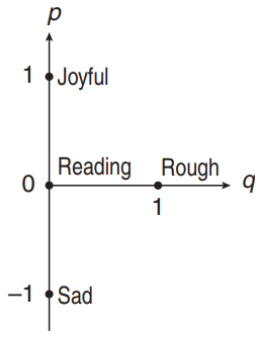
4.3 Ekpresivnost u sintetizovanom govoru

Iako se ekspresivnost u sintetizovanom govoru, u vidu ne toliko naglašavanja konkretnih emocija, ali jasnog izražavanja npr. pozitivnog stava pri saopštavanju dobrih vesti, pokazuje kao izuzetno važna, sinteza ekspresivnog govora zahteva postojanje ekspresivnih govornih baza. Dva su interesantna aspekta u vezi sa sintezom ekspresivnog govora, a to su podešavanje nivoa ekspresivnosti i transplantacija stila. Mogućnost podešavanja nivoa ekspresivnosti u sintetizovanom govoru doprinosi njegovoj prirodnosti. Transplantacija stila omogućava da se izbegne snimanje ekspresivne baze za sve govornike, odnosno da se promena stila sintetizovanog govora za jedan glas ostvari na osnovu znanja izvučenih iz ekspresivnih baza koje postoje za druge glasove.

Izuzetno su retka istraživanja koja se bave podešavanjem nivoa ekspresivnosti sintetizovanog govora. Jednostavnije ideje svode se na adaptaciju modela na željenu bazu koja će biti snimljena određenom jačinom ekpresivnosti govora, ili posmatranjem stilova sa različitim nivoima ekspresivnosti kao zasebnih stilova. Ovakvi pristupi zahtevaju snimanje baza ekspresivnog govora sa različitim nivoom ekspresivnosti i ograničeni su potom upravo na nivoe iz dostupnih baza. U nastavku će biti opisani pristupi korišćeni u modelima baziranim na HMM parametarskoj sintezi, ali bi neke ideje verovatno mogle da se iskoriste i za DNN pristup. U radu [Tachibana, 2004] predloženi su načini za interpolaciju s ciljem dobijanja

kombinacija različitih stilova, što bi se u slučaju dostupnosti baza sa različitim nivoima ekspresivnosti moglo iskoristiti za fino podešavanje nivoa ekspresivnosti sintetizovanog govora. U radu [Miyanaga, 2004] se pominje ideja podešavanja nivoa ekspresivnosti. Ideja je da se dostupni stilovi rasporede u prostoru, odnosno da svaki bude predstavljen određenom tačkom koordinatnog sistema. Međutim, u radu su u obzir uzeti stilovi: *čitanje, grub, radostan* i *tužan*, i podrazumeva se da je *tužan* suprotan od *radostan*, a da se stil *čitanje* nalazi između njih, dok se *grub* nalazi na drugoj dimenziji, slika 4.12. Prilikom promene nivoa ekspresivnosti, korišćene su tačke u prostoru između zadatih pozicija za određene stilove, tj. recimo slabije izražen radosni stil predstavljen je tačkom između stila čitanja i radosnog stila. Iako su u [Miyanaga, 2004] ovakvim pristupom sa oko 45 minuta materijala po stilu dobijeni zadovoljavajući rezultati, ostaje nedefinisano na koji način bi se drugi stilovi mogli dodati u prostor stilova. U [Nose, 2013] je opisana ideja proširena dodavanjem informacije o subjektivnom utisku nivoa ekspresivnosti emocije od strane slušalaca. Naime, 9 slušalaca je ocenjivalo svaku od rečenica iz govorne baze u pogledu nivoa ekpresivnosti, te je takva informacija korišćena pri obuci sistema. Pri evaluaciji obučenog modela pokazano je da je ovakav pristup doprineo intuitivnijoj i preciznijoj kontroli nivoa ekspresivnosti. Ovakav pristup ipak zahteva bazu sa različitim nivoima ekspresivnosti iako je uglavnom pri snimanju TTS baza govornik instruisan da održava konstantan nivo ekspresivnosti, ali ono što je veći problem je činjenica da se zahteva test slušanja za svaku novu bazu kako bi se utvrdili nivoi ekspresivnosti.

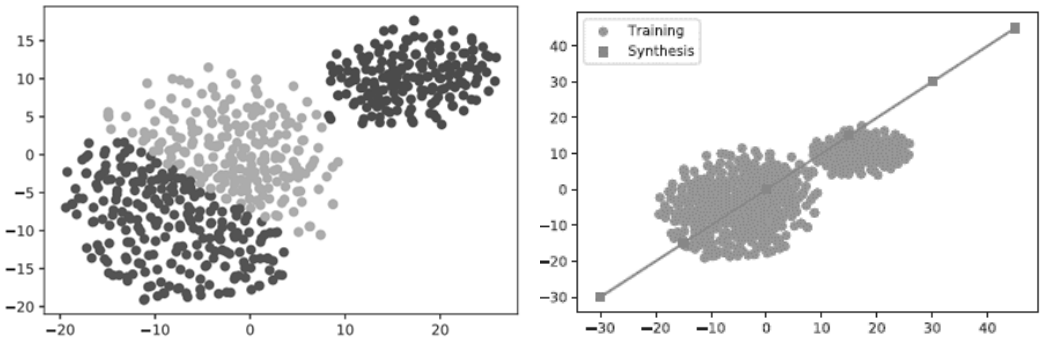
U radovima [An, 2017, Lorenzo-Trueba, 2015, Zhu, 2020] razmotrene su ideje za kontrolu nivoa ekspresivnosti pri DNN sintezi. U [An, 2017] ideja je vrlo slična ideji iz [Miyanaga,



Slika 4.12 Raspored stilova u prostoru [Miyanaga, 2004]

2004] sa istim nedostacima. U radu [Lorenzo-Trueba, 2015] analizirano je pre svega da li pri obuci ekspresivnog modela treba smatrati da je stil onaj koji je govornik imao nameru da izrazi ili onaj koji nezavisni slušaoci smatraju da je izražen. Predstavljeno je nekoliko metoda reprezentacije vektora stila i ispitano je koja reprezentacija je pogodna za kontrolu nivoa ekspresivnosti. I ovaj pristup zahtevao je prvobitno evaluaciju same baze ekspresivnog govora. Čak 266 slušalaca je evaluiralo snimke iz baze davanjem odgovora na pitanje koja od ponuđenih emocija je izražena u snimku, kao i koji je nivo izraženosti te emocije na skali od 1 do 5. Ipak, svaku rečenicu iz baze je ustvari evaluiralo svega dva slušaoca. Oformljene su matrice konfuzije kojima je upoređeno slaganje između govornikove namere koju emociju da izrazi i utiska slušalaca koja je emocija zapravo izražena i na osnovu njih data je informacija pri obuci za svaku rečenicu o vrsti izražene emocije. Urađeni su eksperimenti sa dodavanjem informacije o nivou izraženosti emocije percipiranom od strane slušaoca. Utvrđeno je da je korišćenje informacije o percipiranoj vrsti emocije doprinelo jasnijoj ekspresivnosti sintetizovanog govora, ali eksplicitno podešavanje nivoa ekspresivnosti se nije pokazalo dovoljno efikasnim.

U radu [Zhu, 2020] ideja je da se klasteri rečenica u bazi sa različitim nivoom ekspresivnosti uoče automatski primenom algoritma *k* srednjih vrednosti. Naime, korišćenjem OpenSmile [Eyben, 2010] alata za izdvajanje akustičkih obeležja za svaki od snimaka iz baze izdvojena su obeležja za koja se smatra da su korisna pri automatskom prepoznavanju emocija. Potom je izvršena klasterizacija na 3 klastera i utvrđeno je koji od klastera predstavlja najizraženiju, a koji najmanje izraženu emociju, sprovođenjem testa slušanja sa 3 slušaoca. Konačno je izvršeno smanjenje dimenzionalnosti prostora obeležja korišćenjem t-SNE [Hinton, 2002] algoritma. Dvodimenzionalni t-SNE vektor prosleđivan je kao dodatni ulaz pri obuci modela. Utvrđeno je da povećanje vrednosti po obe dimenzije t-SNE vektora dovodi do veće izraženosti emocije, odnosno visoke vrednosti t-SNE komponenti odgovaraju klasteru sa najizraženijom emocijom, odnosno, smanjenje po obe dimenzije dovodi do manje izraženosti emocije, odnosno niske vrednosti odgovaraju klasteru sa najmanje izraženom emocijom (slika 4.13), te je odabir odgovarajućih vrednosti pri sintezi služio za jednostavnu i efikasnu kontrolu nivoa ekspresivnosti. Ovakav pristup je u [Zhu, 2020] primenjen na bazama dva govornika, svaki po dve emocije. Međutim, pokušaj reprodukcije ovog rada u [Vujović, 2020] na nekoliko baza ekspresivnog govora nije uspeo. Utvrđeno je da klasteri koji se izdvajaju nisu uvek



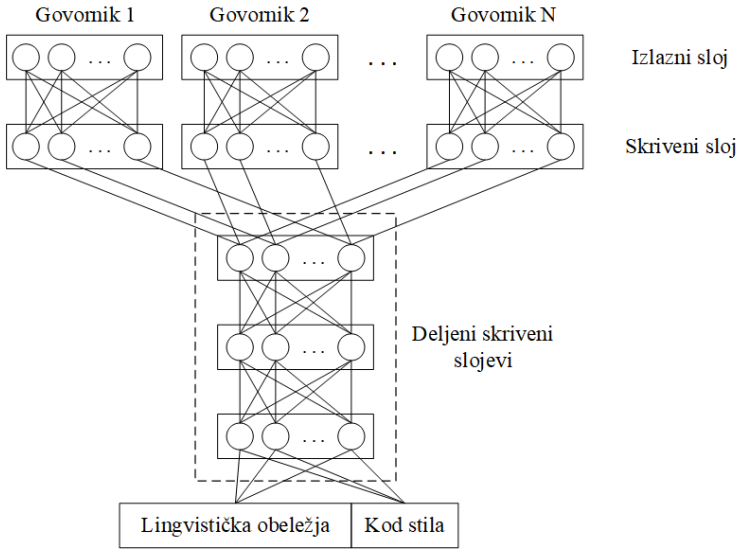
Slika 4.13 Prikaz rezultata klasterovanja u 2-D prostoru [Zhu, 2020]

orijentisani tako da visoke vrednosti komponenti t-SNE vektora odgovaraju ekspresivnijem govoru, odnosno niske vrednosti odgovaraju manje ekspresivnom govoru, što bi značilo da i ovaj pristup, kao i mnogi do sada navedeni, zahteva testove slušanja za svaku novu bazu. Osim toga, nije potvrđeno da se slušaoci jasno slažu oko toga koji klaster sadrži najslabije, a koji najjače izraženu ekspresivnost. I najbitnije, nije utvrđeno da bi se mogla postaviti jasna veza između vrednosti t-SNE vektora i nivoa ekspresivnosti, što bi bio preduslov za primenu ovakvog pristupa.

Kao što je pomenuto, transplantacija stila omogućava produkciju glasa govornika A u stilu X, iako ne postoji govorni materijal govornika A u stilu X, nego samo u neutralnom stilu, ali postoji materijal govornika B, kako u neutralnom, tako i u X stilu. U [Inoue, 2017] predstavljeno je nekoliko arhitektura DNN koje omogućavaju transplantaciju stila. Paralelna arhitektura sadrži odvojene izlazne slojeve za svakog govornika i svaki stil, te se kombinacijom odgovarajućih izlaznih slojeva mogu dobiti kombinacije govornik-stil koje nisu viđene u obuci. Serijska arhitektura poseduje deo sa slojevima koji odgovaraju govornicima i deo sa slojevima koji odgovaraju stilovima. Aktivacijom samo određenih slojeva omogućuje se produkcija željene kombinacije govornik-stil. Razmotrena je i ideja zasnovana na kodovima stila, gde se informacije o govorniku i stilu prosleđuju kao dodatni ulazi. Iako sve tri arhitekture omogućuju sintezu u kombinaciji govornik-stil koja nije viđena u obuci, pokazuje se da je paralelna arhitektura najbolja. Svi pomenuti modeli podrazumevaju velik broj govornika i dosta velike baze ekspresivnog govora. U [Parker, 2018] predlaže se pristup u kom se adaptacija vrši korišćenjem modifikovanog LHUC pristupa (engl. *Learning Hidden Layer Contribution*) [Swietojanski, 2016]. U ovom pristupu koristi se baza samo jednog govornika u

neutralnom stilu za obuku mreže, dok se adaptacija vrši na materijalu nekog novog govornika. U [Suzić, 2019] predložena je arhitektura koja omogućava transplantaciju stila, kada polazna baza za obuku sadrži mali broj govornika, a količina dostupnog ekspresivnog govora je izuzetno mala (slika 4.14). Ova arhitektura inspirisana je metodom kodova stila i adaptacije TTS modela. Informacija o stilu prosleđuje se kao dodatni ulaz, dok za svakog govornika postoji zaseban izlazni deo mreže. Rezultati su pokazali da se bolji rezultati postižu uvođenjem tzv. „uskog grla“ (engl. *bottleneck*), odnosno, skrivenog sloja značajno manjeg u odnosu na ostale (često nazvanog *embedding* sloj). Ovakav pristup omogućava ekstrahovanje obeležja, odnosno kreiranje kompaktnijih transformacija. Pokazano je da ovaj pristup nadmašuje rezultate ostalih predloženih arhitektura i da čak i sa samo dva govornika suprotnog pola postiže tačnost klasifikacije sintetizovanog govora prilikom transplantacije stila približno istu kao u slučaju prirodnog govora.

U radu [Neekhara, 2021] predložen je metod za transplantaciju stila na novog govornika za kog je dostupno vrlo malo materijala, a uspešnost predloženih metoda demonstrirana je na *end-to-end* sistemima. Ova metoda omogućava vrlo precizno podešavanje stila sintetizovanog govora eksplicitnim uslovljavanjem koda govornika, konture visine glasa i ritma tokom obuke modela. Inicijalni TTS model predstavlja jednojezični model sa više govornika, uglavnom u neutralnom stilu. Za adaptaciju modela na glas novog govornika (koji nije viđen tokom obuke)



Slika 4.14 Arhitektura koja omogućava transplantaciju stila iz [Suzić, 2019]

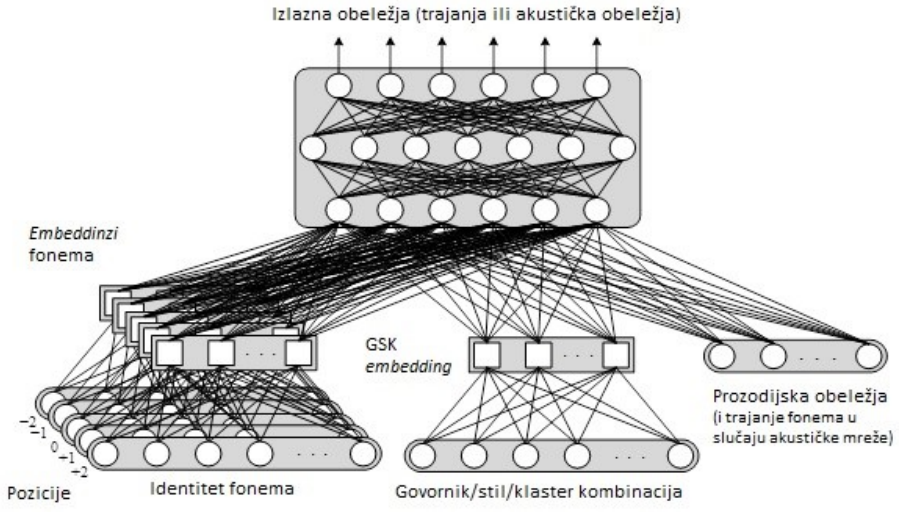
upotrebljena su tri različita pristupa. Prvi ne zahteva transkripciju, samo se iz nekoliko audio snimaka ciljnog govornika (1-20 rečenica) određuje kod (*embedding* vektor) govornika. Drugi pristup zahteva parove tekst-audio i podrazumeva adaptaciju svih parametara TTS modela, dok treći pristup podrazumeva adaptaciju samo govorničkog *embedding* nivoa. Pokazuje se da adaptacija modela značajno doprinosi sličnosti sintetizovanog i originalnog glasa. Transfer stila u ovom radu omogućen je direktnim preuzimanjem visine glasa i ritma iz ekspresivne rečenice nekog drugog govornika (tzv. referentne rečenice). Kontura visine glasa izdvojena iz referentne rečenice skalira se linearno da ima istu srednju vrednost kao srednja vrednost piča izdvojenog iz nekolicine dostupnih audio signala ciljnog govornika. Materijal ciljnog govornika koristi se i za dobijanje govorničkog *embeddinga*. Sličnost sintetizovanog glasa sa originalom testirana je formiranjem klasifikatora govornika i utvrđeno je da što je dostupno više materijala ciljnog govornika, postižu se bolji rezultati. Najbolji MOS skor u subjektivnoj evaluaciji za sličnost sintetizovanog glasa sa originalnim iznosi 3,4 i dobijen je u slučaju kada je celokupan TTS model adaptiran na materijal ciljnog govornika. U istom slučaju dobijena je i najviša ocena za kvalitet, a iznosi 3,6.

5. Ekspresivni višejezični model

U ovom poglavlju biće predstavljen model koji je glavni rezultat ove doktorske disertacije. Ovaj TTS model podržava više stilova i govornika, kao i više jezika, a omogućava sintezu kako kombinacije govornik-jezik koja je viđena u skupu za obuku, tako i onih kombinacija koje nisu viđene. Model predstavlja proširenje modela predstavljenog u [Sečujski, 2020], a neke od ideja i pristupa poklapaju se sa nekoliko ideja izloženih u poglavlju 4. Ceo model se zasniva na najjednostavnijem TTS modelu (slika 3.8), koji je prvo proširen sa *embedding* slojem za jedinstvene kombinacije govornik-stil-klaster [Sečujski, 2020], a sada i sa fonetskim *embedding* slojevima, a u nekim slučajevima može se dodati i prozodijski *embedding*.

Ulazna obeležja u originalni TTS sistem predstavljaju binarne odgovore na mnoštvo pitanja. Deo tih pitanja odnosi se na identitet fonema, a deo na prozodijska obeležja. Kako uvođenje dodatnih jezika može da znači uvođenje novih fonema, pa i novih prozodijskih obeležja, neophodne su određene izmene postojećeg ulaznog sloja. Ideja u [Sečujski, 2020] bila je da se svaka jedinstvena kombinacija govornik-stil-klaster (deo govorne baze za koji se smatra da su visina glasa, brzina govora i druge karakteristike ujednačeni) koduje jedinstvenim *one-hot* vektorom, ali i da se potom izvrši smanjenje dimenzionalnosti, odnosno kreiranje *embedding* prostora u kom bi svaka tačka predstavljala jedinstvenu kombinaciju govornik-stil-klaster. Na taj način je mreži prepušteno da samostalno uvidi sličnosti i razlike između govornika/stilova, te da pojedinim kombinacijama dodeli odgovarajuće tačke u kreiranom *embedding* prostoru. Ista logika primenjena je i u ovom modelu, gde je svaki fonem svakog od jezika kodovan jedinstvenim *one-hot* vektorom, a potom je izvršeno smanjenje dimenzionalnosti, odnosno kreiranje fonetskog *embedding* prostora. Na taj način je mreži prepušteno da samostalno utvrdi stepen sličnosti pojedinih fonema u različitim jezicima i dodeli odgovarajuće tačke u kreiranom *embedding* prostoru fonema. Osim toga, kako originalni TTS model koji se koristi kao baza ovog modela koristi ne samo informaciju o identitetu trenutnog fonema, nego i prvog i drugog prethodnog i narednog fonema, formirani su zasebni *embedding* prostori i za identitete tih fonema. Dakle, pored pomenutih 5 fonetskih *embedding* prostora, i jednog *embedding* prostora govornika, preostaju pitanja koja se odnose na prozodiju. Ukoliko se za sve jezike u modelu koristi ista šema prozodijske anotacije, npr.

ToBI, nije neophodno formirati prozodijski *embedding* prostor, već je dovoljno koristiti pitanja koja se tiču anotacije kao zajedničke za sve jezike. Šematski prikaz predloženog sistema prikazan je na Slici 5.1. Ovakav pristup omogućuje da jezik sa manje materijala potpuno iskoristi znanja koja mreža stekne iz drugog jezika u vezi sa prozodijom. To doprinosi da se prozodija sintetizovanog govora na ciljnom jeziku za koji je dostupno malo materijala značajno poboljša zahvaljujući jeziku za koji je dostupno mnogo materijala, u odnosu na slučaj kada bi mreža prozodiju učila samo na osnovu dostupnog materijala za jedan od ta dva jezika. U slučaju da se koriste jezici za koje je prozodijska anotacija rađena na različite načine, postoji nekoliko načina na koje se te informacije mogu predstaviti mreži na ulazu. Najjednostavnija jeste da se prozodijska obeležja svih jezika spoje u jedan vektor i tako pošalju u skrivene slojeve. U tom pristupu, u zavisnosti koji je jezik u pitanju, deo vektora koji se odnosi na prozodijska obeležja tog jezika sadržao bi 0 i 1 (shodno vrednostima obeležja), dok bi deo vektora koji se odnosi na prozodiju drugih jezika obavezno sadržao sve nule. Pretpostavka je da bi skriveni slojevi mogli da ekstrahuju željenu informaciju, te bi znanja iz različitih jezika o prozodiji mogla da posluže kao deljena infromacija. Ipak, na ovaj način, veliki broj jezika bi doveo do veoma velikog ulaznog vektora (ako neki od jezika nemaju ista prozodijska obeležja). Pretpostavka je da bi formiranje prozodijskog *embeddinga* moglo da pomogne u boljoj generalizaciji i omogući kvalitetniju sintezu u kombinaciji govornik-jezik koja nije viđena u obuci. Međutim, kako se za prozodiju ne koristi *one-hot* vektor, nije očekivano da



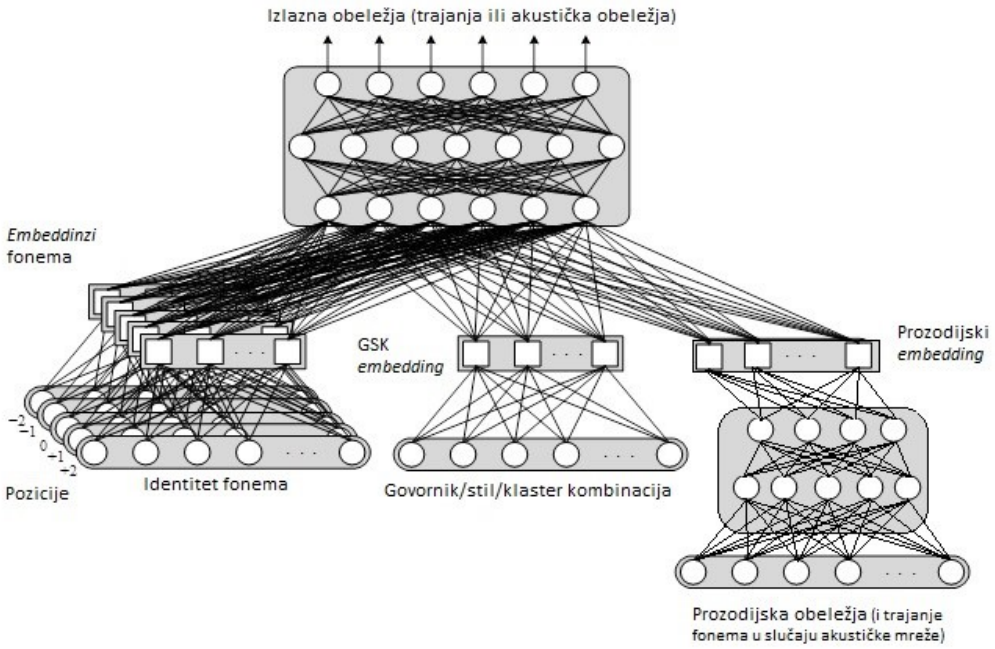
Slika 5.1 Arhitektura predloženog modela za višejezični ekspresivni TTS

jedan *embedding* sloj može da ekstrahuje informaciju o prozodiji. Iz tog razloga je, u slučaju različitih pristupa prozodijske anotacije za različite jezike, ideja da postoji nekoliko slojeva, praktično podmreža, koje bi ekstrahovale informacije o prozodiji bez obzira u kom je formatu ona inicijalno data. Šematski prikaz ovakvog sistema dat je na Slici 5.2.

5.1 Višejezični model sa ujednačenom prozodijskom anotacijom

5.1.1 Model sa dva jezika

Prvi eksperiment višejezičnog sintetizatora urađen je za američki engleski i meksički španski. Baze za pomenuta dva jezika se značajno razlikuju po veličini i strukturi govornika. Engleska baza sadrži snimke 23 različita govornika, ukupno skoro 26 sati snimljenog materijala. Španska baza sadrži 111 različitih govornika, ukupno nešto više od 5 sati snimljenog materijala. Osim navedene značajne razlike u broju govornika i trajanju baze, ove



Slika 5.2 Arhitektura predloženog modela za višejezični ekspresivni TTS sa prozodijskim *embeddingom*

baze se razlikuju i po kvalitetu. Većina materijala na engleskom jeziku (blizu 23 sata) snimljena je u profesionalnom studiju, dok je samo nešto više od polovine baze na španskom snimljeno u profesionalnom studiju. Preostali materijal prikupljen je sa interneta, iz audio knjiga, javnih govora, i sl., te predstavlja materijal lošijeg kvaliteta. Detaljnije informacije o bazama prikazane su u tabeli 5-1. Sav materijal je odabiran frekvencijom 22.05kHz i kodovan sa 16 bita po odbirku.

Implementiran je višejezični model sa više govornika i stilova baziran na *embeddingu* (engl. *Cross-lingual Text-To-Speech model based on Embedding* – CTTSE), odnosno dve neuronske mreže, jedna za predviđanje trajanja fonema, a druga za predviđanje akustičkih obeležja. Model za trajanja sastoji se od ulaznog sloja dimenzije 908, četiri skrivena sloja dimenzije 1024 i izlaznog sloja dimenzije 5. Dimenzija ulaznog sloja određena je brojem ulaznih obeležja, a to su: *one-hot* vektor dužine 205 za identifikaciju GSK, *one-hot* vektor dužine 137 za identifikaciju trenutnog fonema (pri čemu fonem može imati dobru i oštećenu verziju), 4 puta po *one-hot* vektor dužine 70 za identifikaciju fonema na pozicijama ±1 i ±2 u odnosu na trenutni fonem, i 286 binarnih prozodijskih obeležja. Veličina izlaznog sloja definisana je brojem izlaznih obeležja, što su u ovom slučaju pojedinačna trajanja svakog od 5 HMM stanja kojima je fonem modelovan. Svaki *one-hot* vektor se dovodi na sopstveni

Tabela 5-1 Govorne baze korišćene u inicijalnom eksperimentu

			Američki engleski	Meksički španski
Broj govornika	Ženski		10	54
	Muški		13	57
	Ukupno		**23**	**111**
Govornik/stil/klaster kombinacija (GSK)	Ukupan broj		82	123
	Trajanje	Min.	0:01:10	0:00:50
		Maks.	0:59:59	0:28:49
		Sr.vr.	0:18:58	0:02:32
		Medijana	0:13:48	0:01:13
Trajanje	Studijski kvalitet		22:47:39	2:48:05
	Slabiji kvalitet		3:07:04	2:24:19
	Ukupno		**25:54:43**	**5:12:24**

embedding sloj (slika 5.2), gde je *embedding* za GSK veličine 15, *embedding* za identitet trenutnog fonema je veličine 10, dok su preostala četiri, svaki veličine 5. Izlazi *embedding* slojeva se nadovezuju (konkateniraju) međusobno i sa prozodijskim obeležjima, te se prosleđuju u prvi skriveni sloj. Arhitektura mreže za predviđanje akustičkih obeležja je veoma slična, samo se na ulaz dovodi još 9 obeležja koja bliže određuju poziciju trenutnog vremenskog prozora (frejma) u fonemu i njegovo trajanje, tako da je ulazni sloj veličine 917, a izlazni sloj predstavlja broj akustičkih obeležja koja se predviđaju, i njegova veličina je 130. Tih 130 akustičkih obeležja su: osnovna frekvencija, 2 koeficijenta za modelovanje aperiodičnosti i 40 mel-frekvencijskih kepstralnih koeficijenata za modelovanje spektralne obvojnice, prvi i drugi izvodi svih pomenutih obeležja i 1 obeležje koje definiše da li je frejm zvučni ili bezvučni. U obe mreže, prva 3 skrivena sloja sadrže obične neurone (*feedforward*) sa tangens hiperboličnom aktivacionom funkcijom, dok poslednji skriveni sloj sadrži LSTM neurone i istu aktivacionu funkciju. Pronalazak optimalnih parametara mreže vrši se algoritmom opadanja gradijenta kroz *batch*-eve veličine 8x50 u slučaju mreže za trajanja i 4x400 u slučaju mreže za akustiku. Format *batch*-a „n x m“ znači da se jedan batch sastoji od n tokova (nizova rečenica) sa m uzoraka po toku, gde su uzorci u slučaju mreže za trajanja fonemi, a u slučaju mreže za akustiku, frejmovi trajanja 5 ms. Obuka modela za predviđanje trajanja trajala je 100 epoha, a modela za predviđanje akustičkih obeležja 45 epoha. Za optimizaciju je korišćen stohastički gradijent opadanja, a početna brzina učenja postavljena je na 0,008 za model za predviđanje trajanja i 0,01 za model za predviđanje akustičkih obeležja. Brzina učenja je tokom obuke smanjivana ukoliko je detektovan porast vrednosti L_{ukupno}, kombinacije vrednosi funkcije cene na validacionom $L_{\text{validacioni}}$ i skupu za obuku L_{obuka}, prema izrazu

$$L_{\text{ukupno}} = 0{,}2 \cdot L_{\text{validacioni}} + 0{,}8 \cdot L_{\text{obuka}}, \tag{5.1}$$

dok je funkcija cene srednja kvadratna greška, ali su u obzir uzete težine za svaki uzorak kako bi se donekle ujednačio doprinos manje i više zastupljenih GSK u bazi, prema

$$J(\boldsymbol{\theta})_b = \frac{1}{N_b} \sum_{j=1}^{N_b} w_j \sum_{i=1}^{N_{out}} \left(y_{ij} - t_{ij}\right)^2, \tag{5.2}$$

gde je N_b broj uzoraka u *batch*-u, N_{out} veličina izlaznog sloja mreže, w_j težinski koeficijent koji odgovara GSK iz j-tog uzorka iz *batch*-a, a y_{ij} i t_{ij} su vrednost koju predviđa model i ciljna vrednost i-tog izlaza iz mreže za j-ti uzorak iz *batch*-a, respektivno. Težinski koeficijent w_k za k-tog GSK računa se prema izrazu

$$w_k = \alpha\sqrt{N_k}\,, \tag{5.3}$$

gde je N_k ukupan broj rečenica iz baze od k-tog GSK, a α je normalizacioni faktor dat sa

$$\alpha = \sum_{k=1}^{N_{SSC}} \sqrt{N_k}, \tag{5.4}$$

gde je N_{SSC} ukupan broj različitih GSK u bazi. Vršena je normalizacija ulaznih obeležja na opseg (0,1) i standardizacija izlaznih obeležja. Standardizacija je vršena po govorniku. Generisanje govornog signala vršeno je na osnovu predviđenih akustičkih obeležja korišćenjem WORLD vokodera.

Sprovedena su tri eksperimenta slušanja. Ideja prvog eksperimenta je da uporedi kvalitet sinteze jednojezičnog (engl. *single language* – SL) i višejezičnog (engl. *multilanguage* – ML) sintetizatora kada se sintetizuju rečenice koje postoje u bazi, a koje nisu bile korišćene tokom obuke. S obzirom da postoje originalne rečenice, odnosno tačno je poznato kako sinteza treba da zvuči, moguće je bilo sprovesti i objektivnu i subjektivnu evaluaciju rezultata. Objektivna evaluacija podrazumeva računanje odstupanja predviđenih akustičkih obeležja i trajanja u odnosu na akustička obeležja i trajanja izdvojena iz originalnih snimaka. U pitanju su sledeće mere: srednja kvadratna greška između originalnih i predviđenih mel-frekvencijskih kepstralnih koeficijenata po frejmu (engl. *Mel Cepstral Distance* – MCD), koren srednje kvadratne greške (engl. *Root Mean Square Error* - RMSE) i korelacija pravih i predviđenih vrednosti osnovne frekvencije (RMSE F0 i CORR F0) po frejmu kao i RMSE i korelacija pravih i predviđenih trajanja po fonemu. U tabeli 5-2 mogu se videti dobijeni rezultati.

Uočava se da su mere vrlo bliske, odnosno da uspešnost predviđanja akustičkih obeležja nije mnogo zavisila od toga da li je model obučavan samo na jednom ili na više jezika. Ipak uočava se blaga prednost jednojezičnog modela u slučaju engleskog jezika za koji je dostupna veća i kvalitetnija govorna baza. Kako se ova prednost ne uočava u slučaju španskog jezika,

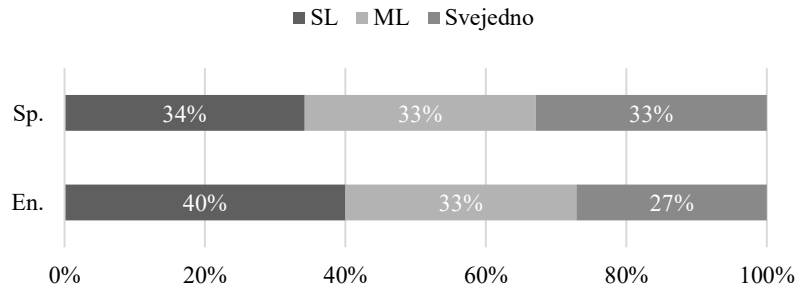
čija je govorna baza manja, može se pretpostaviti da bi novi jezik sa manjom govornom bazom mogao da ima bolje rezultate u okviru višejezičnog modela nego samostalno preuzimajući neka

Tabela 5-2 Objektivne mere odstupanja sintetizovanog od prirodnog govora u inicijalnom eksperimentu

		MCD [dB]	RMSE F0 [Hz]	CORR F0	RMSE DUR [ms]	CORR DUR
engleski	SL	5,26	32,30	0,90	5,79	0,84
	ML	5,39	33,34	0,89	5,58	0,85
španski	SL	5,29	24,39	0,91	5,68	0,77
	ML	5,19	24,39	0,91	5,61	0,78

znanja od drugih jezika. Međutim, u ovom eksperimentu ovakva tvrdnja nije potvrđena, ali ni opovrgnuta.

Kako se u oblasti sinteze govora i dalje smatra da su subjektivni testovi mnogo pouzdaniji od objektivnih za evaluaciju kvaliteta sintetizovanog govora, sproveden je test slušanja. Učestvovao je 31 ispitanik koji je potvrdio da razume obe jezika. Svaki ispitanik imao je 20 zadataka (10 po jeziku), a u svakom zadatku po 2 rečenice istog sadržaja – jedna sintetizovana SL sintetizatorom, a druga ML sintetizatorom. Ispitanik je trebalo da odabere koja rečenica mu zvuči bolje u pogledu razumljivosti i prirodnosti, ili da odabere odgovor da nema preferenciju između ponuđenih rečenica. Rečenice oba jezika su predstavljene sa po 5 govornika, 3 ženska i 2 muška. Rezultati su grafički prikazani na Slici 5.3. Ispostavlja se da su rezultati u skladu sa zaključcima dobijenim na osnovu objektivnih mera, odnosno, da sintetizovane rečenice oba sintetizatora zvuče približno isto. Ipak, uočava se određena prednost za rečenice na engleskom sintetizovane ML sintetizatorom. Kako je u radu [Sečujski, 2020] pokazano da je kvalitet SL sintetizatora, kakav je i ovde implementiran, izuzetno visok, može se smatrati da i ML

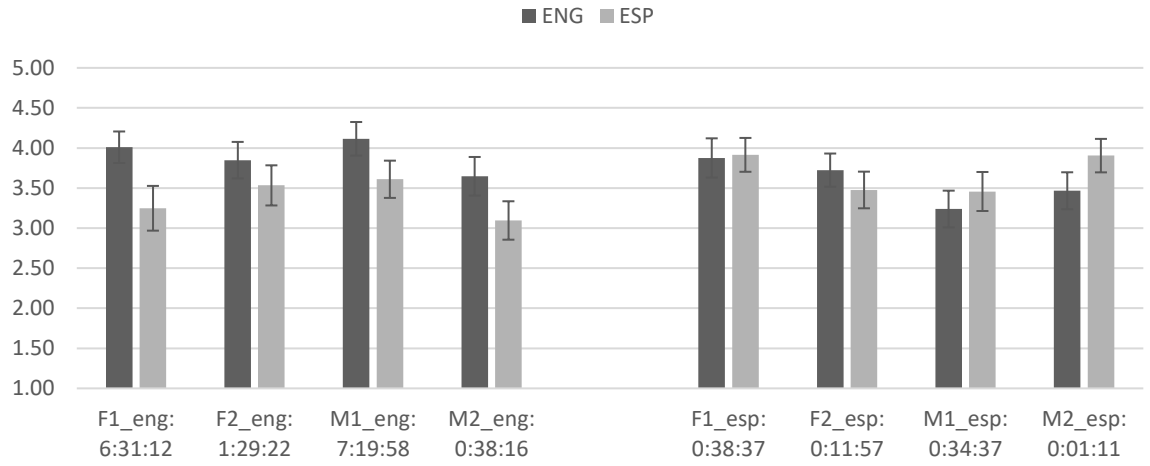


Slika 5.3 Rezultati subjektivnog poređenja kvaliteta sintetizovanog govora dobijenog sa SL i ML

sintetizator daje visok kvalitet sintetizovanog govora u slučaju kombinacije govornik-jezik koja je postojala u skupu za obuku.

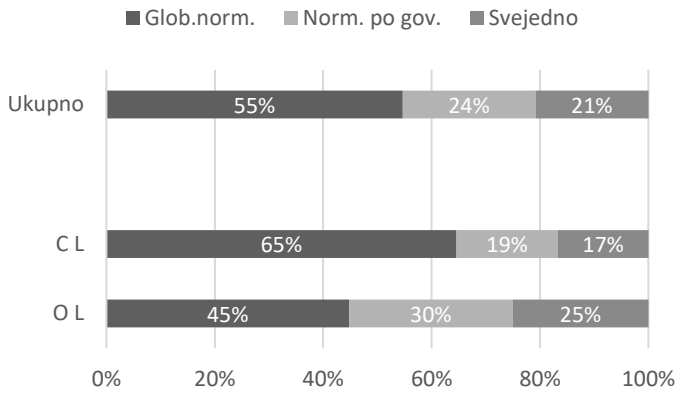
Međutim, interesantno je proveriti kvalitet sintetizovanog govora u tzv. *cross-lingual* (CL) scenariju, odnosno u kombinaciji govornik-jezik koja ne postoji u skupu za obuku. Kako za takve kombinacije govornik-jezik ne postoje originalne rečenice, za proveru kvaliteta sintetizovanog govora osmišljen je i sproveden sledeći subjektivni test. Formirane su 2 grupe pitanja, jedna se tiče engleskog jezika, a druga španskog. U pitanju su testovi, gde se od ispitanika traži da ocenom od 1 do 5 oceni kvalitet sintetizovane rečenice u pogledu razumljivosti i prirodnosti. Svako od 10 pitanja po jeziku sadržalo je po 4 rečenice identičnog sadržaja, ali su dve bile sintetizovane glasovima govornika kojima je to maternji jezik (muški i ženski glas, kobinacija govornik-jezik koja postoji u bazi za obuku), a druge dve su sintetizovane glasovima govornika koji ne govore taj jezik (muški i ženski glas, kombinacija govornik-jezik koja ne postoji u bazi za obuku – CL scenario). Ideja je da se proveri da li će i kolika razlika postojati između sinteze u originalnom i u CL scenariju. U ovom testu slušanja učestvovali su samo ispitanici kojima je jezik koji se koristi u testu maternji ili tvrde da jezik razumeju na izuzetno visokom nivou. Učestvovao je po 21 ispitanik za svaki od jezika. Razlog leži u pretpostavci da neke sitnije razlike slušalac koji ne razume jezik na izuzetno visokom nivou može da previdi.

Rezultati su prikazani grafički na Slici 5.4. Svaki stubić predstavlja jednog govornika, sa F su obeleženi ženski, a sa M muški govornici. Za svakog govornika dat je i podatak o trajanju materijala dostupnog u bazi za obuku modela. Sa leve strane se nalaze engleski govornici (oni koji u bazi za obuku govore engleski), a sa leve španski govornici (oni koji u bazi za obuku govore španski). Tamno sivim stubićima predstavljene su prosečne ocene kada je sinteza na engleskom jeziku, a svetlo sivim kada je sinteza na španskom. Uočava se da je sinteza kombinacija govornik-jezik koja postoji u bazi za obuku gotovo uvek bolja u odnosu na sintezu u kombinaciji govornik-jezik koja ne postoji u bazi za obuku, ali u proseku za svega 0,3. Može se primetiti i da svi engleski govornici za sintezu na engleskom jeziku imaju ocene preko 3,5, dok 2 španska govornika za sintezu na španskom imaju ocene nešto ispod 3,5, što se može objasniti manjim i/ili manje kvalitetnim bazama za obuku za te govornike. Međutim, interesantno je primetiti da španski govornik čija je baza svega nešto veća od 1 minut, ima



Slika 5.4 Rezultati subjektivnog poređenja kvaliteta sinteze u originalnom i *cross-lingual* scenariju po govorniku (sufiksi 'eng' i 'esp' označavaju originalni jezik govornika)

ocenu 0,65 manju u poređenju sa engleskim govornikom čija je baza za obuku preko 7 h, za sintezu u originalnom jeziku. Za španske govornike sinteza na engleskom jeziku je ocenjena do 0,45 manjom ocenom u odnosu na sintezu na španskom, a za engleske govornike je sinteza na španskom ocenjna bar za 0,5 manjom ocenom u odnosu na sintezu na engleskom jeziku. Definitivno se može zaključiti da postoji čujna razlika između sinteze u originalnom i CL scenariju, i iako slušaoci nisu bili svesni kada je koji scenario u pitanju, sigurno je ovoj razlici doprinela neprirodna brzina govora. Naime, uočeno je da španski govornici značajno brže govore engleski, odnosno da engleski govornici prilično usporeno govore španski. Razlog je u normalizaciji izlaznih obeležja mreže po govorniku. Postoji nekoliko načina na koji se ovaj problem može prevezići. U literaturi se pominje ideja da se mreži ne daje informacija koji je jezik u pitanju, ali ovakav pristup bi verovatno uneo druge probleme. Stoga je pokušano sa globalnom normalizacijom izlaznih obeležja. Sproveden je test slušanja u kome je učestvovalo 12 slušalaca. U testu je dato 16 parova rečenica. Od slušalaca je traženo da odaberu koja rečenica iz para im zvuči prirodnije (brzina, izgovor, kvalitet). U svakom paru jedna rečenica je sintetizovana TTS-om koji koristi model za predviđanje trajanja sa normalizacijom po govorniku, a druga sa TTS-om koji koristi model za predviđanje trajanja sa globalnom normalizacijom. Rezultati testa prikazani su grafički na Slici 5.5. Jasno se vidi da su rečenice modela sa globalnom normalizacijom češće birane kao bolje, a to je značajnije izraženo u CL



Slika 5.5 Rezultati subjektivnog poređenja kvaliteta sinteze modelima sa različitom normalizacijom u CL i OL scenariju

scenario u odnosu na OL scenario (scenario u kom se sintetizuje originalna kombinacija govornik-jezik, odnosno ona koja postoji u bazi za obuku). Objektivne mere, prikazane u tabeli 5-3, date za test skup, pokazuju blagu prednost modela koji koristi globalnu normalizaciju za RMSE, dok su korelacije iste za oba modela. Postoji još jedna mogućnost za prevazilaženje pomenutog problema, a to je postavljanje određenog parametra za jezik koji će modifikovati normalizacione koeficijente. Taj parametar utvrđuje se heuristički i predstavlja koeficijent kojim se množi srednja vrednost izlaza mreže za predviđanje trajanja (trajanje HMM stanja), odnosno normalizacioni faktor koji je karakterističan za govornika (dobijen tokom obuke modela), a koji se koristi pri sintezi. Na taj način postiže se ubrzavanje ili usporavanje sintetizovanog govora bez uticaja na varijansu karakterističnu za govornika, čime će se očuvati njegova karakteristična dinamika, kao i raspodela trajanja stanja koje je mreža predvidela. Npr. da bi španski govornici sporije pričali engleski, sinteza engleskih rečenica glasovima španskih govornika vrši se sa parametrom 1,2, dok se koristi parametar 0,9 kako bi engleski govornici brže pričali španski. I ova metoda daje zadovoljavajuće rezultate iako nije sproveden test slušanja na nezavisnim slušaocima.

Tabela 5-3 Objektivne mere za modele trajanja sa različitom normalizacijom izlaza

	RMSE [ms]	CORR
Norm. po gov.	5.39	0.85
Globalna norm.	5.36	0.85

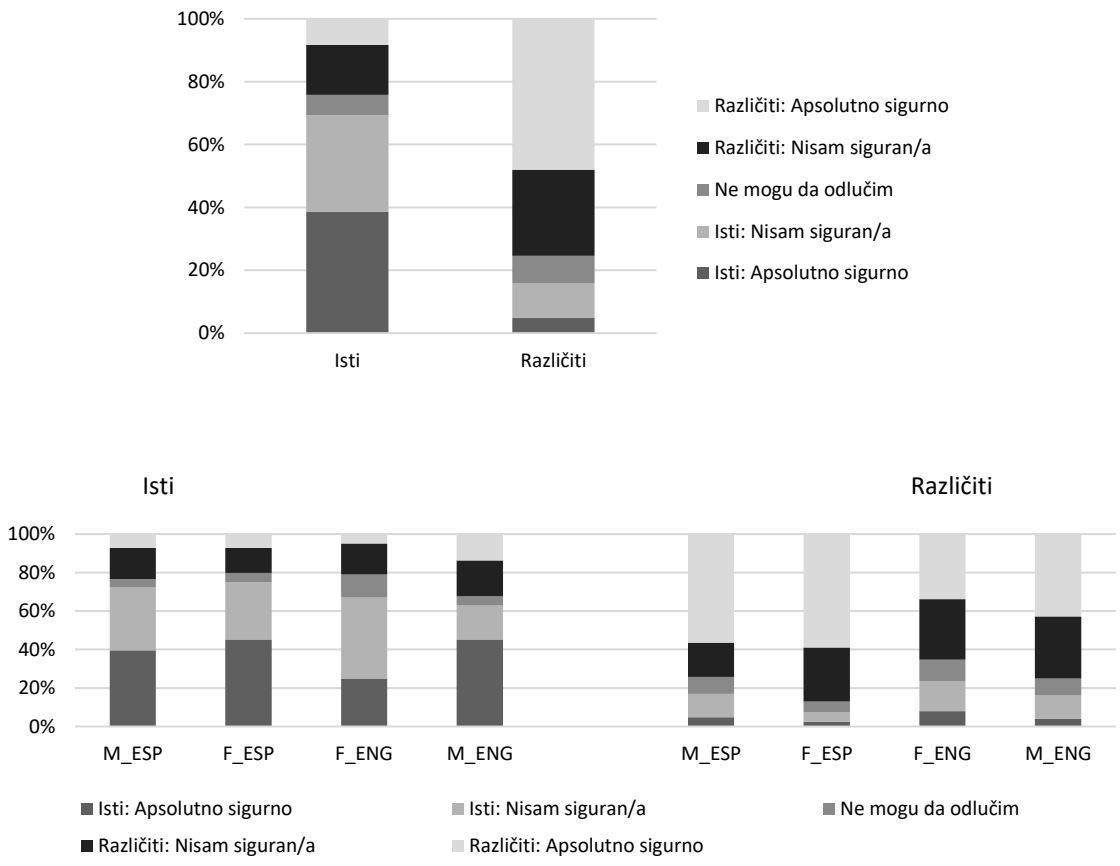
Konačno, trebalo je proveriti da li se i u kojoj meri očuva boja glasa kada se vrši sinteza u CL senariju. Za ovu proveru osmišljen je sledeći test slušanja. Dato je 32 zadatka, svaki sa po dve rečenice – jedna sintetizovana na engleskom, a druga na španskom. U 16 zadataka obe rečenice su sintetizovane glasom istog govornika (jedna u originalnom, druga u CL scenariju), dok su u preostalih 16 zadataka rečenice sintetizovane različitim, ali sličnim govornicima (sličnim u smislu pola i subjektivne percepcije autora). Zadatak ispitanika bio je da za svaki par rečenica na skali od 1 do 5 oceni koliko je siguran da je obe rečenice izgovorio isti govornik, prema sledećem uputstvu:

1 – Siguran sam da su u pitanju različiti govornici;
2 – Mislim da su u pitanju različiti govornici;
3 – Ne znam da li je obe rečenice izgovorio isti govornik;
4 – Mislim da je obe rečenice izgovorio isti govornik;
5 – Siguran sam da je obe rečenice izgovorio isti govornik.

Ovaj zadatak može se smatrati prilično teškim jer doneti odluku o tome da li je dve različite rečenice izgovorio isti govornik, posebno kada su te dve rečenice na različitim jezicima, nije jednostavno. Šta više, često ljudi koji govore više jezika donekle menjaju svoj glas u zavisnosti od toga kojim jezikom govore. U testu je bilo 4 ciljna govornika po jeziku, a za svakog po 4 pitanja, od kojih su 2 bila sa obe rečenice ciljnog govornika, a 2 koja su sadržala jednu rečenicu ciljnog govornika i drugu od nekog govornika iz originalnog modela koji je sadržao mnoštvo govornika. U testu je učestvovao 31 ispitanik koji je potvrdio da razume oba jezika.

Ako bi se ocene 5 i 4 smatrale tačnim u slučaju parova rečenica izgovorenih od strane istog govornika, a 1 i 2 u slučaju rečenica koje su izgovorene od strane različitih govornika, moglo bi se konstatovati da su slušaoci odgovorili tačno u čak 72% slučajeva, dok su u 8% slučajeva bili neodlučni. Detaljnija analiza rezultata eksperimenta grafički je prikazana na Slici 5.6.

Kada su u pitanju parovi rečenica izgovoreni od strane istog govornika, ispitanici su sa visokom sigurnošću (ocena 5) tačno odgovorili u skoro 40% slučajeva. Gledano po ciljnom govorniku, ovo je variralo između 25% i 45%. Ipak, tačno su odgovorili (ocena 5 ili 4) u oko 70% slučajeva, odnosno gledano po govorniku između 65% i 75%. S druge strane, kada su u pitanju rečenice izgovorene od strane različitih govornika, ispitanici su sa visokom sigurnošću (ocena 1) tačno odgovorili u preko 50% slučajeva. Po govorniku, ovaj procenat je varirao



Slika 5.6 Rezultati evaluacije sličnosti glasova u CL scenariju: (gore) ukupni; (dole) za svakog ciljnog govornika ponaosob. Natpisi *'Isti'* i *'Različiti'* označavaju da li su obe rečenice u paru bile izgovorene od strane istog govornika.

između 35 i 60. Ipak, tačno su odgovorili (ocena 1 ili 2) u preko 75% slučajeva, odnosno po ciljnom govorniku u 65% do 90% slučajeva. Potpuno pogrešne odgovore (1 u slučaju kada je u pitanju isti govornik, odnosno 5 u slučaju kada nije) ispitanici su davali u manje od 10% slučajeva iz čega se može zaključiti da se u CL scenariju boja glasa jasno zadržala.

5.1.2 Modeli sa više od dva jezika

Formiran je model korišćenjem četiri jezika: meksički španski, evropski španski, američki engleski i britanski engleski. Ovde se otvara pitanje možemo li da smatramo da se ovde zaista radi o četiri jezika ili samo dva. Iako je velika verovatnoća da mreža može jako puno da nauči za evropski španski od meksičkog španskog i obrnuto (isto i za verzije engleskog), ipak se ovi jezici razlikuju po fonemima, a to je dovoljan razlog da ih razdvojimo kao zasebne jezike.

Naime, postoje jasno definisani fonemi u evropskom španskom koji ne postoje u meksičkom španskom, kao i u britanskom engleskom, a koji ne postoje u američkom engleskom. Čak i oni fonemi koji su jednaki za oba jezika, mogu imati neke minimalne razlike, zbog čega se mreži dopušta da korišćenjem fonetskog *embeddinga* potpuno samostalno uvidi sličnosti i razlike fonema ova četiri jezika, iako se oni na ulazu daju kao potpuno različiti (kodovani *one hot* vektorima). U američkom engleskom razlikujemo 15 fonema za samoglasnike, a u britanskom engleskom 19. I pored navedenog, za potrebe TTS dodatno razlikujemo nenaglašene i dva tipa naglašenih vokala (*secondary stress*) u engleskom. Kada je u pitanju evropski španski u odnosu na meksički španski, dodat je samo fonem /θ/ za izgovor Z i C kada se nađu ispred E ili I.

Sva četiri jezika prozodijski su anotirana na isti način (korišćenjem ToBI anotacije), stoga je sama arhitektura vrlo slična eksperimentu iz poglavlja 5.1.1. Promenjeni su neki hiperparametri koji nisu od ključnog značaja, ali se pokazalo da daju određene prednosti (bolje objektivne mere, brža obuka i dr.). Neki od važnijih hiperparametara koji su izmenjeni jeste korišćenje ReLU aktivacionih funkcija umesto tangens hiperboličkih u skrivenim slojevima, kao i smanjenje broja korišćenih MGC obeležja sa 40 na 32. Naravno, broj ulaza u mrežu se razlikuje jer odgovara broju fonema, što u ovom eksperimentu iznosi 280 (uključujući foneme i njihove oštećene verzije kao različite). Ostali bitni parametri poput broja slojeva, veličine *embeddinga* i broja i vrste obeležja ostali su nepromenjeni.

Baze za jezike korišćene u modelu čiji će rezultati biti analizirani u nastavku se značajno razlikuju po veličini i strukturi govornika. Baza američkog engleskog sadrži snimke 32 različita govornika, ukupno preko 28 h snimljenog materijala. Baza britanskog engleskog je značajno lošija po kvalitetu, a sadrži 26 govornika, ukupno približno 6,5 h materijala. Baza meksičkog španskog sadrži 103 različita govornika, ali ukupno samo oko 6 h snimljenog materijala, dok baza evropskog španskog sadrži samo 12 govornika, lošeg je kvaliteta i ima svega oko 1,5 h materijala. Dakle, postoje značajne razlike u broju govornika, trajanju baza, kao i kvalitetu. Većina materijala na engleskom jeziku (blizu 23 sata) snimljena je u profesionalnom studiju, dok je većina materijala za ostale jezike prikupljena sa interneta, iz audio knjiga, javnih govora, i sl., te često predstavlja materijal značajno lošijeg kvaliteta. Detaljnije informacije o bazama prikazane su u tabeli 5-4. Sav materijal je odabiran frekvencijom 22.05 kHz i kodovan sa 16 bita po odbirku.

Tabela 5-4 Govorne baze korišćene u eksperimentu sa 4 jezika

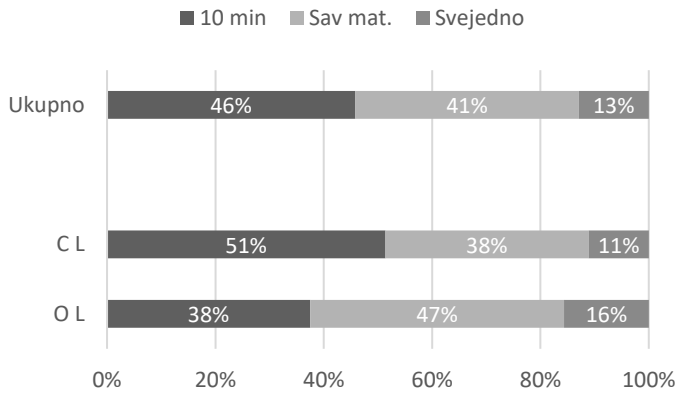
			Američki engleski	Britanski engleski	Meksički španski	Evropski španski
Broj govornika	Ženski		10	13	50	6
	Muški		22	13	53	6
	Ukupno		**32**	**26**	**103**	**12**
Govornik/stil/klaster kombinacija (GSK)	Ukupan broj		92	27	124	12
	Trajanje	Min.	0:01:43	0:04:51	0:01:11	0:02:02
		Maks.	1:30:11	0:53:52	1:44:24	0:14:13
	Ukupno		**28:13:45**	**6:29:45**	**5:54:39**	**1:32:36**

Kako je problem nedovoljne sličnosti sintetizovanog glasa sa originalom, posebno u CL scenariju, u ovom eksperimentu postao izraženiji, pokušano je sledeće. S obzirom da su i male količine materijala po govorniku mogle da daju dobre rezultate, i s obzirom da je dodavanje težinskih koeficijenata za govornike u funkciju cene imalo značajan doprinos kvalitetu sinteze, sada je pokušana drastičnija mera, odnosno ograničavanje količine materijala po govorniku, tačnije po kombinaciji govornik-stil, npr. na 10 minuta. Na taj način se izbegava da u CL sintezi glasovi počnu da liče na govornike čijeg materijala ima drastično više u odnosu na druge govornike. Iako i dalje postoje razlike u količini materijala po govorniku, nije više bilo potrebe za korišćenjem težinskih koeficijenata jer više nisu imali značajnijeg efekta.

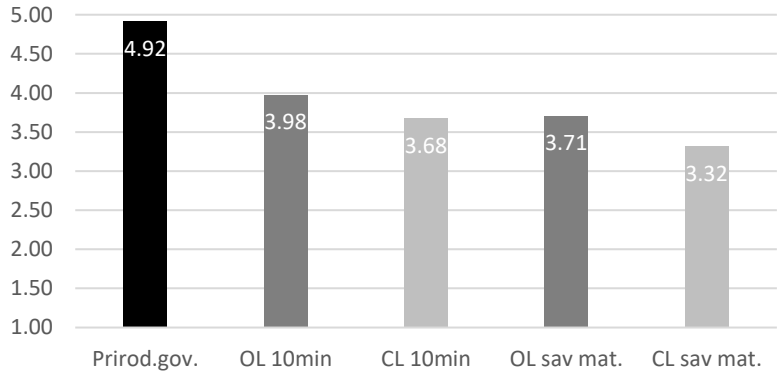
Kako bi se proverili rezultati, sprovedena su tri testa slušanja. U svakom testu je učestvovalo po 12 slušalaca. Prvi test predstavljao je test preferencije i u njemu je dato 20 parova rečenica. Svaki par sadržao je rečenicu na engleskom ili španskom, u CL scenariju ili u scenariju sa originalnim jezikom za govornika. Jedna rečenica u paru sintetizovana je modelom u kom je korišćen celokupan dostupan materijal po govorniku, a korišćeni su i težinski koeficijenti za funkciju cene, dok je druga rečenica u paru sintetizovana modelom u kom je korišćeno do 10 minuta materijala po govorniku, pri čemu nisu korišćeni težinski koeficijenti. Rečenice su sintetizovane glasovima dva govornika koji u bazi govore američki engleski (muško i žensko), kao i dva govornika koji u bazi govore meksički španski (muško i žensko). Po govorniku je sintetizovano dve rečenice na maternjem jeziku (tj. jeziku kojim govore u bazi za obuku), kao i tri rečenice u CL scenariju, za svaki od preostalih jezika po

jedna. Rezultati testa grafički su prikazani na Slici 5.7. Ne vidi se jasna preferencija između modela obučenog na do 10 minuta materijala po govorniku i modela obučenog na celokupnom dostupnom materijalu, generalno gledano. Međutim, ako se posmatra samo CL scenario, jasna je preferencija prema rečenicama sintetizovanim modelom sa ograničenom količinom materijala po govorniku, dok je u slučaju OL scenarija situacija obrnuta. Stoga je u narednom testu proverena konkretna ocena kvaliteta sinteze za različite modele u različitim scenarijima.

Drugi test slušanja sadržao je identične primere iz prethodnog testa, ali samo ženske glasove (radi skraćenja vremena koje slušaoci moraju da izdvoje za test) – ukupno 24 audio snimaka, jer je za svaku govornicu dodato po 2 snimka prirodnog govora iz baze za obuku. Od slušalaca je traženo da svaki od njih (24) ocene na skali od 1 do 5 u smislu kvaliteta sinteze (razumljivost, prirodnost, artefakti). Rezultati testa prikazani su grafički na Slici 5.8. Kao što je očekivano, originalni govor ocenjen je najvišom ocenom, u proseku blizu 5. U oba scenarija dobijene su više ocene za sintezu modelom sa do 10 minuta materijala po govorniku za obuku u odnosu na model obučen na celokupno dostupnom materijalu. Nešto je izraženija razlika u CL scenariju, 3,68 prema 3,32, u odnosu na OL scenario gde su ocene 3,98 prema 3,71. Ocene za CL scenario nisu mnogo niže od ocena za OL scenario. Objektivne mere su nešto bolje za eksperiment koji koristi celokupan dostupan materijal po govorniku, ali kako su one uprosečene po jezicima i govornicima i nisu uvek u saglasnosti sa subjektivnom ocenom, ovde neće biti detaljnije analizirane.



Slika 5.7 Rezultati subjektivnog poređenja kvaliteta sinteze modelima sa različitom količinom materijala za obuku po govorniku u CL i OL scenariju



Slika 5.8 Subjektivne ocene kvaliteta sinteze modelima sa različitom količinom materijala za obuku po govorniku (10 min**uta** i sav dostupan material) u CL i OL scenariju

Poslednji test slušanja sproveden sa višejezičnim modelom sa četiri jezika tiče se sličnosti glasova u sintezi sa glasovima originalnih govornika u CL scenariju. Iskorišćeni su svi CL primeri iz testa preferencije, malopre opisanog. Dakle, test sadrži 24 audio snimka za ocenjivanje, po tri rečenice od četiri različita govornika, sintetizovane modelom koji koristi celokupan dostupan materijal po govorniku i iste te rečenice sintetizovane modelom koji koristi do 10 minuta materijala po govorniku. Uz svaki od njih dat je referentni snimak, odnosno jedna rečenica iz baze, prirodni govor govornika čijim je glasom izvršena sinteza u CL scenariju. Od slušalaca je traženo da za svaki od 24 snimka ocenom od 1 do 5 iskažu u kojoj meri smatraju da su glasovi slični glasu u odgovarajućem referentnom snimku. Ovaj zadatak, kao što je u odeljku 5.1.1 već pomenuto, nije nimalo lak, jer je teško suditi da li se radi o istom govorniku čak i kada je sadržaj rečenice različit, a kamoli jezik. Osim toga, glas govornika varira i od raspoloženja i od zdravstvenog stanja, a neretko govornici i nesvesno izmene svoj glas kada pričaju strani jezik. Stoga je ovakvo poređenje prilično nezahvalno, ali nije pronađen bolji način za proveru ove karakteristike sintetizovanog govora. Rezultati pokazuju da se dobija nešto bolja ocena za sintezu modelom koji koristi do 10 minuta materijala za obuku po govorniku (3,1), u odnosu na sintezu modelom koji koristi sav dostupan materijal (2,9).

5.2 Višejezični model sa neujednačenom prozodijskom anotacijom

5.2.1 Model sa dva jezika

Inicijalni eksperiment za slučaj jezika koji su prozodijski anotirani drugačijim pravilima, te koriste različite prozodijske oznake, urađen je za dva jezika, srpskohrvatski i (američki) engleski. Kao što je opisano u poglavlju 5, postojalo je nekoliko mogućnosti. Jedna mogućnost je da se koristi model sa Slike 5.1, odnosno da se ne formira prozodijski *embedding* već da se prozodijska obeležja, različita za različite jezike, samo konkateniraju i proslede u skrivene slojeve. Druga mogućnost jeste da se formira podmreža, model sa Slike 5.2, koja će formirati prozodijski *embedding* odnosno ekstrahovati nekakva prozodijska obeležja i tako ih kompaktno proslediti u skrivene slojeve, konkatenirajući ih sa preostalim *embedding* slojevima u modelu (5 fonetskih i jedan govornički). Ovo je prva i osnovna stvar koju je bilo neophodno testirati kada je u pitanju višejezični model sa neujednačenom prozodijskom anotacijom.

Za formiranje modela korišćene su baza američkog engleskog koja sadrži 18 govornika, ukupno nešto više od 23 h materijala i srpskohrvatska baza koja sadrži 17 govornika, ukupno nešto više od 24 h materijala. Ova baza sadrži govornike srpskog, hrvatskog i crnogorskog govornog područja, tako da sadrži kako ekavicu i ijekavicu, tako i izražene varijacije u pogledu regionalnog dijalekta, ali je kod svih govornika stil neutralan, pri čemu je kod ponekih govornika zastupljen i srećan stil govora. Postojala je i mogućnost da se ova baza tretira kao baza sa 3 jezika koja imaju istu prozodijsku anotaciju, ali takav slučaj nije analiziran u okviru ove doktorske disertacije. Detaljnija struktura baze prikazana je u tabeli 5-5.

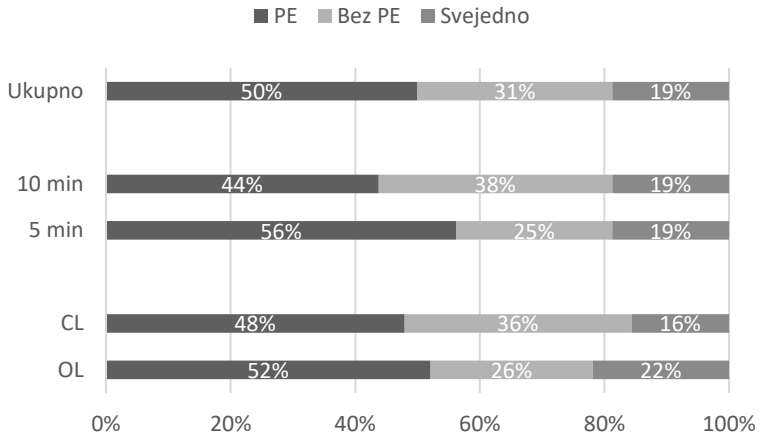
Pri obuci modela korišćeni su isti hiperparametri kao u eksperimentu sa četiri jezika (broj slojeva i neurona po sloju, broj i vrsta obeležja, veličine *embeddinga*, vrste neurona, i dr.). U modelu sa prozodijskim *embeddingom* korišćena je arhitektura podmreže od četiri nerekurzivna sloja sa ReLU aktivacionim funkcijama, a veličine slojeva su redom 512, 512, 128, 32. Veličina slojeva se utvrđuje heuristički. Formirani su modeli sa i bez prozodijskog *embeddinga* za čiju obuku se koristi do 10 minuta po govorniku, kao i modeli za čiju obuku se koristi do 5 minuta po govorniku (tačnije, po kombinaciji govornik-stil).

Tabela 5-5 Govorne baze korišćene u eksperimentu sa 2 jezika neujednačene prozodijske anotacije

			Američki engleski	Srpski
Broj govornika	Ženski		8	12
	Muški		10	5
	Ukupno		**18**	**17**
Govornik/stil/klaster kombinacija (GSK)	Ukupan broj		81	26
	Trajanje	Min.	0:01:43	0:01:38
		Maks.	1:30:11	3:20:03
	Ukupno		**23:12:08**	**24:15:34**

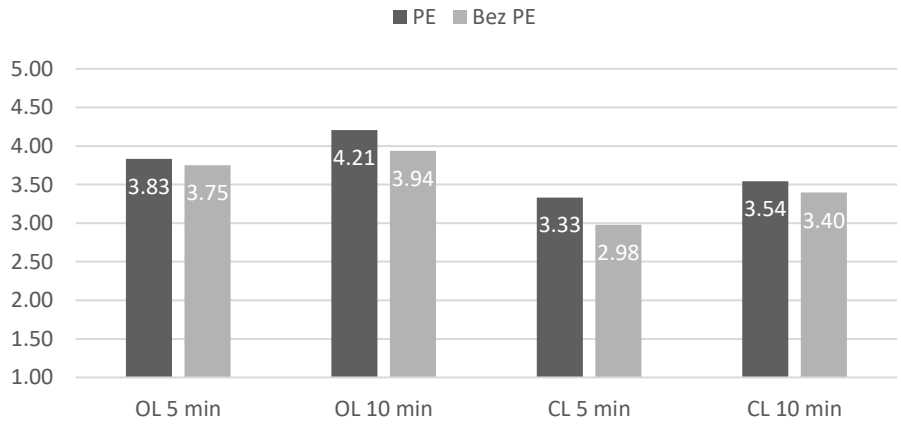
Sprovedena su dva testa slušanja u kojima je učestvovalo 15 slušalaca. Prvi je test preferencije, a dato je 16 parova rečenica. Svaki par sadrži rečenicu sintetizovanu modelom koji koristi prozodijski *embedding* i rečenicu sintetizovanu modelom koji ne koristi prozodijski *embedding*. Od slušalaca se traži da odaberu onu koja im zvuči bolje, odnosno prirodnije, razumljivije, sa manje artefakata. Polovina parova su rečenice na engleskom, a polovina na srpskom, odnosno polovina je sintetizovana modelom koji je koristio za obuku do 10 minuta materijala po govorniku, a polovina modelom koji je koristio do 5 minuta materijala po govorniku. Korišćene su rečenice od dva govornika koji u bazi govore srpski jezik (muško i žensko), kao i dva govornika koji u bazi govore engleski jezik (muško i žensko). Cilj ovog testa je da se otkrije da li je prozodijski *embedding* neophodan. Rezultati testa su grafički predstavljeni na Slici 5.9, i pokazuju da je u oba scenarija i za oba modela (obučeni različitim količinama materijala), preferirana varijanta modela koji koristi prozodijski *embedding*. Pretpostavlja se da bi u slučaju više od dva jezika, važnost formiranja prozodijskog *embeddinga* još više došla do izražaja.

Drugi test sadržao je identične primere kao i prvi test, ali je za svaki od 32 audio snimka od slušalaca traženo da ocene kvalitet sinteze na skali od 1 do 5. Ideja ovog testa je da proceni potrebu za prozodijskim *embeddingom*, ali i da ukaže na razliku u kvalitetu postignutom sa 5 i 10 minuta materijala po govorniku, kao i razliku u kvalitetu sinteze u kombinaciji govornik-



Slika 5.9 Rezultati subjektivnog poređenja kvaliteta sinteze modelima sa i bez prozodijskog *embeddinga*. Pored ukupnih rezultata, dati su i rezultati za modele obučene različitom količinom materijala za obuku po govorniku, kao i rezultati u slučajevima CL i OL scenarija.

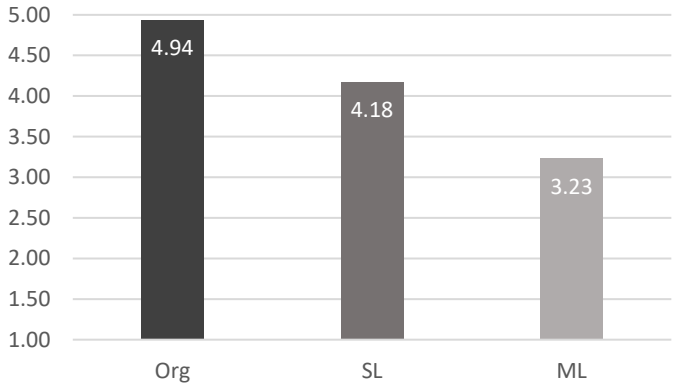
jezik koja je viđena u obuci u odnosu na kombinaciju koja nije viđena, tj. u odnosu na CL scenario. Rezultati su prikazani na Slici 5.10 i pokazuju da su u oba scenarija i za oba modela (obučeni različitim količinama materijala) bolje ocenjene rečenice sintetizovane modelom koji koristi prozodijski *embedding*, iako razlike često nisu velike. Osim toga, bolje su ocenjene rečenice sintetizovane modelom koji koristi do 10 minuta po govorniku za obuku, u odnosu na model koji koristi do 5 minuta materijala. I na kraju, nešto su više ocene dobijene u slučaju OL scenarija u odnosu na CL scenario, što je očekivano.



Slika 5.10 Subjektivne ocene kvaliteta sinteze modelima sa i bez prozodijskog *embeddinga*, modelima sa različitom količinom materijala za obuku po govorniku (10 min i 5 min) u CL i OL scenariju

Nakon što je testovima utvrđeno da model sa 10 minuta materijala po govorniku daje bolje rezultate, odnosno da je formiranje prozodijskog *embeddinga* korisno, na rezultatima takvog modela sprovedena su još dva testa slušanja. Cilj prvog je da uporedi rezultate koje produkuje višejezični model sa rezultatima jednojezičnih modela, kao što je to urađeno u poglavlju 5.1 za višejezični model sa ujednačenim prozodijskim anotacijama za sve jezike. U testovima je učestvovalo 12 slušalaca. U prvom testu je dato 20 rečenica, neke su sintetizovane sa jednojezičnim modelom, neke sa višejezičnim modelom, dok su neke originali (prirodan govor). Korišćen je glas četiri govornika, dva engleska i dva srpska, oba pola. Glasom svakog od govornika sintetizovano je po dve rečenice. Od slušalaca je traženo da svaku rečenicu ocene na skali od 1 do 5 po kvalitetu (prirodnost, razumljivost, artefakti). Rezultat testa je prikazan na Slici 5.11. Prirodan govor ocenjen je ocenom blizu 5, što je očekivano. Govor sintetizovan modelima obučenim na jednom jeziku ocenjen je prosečnom ocenom 4,18, dok je govor sintetizovan višejezičnim modelom ocenjen dosta lošije, 3,23. U slučaju višejezičnog modela koji je bio obučen na dva jezika koji su prozodijski anotirani na isti način, nije postojala jasna preferencija između tog modela i modela obučenih na jednom jeziku. Stoga je neophodno dodatno istražiti šta je u ovom slučaju problem, a jedna od mogućnosti je svakako eksperimentisanje veličinom mreže koja formira prozodijski *embedding*. Projekcija preko 250 prozodijskih obeležja po jeziku na svega 32 je možda previše gruba.

Cilj drugog testa bio je da proveri sličnost sintetizovanog glasa sa originalnim. U testu je dato osam sintetizovanih audio snimka. Rečenice su sintetizovane modelom koji ima



Slika 5.11 Subjektivne ocene kvaliteta prirodnog govora (Org) i sinteze jednojezičnim modelima (SL) i višejezičnim modelom (ML) sa prozodijskim *embeddingom*

prozodijski *embedding,* isključivo u CL scenariju, glasovima četiri govornika (dva po jeziku) i uz svaku je dat referentni snimak, odnosno prirodan govor govornika čijim je glasom vršena sinteza. Od slušalaca je traženo da na skali od 1 do 5 ocene koliko je boja glasa u sintetizovanoj rečenici slična glasu iz njoj referentne. Prosečna ocena iznosi 3,41. Sličan test sproveden je i za model sa 4 jezika koji su prozodijski anotirani na isti način, kada je ocena sličnosti bila nešto niža od ovde dobijene, te se postignuta sličnost sintetizovanog glasa sa originalnim može smatrati zadovoljavajućom.

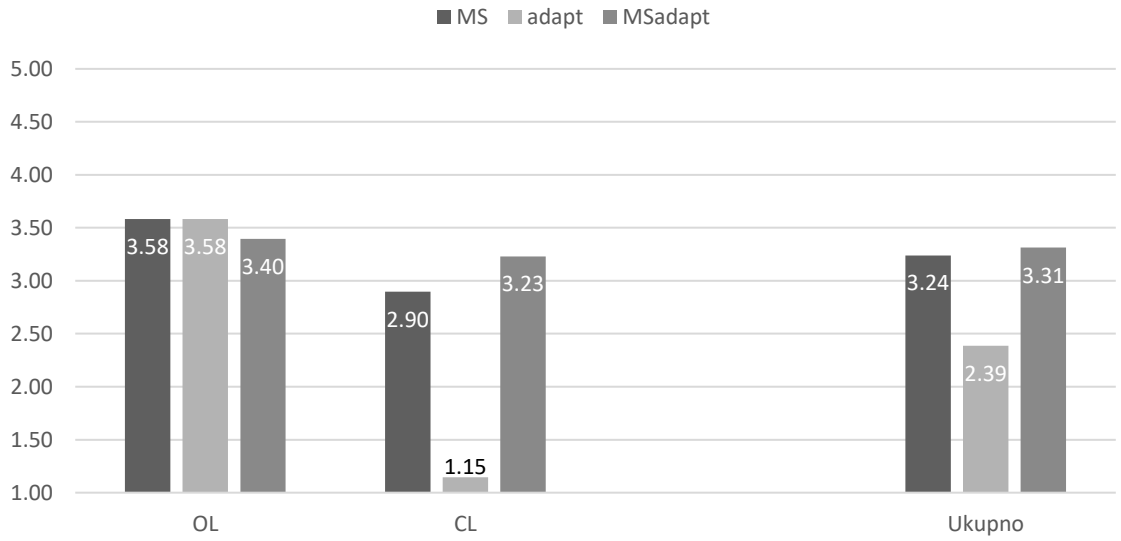
5.2.2 Sinteza novog glasa višejezičnim modelom

Postavlja se pitanje da li je za sintezu novim glasom, odnosno, glasom koji nije viđen pri obuci modela, neophodno formirati model od nule, uključivanjem novog govornika u višejezičnu bazu sa mnoštvom govornika. Ovakav pristup već je razmatran u radovima [Nosek, 2021] i [Sečujski, 2020] i utvrđeno je da je dovoljno da se formirani model adaptira na bazu novog govornika, što daje brže rezultate. Međutim, pokazalo se da CL scenario u tom slučaju za ciljnog govornika ne daje sjajne rezultate.

Prva ideja koja je značajno doprinela poboljšanju jeste zamrzavanje fonetskih *embeddinga*. Dakle, pri doobuci višejezičnog modela sa više govornika smatra se da su fonetski *embeddinzi* formirani na osnovu dovoljnog broja različitih glasova, te da bi izmena parametara koji se odnose na taj deo mreže tokom doobuke mogla biti štetna. Stoga se u prvoj iteraciji dozvoljava izmena samo parametara mreže koji se odnose na govornički *embedding*, odnosno traži se tačka u govorničkom *embedding* prostoru koja najbolje opisuje novog ciljnog govornika. U drugoj iteraciji doobuke, dozvoljavaju se promene parametara samo dela mreže koji se odnosi na prozodiju, kako bi se TTS prilagodio načinu govora ciljnog govornika. Osim navedenog, po ugledu na neke ideje iz literature, pokušana su dva pristupa. U jednom pristupu doobuka se vrši samo na materijal ciljnog govornika, dok se u drugom pristupu koristi po nekoliko govornika jezika kojim govori ciljni govornik i nekoliko govornika ciljnog jezika. Na ovaj način se očekuje da ne dođe do natprilagođenja mreže samo za kombinaciju ciljni govornik i njegov jezik iz baze, već da ostane moguće u visokom kvalitetu sintetizovati rečenicu glasom novog (ciljnog) govornika, na ma kom jeziku podržanom TTS modelom.

Za potrebe analize rezultata organizovana su dva testa slušanja. Cilj testova je da uporedi rezultate dobijene kada je model treniran od starta sa uključenim ciljnim govornikom (model nazvan *MS*), kada je model doobučen samo na bazu ciljnog govornika (model nazvan *adapt*) i kada je model doobučen na po tri govornika ciljnog jezika i jezika ciljnog govornika (model nazvan *MSadapt*). Za konkretan eksperiment, za obuku modela *MSadapt*, pored celokupnog materijala ciljnog govornika, iskorišćena su još dva govornika koji govore istim jezikom kao ciljni govornik i još tri govornika koji govore ciljnim jezikom. Količina materijala tih dodatnih pet govornika ograničena je na do 3 minuta po govorniku. Osim toga, svih pet govornika je odabrano tako da su istog pola kao ciljni govornik.

U testovima je učestvovalo 12 slušalaca. U prvom testu dato je po četiri rečenice od dva govornika sintetizovane svakim od pomenuta tri modela i po dva primera prirodnog govora tih govornika. Po dve rečenice date su u CL, a dve u OL scenariju, od oba govornika. Jedan govornik u bazi za obuku govori engleski, a drugi srpski. Dakle za 28 rečenica je traženo od slušalaca da na skali od 1 do 5 ocene kvalitet, u smislu prirodnosti, razumljivosti i artefakata. Rezultati su prikazani na Slici 5.12. Originalni govor ocenjen je visokom ocenom, 4,79 (nije prikazano na slici). Ukupno gledano, sinteza modelom *MS* i sinteza *MSadapt*, dobile su



Slika 5.12 Subjektivne ocene kvaliteta sinteze modelom koji od starta sadrži ciljnog govornika (MS), modelom adaptiranim samo na ciljnog govornika (adapt) i modelom adaptiranim na nekoliko govornika uključujući i ciljnog (MSadapt). Prikazani su rezultati posebno u CL i OL scenariju, kao i ukupni rezultati.

približno iste ocene, 3,24 i 3,31, respektivno. Značajno nižu ocenu dobila je sinteza modelom *adapt*, 2,39. Ovako niska ocena posledica je činjenice da ovaj model nije zadržao sposobnost sinteze u CL scenariju nakon adaptacije na ciljnog govornika, što potvrđuje izuzetno niska ocena od 1,15. Ovaj model u CL scenariju proizvodio je izuzetno izražene artefakte i produkovao potpuno nerazumljiv govor. S druge strane u OL scenariju davao je podjednako dobre rezultate kao *MS* model, prosečna ocena za oba je 3,58. Model *MSadapt* je u OL scenariju dobio prosečnu ocenu nešto nižu u odnosu na druga dva modela, 3,40, ali se u CL scenariju pokazao značajno boljim u odnosu na preostala dva modela, 3,23, u odnosu na 2,90 i 1,15. Može se zaključiti da je metoda koju koristi model *MSadapt* izuzetno dobra jer se do rezultata dolazi brže nego kompletno novom obukom kada treba uključiti novog govornika, kao što je slučaj sa MS modelom.

Drugi test je sproveden sa CL primerima iz prethodnog testa, ali nisu upotrebljene rečenice modela adaptirane samo na ciljnog govornika jer su artefakti u sintezi ogromni, što je i potvrđeno izuzetno niskim ocenama u prethodnom testu. U ovom testu od slušalaca je traženo da za 8 rečenica na skali od 1 do 5 ocene sličnost sintetizovanog glasa sa referentnim (prirodni govor ciljnog govornika iz baze za obuku). Nešto je bolje ocenjena sličnost sintetizovanog glasa sa originalnim u slučaju modela *MSadapt*, sa prosečnom ocenom 2,42, u odnosu na model *MS* sa prosečnom ocenom 2,23. Ova razlika nije velika, ali se mora primetiti da je ocena dosta niža u poređenju sa ocenom sličnosti glasova u ranije pomenutim testovima tog tipa (u slučaju modela sa dva jezika sa istom i različitom prozodijskom anotacijom). Međutim, tada postignute ocene preko 3, trebalo bi da su u skladu sa ovde postignutom ocenom za MS model (to je identičan model za koji je ovaj test već rađen u poglavlju 5.2.1). Stoga su ovako niske ocene možda samo posledica izbora drugačijih govornika, i/ili strožih slušalaca.

5.3 Mogućnost podešavanja stila u višejezičnom modelu

Model predstavljen u [Sečujski, 2020], a takođe i višejezični model predstavljen kao proširenje modela [Sečujski, 2020], sadrže *embedding* prostor govornika. Kako svaka tačka u tom prostoru predstavlja jedinstvenu kombinaciju govornik-stil-klaster, jasno je da postoji jedinstveni *embedding* vektor za neutralni stil jednog govornika, kao i za bilo koji drugi stil

istog govornika. Stoga je pretpostavka da će tačka u *embedding* prostoru između tačke koja predstavlja neutralni stil, i one koja predstavlja npr. radostan stil, predstavljati tačku koja odgovara manje izraženom radosnom stilu. U okviru [Vujović, 2020] urađeno je inicijalno istraživanje na temu mogućnosti podešavanja nivoa ekspresivnosti u sintetizovanom govoru. Ako je $\vec{a}$ *embedding* vektor za neutralni stil nekog govornika, a $\vec{b}$ *embedding* vektor za tog istog govornika za neki drugi stil govora, tada važi da se novi *embedding* vektor $\vec{c}$ sa slabije izraženim stilom može generisati kao:

$$\vec{c} = \alpha \cdot \vec{a} + \beta \cdot \vec{b} \tag{5.5}$$

gde su α i β težinski faktori stila govora. Važi da je $\alpha + \beta = 1$, te što je β veće, emocija je izraženija. U [Vujović, 2020] potvrđeno je da slušaoci mogu da uoče promene nivoa ekspresivnosti u govoru sintetizovanom opisanim pristupom. Ovakav pristup ima jasnu prednost u odnosu na opisane pristupe za kontrolu nivoa ekspresivnosti iz literature, utoliko što ne zahteva subjektivnu evaluaciju baze, sintetizovani nivoi ekspresivnosti ne zavise direktno od nivoa ekspresivnosti koji postoje u snimljenoj bazi, i može da radi i sa izuzetno malim bazama ekspresivnog govora. Međutim, najizraženiji nivo ekspresivnosti ovim pristupom ograničen je prosečnim nivoom ekspresivnosti snimljene baze, a ponašanje kada je $\alpha + \beta > 1$ pokazalo se nepredvidivim, odnosno dovodilo je do raznih artefakata. Pokušaj analize značenja pojedinih dimenzija *embedding* vektora nije doveo do jasnih zaključaka, te nije bilo moguće utvrditi pravila za promenu *embedding* vektora u cilju dobijanja određenog stila i željenog nivoa ekspresivnosti. Pretpostavka je da bi korišćenje baze koja sadrži veliki broj govornika i paralelan korpus u više stilova, moglo da dovede do jasnijih zaključaka. Takva baza je formirana u okviru projekta S-ADAPT [Delić, 2020], ali njena obrada i anotacija prevazilaze okvire ove doktorske disertacije, a do momenta pisanja nije bila spremna za upotrebu.

6. Zaključak

U okviru disertacije razmatran je sistem za sintezu govora na osnovu teksta koji se sastoji od tri modula. Prvi modul služi za generisanje lingvističkih obeležja na osnovu analize teksta. Drugi modul na osnovu dobijenih lingvističkih obeležja produkuje akustička obeležja, koristeći neuralne mreže. Poslednji modul koristi predviđena akustička obeležja za formiranje talasnog oblika govornog signala. Akcenat disertacije je na drugom modulu. Ovaj modul je uglavnom jezički nezavisan, a njegove varijacije u pogledu arhitekture i algoritama obuke omogućavaju promene govornika, stilova, i dr. Razvijeni su algoritmi koji omogućuju sintezu različitim glasovima, oponašajući različite govornike ili govorne stilove, a koristeći često male količine materijala ciljnog govornika i/ili stila. Ovakve ideje omogućile su široku primenu sintetizovanog govora, ali i izazvale potrebu za sintezom govora na različitim jezicima.

U radu je predstavljena ideja formiranja višejezičnog ekspresivnog modela sa više govornika. Ovakav model omogućava ne samo produkciju ekspresivnog govora glasom različitih govornika na njihovom originalnom jeziku, već i na drugim jezicima iz govorne baze. Dakle, moguće je sintetizovati govor u kombinaciji govornik-jezik koja nije viđena u skupu za obuku. Ovaj model zasniva se na kreiranju različitih *embedding* prostora, kako fonetskih (jedna tačka predstavlja jedan fonem) tako i govorničkih (jedna tačka predstavlja jednu kombinaciju govornik-stil-klaster), a po potrebi i prozodijskih (apstrahovanje različitih prozodijskih obeležja). Ovako kreirani *embedding* prostori omogućavaju opisani tzv. *cross-lingual* scenario. Ove mogućnosti predstavljaju važne novine u modelu sintetizatora govora. Ispitana je mogućnost korišćenja različitih jezika za koje su korišćene iste konvencije prozodijske anotacije, ali i jezika čije su prozodijske anotacije različite, što iziskuje formiranje prozodijskih *embedding* prostora.

Prednosti ovog modela ogledaju se u njegovim mogućnostima da znanja za jedan jezik iskoristi za neki drugi, zahvaljujući *embeddinzima*, što omogućuje korišćenje relativno malih baza za pojedine jezike u modelu. Takođe, model ne zahteva ni velike količine govornog materijala ciljnog govornika, a omogućava sintezu glasom ciljnog govornika na ma kom jeziku podržanom modelom.

Sprovedeni su različiti eksperimenti kojima je utvrđena optimalna arhitektura – veličine i neophodnost *embedding* slojeva, kao i eksperimenti kojima je utvrđen optimalan način adaptacije modela na novog govornika – neophodnost prisustva male količine materijala govornika svih jezika pri adaptaciji kako bi se održala mogućnost višejezične sinteze. Objektivne mere su korišćene kada je za to postojala mogućnost (kada pored sintetizovane rečenice postoji snimak identične rečenice izgovoren od strane govornika čiji je glas korišćen pri sintezi), a svaki eksperiment imao je i subjektivnu evaluaciju kroz test slušanja. Iako mogu biti veoma obimni, pa samim tim zahtevati značajan ljudski napor, testovi slušanja i dalje predstavljaju najpouzdaniji način evaluacije kvaliteta sintetizovanog govora. Visoke ocene potvrdile su kvalitet sintetizovanog govora koji predloženi model postiže, odnosno održavanje kvaliteta u pogledu prirodnosti i razumljivosti u ključnom scenariju sinteze – u kombinaciji govornik-jezik koja nije viđena pri obuci.

6.1 Pravci daljeg istraživanja

Postoji mnogo pravaca za dalje istraživanje i unapređivanje TTS koji nisu razmatrani u doktorskoj disertaciji, a neki od njih su:

- Unapređenje neuralnih vokodera i adaptacije na male baze novog govornika/stila.
- Automatska prozodijska anotacija snimljenog govornog materijala.
- Transplantacija stila sa jednog govornika na drugog uz podešavanje nivoa ekspresivnosti.

Neuralni vokoderi, zasnovani na obuci nauralnih mreža, nisu deteljnije analizirani u okviru disertacije, ali su u poslednjih nekoliko godina gotovo potpuno istisnuli determinističke vokodere, koji vrše sintezu na osnovu akustičkih parametara po principu izvor-filtar model. Iako zahtevaju obuku i značajnije resurse za produkciju sintetizovanog govora, kvalitet koji postižu daleko nadmašuje sintezu determinističkim vokoderima. Inicijalna istraživanja, [Sečujski, 2020] pokazuju da i se sjajni rezultati postižu i za CL scenario. Takođe, doobuke engleskih modela daju sjajne rezultate za srpski jezik [Suzić, 2022]. Ova istraživanja daju odličnu motivaciju, ali zahtevaju detaljnije istraživanje o obuci i upotrebi višejezičnih ili

nekoliko jednojezičnih adaptiranih modela neuralnih vokodera za potrebe realizacije višejezičnog TTS modela.

Deo procesa u stvaranju predloženog TTS koji je najsporiji i najproblematičniji, jeste prozodijska anotacija govornih baza za obuku. Jedan od pravaca daljih istraživanja mogao bi biti upravo prevazilaženje ovog problema automatizacijom prozodijske anotacije. Jedan korak ka ovome je upravo pravljenje prozodijskog *embeddinga*. Ovakav *embedding* mogao bi da otkrije nova prozodijska obeležja koja nisu intuitivna, nisu osmišljena od strane čoveka, ali bi se integracijom ovako obučenog sloja i neke vrste koder-dekoder arhitekture, moglo doći do automatske anotacije baza.

U okviru disertacije pomenut je i značaj transplantacije stila i podešavanja nivoa ekspresivnosti. Postoji i jasna ideja kako bi se mogao iskoristiti kreirani višejezični TTS, ali je neophodna baza sa velikim brojem govornika koji govore različite stilove, kako bi se mogla obučiti mreža za konverziju neutralnog *embeddinga* u *embedding* ciljnog stila. Takva mreža bi omogućila stvaranje *embeddinga* stila za ma kog govornika čiji je materijal dostupan samo u neutralnom stilu, a linearnom interpolacijom između neutralnog *embeddinga* i *embeddinga* stila mogao bi se dozirati i nivo ekspresivnosti.

Literatura

Ai, Y., Wu, H. C., & Ling, Z. H. (2018, April). SampleRNN-based neural vocoder for statistical parametric speech synthesis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5659-5663). IEEE.

An, S., Ling, Z., & Dai, L. (2017, December). Emotional statistical parametric speech synthesis using LSTM-RNNs. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1613-1616). IEEE.

Azizah, K., Adriani, M., & Jatmiko, W. (2020). Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, *8*, 179798-179812.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. Im S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing.*

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., & Chan, W. (2020). WaveGrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

Cho, H., Jung, W., Lee, J., & Woo, S. H. (2022). SANE-TTS: Stable And Natural End-to-End Multilingual Text-to-Speech. *arXiv preprint arXiv:2206.12132*.

Coker, C. H. (1976). A model of articulatory dynamics and control. Proceedings of the IEEE, 64(4), 452-460.

Delić, T., Suzić, S., Ostrogonac, S., Đurić, S., & Pekar, D. (2017). Multi-style Statistical Parametric TTS. *Zbornik radova konferencije Digitalna obrada govora i slike (DOGS)*, 5-8.

Delić, T., Sečujski, M., & Suzić, S. (2017). A review of Serbian parametric speech synthesis based on deep neural networks. *Telfor Journal*, *9*(1), 32-37.

Delić, T., Suzić, S., Sečujski, M., & Pekar, D. (2018, March). Rapid development of new TTS voices by neural network adaptation. In *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1-6). IEEE.

Delić, V., PI (2020-2022). Speaker/Style Adaptation for Digital Voice Assistants Based on Image Processing Methods, AI-SADAPT, #6524560. Project supported by the Science Fund of the Republic of Serbia. *www.ktios.ftn.uns.ac.rs/sadapt/SADAPT.html*

Donahue, J., Dieleman, S., Bińkowski, M., Elsen, E., & Simonyan, K. (2020). End-to-end

adversarial text-to-speech. arXiv preprint arXiv:2006.03575.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis* (Vol. 3). Springer Science & Business Media.

Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459-1462).

Fan, Y., Qian, Y., Xie, F. L., & Soong, F. K. (2014). TTS synthesis with bidirectional LSTM based recurrent neural networks. In Fifteenth annual conference of the international speech communication association.

Fan, Y., Qian, Y., Soong, F. K., & He, L. (2015, April). Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4475-4479). IEEE.

Fan, Y., Qian, Y., Soong, F. K., & He, L. (2016, March). Speaker and language factorization in DNN-based TTS synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5540-5544). IEEE.

Flanagan, J. L., & Golden, R. M. (1966). Phase vocoder. *Bell System Technical Journal*, *45*(9), 1493-1509.

Giles, C. L., & Maxwell, T. (1987). Learning, invariance, and generalization in high-order neural networks. *Applied optics*, *26*(23), 4972-4978.

Godjevac, S. (2005). Transcribing serbo-croatian intonation. *Prosodic typology: The phonology of intonation and phrasing*, 146-171.

Gold, B., & Rader, C. (1967). The channel vocoder. *IEEE Transactions on Audio and Electroacoustics*, *15*(4), 148-161.

Hamza, W., Eide, E., Bakis, R., Picheny, M., & Pitrelli, J. (2004). The IBM expressive speech synthesis system. In Eighth International Conference on Spoken Language Processing.

Himawan, I., Aryal, S., Ouyang, I., Kang, S., Lanchantin, P., & King, S. (2020, May). Speaker adaptation of a multilingual acoustic model for cross-language synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7629-7633). IEEE.

Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. *Advances in neural*

information processing systems, *15*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hojo, N., Ijima, Y., & Mizuno, H. (2016, September). An Investigation of DNN-Based Speech Synthesis Using Speaker Codes. In *INTERSPEECH* (pp. 2278-2282).

Hojo, N., Ijima, Y., & Mizuno, H. (2018). DNN-based speech synthesis using speaker codes. *IEICE TRANSACTIONS on Information and Systems*, *101*(2), 462-472.

Hu, Q., Richmond, K., Yamagishi, J., & Latorre, J. (2013). An experimental comparison of multiple vocoder types. In *Eighth ISCA Workshop on Speech Synthesis*.

Hunt, A. J., & Black, A. W. (1996, May). Unit selection in a concatenative speech synthesis system using a large speech database. In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (Vol. 1, pp. 373-376). IEEE.

Inoue, K., Hara, S., Abe, M., Hojo, N., & Ijima, Y. (2017, December). An investigation to transplant emotional expressions in DNN-based TTS synthesis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1253-1258). IEEE.

Jaeger, H. (2002). Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach.

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018, December). Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 266-273). IEEE.

Kaneko, T., & Kameoka, H. (2017). Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., ... & Kavukcuoglu, K. (2018, July). Efficient neural audio synthesis. In *International Conference on Machine Learning* (pp. 2410-2419). PMLR.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, *27*(6), 349-353.

Kim, M., Cheon, S. J., Choi, B. J., Kim, J. J., & Kim, N. S. (2021). Expressive text-to-speech using style tag. *arXiv preprint arXiv:2104.00436*.

Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial

learning for end-to-end text-to-speech. In *International Conference on Machine Learning* (pp. 5530-5540). PMLR.

King, S. (2010). A beginners' guide to statistical parametric speech synthesis. *The Centre for Speech Technology Research, University of Edinburgh, UK*.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. the Journal of the Acoustical Society of America, 67(3), 971-995.

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., ... & Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, *32*.

Kuo, F. Y., Aryal, S., Degottex, G., Kang, S., Lanchantin, P., & Ouyang, I. (2018, December). Data selection for improving naturalness of tts voices trained on small found corpuses. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 319-324). IEEE.

Li, M., Wu, Z., & Xie, L. (2016, September). On the impact of phoneme alignment in DNN-based speech synthesis. In *SSW* (pp. 196-201).

Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., & Montero, J. M. (2015). Emotion transplantation through adaptation in HMM-based speech synthesis. *Computer Speech & Language*, *34*(1), 292-307.

Lorenzo-Trueba, J., Henter, G. E., Takaki, S., Yamagishi, J., Morino, Y., & Ochiai, Y. (2018). Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Communication*, *99*, 135-143.

Luong, H. T., Takaki, S., Henter, G. E., & Yamagishi, J. (2017, March). Adapting and controlling DNN-based speech synthesis using input codes. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4905-4909). IEEE.

Ma, D., Su, Z., Wang, W., & Lu, Y. (2020, April). FPETS: Fully Parallel End-to-End Text-to-Speech System. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 8457-8463).

Miyanaga, K., Masuko, T., & Kobayashi, T. (2004, October). A style control technique for HMM-based speech synthesis. In *Proc. ICSLP* (Vol. 4).

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, *99*(7), 1877-1884.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech communication, 9(5-6), 453-467.

Nachmani, E., & Wolf, L. (2019, May). Unsupervised polyglot text-to-speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7055-7059). IEEE.

Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., & McAuley, J. (2021, November). Expressive neural voice cloning. In *Asian Conference on Machine Learning* (pp. 252-267). PMLR.

Nekvinda, T., & Dušek, O. (2020). One model, many languages: Meta-learning for multilingual text-to-speech. *arXiv preprint arXiv:2008.00768*.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). San Francisco, CA, USA: Determination press.

Nose, T., & Kobayashi, T. (2013). An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. *Speech Communication*, *55*(2), 347-357.

Nosek, T. V., Suzić, S. B., Pekar, D. J., Obradović, R. J., Sečujski, M. S., & Delić, V. D. (2021). Cross-Lingual Neural Network Speech Synthesis Based on Multiple Embeddings. *International Journal of Interactive Multimedia and Artificial Intelligence, 7(2),* 110-120.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Pantazis, Y., & Stylianou, Y. (2007). On the detection of discontinuities in concatenative speech synthesis. In Progress in nonlinear speech processing (pp. 89-100). Springer, Berlin, Heidelberg.

Parker, J., Stylianou, Y., & Cipolla, R. (2018, April). Adaptation of an expressive single speaker deep neural network speech synthesis system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5309-5313). IEEE.

Paul, D. (1981). The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(4), 786-794.

Ping, W., Peng, K., & Chen, J. (2018). Clarinet: Parallel wave generation in end-to-end text-to-

speech. arXiv preprint arXiv:1807.07281.

Qian, Y., Fan, Y., Hu, W., & Soong, F. K. (2014, May). On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3829-3833). IEEE.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.

Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In *Affective information processing* (pp. 111-126). Springer, London.

Sečujski, M., Delić, V. (2011). Automatska konverzija tekstualnih informacija u govor. Kumulativna naučnotehnička informacija, Monografska serija Vol. XLVI, No. 4.

Sečujski, M., Pekar, D., & Jakovljević, N. (2011). Automatic prosody generation for Serbo-Croatian speech synthesis based on regression trees. In *Twelfth Annual Conference of the International Speech Communication Association*.

Sečujski, M., Ostrogonac, S., Suzić, S., & Pekar, D. (2018). Learning prosodic stress from data in neural network based text-to-speech synthesis. *Информатика и автоматизация*, *4*(59), 192-215.

Sečujski, M., Pekar, D., Suzić, S., Smirnov, A., & Nosek, T. V. (2020). Speaker/Style-Dependent Neural Network Speech Synthesis Based on Speaker/Style Embedding. *J. Univers. Comput. Sci.*, *26*(4), 434-453.

Seeviour, P., Holmes, J., & Judd, M. (1976, April). Automatic generation of control signals for a parallel formant speech synthesizer. In ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 690-693). IEEE.

Shadle, C. H., & Damper, R. I. (2001). Prospects for articulatory synthesis: A position paper. In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. *Computer speech & language*, *15*(3), 287-333.

Sproat, R., & Jaitly, N. (2016). RNN approaches to text normalization: A challenge. *arXiv preprint arXiv:1611.00068*.

Suzić, S., Delić, T., Pekar, D., & Ostojić, V. (2017, September). Novel alignment method for DNN TTS training using HMM synthesis models. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)* (pp. 000271-000276). IEEE.

Suzić, S., Delić, T., Jovanović, V., Sečujski, M., Pekar, D., & Delić, V. (2018). A comparison of multi-style DNN-based TTS approaches using small datasets. In *MATEC Web of Conferences* (Vol. 161, p. 03005). EDP Sciences.

Suzić, S., Delić, T., Pekar, D., Delić, V., & Sečujski, M. (2019). Style transplantation in neural network based speech synthesis. *Acta Polytechnica Hungarica*, *16*(6), 171-189.

Suzić, S. (2019). *Parametarska sinteza ekspresivnog govora* (Doctoral dissertation, University of Novi Sad (Serbia)).

Suzić, S., Delić, T., Pekar, D., Delić, V., & Sečujski, M. (2019). Style transplantation in neural network based speech synthesis. *Acta Polytechnica Hungarica*, *16*(6), 171-189.

Suzić, S., Pekar, D., Sečujski, M., Nosek, T., & Delić, V. (2022, August). HiFi-GAN based Text-to-Speech Synthesis in Serbian. In 2022 30th European Signal Processing Conference (EUSIPCO) (pp. 2231-2235). IEEE.

Swietojanski, P., Li, J., & Renals, S. (2016). Learning hidden unit contributions for unsupervised acoustic model adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(8), 1450-1463.

Tachibana, M., Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2004). HMM-based speech synthesis with various speaking styles using model interpolation. In *Speech Prosody 2004, International Conference*.

Taigman, Y., Wolf, L., Polyak, A., & Nachmani, E. (2017). Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv:1707.06588*.

Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000, June). Speech parameter generation algorithms for HMM-based speech synthesis. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100) (Vol. 3, pp. 1315-1318). IEEE.

Tokuda, K., Zen, H., & Black, A. W. (2002, September). An HMM-based speech synthesis system applied to English. In *IEEE speech synthesis workshop* (pp. 227-230). Santa Monica: IEEE.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech synthesis based on hidden Markov models. Proceedings of the IEEE, 101(5), 1234-1252.

Valentini-Botinhao, C., Wang, X., Takaki, S., & Yamagishi, J. (2016, September). Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks. In Interspeech (Vol. 8, pp. 352-356).

Vujović, M. (2020). *Poređenje sistema za sintezu ekspresivnog govora sa mogućnošću kontrole jačine emocije* (Master, Faculty of Technical Sciences, University of Novi Sad (Serbia)).

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wang, W., Xu, S., & Xu, B. (2016, September). First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention. In Interspeech (pp. 2243-2247).

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, *78*(10), 1550-1560.

Wu, Z., Valentini-Botinhao, C., Watts, O., & King, S. (2015, April). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4460-4464). IEEE.

Wu, Z., Swietojanski, P., Veaux, C., Renals, S., & King, S. (2015). A study of speaker adaptation for DNN-based speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Wu, Z., Watts, O., & King, S. (2016, September). Merlin: An Open Source Neural Network Speech Synthesis System. In *SSW* (pp. 202-207).

Wu, Z., & King, S. (2016, March). Investigating gated recurrent networks for speech synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5140-5144). IEEE.

Xu, J., Fu, G., & Li, H. (2004). Grapheme-to-phoneme conversion for Chinese text-to-speech. In *Eighth International Conference on Spoken Language Processing*.

Yamamoto, R., Song, E., & Kim, J. M. (2020, May). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6199-6203). IEEE.

Yang, S., Wu, Z., & Xie, L. (2016, December). On the training of dnn-based average voice model for speech synthesis. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1-6). IEEE.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*.

Yoshimura, T. (2002). Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems. PhD diss, Nagoya Institute of Technology.

Yu, Q., Liu, P., Wu, Z., Ang, S. K., Meng, H., & Cai, L. (2016, March). Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5545-5549). IEEE.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. speech communication, 51(11), 1039-1064.

Zen, H., Senior, A., & Schuster, M. (2013, May). Statistical parametric speech synthesis using deep neural networks. In 2013 ieee international conference on acoustics, speech and signal processing (pp. 7962-7966). IEEE.

Zhang, Z. R., Chu, M., & Chang, E. (2002). An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese. In *International Symposium on Chinese Spoken Language Processing*.

Zhang, H., Sproat, R., Ng, A. H., Stahlberg, F., Peng, X., Gorman, K., & Roark, B. (2019). Neural models of text normalization for speech applications. *Computational Linguistics*, *45*(2), 293-337.

Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R. J., ... & Ramabhadran, B. (2019). Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*.

Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R. J., ... & Ramabhadran, B. (2019). Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*.

Zhang, J., Pan, J., Yin, X., Li, C., Liu, S., Zhang, Y., ... & Ma, Z. (2020, May). A hybrid text normalization system using multi-head self-attention for mandarin. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6694-6698). IEEE.

Zhu, X., & Xue, L. (2020). Building a controllable expressive speech synthesis system with multiple emotion strengths. *Cognitive Systems Research*, *59*, 151-159.

План третмана података

Назив пројекта/истраживања
Експресивни вишејезични синтетизатор говора (Expressive Multilingual Speech Synthesizer)
Назив институције/институција у оквиру којих се спроводи истраживање
а) Факултет техничких наука, Универзитет у Новом Саду б) АлфаНум доо, Нови Сад в) Speech Morphing System, Inc, Saj Jose, California
Назив програма у оквиру ког се реализује истраживање
/
1. Опис података

1.1 Врста студије

Укратко описати тип студије у оквиру које се подаци прикупљају

У питању су докторске студије, које су се одвијале у паралели са развојем научно-истраживачких пројеката и комерцијалних прозивода за горе наведене компаније.

1.2 Врсте података

а) квантитативни

б) квалитативни

1.3. Начин прикупљања података

а) анкете, упитници, тестови

б) клиничке процене, медицински записи, електронски здравствени записи

в) генотипови: навести врсту _____

г) административни подаци: навести врсту _____

д) узорци ткива: навести врсту_____

ђ) снимци, фотографије: навести врсту_____

е) текст, навести врсту _____

ж) мапа, навести врсту _____

з) остало: мерење растојања између параметара (објективне мере)

1.3 Формат података, употребљене скале, количина података

1.3.1 Употребљени софтвер и формат датотеке:

а) Excel фајл, датотека (више њих)

b) SPSS фајл, датотека _____

c) PDF фајл, датотека _____

d) Текст фајл, датотека _____

e) JPG фајл, датотека _____

f) Остало, датотека _____

1.3.2. Број записа (код квантитативних података)

а) број варијабли: **у зависности од експеримената, мерено неколико параметара (2-7)**

б) број мерења (испитаника, процена, снимака и сл.): **број испитаника се кретао око 20, број снимака до 30**

1.3.3. Поновљена мерења

а) да

б) не

Уколико је одговор да, одговорити на следећа питања:

а) временски размак између поновљених мера је _____

б) варијабле које се више пута мере односе се на _____

в) нове верзије фајлова који садрже поновљена мерења су именоване као _____

Напомене: _____

Да ли формати и софтвер омогућавају дељење и дугорочну валидност података?

а) Да

б) Не

Ако је одговор не, образложити _____

_____

2. Прикупљање података

2.1 Методологија за прикупљање/генерисање података

У тестовима слушања користиле су се MOS (енгл. Mean Opinion Score) и MUSHRA (енгл. MUltiple Stimuli with Hidden Reference and Anchor) методе.

За мерење објективних мера користило се растојање између параметара синтетизованог и оригиналног говора. Мерени параметри су: мел кепстрали, основна учестаност, степен звучности, степен апериодичности по фреквенцијским опсезима.

2.1.1. У оквиру ког истраживачког нацрта су подаци прикупљени?

а) експеримент, навести тип: MOS, MUSHRA.

б) корелационо истраживање, навести тип _____

ц) анализа текста, навести тип _____

д) остало, навести шта _____

2.1.2 Навести врсте мерних инструмената или стандарде података специфичних за одређену научну дисциплину (ако постоје).

У експериментима слушања су се користиле слушалице.

Формат аудио фајлова је био 22kHz, 16bit, PCM.

2.2 Квалитет података и стандарди

Табеле су у стандардном (Excel) формату.

2.2.1. Третман недостајућих података

а) Да ли матрица садржи недостајуће податке? Да **Не**

Ако је одговор да, одговорити на следећа питања:

а) Колики је број недостајућих података? _____

б) Да ли се кориснику матрице препоручује замена недостајућих података? Да Не

в) Ако је одговор да, навести сугестије за третман замене недостајућих података

_____

2.2.2. На који начин је контролисан квалитет података? Описати

Углавном ручно, односно праћењем тока експеримента.

2.2.3. На који начин је извршена контрола уноса података у матрицу?

Софтвер прилагођен за ове сврхе је вршио ту функцију, уз накнадну ручну контролу.

3. Третман података и пратећа документација

3.1. Третман и чување података

*3.1.1. Подаци ће бити депоновани у **компанијски** репозиторијум.*

3.1.2. URL адреса _____

3.1.3. DOI _____

3.1.4. Да ли ће подаци бити у отвореном приступу?

а) Да

б) Да, али после ембарга који ће трајати до _____

*в) **Не***

*Ако је одговор не, навести разлог **– јер су власништво компаније.***

3.1.5. Подаци неће бити депоновани у репозиторијум, али ће бити чувани.

Образложење

_____

_____

3.2 Метаподаци и документација података

3.2.1. Који стандард за метаподатке ће бити примењен? **Слободна форма, у оквиру Excel докумената.**

3.2.1. Навести метаподатке на основу којих су подаци депоновани у репозиторијум.

У оквиру метаподатака је наведен број експеримената, субјеката и снимака који су оцењивани, као и њихове најважније карактеристике.

Ако је потребно, навести методе које се користе за преузимање података, аналитичке и процедуралне информације, њихово кодирање, детаљне описе варијабли, записа итд.

_____

_____

_____

_____

3.3 Стратегија и стандарди за чување података

3.3.1. До ког периода ће подаци бити чувани у репозиторијуму? **неограничено**

3.3.2. Да ли ће подаци бити депоновани под шифром? Да **Не**

3.3.3. Да ли ће шифра бити доступна одређеном кругу истраживача? Да Не

3.3.4. Да ли се подаци морају уклонити из отвореног приступа после извесног времена?

Да Не

Образложити

_____

_____

4. Безбедност података и заштита поверљивих информација

Овај одељак МОРА бити попуњен ако ваши подаци укључују личне податке који се односе на учеснике у истраживању. За друга истраживања треба такође размотрити заштиту и сигурност података.

4.1 Формални стандарди за сигурност информација/података

Истраживачи који спроводе испитивања с људима морају да се придржавају Закона о заштити података о личности *(https://www.paragraf.rs/propisi/zakon_o_zastiti_podataka_o_licnosti.html)* и одговарајућег институционалног кодекса о академском интегритету.

4.1.2. Да ли је истраживање одобрено од стране етичке комисије? Да **Не**

Ако је одговор Да, навести датум и назив етичке комисије која је одобрила истраживање

_____

4.1.2. Да ли подаци укључују личне податке учесника у истраживању? Да **Не**

Ако је одговор да, наведите на који начин сте осигурали поверљивост и сигурност информација везаних за испитанике:

а) **Подаци нису у отвореном приступу**

б) Подаци су анонимизирани

ц) Остало, навести шта

_____

_____

5. Доступност података

5.1. Подаци ће бити

а) јавно доступни

б) доступни само уском кругу истраживача у одређеној научној области

ц) ***затворени***

Ако су подаци доступни само уском кругу истраживача, навести под којим условима могу да их користе:

_____

_____

Ако су подаци доступни само уском кругу истраживача, навести на који начин могу приступити подацима:

_____

_____

5.4. Навести лиценцу под којом ће прикупљени подаци бити архивирани.

_____

6. Улоге и одговорност

6.1. Навести име и презиме и мејл адресу власника (аутора) података

Компанија АлфаНум доо, Нови Сад, Србија (office@alfanum.co.rs)

6.2. Навести име и презиме и мејл адресу особе која одржава матрицу с подацима

_____

6.3. Навести име и презиме и мејл адресу особе која омогућује приступ подацима другим истраживачима

_____