UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA U NOVOM SADU

Siniša Suzić

PARAMETARSKA SINTEZA EKSPRESIVNOG GOVORA

DOKTORSKA DISERTACIJA

Нови Сад, 2019.

УНИВЕРЗИТЕТ У НОВОМ САДУ ● **ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА**
21000 НОВИ САД, Трг Доситеја Обрадовића 6

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, **РБР**:	
Идентификациони број, **ИБР**:	
Тип документације, **ТД**:	Монографска документација
Тип записа, **ТЗ**:	Текстуални штампани материјал
Врста рада, **ВР**:	Докторска дисертација
Аутор, **АУ**:	Синиша Сузић
Ментор, **МН**:	др Владо Делић, редовни професор
Наслов рада, **НР**:	Параметарска синтеза експресивног говора
Језик публикације, **ЈП**:	Српски
Језик извода, **ЈИ**:	Српски / Енглески
Земља публиковања, **ЗП**:	Република Србија
Уже географско подручје, **УГП**:	Војводина
Година, **ГО**:	2019
Издавач, **ИЗ**:	Ауторски репринт
Место и адреса, **МА**:	Факултет техничких наука, Трг Доситеја Обрадовића 6, 21000 Нови Сад
Физички опис рада, **ФО**: (поглавља/страна/ цитата/табела/слика/графика/прилога)	8/107/136/10/41/0/0/
Научна област, **НО**:	Електротехничко и рачунарско инжењерство
Научна дисциплина, **НД**:	Телекомуникације и обрада сигнала
Предметна одредница/Кључне речи, **ПО**:	синтеза говора, експресивни говор, неуронске мреже, скривени Марковљеви модели
УДК	
Чува се, **ЧУ**:	Библиотека ФТН, Нови Сад
Важна напомена, **ВН**:	
Извод, **ИЗ**:	У дисертацији су описани поступци синтезе експресивног говора коришћењем параметарских приступа. Показано је да се коришћењем дубоких неуронских мрежа добијају бољи резултати него коришћењем скривених Марковљевих модела. Предложене су три нове методе за синтезу експресивног говора коришћењем дубоких неуронских мрежа: метода кодова стила, метода додатне обуке мреже и архитектура заснована на дељеним скривеним слојевима. Показано је да се најбољи резултати добијају коришћењем методе кодова стила. Такође је предложана и нова метода за трансплантацију емоција/стилова базирана на дељеним скривеним слојевима. Предложена метода оцењена је боље од референтне методе из литературе.
Датум прихватања теме, **ДП**:	27.12.2018.
Датум одбране, **ДО**:	

Чланови комисије, **КО**:		
Председник:	др Милан Сечујски	
Члан:	др Жељен Трповски	
Члан:	др Татјана Грбић	Потпис ментора
Члан:	др Зоран Перић	
Члан:	др Никша Јаковљевић	
Члан, ментор:	др Владо Делић	

Accession number, **ANO**:		
Identification number, **INO**:		
Document type, **DT**:		Monograph publication
Type of record, **TR**:		Textual printed material
Contents code, **CC**:		PhD thesis
Author, **AU**:		Siniša Suzić
Mentor, **MN**:		Vlado Delić, PhD
Title, **TI**:		Parametric synthesis of expressive speech
Language of text, **LT**:		Serbian
Language of abstract, **LA**:		Serbian / English
Country of publication, **CP**:		Republic of Serbia
Locality of publication, **LP**:		Vojvodina
Publication year, **PY**:		2019
Publisher, **PB**:		Author's reprint
Publication place, **PP**:		Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad
Physical description, **PD**: (chapters/pages/ref./tables/pictures/graphs/appendixes)		8/107/136/10/41/0/0/
Scientific field, **SF**:		Electrical and computer engineering
Scientific discipline, **SD**:		Telecommunications and signal processing
Subject/Key words, **S/KW**:		text-to-speech synthesis, expressive speech, neural networks, hidden Markov models
UC		
Holding data, **HD**:		Library of Faculty of technical sciences, Novi Sad
Note, **N**:		
Abstract, **AB**:		In this thesis methods for expressive speech synthesis using parametric approaches are presented. It is shown that better results are achived with usage of deep neural networks compared to synthesis based on hidden Markov models. Three new methods for synthesis of expresive speech using deep neural networks are presented: style codes, model re-training and shared hidden layer architecture. It is shown that best results are achived by using style code method. The new method for style transplantation based on shared hidden layer architecture is also proposed. It is shown that this method outperforms referent method from literature.
Accepted by the Scientific Board on, **ASB**:		December 27th, 2018
Defended on, **DE**:		
Defended Board, **DB**:	President:	Milan Sečujski, PhD
	Member:	Željen Trpovski, PhD
	Member:	Tatjana Grbić, PhD
	Member:	Zoran Perić, PhD
	Member:	Nikša Jakovljević, PhD
	Member, Mentor:	Vlado Delić, PhD

Menthor's sign

Sažetak

Sinteza govora je tehnologija koja omogućava računarima da tekst pretvaraju u ljudski govor. Moderne primene ove tehnologije, osim razumljivosti i prirodnosti generisanog govora, zahtevaju da govor bude generisan u odgovarajućem stilu ili emociji, koji treba da odgovaraju konkretnom domenu u kojem se sinteza primenjuje. Predmet istraživanja disertacije je sinteza ekspresivnog govora korišćenjem parametarskih pristupa sintezi.

Prvi cilj istraživanja je razvoj pristupa koji omogućavaju sintezu ekspresivnog govora, čak i u slučaju korišćenja ograničene količine ekspresivnog govora u obuci. Osnovni predloženi pristup za sintezu ekspresivnog govora, nazvan metoda kodova stila, uspešno je primenjen u dva najčešće korišćena parametarska pristupa: u sintezi na bazi skrivenih Markovljevih modela i u sintezi na bazi dubokih neuronskih mreža. Objektivnim merama i testovima slušanja pokazano je da govor generisan korišćenjem dubokih neuronskih mreža po kvalitetu i izraženosti stila nadmašuje govor generisan korišćenjem skrivenih Markovljevih modela. Pored metode kodova stila, predložena su još dva pristupa za sintezu ekspresivnog govora u okviru sinteze bazirane na primeni dubokih neuronskih mreža: arhitektura sa deljenim skrivenim slojevima i dodatna obuka mreže. Ovim pristupima dobija se govor lošijeg kvaliteta nego govor dobijen korišćenjem osnovnog predloženog pristupa. Svi predloženi pristupi testirani su korišćenjem baza u kojem je količina ekspresivnog materijala nekoliko puta manja od količine neutralnog govora.

Drugi cilj istraživanja je razvoj postupka koji omogućava transplantaciju stila, odnosno sintezu govora u stilu koji postoji u bazi jednog govornika, ali ne i u bazi nekog drugog govornika. Za ove potrebe predložena je metoda koja je takođe bazirana na arhitekturi sa deljenim skrivenim slojevima. Objektivni i subjektivni testovi pokazali su da predloženi pristup nadmašuje referenti pristup iz literature.

Abstract

Speech synthesis is a technology which enables computers to transform text into human speech. Modern applications of this technology demand, besides intelligibility and naturalness of the generated speech, that the speech is generated in a suitable style or emotion which need to comply with the exact environment in which the synthesis is being applied. The object of the research presented in this dissertation is the synthesis of expressive speech using parametric approaches.

The first goal of the research is to develop approaches which enable expressive speech synthesis even when there is a limited amount of expressive speech available. The basic proposed approach to expressive speech synthesis, named style code method, is successfully applied in two of the most parametric approaches: the synthesis based on hidden Markov models and in the synthesis based on deep neural networks. Objective measures and listening tests show that the speech generated by utilizing deep neural networks surpasses in both quality and style expression the speech generated by hidden Markov models. Besides the style code method, two more expressive speech synthesis approaches are proposed within the deep neural network synthesis framework: shared-hidden-layer architecture and network retraining. However, these approaches produce speech of poorer quality than the speech generated by using the basic proposed approach. All the proposed methods are tested using databases which contain few times less expressive speech then neutral speech.

The second goal of the research is developing a procedure which enables style transplantation, namely the synthesis in the style of speech which exists in the database of one speaker but not in the database of another speaker. For these purposes, a method also based on the shared-hidden-layer architecture is proposed. Objective and subjective tests have shown that the proposed approach surpasses the approach referenced in the literature.

Zahvalnica

Zahvaljujem se svom mentoru, prof. dr Vladi Deliću, koji mi je omogućio da se bavim naučnim radom, kao i na svim korisnim savetima i podršci koju mi je pružio u profesionalnoj karijeri.

Veliku zahvalnost dugujem saradnici Tijani Delić, koja je sa mnom delila sve uspone i padove u istraživačkom radu čiji rezultati su prezentovani u ovoj disertaciji.

Takođe, želim da se zahvalim preduzeću „AlfaNum" iz Novog Sada na ustupljenim govornim bazama i računarskim resursima.

Zahvaljujem se i svim članovima komisije koji su svojim sugestijama doprineli kvalitetu ove disertacije.

Posebnu zahvalnost dugujem svojoj porodici koja mi je uvek pružala bezrezervnu podršku i ljubav.

Sadržaj

Spisak slika

Spisak tabela

Spisak skraćenica

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

HMM - *engl.* Hidden Markov Models

DNN - *engl.* Deep Neural Networks

CART - *engl.* Classification and Regression Trees

MP (neuron) - engl. *McColough-Pitts*

ADALINE (neuron) - *engl.* Adaptive Linear

RNN - *engl.* Recurrent Neural Networks

LSTM (neuron) - *engl.* Long Short-Term Memory

VUV (obeležje) - *engl.* Voiced UnVoiced

RMSE - *engl*. Root Mean Square Error

MOS (test) - *engl.* Mean Opinion Score

MUSHRA (test) - *engl.* Multiple Stimuli Hidden Reference and Anchor

ADSS - arhitektura sa deljenim skrivenim slojevima

EAM - emotivni aditivni model

LHUC (pristup) - *engl.* Learning Hidden Layer Contribution

ADU - arhitektura sa dodatnim ulazima

1. Uvod

Komunikacija je proces u kojem ljudi međusobno vrše interakciju i putem simbola kreiraju i interpretiraju značenje [1]. Govor je jedan od osnovnih medijuma preko kojeg se odvija komunikacija. Prvi pokušaji da se veštački generiše govor potiču iz XVIII veka [2]. Naime, 1779. godine Kristijan Gotlib Kracenštajn (*Christian Gotlieb Kratzenstein*) osvojio je prvu nagradu na konkursu koji je raspisala Kraljevska akademija nauka i umetnosti u Sankt Petersburgu za rad u kojem je opisao razlike između vokala sa fiziološkog stanovišta. On je tada predstavio i mehanički uređaj koji je mogao da reprodukuje te glasove. Od predstavljanja Kracenštajnovog uređaja do danas generisanje ljudskog govora veštačkim putem prevalilo je dugačak put. Savremeni pristupi ovoj problematici bazirani su na upotrebi računara. Imajući to u vidu, sinteza govora na osnovu teksta (engl. *Text-to-Speech Synthesis*, TTS) definiše se kao tehnologija koja omogućava računarima da tekst pretvaraju u ljudski govor.

Ovakva tehnologija ima širok spektar primena. Prvobitno je bila korišćena za čitanje tekstualnog sadržaja slepim osobama, a danas ova tehnologija može biti od velikog značaja i osobama kod kojih, zbog različitih uzroka, dolazi do gubitka glasa, kao što su npr. laringektomisani pacijenti. TTS se koristi i u pozivnim centrima za saopštavanje različitih informacija korisnicima. Sa porastom upotrebe pametnih telefona ova tehnologija je pronašla svoje mesto u okviru različitih virtualnih asistenata i aplikacija za navigaciju. U poslednje vreme sve više raste i popularnost audio knjiga. TTS omogućava znatno brže i jednostavnije generisanje ovakvih materijala korišćenjem računara umesto dugotrajnog i napornog snimanja profesionalnih govornika.

Trenutno su u sintezi govora dominantna dva pristupa: konkatenativni i parametarski. Sintetizatori govora koji koriste konkatenativni pristup vrše odabir i spajanje govornih segmenata iz odgovarajuće govorne baze. Parametarski pristupi sintezi govora zasnovani su na parametrizaciji govora i razvoju modela koji mogu uspešno da predvide korišćene parametre. Iako nešto lošiji od konkatenativnih pristupa po pitanju ukupnog kvaliteta

sintetizovanog govora, parametarski pristupi su našli svoju primenu u praksi zbog mogućnosti lakše manipulacije karakteristikama generisanog govora.

1.1. Predmet i ciljevi istraživanja

Dve glavne karakteristike koje sintetizovani govor treba da ispuni jesu razumljivost i prirodnost [3]. Većina istraživača saglasna je sa činjenicom da moderni sintetizatori govora postižu dobre rezultate prema ovim kriterijumima, ali se često ističe da sintetizovani glas zvuči isuviše monotono, odnosno da se, bez obzira na domen primene, govor generiše u samo jednom dostupnom stilu, a koji odgovara neutralnom govoru.

Ljudski govor ne služi samo da se prenese određena informacija. U njemu su sadržane i informacije kao što je npr. emocionalno stanje govornika, starost, pol i druge. U svakodnevnoj komunikaciji često primenjujemo i govor u određenom stilu kako bismo postigli neki cilj. Slušaoci bolje pamte emotivne reklame [4]. Duhovit glas i pozitivno raspoloženje mogu uticati na raspoloženje slušalaca [5], što ima terapeutske implikacije.

Potreba za generisanjem govora u određenim stilovima i emocijama postoji i u komercijalnim primenama TTS tehnologije. Različiti stilovi se mogu primeniti u zavisnosti od toga da li pročitani sadržaj predstavlja servisnu informaciju, upozorenje ili reklamu [6]. Istraživanja su takođe pokazala da ljudi interakciju sa računarima doživljavaju na isti način kao i kada u stvarnom životu imaju komunikaciju sa drugim ljudima. Takođe je pokazano da su korisnici TTS sistema ljubazni dok komuniciraju sa kompjuterima, kao i da drugačije reaguju na sintetizovani muški, odnosno ženski glas [7]. U [8] je navedeno da ljudi veću sklonost iskazuju ka robotima koji pokazuju određenu ekspresivnost nego ka onima koji su efikasni. Takođe, neki istraživači avionskih nesreća veruju da neutralni stil govora korišćen prilikom generisanja poruka upozorenja utiče na putnike da ne shvate dovoljno dobro potencialnu opasnost [9].

Jedan od glavnih problema prilikom kreiranja novog TTS glasa jeste snimanje govorne baze. Procedura snimanja i naknadnog anotiranja baze je dugotrajan proces koji zahteva velike resurse. Potreba za proširenjem ovakve baze sa ekspresivnim delovima samo dodatno

komplikuje celi proces. Stoga je i količina ekspresivnog govora (najčešće se radi o emotivnom govoru) u bazi za kreiranje TTS govornika obično značajno manja od količine materijala snimljenog u neutralnom stilu.

Cilj ovog istraživanja jeste da ispita mogućnost sintetizovanja ekspresivnog govora korišćenjem parametarskih pristupa, sa posebnim osvrtom na sintezu baziranu na upotrebi dubokih neuronskih mreža (engl. *Deep Neural Networks,* DNN). Svi predloženi pristupi biće testirani na govornoj bazi koja sadrži relativno malu količinu ekspresivnog materijala. Hipoteza koja će biti testirana jeste da se korišćenjem ograničene količine ekspresivnog govornog materijala u obuci može dobiti zadovoljavajući kvalitet sintetizovanog ekspresivnog govora. U procesu provere kvaliteta ekspresivnog govora pažnja će biti usmerena na rezultate prepoznavanja odgovarajućeg stila u sintetizovanom govoru, pri čemu bi trebalo da se dobiju približni rezultati kao u slučaju prepoznavanja u originalnom materijalu korišćenom u obuci. Neophodno je uporediti i kvalitet neutralnog govora, za koji je dostupno mnogo više materijala za obuku, sa kvalitetom dobijenog ekspresivnog govora.

Takođe, biće predloženi i testirani pristupi koji omogućavaju da se emocija (ili govorni stil) sadržani u bazi jednog govornika sintetišu u govoru drugog govornika kod kojeg ta emocija (ili govorni stil) ne postoji u bazi za obuku. Ovakva procedura se u literaturi naziva transplantacija emocija. Pretpostavka je da se korišćenjem predloženih pristupa može dobiti govor koji će posedovati karakteristike transplantovanog stila, ali da će ukupan kvalitet takvog govora ipak biti nešto lošiji nego u slučaju da emocija (ili stil) zaista postoje u bazi za obuku.

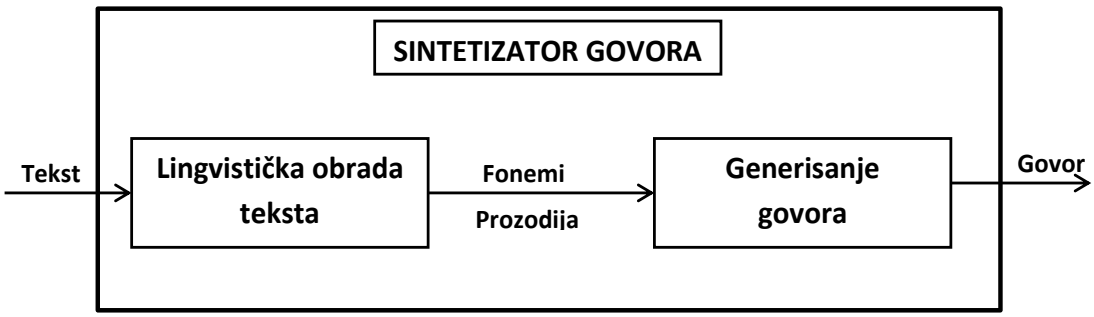
1.2. Organizacija disertacije

Disertacija se sastoji iz osam poglavlja. Nakon uvodnog dela u kojem su predstavljeni predmet i ciljevi istraživanja sledi poglavlje u kojem su detaljnije opisani različiti pristupi u sintezi govora. U drugom poglavlju je takođe dat pregled dosadašnjih rezultata u sintezi ekspresivnog govora. U trećem poglavlju detaljno je opisana sinteza korišćenjem skrivenih Markovljevih modela (engl. *Hidden Markov Models,* HMM), dok je u četvrtom opisana sinteza govora primenom dubokih neuronskih mreža. Na kraju ovog poglavlja dato je

poređenje kvaliteta govora generisanog DNN i HMM pristupima za sintetizator na srpskom jeziku. U petom poglavlju opisana je osnovna metoda za sintezu ekspresivnog govora i dato je poređenje HMM i DNN pristupa po pitanju kvaliteta sintetizovanog ekspresivnog govora. U šestom poglavlju su pored osnovne metode kodova stila predložene dve nove metode za sintezu ekspresivnog govora u okviru DNN pristupa i predstavljeni su rezultati međusobnog poređenja ovih metoda za slučaj male količine ekspresivnog materijala. Metoda za transplantaciju stilova predložena je u sedmom poglavlju. U osmom poglavlju dati su osnovni zaključci i definisani pravci daljeg istraživanja.

2. Uvod u sintezu ekspresivnog govora

Izgled sistema za sintezu govora prikazan je na slici 2.1. Proces sinteze sastoji se od dve faze [3]. U prvoj fazi se iz teksta izdvajaju obeležja koja omogućavaju sintezu prirodnog govora, tj. vrši se tzv. lingvistička obrada ulaznog teksta. Kao rezultat lingvističke obrade teksta dobija se informacija o fonemima koje treba sintetizovati. Ovaj proces je složen, jezički zavisan i sastoji se od nekoliko potprocesa, ali njihov detaljan opis izlazi van obima ove disertacije. Za opis skupa alata i metoda koji su obuhvaćeni ovom fazom obično se koristi engleska reč *frontend.* Više o postupcima uključenim u ovaj deo sinteze može se pronaći u [10]. Izlazi bloka za lingvističku obradu teksta prosleđuju se u blok na čijem izlazu se dobija sintetizovan govor.



Slika 2.1 Arhitektura TTS sistema

2.1. Pristupi sintezi govora

Kada se govori o različitim pristupima za sintezu govora u stvari se misli na različite realizacije bloka za generisanje govora sa slike 2.1. U uvodnom poglavlju je navedeno da su konkatenativni i parametarski pristup dva najpopularnija pristupa za sintezu govora. U nastavku teksta ovi pristupi će biti detaljnije predstavljeni. Biće pomenuti i neki dodatni

pristupi, koji nisu značajni u komercijalnim primenama, ali se ponekad i dalje koriste u sintezi ekspresivnog govora.

2.1.1. Konkatenativna sinteza govora

Konkatenativni pristupi zasnivaju se na povezivanju odgovarajućih govornih segmenata iz ranije snimljene govorne baze. Najjednostavniji sistemi bazirani na povezivanju govornih segmenata i dalje se koriste prilikom obaveštavanja putnika u vozovima ili tramvajima o trenutnoj ili sledećoj stanici [11]. U ovakvim sistemima skoro čitave rečenice su unapred snimljene. Menjaju se samo informacije o nazivima stanica. Problem sa ovim pristupom predstavlja situacija u kojoj je potrebno dodati naziv nove stanice. Pošto se dešava da originalni govornik nije dostupan za dodatno snimanje, potreban materijal mora da snimi neki novi govornik. Takođe, rečenice sintetizovane na ovaj način zvuče dosta neprirodno u delovima na kojima se vrši spajanje dva segmenta (rečenice i željenog naziva lokacije).

Realizacija konkatenativne sinteze opisana u prethodnom pasusu ima primenu samo u ograničenim domenima. Najjednostavniji način za generisanje govora konkatenativnim pristupom, a koji bi omogućavao generisanje govora bez obzira na željeni domen primene, podrazumevao bi povezivanje odgovarajućih fonema. Međutim, na ovaj način dobio bi se govor izuzetno lošeg kvaliteta pošto se u obzir ne bi uzimala koartikulacija, pojava uzrokovana kontinualnim kretanjem artikulacionih organa prilikom realizacije dva uzastopna fonema. U [12] pokazano je da su upravo prelazni regioni veoma bitni za razumljivost govora. Problem koartikulacije može se rešiti upotrebom difonske sinteze [13]. Difonski sintetizatori generišu govor spajanjem difona, govornih segmenata koji obuhvataju deo govora od sredine jednog fonema do sredine narednog fonema. Ukoliko u nekom jeziku postoji N različitih fonema tada bi maksimalan broj difona iznosio N^2. U praksi je broj difona manji od maksimalnog broja pošto u govornom jeziku obično ne postoje kombinacije svih mogućih fonema.

Difoni koji postoje u bazi obično ne odgovaraju difonima koje treba upotrebiti u nekom kontekstu po pitanju dužine i osnovne učestanosti, stoga se pristupa postupcima kojima se vrši manipulacija ovih karakteristika govora kako bi se dobile željene vrednosti. Najčešće

korišćena metoda naziva se TD-PSOLA (engl. *Time Domain Pitch Synchronous Overlap and Add*) [14]. Ukoliko se želi promeniti trajanje nekog difona briše se, odnosno dodaje, odgovarajući broj osnovnih perioda izabranog segmenta. Manipulacija osnovnom učestanošću zasniva se na pomeranju pozicija centara osnovnih perioda željenog segmenta. TD-PSOLA je jednostavan algoritam koji daje dobre rezultate. Međutim, ukoliko su modifikacije velike javljaju se artefakti u govoru.

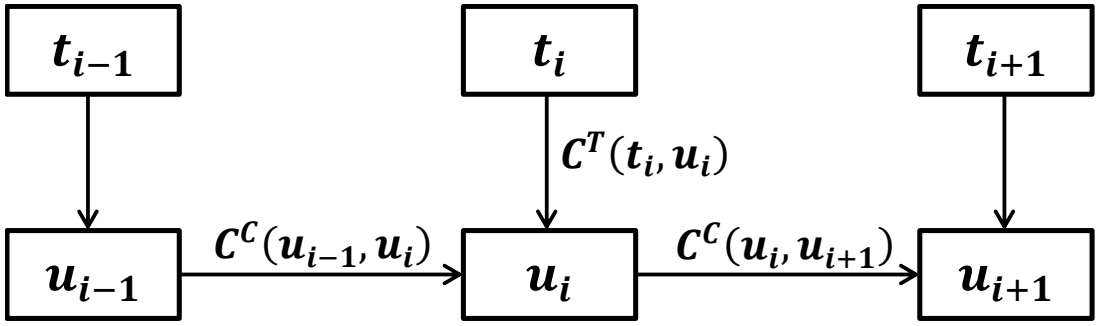
Segmenti koji se koriste u konkatenativnoj sintezi treba da ispunjavaju neke od sledećih uslova [15]:

- obuhvatanje što je moguće većeg skupa koartikulacionih efekata,
- jedan segment treba da bude dostupan u više različitih dužina i osnovnih učestanosti,
- efekti spajanja treba da budu što je moguće manje izraženi,
- broj različitih segmenata bi trebalo da bude minimalan (pre svega zbog upotrebe na platformama sa ograničenim memorijskim resursima).

Iako se primenom difonske sinteze dobija razumljiv govor, korišćenje difona nije dovoljno da se pokriju svi mogući koartikulacioni efekti pošto se oni često protežu i tokom čitavog trajanja fonema [15], stoga je postojala potreba za razvojem boljih pristupa baziranih na spajanju segmenata. Pristup najčešće korišćen u praksi naziva se selekcija segmenata [16], [17]. Selekcija segmenata zasniva se na odabiru optimalne sekvence segmenata iz velike govorne baze poređenjem prozodijskih i akustičkih parametara.

Algoritam selekcije segmenata za ciljnu sekvencu segmenata, $t^n = (t_1, t_2 \dots t_n)$, pronalazi sekvencu segmenata iz baze, $u^n = (u_1, u_2 \dots u_n)$, koja je najsličnija ciljnoj sekvenci. Na taj način umanjuje se potreba za dodatnom obradom dobijene sekvence. U procesu traženja najsličnije sekvence definišu se dve funkcije cene: cena sličnosti segmenata i cena povezivanja segmenata, kao što je prikazano na slici 2.2.

Svaki ciljni segment i segment u bazi opisani su odgovarajućim vektorom obeležja koji sadrži informacije kao što su osnovna učestanost, trajanje i energija segmenta. Cena sličnosti segmenata definiše se kao suma razlika između pojedinačnih elemenata vektora obeležja ciljnog segmenta i segmenta iz baze, $C_j^T(t_i, u_i)$,



Slika 2.2 Opis cena segmenata u proceduri selekcije segmenata

$$C^T(t_i, u_i) = \sum_{j=1}^{p} w_j^T \, C_j^T(t_i, u_i), \tag{2.1}$$

pri čemu w_j^T predstavlja težinu pridruženu datom elementu vektora obeležja, a p dimenzionalnost vektora obeležja.

Cena povezivanja segmenata takođe se može predstaviti kao suma razlika više pojedinačnih cena koje opisuje povezivanje segmenata, $C_j^C(u_{i-1}, u_i)$. Pojedinačne cene koje opisuju povezivanje mogu npr. biti: kepstralno rastojanje u tački povezivanja, apsolutne razlike u osnovnoj učestanosti ili intenzitetu segmenata koji se spajaju. Prema tome, cena povezivanja segmenata definiše se na sledeći način

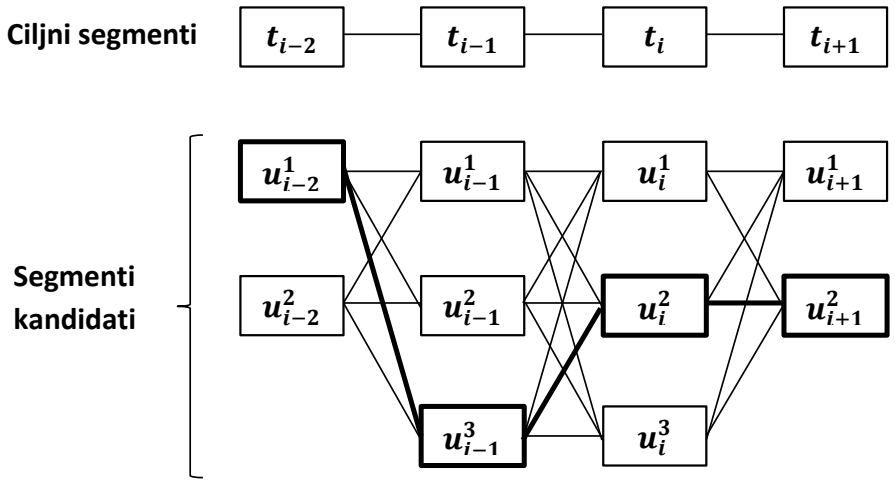
$$C^C(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^C \, C_j^C(u_{i-1}, u_i), \tag{2.2}$$

gde je q broj pojedinačnih cena povezivanja, a w_j^C težina pridružena svakoj pojedinačnoj ceni.

Ukupna mera sličnosti ciljne sekvence i sekvence iz baze računa se prema jednačini

$$C(t^n, u^n) = \sum_{i=1}^{n} \sum_{j=1}^{p} w_j^T \, C_j^T(t_i, u_i) + \sum_{i=2}^{n} \sum_{j=1}^{q} w_j^C \, C_j^C(u_{i-1}, u_i). \tag{2.3}$$

Pošto svakom ciljnom segmentu odgovara nekoliko realizacija u bazi, zadatak algoritma selekcije segmenata jeste pronalaženje sekvence $\bar{u}^n$ za koju je vrednost izraza (2.3) minimalna, tj.



Slika 2.3 Ilustracija algoritma selekcije segmenata

$$\bar{u}^n = arg \min_{u_i} C(t^n, u_i^n). \tag{2.4}$$

Postupak selekcije optimalne sekvence segmenata prikazan je na slici 2.3. Za rešavanje problema opisanog u jednačini (2.4) koristi se Viterbijev algoritam [18]. Budući da broj svih mogućih sekvenci u bazi može biti veoma veliki, rešavanje jednačine (2.4) može da traje prilično dugo te se pristupa postupcima optimizacije. Na primer, u procesu pretrage za određeni ciljni segment koriste se samo segmenti u bazi kod kojih je cena sličnosti segmenata iznad određenog praga. Još jedan primer optimizacije opisan je u [19]. Predloženo je da se u procesu pripreme sistema izvrši klasterizacija segmenata na osnovu njihovog prozodijskog i fonetskog sadržaja. U procesu traženja optimalne sekvence računa se rastojanje ciljnog segmenta od centra odgovarajućeg klastera, a ne od svakog segmenta pojedinačno.

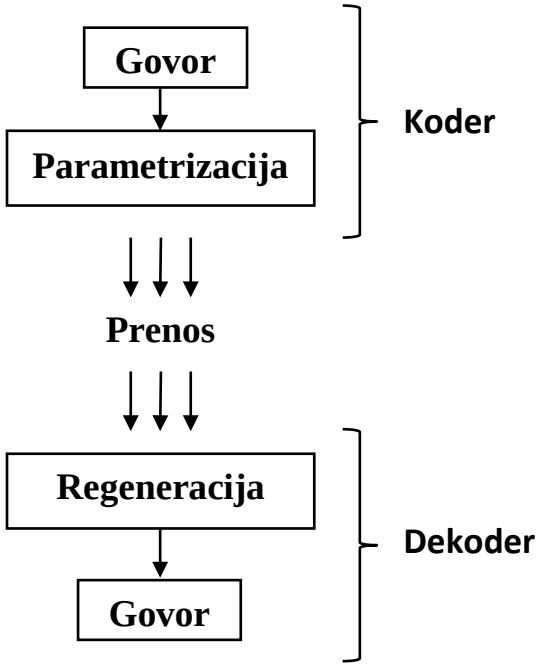
U procesu selekcije mogu se koristiti segmenti različite veličine, počev od segmenata dužine jednog frejma [20], preko polufona [21] i difona [22], sve do dužih neuniformnih segmenata [23], [24].

Uzimajući u obzir sve pristupe za sintezu govora, smatra se da se upotrebom sinteze na osnovu selekcije segmenata dobija sintetizovani govor najboljeg kvaliteta. Međutim, da bi to zaista bilo ispunjeno, odgovarajuće baze moraju biti dovoljno velike, a sadržaj rečenica koje se snimaju pažljivo izabran, kako bi bio pokriven što je moguće veći broj različitih fonetskih konteksta. Problem koji se javlja prilikom korišćenja ovog pristupa jesu artefakti koji se

javljaju na mestima spajanja dva segmenta, kada dovoljno slični segmenti ne mogu biti pronađeni u bazi [25], [26]. Pored toga, ograničene su manipulacije parametrima govornog signala. U [27] navodi se da se moguće manipulacije sastoje od promene trajanja izabranih segmenata, odnosno osnovne učestanosti.

2.1.2. Parametarska sinteza govora

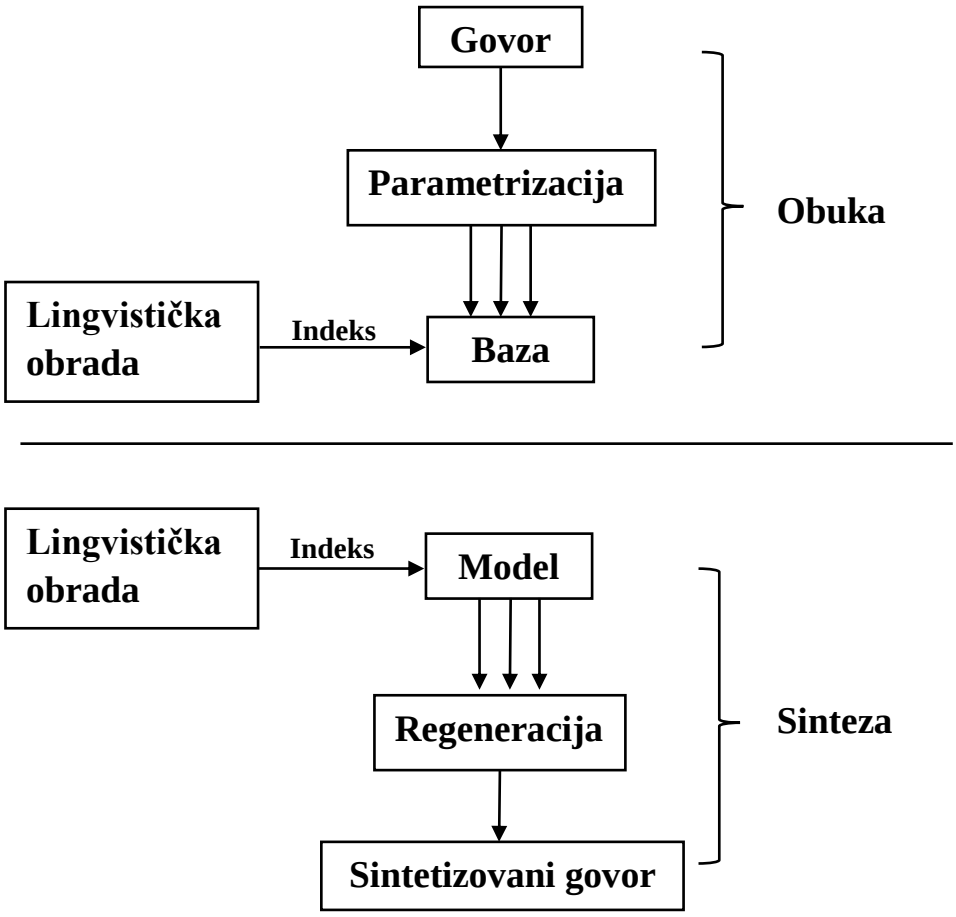
Parametarski pristupi sintezi govora zasnovani su na upotrebi vokodera. Vokoderi su se počeli upotrebljavati tokom tridesetih godina XX veka sa ciljem kompresije govornog signala radi efikasnog iskorišćenja dostupnog prenosnog opsega. U tom periodu vokoderi su u stvari predstavljali elektronske sklopove koji su omogućavali kodovanje i dekodovanje govornog signala [28]. U savremenoj terminologiji pod pojmom vokoder podrazumeva se skup postupaka koji omogućavaju parametarsku predstavu govornog signala, kao i rekonstrukciju govora na osnovu datih parametara kao što je prikazano na slici 2.4. Broj parametara koji se formiraju u toku kodovanja obično je manji od odgovarajućeg broja odbiraka govornog signala, tj. primenom vokodera postiže se kompresija.



Slika 2.4 Princip funkcionisanja vokodera

Jedan način generisanja govora u procesu sinteze, baziran na principu funkcionisanja vokodera, dat je na slici 2.5 [3]. Ovakvo predstavljanje podrazumeva da je postupak generisanja govora podeljen na dve faze. U prvoj fazi, takozvanoj fazi obuke, vrši se izdvajanje obeležja iz govornog signala korišćenjem kodera (vokodera). Izdvojena obeležja potom se čuvaju u odgovarajućem obliku. Obeležja se mogu čuvati u neizmenjenom obliku, što bi bilo memorijski prilično neefikasno, ali se isto tako mogu formirati i modeli koji ih opisuju. U fazi sinteze, na osnovu lingvističke specifikacije, pristupa se unosu u bazi, odnosno direktno sačuvanim obeležjima ili modelu koji se koristi da bi generisao odgovarajuća obeležja. Generisana obeležja se pomoću vokodera (dekodera) pretvaraju u odbirke govornog signala.

U sklopu izrade ove disertacije korišćen je pristup sintezi govora predstavljen na slici 2.5



Slika 2.5 Sinteza govora na osnovu funkcionisanja vokodera

koji je baziran na kreiranju modela koji opisuju parametre vokodera. Takvi modeli nazivaju se statističkim ukoliko su parametri opisani statističkim pojmovima, kao štu su npr. srednja vrednost i varijansa Gausove raspodele. Najpopularniji statistički pristup sintezi govora baziran je na upotrebi skrivenih Markovljevih modela (engl. *Hidden Markov Models*, HMM) i smeša Gausovih funkcije gustine verovatnoće (engl. *Gaussian Mixture Models*, GMM) [29]. Ovaj pristup u nastavku disertacije nazivaće se kratko HMM sinteza. HMM sinteza opisuje nekoliko srodnih lingvističkih obeležja jednim modelom u toku obuke. Dakle, u sistemu zapravo postoji veći broj modela.

HMM sinteza se u literaturi često koristi i kao sinonim za opisivanje parametarskih statističkih metoda sinteze [30]. Međutim, u literaturi se mogu pronaći i drugi statistički pristupi sintezi kao što je autoregresiono modelovanje [31] ili linearno dinamičko modelovanje [32].

HMM sinteza je skoro dvadesetak godina bila dominantan pristup sintezi govora na osnovu parametarskih modela. Međutim, u poslednjih nekoliko godina predstavljeni su modeli bazirani na korišćenju dubokih neuronskih mreža (engl. *Deep Neural Networks*, DNN) [33], koji uspevaju da prevaziđu određene probleme koji se javljaju kod HMM sinteze i generišu govor boljeg kvaliteta. U nastavku disertacije sinteza govora korišćenjem dubokih neuronskih mreža će se kratko nazivati DNN sinteza. DNN sinteza se takođe uklapa u prikaz sa slike 2.5. Međutim, u ovom slučaju u stvari postoji jedan model koji opisuje sve date lingvističke specifikacije.

Detaljan prikaz DNN i HMM sinteze biće predstavljen u narednim poglavljima disertacije.

Govor dobijen korišćenjem parametarskih pristupa zvuči ublaženo, pošto se zbog različitih procesa uprosečavanja gube neki bitni detalji u spektru ali za razliku od konkatenativnih pristupa parametarski pristupi omogućavaju različite manipulacije sa parametrima govora, kao i pojednostavljenu adaptaciju modela na glas novog govornika. Takođe, parametarski pristupi su manje zahtevni po pitanju potrebnih memorijskih resursa, pošto ne zahtevaju čuvanje čitave govorne baze, već samo parametara koji opisuju model. Parametarski pristupi omogućavaju da se govor zadovoljavajućeg kvaliteta dobije

korišćenjem manje baze za obuku u poređenju sa bazom koju koriste konkatenativni pristupu. Zbog prethodno opisanih prednosti odlučeno je da se u okviru disertacije istraže postupci sinteze ekspresivnog govora pokuša korišćenjem parametarskih pristupa.

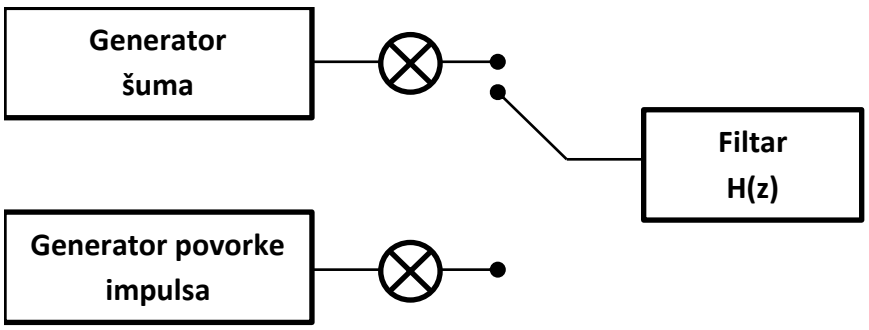
2.1.3. Ostali pristupi sintezi govora

Pored dva prethodno opisana pristupa sintezi govora, koji su dominantni, kako u istraživačkoj zajednici, tako i u komercijalnim primenama, postoje i drugi pristupi o kojima će biti reči u nastavku ovog odeljka.

Formantni sintetizatori zasnovani su na akustičkom modelu produkcije govora [34]. Ovaj model produkciju govora predstavlja kao filtriranje pobudnog signala vremenski promenljivim filtrom koji opisuje vokalni trakt kao što je prikazano na slici 2.6. Pobuda koja odgovara zvučnim glasovima predstavljena je periodičnom povorkom impulsa, dok se bezvučna pobuda modeluje šumom.

Formantni sintetizatori opisuju prenosnu karakteristiku filtra koji predstavlja vokalni trakt položajem formanata i njihovom širinom, a dodatni ulazni parametar predstavljaju parametri pobude. Detaljnija analiza postupaka za kreiranje formantnih sintetizatora može se pronaći u [35], [36].

Artikulatorni sintetizatori takođe koriste model predstavljen na slici 2.6. Za razliku od formantnih sintetizatora, artikulatorni sintetizatori opisuju karakteristike vokalnog trakta korišćenjem geometrijskih veličina koje odgovaraju pojedinim njegovim delovima.



Slika 2.6 Model produkcija govora pobuda-filtar

Geometrijska struktura vokalnog trakta se potom transformiše u jednačine koje opisuju filtar vokalnog trakta. Opisi ovakvih sintetizatora mogu se pronaći u [37]–[39]. Ovakvi sintetizatori su uglavnom ograničeni na sintezu pojedinačnih vokala. Međutim, postoje i pokušaji generisanja povezanog govora [40], [41].

Kvalitet govora dobijenog korišćenjem formantnih i artikulatornih sintetizatora je generalno loš. Međutim, oni se i dalje koriste u istraživačkoj zajednici, između ostalog i za proučavanje sinteze ekspresivnog govora.

2.2. Evaluacija TTS sistema

Postoji više različitih pristupa koji se koriste da bi se ocenio kvalitet nekog TTS sistema. Sve pristupe možemo podeliti u sledeće grupe [42]:

- objektivne mere,
- subjektivne mere,
- poređenje karakteristika sistema.

Objektivne mere dobijaju se direktnim poređenjem parametara izdvojenih iz prirodnog govora sa generisanim parametrima. Ovaj tip mera može biti naročito koristan u toku razvoja TTS sistema. Naime, veliko odstupanje vrednosti generisanih parametara od parametara izdvojenih iz prirodnog govora, obično ukazuje na postojanje problema u procesu razvoja. Rano otkrivanje problema je značajno pošto može da skrati vreme neophodno da bi se novi sistem konstruisao. Međutim, problem sa ovim pristupom predstavlja činjenica da vrednosti objektivnih mera ne moraju uvek da se poklapaju sa ocenom slušalaca. Moguće je, da određeno pogoršanje u objektivnim merama, slušaoci uopšte ne prepoznaju kao smanjenje kvaliteta govora, ali i da bolje objektivne mere uopšte ne dovode do povećanja subjektivnih ocena. U sklopu rada na ovoj disertaciji korišćene su sledeće objektivne mere:

1. srednja kepstralna distorzija (engl. *Mean Cepstral Distortion,* MCD) izražena jednakošću [43]

$$MCD = \frac{1}{T}\frac{10\sqrt{2}}{\ln 10}\sum_{t=0}^{T-1}\sqrt{\sum_{d=0}^{D-1}\left(v_d^{tar}(t) - v_d^{ref}(t)\right)}, \quad (2.5)$$

gde je T ukupan broj frejmova u referentnoj v^{ref}, i generisanoj v^{tar}, sekvenci i D broj kepstralnih koeficijenata izdvojenih u svakom frejmu,

2. distorzija koeficijenata aperiodičnosti (ukoliko se koriste), a koja se računa na isti način kao u jednakosti (2.5),

3. srednja kvadratna greška osnovne učestanosti (engl. *Root Mean Square Error,* RMSE),

4. greška procene zvučnosti (engl. *Voiced Unvoiced,* VUV), koja se računa kao količnik broja frejmova za koje je ovo obeležje pogrešno određeno i ukupnog broja frejmova,

5. korelacija za obeležja zvučnosti i osnovne učestanosti.

Subjektivne mere zasnivaju se na ocenama koje za pojedine karakteristike govora daju slušaoci. Ova grupa testova može da se podeli na dve podgrupe:

1. ocena razumljivosti,
2. ocena kvaliteta sintetizovanog govora.

Pošto je prihvaćeno da savremeni sintetizatori imaju dobru razumljivost [44], metode predstavljene u disertaciji evaluirane su pre svega pristupima za ocenu kvaliteta sintetizovanog govora. Korišćene su metode koje se baziraju na MOS (engl. *Mean Opinion Score*) testovima i MUSHRA (engl. *Multiple Stimuli Hidden Reference and Anchor*) testovima [45].

U okviru MOS testa učesnici treba neku karakteristiku govora da ocene od 1 (loše) do 5 (odlično). Krajnji rezultat ovog testa je prosečna vrednost ocena svih učesnika.

MUSHRA testovi su prvobitno korišćeni za procenu kvaliteta različitih kodera govora. U sklopu jednog zadatka učesnici prvo treba da preslušaju originalnu rečenicu (jasno naznačenu), a potom i nekoliko rečenica čiji kvalitet se ocenjuje. Među rečenicama čiji kvalitet se ocenjuje data je i originalna rečenica (pri čemu je ova činjenica skrivena od

slušalaca), kao i jedna rečenica znatno lošijeg kvaliteta od ostalih. Zadatak slušalaca je da prvo odrede rečenicu najboljeg kvaliteta u poređenju sa originalnom, a zatim u odnosu na nju da ocene kvalitet ostalih rečenica brojem od 1 do 100.

Subjektivni testovi predstavljaju najbolji način ocene kvaliteta sintetizovanog govora. Njihove glavne nedostatke predstavljaju vreme i resursi neophodni da bi se ovi testovi sproveli.

Treća navedena grupa za ocenu TTS sistema nije široko prihvaćena u literaturi i podrazumeva proveru da li sistem podržava neke od unapred definisanih funkcionalnosti.

2.3. WORLD vokoder

U svim eksperimentima koji su opisani u okviru disertacije korišćen je WORLD vokoder [46]. Funkcionisanje ovog vokodera može se podeliti u tri faze. U prvoj fazi vrši se određivanje osnovne učestanosti. Tokom vremena autori ovog vokodera razvili su nekoliko algoritama za određivanje osnovne učestanosti. Eksperimenti u disertaciji bazirani su na korišćenju DIO algoritma [47]. Druga faza rada vokodera sastoji se u određivanju spektralne obvojnice koristeći prethodno izračunatu osnovnu učestanost [48]. U poslednjoj fazi vrši se određivanje koeficijenata aperiodičnosti korišćenjem D4C algoritma [49].

Spektralna obvojnica koja se dobija u drugoj fazi procesiranja obično je predstavljena sa 513 ili 1025 koeficijenata i zbog visoke dimenzionalnosti nije pogodna za upotrebu u obuci sistema za sintezu govora, stoga se umesto spektralnh obvojnica koristi njihova efikasnija reprezentacija pomoću mel-generalizovanih kepstralnih koeficijenata (engl. *Mel-generalized Cepstral Coefficients,* MGC) [50].

Poređenje WORLD vokodera sa drugim savremenim vokoderima dato je u [51], gde je potvrđeno da verzija vokodera korišćena u eksperimentima opisanim u disertaciji postiže visok kvalitet resintetizovanog govora.

2.4. Definicija ekspresivnog govora

U literaturi ne postoji jedinstvena definicija koja se koristi za označavanje pojma ekspresivnog govora. Ovaj pojam se često upotrebljava i kao sinonim za pojam emotivnog govora.

U [52] predlaže se postojanje 6 osnovnih tipova emocija koje su univerzalne i odgovaraju izrazima lica u različitim kulturama. Navedene emocije su: sreća, ljutnja, tuga, gađenje, iznenađenje i strah. U [53] dalje se diskutuje da svaka od osnovnih emocija takođe ima svoje osobenosti izražene i u karakteristikama govornog signala. Ova činjenica omogućava slepim osobama da prepoznaju emocije u glasu sugovornika. Međutim, postoje autori koji tvrde da se emocije ne mogu jednostavno svrstati u neke diskretne kategorije. Takođe, prethodna kategorizacija onemogućava da se za datu emociju definiše stepen njene izraženosti.

U [54] emocije se definišu kao tačke u trodimenzionalnom prostoru emocija pri čemu se takođe pokušava objasniti veza između dimenzija i određenih karakteristika govora. Dimenzije koje određuju prostor emocija su: aktivacija, dominantnost i valenca. Aktivacija se povezuje sa brzinom govora. Povećana brzina govora odgovara povećanom uzbuđenju govornika. Dominantnost se odnosi na izraženost emocije i povezuje se sa intenzitetom govora i osnovnom učestanošću. Valenca se odnosi na pozitivan, odnosno negativan stav posmatrane emocije. Pozitivan stav može se povezati sa određenim prozodijskim konturama dok za negativni stav ne postoji jasna veza sa određenim prozodijskim elementima.

Prethodno navedena dva primera samo su neki od mnogobrojnih načina klasifikacije emocija. Detaljna analiza klasifikacije emocija može se pronaći u [55].

Pojam ekspresivnog govora je širi od pojma emotivnog govora. Često se kao primeri ekspresivnog govora navode nastupi na javnim skupovima, kao npr. govori političara. Definitivno se i u takvom govoru pronalazi odstupanje prozodijskih i akustičkih elemenata od uobičajenog govora, ali takve promene nisu prouzrokovane emotivnim stanjem govornika nego konkretnim ciljem koji on želi da postigne svojim govorom. Ovde se umesto konkretne emocije izražene govorom može upotrebiti pojam govorni stil. Primeri ekspresivnog govora su i izražavanje sarkazma ili ironije. Različita emotivna stanja govornika mogu da dovedu do

upotrebe ovakvog načina govora, ali će nezavisno od emocije koja ga prouzrokuje manifestacija u govoru biti ista.

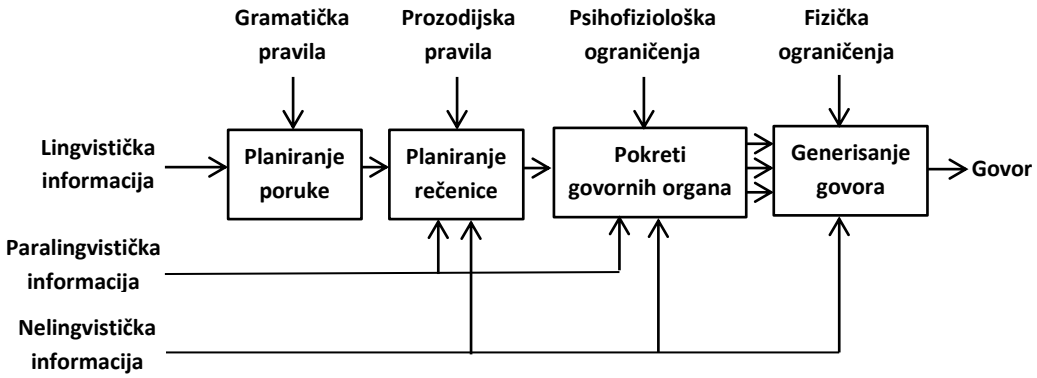
U okviru ovog rada koristiće se definicija predložena u [56] koja pod ekspresivnim govorom smatra sve što nije neutralni govor. Neutralni govor je u [57] definisan kao odsustvo bilo kakve informacije o stanju govornika. U okviru ranije pomenute klasifikacije emocija u trodimenzionalnom prostoru neutralni govor bi predstavljao tačku koja odgovara koordinatnom početku ili jednoj tački iz skupa tačaka čiji su svi elementi u blizini koordinatnog početka prostora emocija.

2.5. Ekspresivni govor u sintezi govora

U [58] navodi se da se informacija sadržana u govoru može posmatrati sa tri stanovišta: lingvističkog, paralingvističkog i nelingvističkog. Lingvistička informacija može se opisati diskretnim skupom simbola i pravilima za njihovo kombinovanje. Ona je najčešće predstavljena pisanim jezikom. Paralingvistička informacija predstavlja sadržaj koji se ne može zaključiti na osnovu teksta i koji dodaje sam govornik. Na primer, jedna rečenica se može izgovoriti na nekoliko različitih načina. Nelingvističke informacije obuhvataju starost, pol i fizičko stanje govornika. Na njih govornik ne može neposredno da utiče. Povezanost ovih tipova informacija i uticaj na prozodijske karakteristike govora prikazani su na slici 2.7 [58]. Sveobuhvatna analiza sinteze ekspresivnog govora bi trebalo u obzir da uzme sve navedene informacije. U okviru ove disertacije razmatrane su, pre svega, paralingvističke informacije.

U [27] navodi se da postoje tri pristupa sintezi ekspresivnog govora: eksplicitna kontrola, reprodukcioni pristup i implicitna kontrola.

U pristupu baziranom na eksplicitnoj kontroli ekspresivni govor se generiše tako što se na neutralnim rečenicama primenjuju prozodijske transformacije koje su formirane na osnovu analize određene ekspresivne baze. U [59] predložena su tri pristupa za transformaciju prozodije neutralne rečenice u prozodiju iste rečenice izgovorene u nekoj drugoj emociji. Prvi predloženi pristup zasniva se na linearnom preslikavanju i može se predstaviti



Slika 2.7 Uticaj različitih izvora informacija na generisanje govora

jednačinom

$$y_{n,i} = \alpha_{n,i} x, \tag{2.6}$$

pri čemu x predstavlja prozodijski element neutralne rečenice, α koeficijent preslikavanja, a y vrednost prozodijskog elementa govora u odgovarajućoj emociji. Indeks n označava konkretnu emociju, a i intenzitet emocije koji može biti visok, srednji i nizak. Autori su pokazali da ovaj pristup nije dovoljno dobar u preslikavanju intenziteta emocije.

Stoga su predložene dve dodatne metode za preslikavanje prozodije. Prva je preslikavanje bazirano na upotrebi smeša Gausovih funkcija raspodele kao što je ranije već predloženo u polju konverzije govornika [60]. U polju konverzije emocija za svaki fonem obučavaju se tri GMM modela preslikavanja koji prozodiju neutralnog fonema preslikavaju u prozodiju emotivnog fonema sa različitim nivoima izraženosti emocije. Ovaj model direktno koristi akustičke parametre bez uključivanja lingvističkih informacija. Druga, dodatna metoda za preslikavanje u proces konverzije uključuje i lingvističke informacije i bazirana je na korišćenju klasifikacionih i regresionih stabala odluke (engl. *Classification and Regression Trees,* CART). U predloženom pristupu ulazni parametri CART modela su lingvistička obeležja, a izlazni parametri su razlike u prozodijskim elementima između neutralne i emotivne rečenice. Eksperimentalno je pokazano da su GMM i CART pristup bolji od linearnog, kao i da je GMM bolji za manje baze ekspresivnog govora, a CART za veće.

U [61] predstavljen je metod za promenu osnovne učestanosti i trajanja baziran na vremenskom skaliranju signala reziduuma dobijenog nakon linearne predikcije, a koji može

uspešno da se primeni i kod transformacije prozodijskih elemenata neutralnog govora za dobijanje ekspresivnog govora.

Formantni sintetizatori su najprikladniji za primenu procenjenih (konvertovanih) prozodijskih elemenata. Neki primeri formantnih sintetizatora primenjenih u sintezi ekspresivnog govora opisani su u [62]–[64]. Kod difonskih sintetizatora primenom odgovarajućih metoda moguće je promeniti trajanje fonema i njegovu osnovnu učestanost. Međutim, nije u potpunosti dokazano da se samo primenom ovih parametara može postići izraženost odgovarajućih emocija. U [65] dat je primer difonskog sintetizatora baziranog na manipulaciji prozodijskim parametrima.

Detaljan pregled sinteze ekspresivnog govora korišćenjem eksplicitnog pristupa može se pronaći u [27], [66].

U reprodukcionom pristupu ekspresivni govor generiše se direktnim korišćenjem odgovarajuće baze ekspresivnog govora. U [67] predstavljen je sintetizator koji koristi selekciju segmenata, a koji, pored neutralnog govora, može da generiše govor u jednoj od tri emocije: uživanje, ljutnja i tuga. Za generisanje govora u datoj emociji koristi se odgovarajuća baza emotivnog govora koja sadrži otprilike sat vremena govornog materijala po emociji. Kada frontend detektuje da neka rečenica odgovara određenoj emociji, tada se u procesu selekcije optimalne sekvence segmenata koristi samo odgovarajući emotivni deo baze. Za razliku od ovog pristupa koji za sintezu govora u određenoj emociji koristi posebnu bazu, u [68] opisan je sintetizator koji u procesu generisanja govora koristi bazu u kojoj su pomešani neutralni i emotivni govor. Autori ove metode polaze od pretpostavke da u rečenici koja treba da bude generisana u nekoj emociji samo određene reči moraju odgovarati datom stilu (emociji), a ostale mogu pripadati i neutralnom stilu. U algoritmu pretrage baze ne koristi se samo informacija da neka rečenica mora biti sintetizovana u određenoj emociji (i da se shodno tome pretražuje samo taj deo govorne baze), već se svakoj reči u ciljnoj rečenici pridružuje i vrednost koja se iščitava iz rečnika afekta, a koja predstavlja verovatnoću da bi data reč trebalo da bude sintetizovana u nekoj emociji. Sa pomenutom vrednošću proširuje se cena nekog segmenta, odnosno ako je frontend definisao rečenicu kao srećnu, a neka reč ima visoku vrednost pomenute veličine iz rečnika, tada ona dobija manju cenu, za razliku od situacije kada se emocija rečenice i emocija reči ne poklapaju. Ako se emocije reči i rečenice

ne poklapaju, algoritam selekcije forsira odabir neutralnih segmenata. U [69] opisan je pristup koji takođe koristi mešovitu bazu neutralnog i ekspresivnog govora. U ovom pristupu prvo se kreiraju stabla odluke korišćenjem samo dela baze koji odgovara određenom stilu. U procesu pretrage cena segmenta se dodatno povećava ukoliko se ciljni i segment kandidati ne slažu po pitanju emotivnog sadržaja.

Generisanje ekspresivnog govora primenjeno je i u statističko-parametarskim pristupima sintezi. U [70] opisane su dve metode zasnovane na korišćenju HMM sinteze. Prva metoda podrazumeva da se za svaki stil formira zaseban akustički model. U drugoj opisanoj metodi informacija o stilu zadaje se kao proširenje ulazne lingvističke informacije. Pokazano je da oba metoda postižu približno iste rezultate.

U [71] dato je poređenje konkatenativnog i HMM pristupa u sintezi ekspresivnog govora. Pokazano je da oba pristupa postižu jednak kvalitet u sintezi emocija. Konkatenativni pristup postiže bolji rezultat po pitanju izraženosti emocije, dok je HMM pristup, što je i očekivano, prikladniji za manipulisanje nivoom izraženosti neke emocije. Jedan od zaključaka u ovom istraživanju jeste i nemogućnost oba pristupa da kvalitetno sintetizuju govor koji pripada određenim emocijama.

Implicitna sinteza ekspresivnog govora koristi se u statističkim pristupima, a podrazumeva interpolaciju između više modela. U [72] opisan je postupak kreiranja novog TTS glasa ili glasa u određenom stilu primenom adaptacionih tehnika koje za polazni model usvajaju model prosečnog govornika. Pristup baziran na adaptaciji modela predstavljen je i u [73], pri čemu su opisani i postupci koji omogućavaju kontrolu nivoa ekspresivnosti korišćenjem vektora stila.

U vreme početka rada na ovoj disertaciji sinteza ekspresivnog govora korišćenjem DNN pristupa skoro da i nije bila zastupljena u literaturi.

2.6. Baze ekspresivnog govora

Kao što je spomenuto u uvodnom poglavlju, jedan od problema u sintezi ekspresivnog govora predstavlja i dostupnost odgovarajućih baza. U [74] navode se pristupi koji se koriste

pri kreiranju emotivnih govornih baza. Materijal se najčešće dobija snimanjem govora profesionalnih glumaca. Ovakav način skupljanja govornog materijala omogućava da se rečenica sa istim sadržajem snima u različitim emocijama što može biti izuzetno značajno u određenim istraživanjima. Takođe, mogu se snimiti i rečenice sa ekstremnim manifestacijama određene emocije. Nedostatak ovakvog pristupa predstavlja činjenica da će glumci često iskoristiti stereotipne realizacije određene emocije i da se u snimljenom materijalu neće reflektovati sve karakteristike te emocije kao u slučaju spontanog govora.

Varijaciju prethodno opisanog pristupa predstavlja snimanje emotivnog govora pri čemu glumci čitaju tekst čiji sadržaj odgovara željenoj emociji. Istraživanja su pokazala da je ljudima u takvoj situaciji lakše da predstave odgovarajuću emociju. Ovo istovremeno onemogućava snimanje rečenica različitog emotivnog, a istog fonetskog sadržaja.

Pobuđivanje emocija u laboratorijskim uslovima može se koristiti da bi se kod ispitanika indukovale određene emocije, ali nije moguće pobuđivanje svih emocija ili ekstremnih manifestacija određenih emocija zbog etičkih razloga. Pregled različitih pristupa u pobuđivanju emocija može se pronaći u [74].

Emocije iskazane na najprirodniji način dobijaju se snimanjem interakcija između ljudi, kao što su npr. razgovori u pozivnim centrima ili gostovanja u različitim radio ili televizijskim emisijama. Iako koristan zbog prirodnosti izraženih emocija, problem sa ovakvim pristupom predstavlja činjenica da je ipak snimljen u nekontrolisanim uslovima, što otežava obradu prikupljenih podataka.

Pregled dostupnih emotivnih baza govora dat je u [75], dok su u [65] i [66] opisani postupci u snimanju ekspresivnih baza spontanog govora.

Govor sakupljen prethodno opisanim pristupima može uspešno biti primenjen u analizi akustičkih i artikulacionih karakteristika ekspresivnog govora. Međutim, za dobijanje sintetizovanog govora visokog kvaliteta koji bi uspešno mogao biti primenjen u različitim domenima nijedan opisani pristup nije dobar. Za potrebe sinteze govora najčešće se koriste odglumljene govorne baze, ali često se dešava da je u takvim bazama jedan govornik izgovarao samo nekoliko rečenica pridruženih određenom stilu i da se takve baze ne mogu koristiti za potrebe TTS-a.

Primer emotivne baze koja je istovremeno dobra i za izučavanje karakteristika emotivnog govora, ali i za sintezu, jeste španska emotivna govorna baza koja se naziva SEV (engl. *Spanish Emotional Voices*) [78]. Ova multimodalna baza, koja pored govora sadrži i video materijal, sastoji se od materijala koji su snimili muški i ženski govornik u neutralnom stilu i 6 osnovnih emocija prema Ekmanu [53]. Svakoj emociji odgovara približno dva sata materijala. Baza je automatski prozodijski anotirana i fonetski poravnata. Izvršena je i perceptualna evaluacija baze. Ispitanici su imali zadatak da prepoznaju emociju rečenice iz baze i da svakoj rečenici dodele oznaku nivoa izraženosti emocije koja može da ima vrednosti: veoma nizak, nizak, prosečan, visok i veoma visok. Rezultati prepoznavanja su pokazali da se u 85% rečenica prepoznata emocija poklapa sa emocijom koja je zaista trebalo da bude predstavljena u testiranoj rečenici. U 65% rečenica nivo emocije označen je kao visok ili veoma visok. Baza je uspešno primenjena za potrebe kreiranja ekspresivnog TTS-a [71].

U [67] opisano je kreiranje govorne baze za potrebe konkatenativne sinteze emotivnog govora u japanskom jeziku. U bazi postoji materijal snimljen za muškog i ženskog govornika. Izabrani su amaterski govornici, jer su istraživanja pokazala da radio govornici ili profesionalni glumci teže ka preteranoj ekspresivnosti. Baza se sastoji od 50-ak minuta materijala za svaku od tri emocije: užitak, ljutnja i tuga.

Snimanje govorne baze za potrebe generisanja odgovarajućeg govornog stila predstavljeno je i u [9]. Stil koji se koristi zasniva se na potrebi generisanja govornih poruka u kritičnim situacijama. Autori predlažu snimanje iste rečenice u nekoliko različitih nivoa izraženosti stila, što bi omogućilo korišćenje odgovarajućeg nivoa jačine poruke u zavisnosti od ozbiljnosti situacije u kojoj se poruka koristi. Adekvatnost snimljene baze za potrebe sinteze govora demonstrirana je konstruisanjem HMM sintetizatora. Jedan od zaključaka u istraživanju je bio da HMM ne može identično da modeluje sve razlike u nivou izraženosti stila u originalnom govoru u poređenju sa sintetizovanim govorom.

Jedan od načina za kreiranje govornih baza za potrebe ekspresivne sinteze govora jeste i korišćenje audio knjiga. U [79] opisan je postupak kreiranja takve baze koja se koristi za formiranje konkatenativnog sintetizatora. Prvo se u govornom materijalu označe uloge koje tokom čitanja govornik tumači. Nakon poravnanja baze korišćenjem pristupa iz automatskog

prepoznavanja govora, za svaku ulogu formira se GMM model. Pošto nisu sve uloge predstavljene istom količinom materijala, a i neke uloge su međusobno slične, vrši se klasterizacija uloga korišćenjem algoritma sličnog *k-means* algoritmu [80]. Dobijeni klasteri se potom smatraju različitim govornim stilovima. Svaki govorni stil je nakon preslušavanja klasterizovanih primeraka opisan odgovarajućim karakteristikama govora. Perceptualni testovi sa rečenicama sintetizovanim u datim stilovima pokazali su dobre rezultate u prepoznavanju klasterizovanih stilova na osnovu opisanih karakteristika.

Za razliku od prethodno opisane metode koja je bazirana na korišćenju audio knjiga snimanih u profesionalnom okruženju u [81] prikazano je kreiranje baze za potrebe HMM sinteze na bazi audio knjiga koje su snimali amaterski govornici. Materijal snimljen od strane amaterskih govornika ima nekoliko nedostataka: govor je često nekonzistentan, kvalitet korišćenih mikrofona je prilično nizak, a često postoji i određena pozadinska buka u okruženju u kojem je materijal sniman. Klasterizacija uloga odrađena je korišćenjem glotalnih karakteristika govora. Obuka HMM sintetizatora bazira se na adaptaciji modela prosečnog govornika na materijal dobijen nakon postupka klasterizacije. Testovi slušanja su pokazali da su slušaoci bili u mogućnosti da razlikuju dobijene stilove, kao i da stilovi odgovaraju sadržaju rečenica.

3. HMM sinteza

Modeli koji se koriste za opisivanje signala mogu se podeliti u dve grupe: determinističke i statističke [82]. Primer determinističkog modela jeste predstavljanje signala u obliku sume sinusa i kosinusa. Statistički modeli polaze od pretpostavke da se signali mogu predstaviti kao slučajni procesi. U ovu grupu modela spadaju i skriveni Markovljevi modeli (HMM).

3.1. Teorijske osnove skrivenih Markovljevih modela

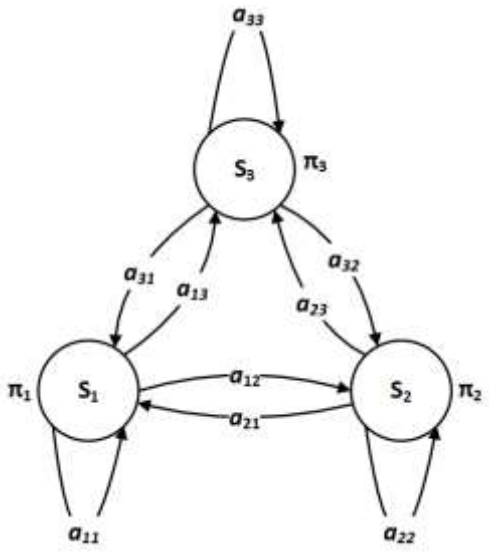
Na slici 3.1 prikazan je sistem koji u diskretnim vremenskim trenucima $t = 0, 1, 2 \ldots$ prelazi iz jednog stanja u neko drugo pri čemu je skup mogućih stanja konačan, $S = \{s_1, s_2, \ldots, s_N\}$. Neka je sa q_t označena slučajna promenljiva koja predstavlja stanje u kojem se nalazi sistem u momentu $t = 0,1,2 \ldots$. Ukoliko stanje sistema u budućnosti zavisi samo od stanja u kojem se sistem trenutno nalazi, a ne i od toga kako je sistem došao u to stanje, stohastički proces se naziva lancem Markova. Dakle, za lanac Markova važi

$$P\left(q_t = s_j \middle| q_0 = s_{i_0}, q_1 = s_{i_1,\ldots,} \, q_{t-1} = s_i\right) = P\left(q_t = s_j \middle| q_{t-1} = s_i\right). \qquad (3.1)$$

Verovatnoća iz jednačine (3.1) naziva se verovatnoćom prelaza i označava sa a_{ij}[1], pri čemu važi

$$\sum_{j=1}^{N} a_{ij} = 1. \qquad (3.2)$$

[1] Pretpostavka je da je sistem homogen odnosno da su verovatnoće prelaza a_{ij} invarijantne u odnosu na translaciju vremena



Slika 3.1 Primer Markovljevog modela

Pored matrice prelaza $\boldsymbol{A}$ za definisanje sistema potrebno je i poznavanje vektora verovatnoća, čiji elementi predstavljaju verovatnoće nalaženja sistema u jednom od datih stanja

$$\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_N], \tag{3.3}$$

pri čemu je

$$\pi_i = P(q_0 = s_i),\ i = 1, 2, \dots, N, \tag{3.4}$$

i važi

$$\sum_{i=1}^{N} \pi_i = 1. \tag{3.5}$$

Prethodno opisani sistem naziva se vidljivi Markovljev model, s obzirom na to da je u svakom trenutku poznato u kojem se stanju sistem nalazi. Ovakav model je koristan kada je potrebno izračunati verovatnoću sekvence nekih događaja.

U primenama se često dešava da nisu poznata stanja u kojem se nalazi sistem već opservacije koje u svakom stanju bivaju emitovane sa određenom verovatnoćom koja se naziva emisionom verovatnoćom

$$b_i(\boldsymbol{o}) = P(\boldsymbol{o} \text{ u trenutku } t | q_t = s_i), \quad \boldsymbol{o} = [o_1, o_2, \dots, o_M]^T. \tag{3.6}$$

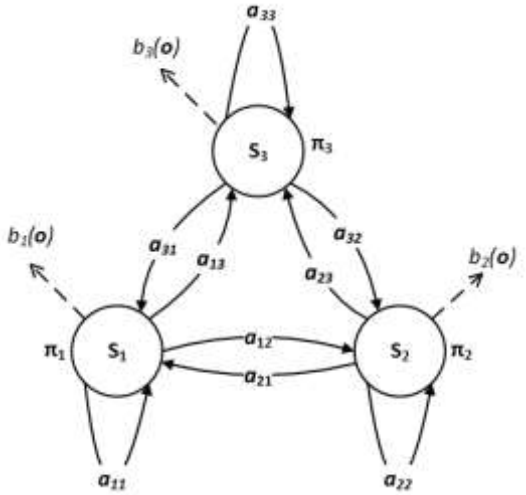
Na primer, prilikom prepoznavanja govora poznata su određena akustička obeležja govora dok su fonemi koji su opisani tim obeležjima skriveni. Ovakvi sistemi mogu se opisati skrivenim Markovljevim modelima. Pored matrice verovatnoća prelaza **A**, vektora početnih verovatnoća stanja $\boldsymbol{\pi}$, skriveni Markovljev model opisan je i zakonom raspodele verovatnoća pridruženom svakom stanju

$$\boldsymbol{B} = \{b_i(\boldsymbol{o})\}, i = 1,2, \dots, N, \tag{3.7}$$

gde je *N* broj stanja u sistemu. Radi jednostavnosti koristi se sledeća notacija za označavanje skrivenog Markovljevog modela

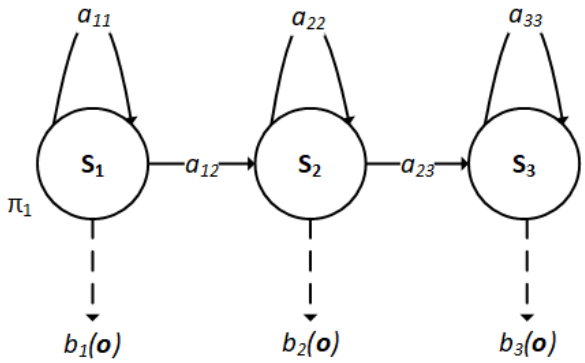
$$\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}). \tag{3.8}$$

Na slici 3.2 dat je jedan primer skrivenog Markovljevog modela. Kod ovog modela iz jednog stanja se može preći u sva ostala stanja (ili ostati u trenutnom stanju). Kod ovog modela za sve elemente a_{ij} važi da su različiti od nule[2]. Međutim, često se dešava da su pojedini elementi a_{ij} jednaki nuli, odnosno da sistem iz jednog stanja ne može preći u neka druga stanja u samo jednom koraku. Primer takvog sistema, a koji je pogodan za



Slika 3.2 Ergodičan skriveni Markovljev model

[2] Ovo je primer ergodičnog sistema. Kod ergodičnih sistema se u konačnom broju prelaza iz jednog stanja može doći u sva ostala stanja.



Slika 3.3 Primer skrivenog Markovljevog modela korišćenog u modelovanju govora

modelovanje govora, dat je na slici 3.3.

Opservacije koje HMM emituje u svakom stanju mogu da pripadaju diskretnom skupu, ali su u modelovanju govora predstavljene funkcijama raspodele neprekidnog tipa. Odgovarajuća funkcija gustine verovatnoće najčešće se opisuje kao suma Gausovih funkcija raspodele gustine verovatnoće

$$b_i(\boldsymbol{o}) = \sum_{j=1}^{M} w_{ij}\mathcal{N}(\boldsymbol{o}; \mu_{\boldsymbol{ij}}, \sum_{ij}), \tag{3.9}$$

gde je M broj smeša, w_{ij} težina smeše, $\mu_{\boldsymbol{ij}}$ vektor srednjih vrednosti, a $\sum_{ij}$ kovarijansna matrica M-dimenzionalne Gausove gustine verovatnoće

$$\mathcal{N}(\boldsymbol{o}; \mu_{\boldsymbol{ij}}, \sum_{ij}) = \frac{1}{\sqrt{(2\pi)^M |\sum_{ij}|}} e^{-\frac{1}{2}(\boldsymbol{o}-\mu_{\boldsymbol{ij}})^T \sum_{ij}^{-1} (\boldsymbol{o}-\mu_{\boldsymbol{ij}})}. \tag{3.10}$$

Težine smeša treba da zadovolje sledeće uslove [83]

$$\sum_{j=1}^{M} w_{ij} = 1, \tag{3.11}$$

$$w_{ij} \geq 0. \tag{3.12}$$

3.2. Tri osnovna problema u primeni HMM

Primena HMM modela zahteva poznavanje rešenja za tri problema [82]:

1. određivanje verovatnoće sekvence opservacija $\boldsymbol{O}$ ukoliko je poznat HMM model $\boldsymbol{\lambda}$,
2. određivanje optimalne sekvence stanja ako je data sekvenca opservacija i poznat je model $\boldsymbol{\lambda}$,
3. određivanje optimalnih parametara modela $\boldsymbol{\lambda}$ ukoliko je poznata sekvenca opservacija.

3.2.1. Računanje verodostojnosti

Postupak određivanja verovatnoće sekvence opservacija $\boldsymbol{O} = \boldsymbol{o_1 o_2} \dots \boldsymbol{o_T}$ ukoliko su poznati parametri modela $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ u literaturi se još naziva i računanje verodostojnosti [84]. Da bi se ovaj problem rešio obično se polazi od jednostavnije pretpostavke da je pored sekvence opservacija poznata i sekvenca stanja iz kojih su emitovane opservacije, $\boldsymbol{Q} = q_1 q_2 \dots q_T$. U tom slučaju verovatnoća sekvence opservacija računa se na sledeći način

$$P(\boldsymbol{O}|\boldsymbol{Q}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} P(\boldsymbol{o_t}|q_t, \boldsymbol{\lambda}) = \prod_{t=1}^{T} b_{q_t}(\boldsymbol{o_t}) . \tag{3.13}$$

.

Verovatnoća sekvence stanja, ukoliko su poznati parametri modela, definisana je pomoću verovatnoća prelaza i vektora početnih verovatnoća stanja

$$P(\boldsymbol{Q}|\,\boldsymbol{\lambda}) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T}. \tag{3.14}$$

Korišćenjem Bajesove teoreme, združena verovatnoća sekvence stanja i sekvence opservacija data je izrazom

$$P(\boldsymbol{O}, \boldsymbol{Q}|\,\boldsymbol{\lambda}) = P(\boldsymbol{O}|\boldsymbol{Q}, \boldsymbol{\lambda}) P(\boldsymbol{Q}|\,\boldsymbol{\lambda}). \tag{3.15}$$

Sumiranjem jednakosti (3.15) po svim mogućim sekvencama $\boldsymbol{Q}$ dobija se verovatnoća sekvence opservacija ako su poznati parametri modela

$$P(\boldsymbol{O}|\boldsymbol{\lambda}) = \sum_{\forall \boldsymbol{Q}} P(\boldsymbol{O}|\boldsymbol{Q},\boldsymbol{\lambda})(P(\boldsymbol{Q}|\,\boldsymbol{\lambda}) = \sum_{\forall q_1 q_2 \dots q_T} \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T} \prod_{t=1}^{T} b_{q_t}(\boldsymbol{o}_t)\,. \tag{3.16}$$

Računanje vrednosti date izrazom (3.16) zahteva korišćenje približno $2TN^T$ računskih operacija [82], što bi i za neke male vrednost N i T bilo nemoguće izračunati u realnom vremenu. Pojednostavljenje u izračunavanju postiže se definisanjem promenjive $\alpha_t(i)$ koja predstavlja verovatnoću da se sistem u trenutku *t* nađe u stanju s_i i da do tog trenutka emituje sekvencu opservacija $\boldsymbol{o_1 o_2} \dots \boldsymbol{o_t}$, ukoliko su poznati parametri modela

$$\alpha_t(i) = P(\boldsymbol{o_1 o_2} \dots \boldsymbol{o_t}, q_t = s_i|\boldsymbol{\lambda}). \tag{3.17}$$

Korišćenjem izraza definisanog u (3.17) izračunavanje verovatnoće neke sekvence opservacija vrši se korišćenjem algoritma unapred (engl. *forward algorithm*) koji se sastoji od tri dela:

1. **inicijalizacija**

$$\alpha_1(i) = \pi_i b_i(\boldsymbol{o_1}),\ \ 1 \le i \le N, \tag{3.18}$$

2. **indukcija**

$$\alpha_{t+1}(i) = b_i(\boldsymbol{o_{t+1}}) \sum_{j=1}^{N} \alpha_t(i) a_{ji}\,,\quad 1 \le i \le N,\ \ 2 \le t \le T, \tag{3.19}$$

3. **terminacija**

$$P(\boldsymbol{o_1 o_2} \dots \boldsymbol{o_T}|\boldsymbol{\lambda}) = \sum_{i=1}^{N} \alpha_T(i). \tag{3.20}$$

Približan broj izračunavanja koje treba izvršiti prethodno opisanim algoritmom je N^2T [82], što je znatno manje nego u direktnom pristupu.

3.2.2. Računanje najverovatnije sekvence stanja

Procedura traženja optimalne sekvence stanja kroz koju je sistem prošao prilikom emitovanje sekvence opservacija $\boldsymbol{O}$ naziva se i dekodovanje. Ovo procedura naročito je značajna u procesu prepoznavanja govora. Postoji više različitih kriterijuma koji definišu optimalnu sekvencu stanja, ali se najčešće optimalna sekvenca stanja $\boldsymbol{Q}^*$ definiše kao sekvenca koja daje najveću vrednost združene verovatnoće sekvence stanja i sekvence opservacija

$$\boldsymbol{Q}^* = arg \max_{\boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q}|\boldsymbol{\lambda}). \tag{3.21}$$

Optimizacioni problem predstavljen u jednakosti (3.21) rešava se primenom dinamičkog algoritma koji se naziva Viterbijev (*Andrew Viterbi*) algoritam, a čiji koraci su dati u nastavku:

1. inicijalizacija

$$\delta_1(i) = \pi_i b_i(\boldsymbol{o_1}), \;\; 1 \le i \le N, \tag{3.22}$$

$$\varphi_1(i) = 0, \;\; 1 \le i \le N, \tag{3.23}$$

2. rekurzija

$$\delta_t(i) = \max_{1 \le j \le N} \delta_{t-1}(j) a_{ji} b_i(\boldsymbol{o_t}), \;\; 1 \le i \le N, \;\; 2 \le T \le T, \tag{3.24}$$

$$\varphi_t(i) = arg \max_{1 \le j \le N} \delta_{t-1}(j) a_{ji} b_i(\boldsymbol{o_t}), \;\; 1 \le i \le N, \tag{3.25}$$

3. terminacija

$$P^* = \max_{1 \le i \le N} \delta_t(i), \;\; 1 \le i \le N, \tag{3.26}$$

$$q_T^* = \max_{1 \le i \le N} \delta_t(i), \;\; 1 \le i \le N, \tag{3.27}$$

4. nalaženje optimalne sekvence

$$q_t^* = \varphi_{t+1}(q_{t+1}^*), \;\; t = T-1, T-2, \dots, 1. \tag{3.28}$$

Veličina $\delta_t(i)$ predstavlja maksimalnu združenu verovatnoću neke sekvence stanja do trenutka *t* u kojem se sistem nalazi u stanju s_i i sekvence opservacija

$$\delta_t(i) = \max_{q_1 q_2 \dots q_{t-1}} P(q_1 q_1 \dots q_t = s_i, \boldsymbol{o_1 o_2} \dots \boldsymbol{o_t}|\boldsymbol{\lambda}). \quad (3.29)$$

3.2.3. Nalaženje optimalnih parametara HMM modela

Proces traženja optimalnih parametara modela koji je emitovao neku sekvencu opservacija naziva se i obuka HMM modela. Ova procedura značajna je i za prepoznavanje govora, ali i za sintezu. Optimalni parametri modela obično se definišu na osnovu kriterijuma maksimalne verodostojnosti

$$\boldsymbol{\lambda}^* = arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{o}|\boldsymbol{\lambda}) = arg\max_{\boldsymbol{\lambda}} \sum_{\forall \boldsymbol{q}} P(\boldsymbol{o}, \boldsymbol{q}|\boldsymbol{\lambda}). \quad (3.30)$$

Nažalost, postupak za pronalaženje globalnog maksimuma optimizacionog problema predstavljenog u jednačini (3.30) ne postoji, ali je definisan postupak koji pronalazi lokalne maksimume koji je poznat pod nazivom Baum-Velč (*Baum-Welch*) algoritam[3] [85]. Za definisanje ovog algoritma neophodno je uvođenje pomoćnih promenljivih. Prva od njih je verovatnoća da je sistem emitovao deo sekvence $\boldsymbol{o_{t+1} o_{t+2}} \dots \boldsymbol{o_T}$, ako se u trenutku *t* nalazio u stanju s_i i ako su poznati parametri modela

$$\beta_t(i) = P(\boldsymbol{o_{t+1} o_{t+2}} \dots \boldsymbol{o_T}, |q_t = s_i, \boldsymbol{\lambda}). \quad (3.31)$$

Zatim se definiše verovatnoća da se sistem u trenutku *t* nalazi u stanju s_i, a u trenutku *t*+1 u stanju s_j, ako su poznati cela sekvenca opservacija i parametri modela

$$\xi_t(i,j) = P\big(q_t = s_i, q_{t+1} = s_j \big| \mathbf{o}, \boldsymbol{\lambda}\big). \quad (3.32)$$

Verovatnoća $\xi_t(i,j)$ se pomoću α_t i β_t može predstaviti u obliku [82]

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(\boldsymbol{o_{t+1}}) \beta_t(j)}{\sum_{k=1}^{N}\sum_{l=1}^{N} \alpha_t(k) a_{kl} b_l(\boldsymbol{o_{t+1}}) \beta_t(l)}. \quad (3.33)$$

Poslednja pomoćna veličina koja se definiše jeste verovatnoća da se sistem u trenutku *t* nalazi u stanju s_i ako su dati parametri modela i opservaciona sekvenca

[3] Može se pokazati da se reestimacione formule koje će biti date u nastavku takođe mogu dobiti primenom algorima maksimizacije očekivanja (EM algoritam) [136]

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j). \tag{3.34}$$

Pokazuje se da se formule za računanje parametara modela mogu iskazati u obliku [82]

$$\pi_i = \gamma_1(i), \tag{3.35}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{3.36}$$

$$b_i(k) = \frac{\sum_{t=1}^{T} \gamma_t^k(i)}{\sum_{t=1}^{T} \gamma_t(i)}. \tag{3.37}$$

Jednačina (3.37) predstavlja verovatnoću da sistem u stanju s_i emituje diskretni simbol *k*. Sa $\gamma_t^k(i)$ označena je verovatnoća da se sistem u trenutku *t* nalazi u stanju s_i i da emituje simbol *k*. Međutim, za korišćenje HMM modela u obradi govora mnogo je značajniji slučaj kada su izlazne verovatnoće opisane Gausovom raspodelom[4] i tada se srednja vrednost i kovarijansna matrica računaju prema formulama [83]

$$\boldsymbol{\mu_i} = \frac{\sum_{t=1}^{T} \gamma_t(i)\, \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{3.38}$$

$$\sum_i = \frac{\sum_{t=1}^{T} \gamma_t(i)\, (\boldsymbol{o}_t - \boldsymbol{\mu_i})(\boldsymbol{o}_t - \boldsymbol{\mu_i})^T}{\sum_{t=1}^{T} \gamma_t(i)}. \tag{3.39}$$

3.3. Problem modelovanja osnovne učestanosti

U odeljku 3.1 objašnjeno je da se verovatnoće emitovanja opservacija prilikom korišćenja HMM modela u obradi govora najčešće predstavljaju korišćenjem sume Gausovih funkcija gustine verovatnoće. Međutim, za modelovanje govora za potrebe sinteze neophodno je i korišćenje osnovne učestanosti koja u bezvučnim regionima nije definisana, što predstavlja problem u primeni konvencionalnih HMM modela. Ovaj problem, u domenu HMM sinteze, prevaziđen je modelovanjem osnovne učestanosti pomoću verovatnoće na uniji prostora verovatnoća [86]

[4] Radi jednostavnosti je pretpostavljeno da su verovatnoće $b_i(\boldsymbol{o})$ modelovane samo jednom Gausovom raspodelom, iako analogni izrazi postoje i za slučaj sume Gausovih raspodela.

$$\Omega = \bigcup_{g=1}^{G} \Omega_g, \tag{3.40}$$

pri čemu je Ω_g n_g-dimenzionalan realan prostor R^{n_g}. Ukoliko je n_g jednako nuli smatra se da u datom prostoru postoji samo jedna tačka.

U novodobijenom prostoru verovatnoća čiji skup elementarnih događaja je skup Ω iz (3.41) i skupovi Ω_g su slučajni događaji i važi

$$w_g = P(\Omega_g) \tag{3.41}$$

pri čemu važi

$$\sum_{g=1}^{G} w_g = 1 \tag{3.42}$$

i P iz (3.41) je verovatnoća na novodobijenom prostoru verovatnoća.

Svaki događaj koji pripada prostoru Ω definiše se pomoću uređenog para $\boldsymbol{o}$ skupa indeksa X kojim je određen skup $\{\Omega_g\}$ kojima pripada posmatrani događaj i neprekidne slučajne promenjive $\boldsymbol{x} \in R^n$, tj.

$$\boldsymbol{o} = (X, \boldsymbol{x}), \tag{3.43}$$

pri čemu je n dimenzionalnost najvećeg potprostora koji pripada skupu X.

Verovatnoća događaja $\boldsymbol{o}$ definiše se na sledeći način

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g \mathcal{N}_g\big(V(\boldsymbol{o})\big), \qquad S(\boldsymbol{o}) = X, V(\boldsymbol{o}) = \boldsymbol{x}. \tag{3.44}$$

U izrazu (3.44) $\mathcal{N}_g$ predstavlja funkciju gustine verovatnoće definisanu za odgovarajući potprostor Ω_g.

U [86] dati su izrazi koji se koriste za rešavanje osnovnih problema HMM modela (opisanih u odeljcima 3.2.1, 3.2.2 i 3.2.3) u slučaju kada se verovatnoća emitovanja opservacija opisuje verovatnoćom nad unijom prostora verovatnoća.

Za modelovanje sekvence opservacija osnovne učestanosti koristi se verovatnoća nad unijom prostora verovatnoća. Opservacije osnovne učestanosti koje odgovaraju zvučnim regionima pripadaju prostorima koji su opisani funkcijom gustine jednodimenzionalne slučajne promenjive, a „opservacije" pridružene bezvučnim regionima pripadaju potprostoru za koji važi $n_g = 0$, odnosno u kom postoji samo jedna tačka. Tada važi

$$S(\boldsymbol{o_t}) = \begin{cases} \{1,2,\dots G-1\}, & \text{zvučno} \\ \{G\}, & \text{bezvučno} \end{cases}. \tag{3.45}$$

U modelovanju osnovne učestanosti za zvučne regione obično se koristi samo jedan potprostor. Tada u jednačini (3.45) *G* ima vrednost dva.

3.4. Generisanje parametara

HMM modeli prvobitno su se koristili za prepoznavanje govora, za šta se koristi algoritam koji je opisan u odeljku 3.2.2. Međutim, u domenu primene HMM modela za sintezu govora javlja sa nova vrsta problema – za poznate parametre modela neophodno je odrediti najverovatniju sekvencu opservacija dužine *T*, $\widehat{\boldsymbol{O}} = [\boldsymbol{o_1}, \boldsymbol{o_2} \dots \boldsymbol{o_T}]^T$. U nastavku će biti opisan algoritam koji se koristi za ove potrebe pod pretpostavkom da je poznata sekvenca stanja $\boldsymbol{q} = [q_1, q_2 \dots q_T]$. Tada se problem nalaženja optimalne sekvence može definisati na sledeći način

$$\widehat{\boldsymbol{O}} = arg \max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{q}, \boldsymbol{\lambda}). \tag{3.46}$$

Verovatnoća $P(\boldsymbol{O}|\boldsymbol{q}, \boldsymbol{\lambda})$ može se izračunati korišćenjem izraza

$$P(\boldsymbol{O}|\boldsymbol{q}, \boldsymbol{\lambda}) = b_{q_1}(\boldsymbol{o_1}) b_{q_2}(\boldsymbol{o_2}) \dots b_{q_T}(\boldsymbol{o_T}). \tag{3.47}$$

Radi jednostavnosti pretpostavlja se da su gustine verovatnoće $b_j(\boldsymbol{o})$ predstavljene samo jednom Gausovom raspodelom, odakle sledi

$$b_{q_i}(\boldsymbol{o_i}) = \frac{1}{\sqrt{(2\pi)^M \left|\sum_{q_i}\right|}} e^{-\frac{1}{2}(\boldsymbol{o_i} - \mu_{\boldsymbol{q_i}})^T \sum_{\boldsymbol{q_i}}^{-1} (\boldsymbol{o_i} - \mu_{\boldsymbol{q_i}})}. \tag{3.48}$$

U tom slučaju verovatnoća $P(\boldsymbol{O}|\boldsymbol{q},\boldsymbol{\lambda})$ može se predstaviti kao

$$P(\boldsymbol{O}|\boldsymbol{q},\boldsymbol{\lambda}) = \frac{1}{\sqrt{(2\pi)^M \prod_{t=1}^{T}|\sum_{q_t}|}} e^{-\frac{1}{2}(\boldsymbol{O}-\boldsymbol{U})^T \boldsymbol{G}^{-1}(\boldsymbol{O}-\boldsymbol{U})}, \tag{3.49}$$

gde je

$$\boldsymbol{U} = [\mu_{\boldsymbol{q_1}}, \mu_{\boldsymbol{q_2}} \dots, \mu_{\boldsymbol{q_T}}]^{\mathrm{T}}, \tag{3.50}$$

$$\boldsymbol{G} = diag[\sum_{q_1}, \sum_{q_2,\dots,}.\sum_{q_T,}]. \tag{3.51}$$

Kada se na jednakost (3.49) primeni operacija logaritma, a što pojednostavljuje nalaženje maksimuma, dobija se

$$\ln P(\boldsymbol{O}|\boldsymbol{q},\boldsymbol{\lambda}) = -\frac{1}{2}(\boldsymbol{O}-\boldsymbol{U})^T \boldsymbol{G}^{-1}(\boldsymbol{O}-\boldsymbol{U}) - \frac{1}{2}MT\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln|\boldsymbol{G}|. \tag{3.52}$$

Očigledno da prilikom pronalaženja sekvence $\boldsymbol{O}$ koja maksimizuje vrednost datu izrazom (3.52) drugi i treći član neće uticati na krajnji rezultat, jer su i L i M fiksirani, kao i sekvenca stanja, a samim tim i kovarijansne matrice $\sum_{q_t}$. Maksimalna vrednost izraza (3.52) dobija se kada je sekvenca $\boldsymbol{O}$ jednaka sekvenci srednjih vrednosti Gausovih raspodela odgovarajućih stanja $\boldsymbol{U}$. Ovakav izbor parametara loše utiče na kvalitet sintetizovanog govora zbog naglih skokova prilikom prelaska iz jednog stanja u drugo. Stoga je u [87] predloženo da se opservacijama koje emituju pojedina stanja HMM pored statičkih obeležja $\boldsymbol{c}_t$ dodeljuju i dinamička obeležja, koja se još nazivaju i delta obeležjima, a koja se računaju na osnovu jednakosti

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L}^{L} w_\tau \boldsymbol{c}_{t+\tau}. \tag{3.53}$$

Vektor opservacija sada se može predstaviti kao

$$\boldsymbol{o_t} = [\boldsymbol{c}_t, \Delta\boldsymbol{c}_t]^T. \tag{3.54}$$

Veza između niza svih opservacija i niza statičkih obeležja može se prikazati izrazom

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{c}, \tag{3.55}$$

gde je

$$\boldsymbol{W} = [\boldsymbol{w_1}, \boldsymbol{w_2}, \dots \boldsymbol{w_T}]^{\boldsymbol{T}}, \tag{3.56}$$

$$\boldsymbol{w_i} = \left[\boldsymbol{w_i^0}, \boldsymbol{w_i^1}\right]^{\boldsymbol{T}}, \tag{3.57}$$

$$\boldsymbol{w_i^0} = \left[\underbrace{\boldsymbol{0}, \boldsymbol{0}, \dots, \boldsymbol{0}}_{1,2,\dots,t-1}\ \boldsymbol{I}\ \underbrace{\boldsymbol{0}, \boldsymbol{0}, \dots, \boldsymbol{0}}_{\boldsymbol{t+1},\boldsymbol{t+2},\dots,\boldsymbol{T}}\right], \tag{3.58}$$

$$\boldsymbol{w_i^1} = \left[\underbrace{\boldsymbol{0}, \boldsymbol{0}, \dots, \boldsymbol{0}}_{1,2,\dots,t-L-1}\ \underbrace{\boldsymbol{A_i^{-L}}, \boldsymbol{A_i^{-L+1}}, \dots, \boldsymbol{A_i^{L-1}}, \boldsymbol{A_i^{L}}}_{\boldsymbol{t-L},\boldsymbol{t-L+1},\dots,\boldsymbol{t+L}}\ \underbrace{\boldsymbol{0}, \boldsymbol{0}, \dots, \boldsymbol{0}}_{\boldsymbol{t+L+1},\boldsymbol{t+2},\dots,\boldsymbol{T}}\right], \tag{3.59}$$

$$\boldsymbol{A_i^k} = w_k \boldsymbol{I}. \tag{3.60}$$

U izrazima (3.56)-(3.60) $\boldsymbol{I}$ predstavlja jediničnu matricu dimenzija $M \times M$, a $\mathbf{0}$ nula matricu dimenzija $M \times M$. Ilustracija veze između statičkih obeležja i vektora opservacija za slučaj kada je L jednako 1 data je na slici 3.4.

Ako se u jednačinu (3.47) ubaci vektor obeležja baziran na proširenju statičkih obeležja dinamičkim, dobija se jednačina analogna jednačini (3.60), a koja ima oblik

$$\ln P(\boldsymbol{O}|\boldsymbol{q}, \boldsymbol{\lambda}) = -\frac{1}{2}(\boldsymbol{Wc} - \boldsymbol{U})^T \boldsymbol{G}^{-1}(\boldsymbol{Wc} - \boldsymbol{U}) - \frac{1}{2} 2MT \ln 2\pi \ - \frac{1}{2}\sum_{t=1}^{T} \ln\left|\sum_{q_t}\right|. \tag{3.61}$$

U jednačini (3.61) vektori opservacija i vektori srednjih vrednosti imaju dužine $2MT$, gde je M veličina vektora statičkih obeležja, a T posmatrani broj frejmova.

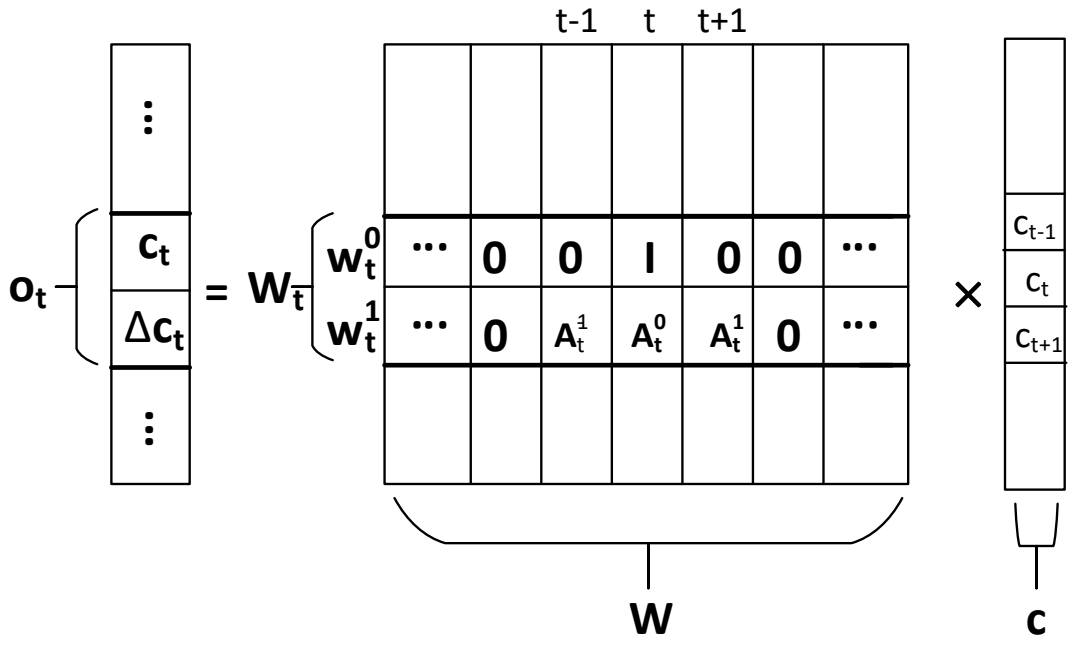
U [87] pokazano je da se maksimizacijom izraza (3.61) dobija izraz

$$\boldsymbol{W}^T \boldsymbol{G}^{-1} \boldsymbol{Wc} = \boldsymbol{W}^T \boldsymbol{G}^{-1} \boldsymbol{U}. \tag{3.62}$$

Iz (3.62) dobija se jednakost za računanje statičkih obeležja

$$\boldsymbol{c} = (\boldsymbol{W}^T \boldsymbol{G}^{-1} \boldsymbol{Wc})^{-1} \boldsymbol{W}^T \boldsymbol{G}^{-1} \boldsymbol{U}. \tag{3.63}$$

U [87] navedeni su takođe postupci kako je moguće efikasnog rešavanja jednakosti (3.63).



Slika 3.4 Formiranje vektora opservacija na osnovu statičkih obeležja

Treba napomenuti da se vektor opservacija pored delta obeležja često proširuje i delta-delta obeležjima, koja se računaju na identičan način kao u jednakosti (3.53), uz korišćenje drugačijih vrednosti koeficijenata w_τ i parametra *L*. U tom slučaju vektor opservacija ima oblik

$$\boldsymbol{o_t} = [\boldsymbol{c}_t, \Delta \boldsymbol{c}_t, \Delta^2 \boldsymbol{c}_t]^T. \tag{3.64}$$

Izvođenje pokazano kroz jednakosti (3.55)-(3.61) potpuno je analogno i za vektor opservacija definisan u (3.64) pa u daljem tekstu neće biti posebno obrađivano.

Prethodno opisani algoritam za računanje obeležja naziva se generisanje parametara na osnovu maksimalne verodostojnosti (engl. *Maximum Likelihood Parameter Generation,* MLPG).

U [87] predstavljena su i dva dodatna pristupa za traženje optimalne sekvence opservacija definisana na sledeći način:

1. $\widehat{\boldsymbol{O}} = arg \max_{\boldsymbol{O}} P(\boldsymbol{O}, \boldsymbol{q}|\boldsymbol{\lambda})$, (3.65)

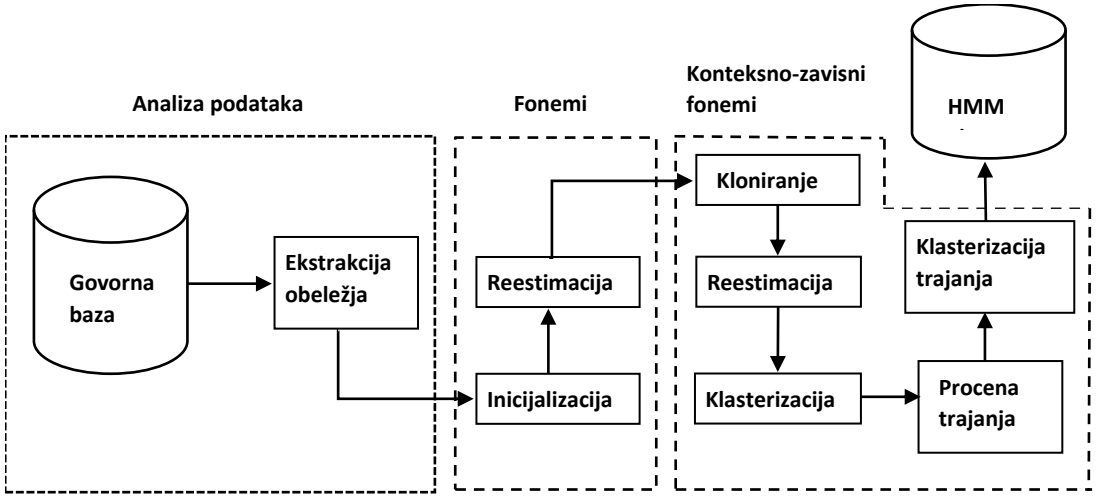
2. $\widehat{\boldsymbol{O}} = arg \max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{\lambda})$. (3.66)

S obzirom na to da su dati postupci dosta ređe korišćeni u praksi ovde će biti izostavljeno objašnjenje pomenutih algoritama.

3.5. HMM obuka u sintezi govora

Akustička realizacija jednog fonema zavisi od konteksta u kojem se on nalazi. Stoga postoji mnoštvo različitih realizacija istog fonema. Ovu varijabilnost potrebno je uzeti u obzir i prilikom modelovanja govora. Za potrebe prepoznavanja govora modeluju se trifoni, tj. fonemi koji uzimaju u obzir informaciju o fonemima koji se nalaze ispred i iza posmatranog fonema. U sintezi govora kontekst koji se posmatra je mnogo širi i pored podataka o prethodnom i sledećem fonemu obuhvata i informacije kao što su broj slogova u trenutnoj reči, tip akcenta i druge. Međutim, zbog tako širokog opsega informacija o kontekstu nemoguće je u podacima za obuku imati primere realizacije fonema za svaki mogući kontekst, niti je broj realizacija za pojedine kontekste dovoljan za formiranje pouzdanih modela. Na ove činjenice treba obratiti posebnu pažnju prilikom obuke.

Celokupna trening faza u okviru HMM sinteze predstavljena je na slici 3.5 [88]. Prvo se na osnovu informacija o granicama fonema i izdvojenih akustičkih obeležja obučavaju modeli monofona. Parametri modela monofona koriste se kao početne vrednosti parametara modela kontekstno zavisnih fonema. Kontekstno zavisni modeli se potom reestimiraju pomoću Baum-Velč algoritma predstavljenog u odeljku 3.2.3. Problem u vezi sa malim brojem opservacija za pojedine kontekstno zavisne modele rešava se primenom algoritma klasterizacije na bazi stabala odluke [89]. Proces kreiranja stabla je iterativna procedura. U korenu stabla smešteni su svi modeli u sistemu. U svakom čvoru stabla vrši se odabir pitanja iz skupa predefinisanih pitanja koje skup modela pridruženih posmatranom čvoru deli na dva podskupa tako da se smanji ukupna verodostojnost. Ovaj proces se prekida kada bude ispunjen zadati uslov za zaustavljanje. Uslov zaustavljanja algoritma zavisi od izbora kriterijuma minimizacije. Najčešće se koristi MDL (engl. *Minimum Description Length*) kriterijum [90]. Na kraju procedure klasterizacije svi modeli koji se nalaze u istom listu stabla će imati iste parametre. Ti parametri predstavljaju uprosečenu vrednost svih

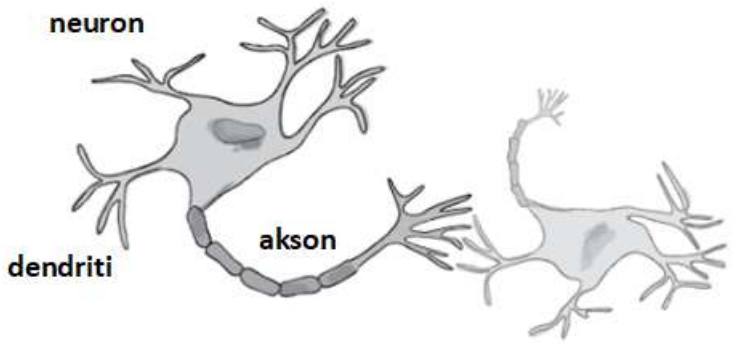


Slika 3.5 HMM trening u sintezi govora

opservacija koje pripadaju odgovarajućem listu. Nakon klasterizacije modeli se ponovo reestimiraju korišćenjem Baum-Velč algoritma i potom se vrši poravnavanje podataka za obuku sa dobijenim konačnim modelima preko Viterbijevog algoritma. Na ovaj način dobijaju se raspodele verovatnoće trajanja pojedinih stanja HMM, a trajanje svakog stanja modelovano je jednom Gausovom raspodelom.

4. DNN sinteza

Veštačke neuronske mreže predstavljaju algoritam mašinskog učenja zasnovan na principu funkcionisanja ljudskog nervnog sistema. Nervni sistem sastoji se od ogromnog broja međusobno povezanih neurona. Svaki neuron ima ulazni sloj koji se sastoji od dendrita i izlazni sloj koji čine aksoni. Dendriti omogućavaju neuronu da prima signale od drugih neurona, dok se aksoni koriste da bi neuron obrađene signale prosledio drugim neuronima, kao što je prikazano na slici 4.1.

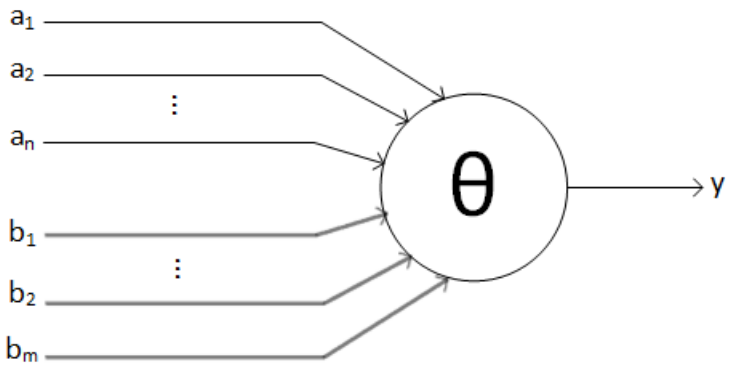


Slika 4.1 Pojednostavljen prikaz neurona

4.1. Kratak istorijski pregled razvoja neuronskih mreža

Prvi matematički model inspirisan funkcionisanjem neurona predstavljen je u [91]. Ovaj model neurona, koji ćemo u daljem tekstu nazivati MP (engl. *McColough-Pitts*) neuron, kao izlazni signal daje vrednosti nula ili jedan. Ulazni signali u ovom modelu dele se na aktivacione i inhibitorne, pri čemu je dovoljno da vrednost samo jednog inhibitornog signala bude jedan kako bi izlaz MP neurona postao nula. Ako nijedan inhibitorni signal nije aktivan, izlaz neurona zavisi od poređenja sume vrednosti aktivacionih signala sa određenim pragom. Prethodni opis formalizovan je u narednoj jednakosti

$$y = \begin{cases} 1, & \sum_{i=1}^{n} a_i \geq \theta \ \wedge \ b_1 = b_2 = \cdots = b_m = 0 \\ 0, & \text{inače} \end{cases}. \tag{4.1}$$

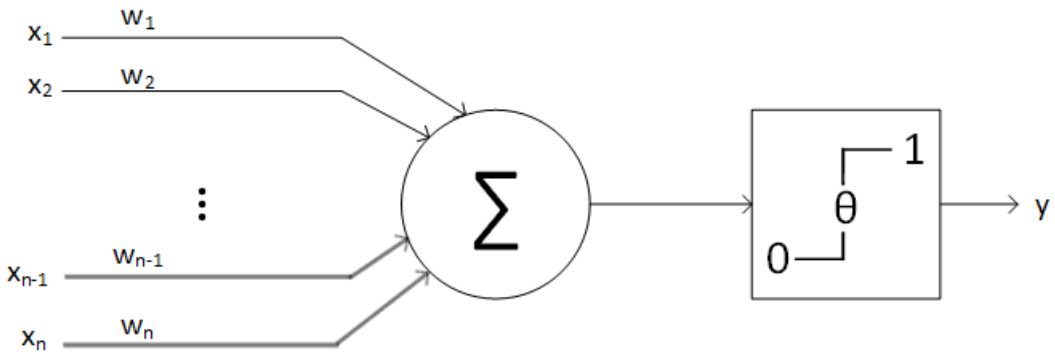


Slika 4.2 Model MP neurona

U [91] takođe je pokazano da se ovakav model može iskoristiti za prikazivanje svih logičkih operacija. Nedostatak MP modela neurona predstavlja nemogućnost učenja, tj. za ovaj model nije definisan algoritam za određivanje optimalne vrednosti praga θ. Dodatni problem predstavlja i činjenica da su ulazne vrednosti binarne, odnosno da a_i i b_j mogu da imaju samo vrednosti nula ili jedan.

Prethodno opisani problemi MP neurona rešeni su u modelu neurona koji je predstavljen u [92] i koji se naziva perceptron, a koji je prikazan na slici 4.3.

Pretpostavka je da su u modelu perceptrona svakom ulazu dodeljene odgovarajuće težine,



Slika 4.3 Model perceptrona

koje, kao i vrednosti ulaznih signala, pripadaju skupu realnih brojeva. Princip funkcionisanja perceptrona određen je izrazom

$$y = \begin{cases} 1, & \sum_{i=1}^{n} w_i x_i \geq \theta \\ 0, & \sum_{i=1}^{n} w_i x_i < \theta \end{cases} . \tag{4.2}$$

Ukoliko se uvedu oznake $\boldsymbol{x} = [1, x_1, x_2 \dots x_n]^{\boldsymbol{T}}$ i $\boldsymbol{w} = [-\theta, w_1, w_2 \dots w]^{\boldsymbol{T}}$ tada se jednakost (4.2) može predstaviti u obliku

$$y = \begin{cases} 1, & \boldsymbol{w}^T \boldsymbol{x} \geq \boldsymbol{1} \\ 0, & \boldsymbol{w}^T \boldsymbol{x} < \boldsymbol{0} \end{cases} . \tag{4.3}$$

Za poznati skup uzoraka $D = \{(\boldsymbol{x}_1, d_1), (\boldsymbol{x}_2, d_2) \dots (\boldsymbol{x}_2, d_m)$, gde je d_i željeni izlaz perceptrona za *i*-ti uzorak (čija je vrednost nula ili jedan), algoritam određivanja težina perceptrona dat je u nastavku.

Algoritam 1 Perceptronsko učenje

Ulazi

$$D = \{(\boldsymbol{x}_1, d_1), (\boldsymbol{x}_2, d_2) \dots (\boldsymbol{x}_2, d_m)\}$$

Izlazi

Vektor težina pereceptrona $\boldsymbol{w}$

Postavi početne vrednosti težina na slučajan način

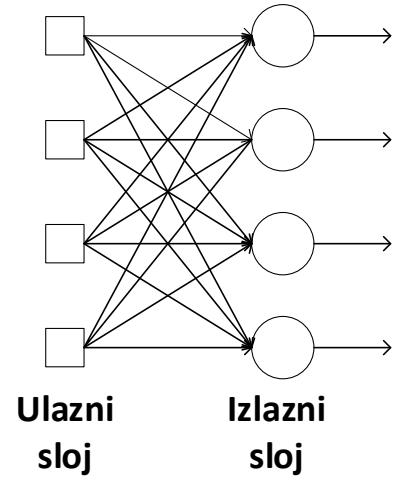
for i=1:m

$$y_i = \begin{cases} 1, & \boldsymbol{w}_{i-1}^T \boldsymbol{x}_i \geq \boldsymbol{1} \\ 0, & \boldsymbol{w}_{i-1}^T \boldsymbol{x}_i < \boldsymbol{0} \end{cases}$$

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + (d_i - y_i)\boldsymbol{x}_i$$

Iz opisa algoritma može se zaključiti da se promena težinskih koeficijenata neće desiti ukoliko je trenutni uzorak ispravno klasifikovan. Ukoliko je greška klasifikacije pozitivna, tj. $d_i - y_i = 1$, tada se vektor trenutnih vrednosti dodaje težinskim koeficijentima, a ukoliko je greška klasifikacije negativna, tj. $d_i - y_i = -1$, trenutni ulazni vektor se oduzima od težinskih koeficijenata.

Perceptron se može uspešno primeniti pri klasifikaciji podataka koji su međusobno linearno razdvojivi. Ukoliko se koristi samo jedan neuron tada se može odrediti samo pripadnost jednoj klasi od dve moguće. Klasifikacija u slučaju postojanja više izlaznih klasa vrši se kombinovanjem većeg broja pojedinačnih neurona u jednom sloju. Na taj način se dobija tzv. jednoslojni perceptron kao što je prikazano na slici 4.4. U ovakvom modelu svi perceptroni u izlaznom sloju imaju iste ulaze, a svaki od njih predstavlja jednu moguću klasu.



Slika 4.4 Jednoslojni perceptron

Još jedan model neurona koji je korišćen naziva se ADALINE neuron (engl. *Adaptive Linear Neuron*). Njegov način funkcionisanja takođe se može opisati korišćenjem jednakosti (4.2). U odnosu na model perceptrona razlikuje se po algoritmu učenja, odnosno definisanju greške, kao što je prikazano na slici 4.5. Kod perceptrona greška predstavlja razliku vrednosti nakon bloka za binarno odlučivanje i može da ima vrednosti 0, 1 ili -1. Kod ADALINE neurona greška se definiše kao razlika odgovarajućih vrednosti nakon bloka koji računa linearnu kombinaciju ulaznih elemenata, tj. skup za obuku može se predstaviti u obliku, $D = \{(\boldsymbol{x}_1, \hat{y}_1), (\boldsymbol{x}_2, \hat{y}_2) \dots (\boldsymbol{x}_m, \hat{y}_m)\}$, gde je $\hat{y}_i$ izlaz bloka za računanje linearne kombinacije. Algoritam za učenje, dat u nastavku, još se naziva i Vidrou-Hof (engl. *Widrow-Hof*) delta pravilo prema autorima ADALINE modela. Može se pokazati da ovaj algoritam u stvari odgovara algoritmu najbržeg opadajućeg gradijenta (engl. *steepest gradient descent*) kod kojeg je greška definisana kao srednja kvadratna greška [93].

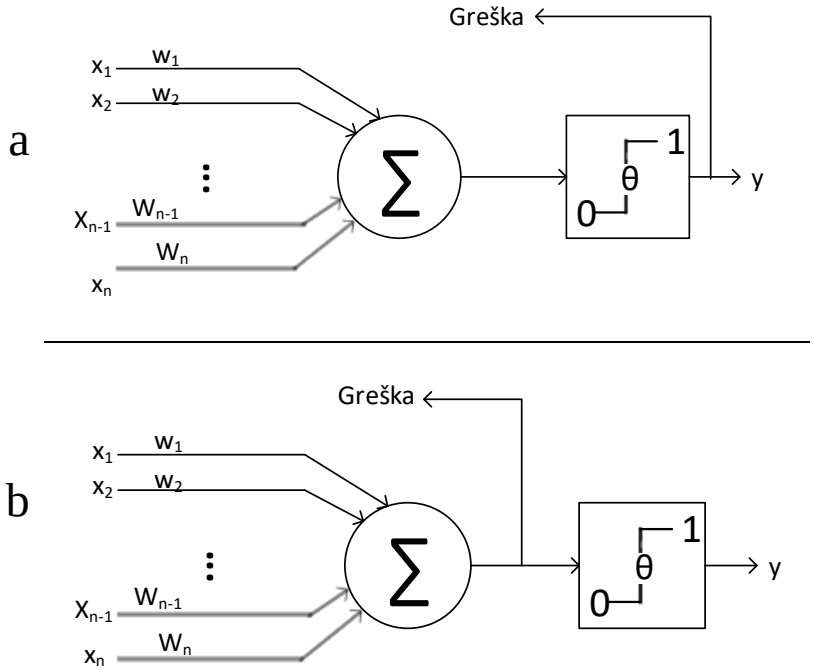
Algoritam 2 Vidrou-Hof delta pravilo

Ulazi
$D = \{(\boldsymbol{x}_1, \hat{y}_1), (\boldsymbol{x}_2, \hat{y}_2) \dots (\boldsymbol{x}_2, \hat{y}_m),$
ξ - stepen učenja

Izlazi
Vektor težina pereceptrona $\boldsymbol{w}$

Postavi početne vrednosti težina na slučajan način
for i=1:m
$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \xi(\hat{y}_i - \boldsymbol{w}_{i-1}^T \boldsymbol{x}_i)\boldsymbol{x}_i$

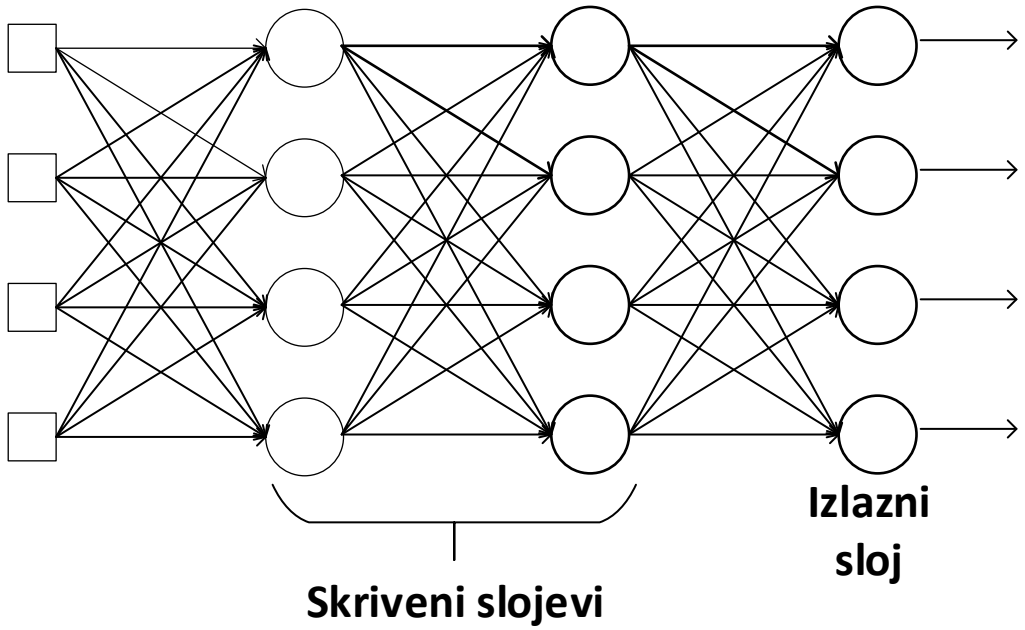
Za algoritam perceptronskog učenja dokazano je da će sigurno konvergirati u slučaju da je problem koji se analizira linearno razdvojiv, dok za algoritam učenja primenjen kod ADALINE modela ne postoji garancija da će uspeti uspešno da razdvoji sve uzorke, čak iako su oni zaista linearno razdvojivi (iako obično bude blizu ispunjenja ovog uslova) [93]. Takođe, vrednosti težinskih koeficijenata kod perceptronskog učenja nisu ograničene. Kada posmatrani problem nije linearno razdvojiv težinski koeficijenti obično teže nuli. U tom slučaju se prilikom klasifikacije dobija velika greška.



Slika 4.5 Poređenja perceptronskog (a) i ADALINE (b) učenja

U [94] data je opsežna kritika modela perceptrona bazirana pre svega na njegovoj nemogućnosti korišćenja za nelinearne probleme. Da bi se perceptroni mogli koristiti za klasifikaciju nelinearnih problema neophodno je njihovo grupisanje u višeslojne strukture, kao što je prikazano na slici 4.6, a za koje u tom trenutku nisu postojali algoritmi učenja. To je bio jedan od glavnih razloga zastoja istraživanja u vezi sa veštačkim neuronskim mrežama sve do 80-ih godina XX veka i definisanja algoritma propagacije unazad.

Struktura predstavljena na slici 4.6 u literaturi se naziva dubokom neuronskom mrežom –



Slika 4.6 Višeslojni perceptron

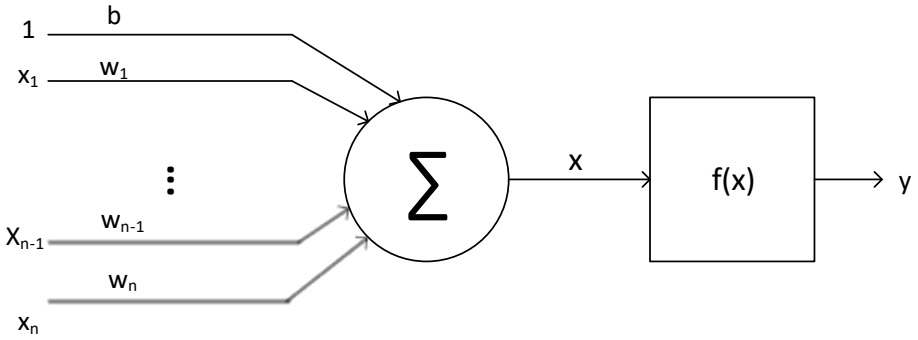
to je mreža koja sadrži barem dva skrivena sloja neurona. Kod savremenih neuronskih mreža gradivne jedinice nisu modeli klasičnog perceptrona opisanog u ovom odeljku nego modeli izmenjene strukture, čiji će princip funkcionisanja biti opisan u odeljku 4.2.

4.2. Algoritam propagacije unazad

Algoritmi perceptronskog učenja i Vidrou-Hof delta pravilo omogućavali su učenje težinskih koeficijenata za pojedinačne neurone. Takvi algoritmi nisu mogli da se primene u slučaju višeslojnih struktura sa slike 4.6 za učenje težinskih koeficijenata koji odgovaraju

neuronima u skrivenim slojevima. Za te neurone nisu poznati njihovi stvarni izlazi koji bi omogućavali poređenje sa njihovim trenutnim izlazima.

Rešenje ovog problema omogućeno je upotrebom algoritma propagacije unazad [95]. Korišćenje ovog algoritma zahteva nešto drugačiju strukturu neurona u poređenju sa perceptronom, a koja je prikazana na slici 4.7.



Slika 4.7 Model neurona sa nelinearnom aktivacionom funkcijom

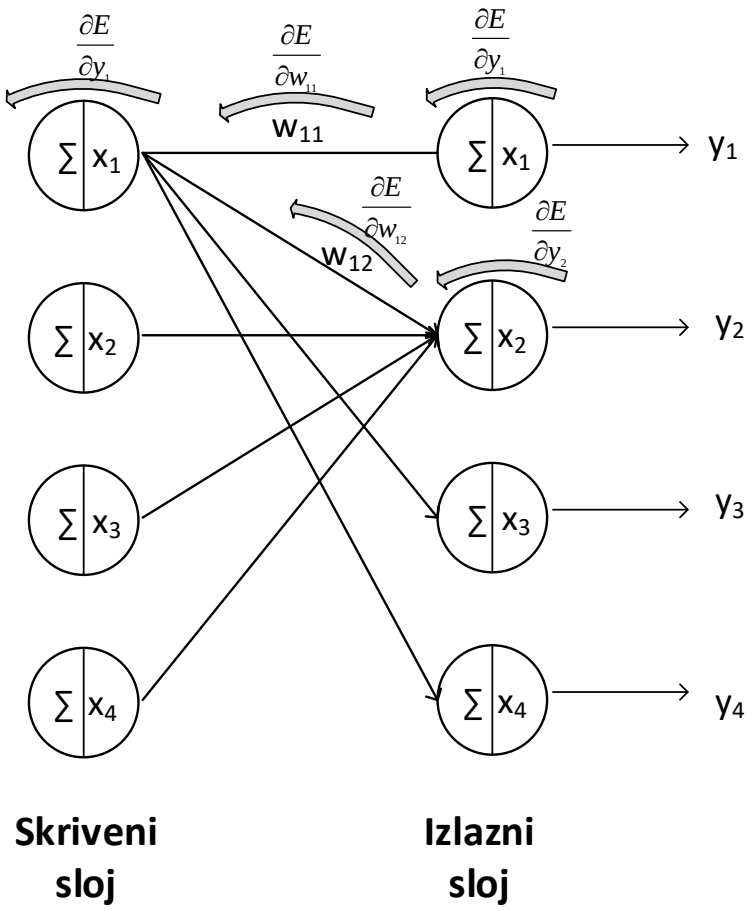
Nelinearna funkcija koja linearnu kombinaciju ulaznih pobuda neurona *x*, preslikava u izlaz neurona *y*, naziva se aktivaciona funkcija. Primer jedne često korišćene aktivacione funkcije jeste sigmoidalna funkcija

$$y = \frac{1}{1 + e^{-x}}. \tag{4.4}$$

Cilj algoritma propagacije unazad jeste pronalaženje svih gradijenata težinskih koeficijenata u procesu minimizacije određene funkcije greške. Najčešće korišćena funkcija greške jeste srednja kvadratna greška

$$E = \frac{1}{2}\sum_{j=1}^{K}(y_j - t_j)^2, \tag{4.5}$$

gde K predstavlja ukupan broj neurona u izlaznom sloju, y_j trenutnu izlaznu vrednost neurona na poziciji j i t_j željenu izlaznu vrednost.



Slika 4.8 Ilustracija algoritma propagacije unazad

Jedan deo mreže prikazan je na slici 4.8 na kojoj su zbog preglednosti izostavljene neke veze između neurona. Ukoliko je potrebno proveriti koliko svaki pojedinačni izlaz y_j utiče na ukupnu funkciju greške neophodno je pronaći parcijalne izvode

$$\frac{\partial E}{\partial y_j} = \left(y_j - t_j\right). \tag{4.6}$$

Ono što zaista treba da se izračuna jeste kako promena nekog koeficijenta utiče na vrednost greške. Da bi se definisala ova veličina, prvo se definiše na koji način promena ulaza neurona *x* utiče na vrednost greške

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j}\frac{dy_j}{dx_j}. \tag{4.7}$$

Vrednost izvoda $\frac{dy_j}{dx_j}$ zavisi od aktivacione funkcije. Korišćenjem prethodnih jednakosti, uticaj na vrednost greške težinskog koeficijenta w_{ij}, koji povezuje *i*-ti neuron u jednom sloju sa *j*-tim neuronom u sledećem sloju dat je izrazom

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial x_j}\frac{\partial x_j}{\partial w_{ij}}. \tag{4.8}$$

Kako je ulaz x_j neurona j jednak linearnoj kombinaciji izlaza neurona iz prethodnog sloja, tj. $x_j = \sum_k w_{kj} y_j$, tada je $\frac{\partial x_j}{\partial w_{ij}} = y_i$ pa se jednačina (4.8) može napisati u obliku

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial x_j} y_i. \tag{4.9}$$

Ako posmatramo samo prethodni skriveni sloj tada se uticaj izlaza neurona *i* u skrivenom sloju na vrednost celokupne greške može opisati sumom odgovarajućih vrednosti datih jednačinom (4.9)

$$\frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial w_{ij}}. \tag{4.10}$$

Vrednosti definisane izrazom (4.9) nazivamo gradijentima i obično označavamo sa Δw_{ij}, a na osnovu njih se nova vrednost težinskih koeficijenata računa korišćenjem jednakosti

$$w_{ij} = w_{ij} - \xi \Delta w_{ij}, \tag{4.11}$$

pri čemu se ξ naziva koeficijentom učenja.

U najjednostavnijem obliku algoritma računanje novih vrednosti težinskih koeficijenata na osnovu vrednosti Δw vrši se nakon svakog prolaska jednog uzorka za obuku kroz mrežu. Često se koristi pristup u kom se novi koeficijenti ne računaju pri svakom novom uzorku nego se vrednosti Δw akumuliraju pri prolasku nekoliko uzoraka kroz mrežu i tek nakon toga se vrši računanje novih vrednosti težinskih koeficijenata.

Algoritam 3 Optimizacija težinskih koeficijenata neuronske mreže

Ulazi
Skup trening uzoraka $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots (\boldsymbol{x}_m, y_m)$
Koeficijent učenja

Izlazi
Optimizovani težinski koeficijenti

Postavi početne vrednosti težinskih koeficijenata
Za svaki trening uzorak
1. *Izračunaj izlaze svih neurona u mreži (engl. forward pass)*
2. *Izračunaj grešku kao razliku stvarne vrednosti i izlaza iz mreže*
3. *Izračunaj gradijente za sve težinske koeficijente u mreži (jednakosti 4.6-4.10)*
4. *Izračunaj nove vrednosti težinskih koeficijenata korišćenjem gradijenata (jednakost 4.11)*

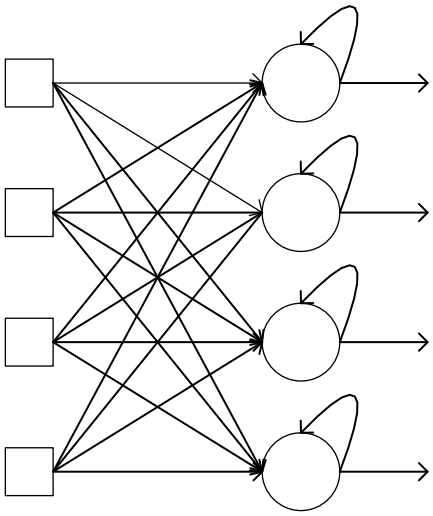
Problem sa opisanim algoritmom traženja novih vrednosti može da predstavlja spora konvergencija težinskih koeficijenata ka optimalnim vrednostima. Jedan od načina za ubrzavanje konvergencije jeste nešto drugačiji postupak za računanje gradijenta, a koji je dat jednakošću

$$\Delta w(t) = -\xi \frac{\partial E}{\partial w} + \alpha \Delta w(t-1), \tag{4.12}$$

pri čemu se veličina α naziva momentum.

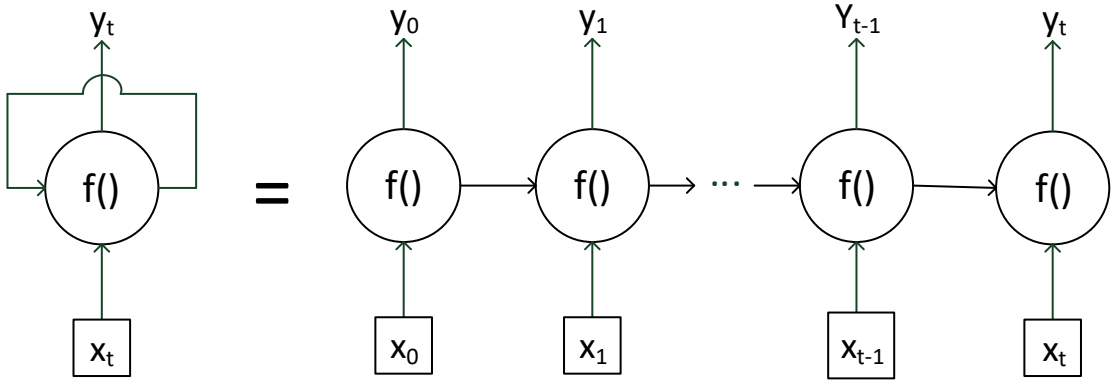
4.3. Rekurentne neuronske mreže

U izvođenju datom u prethodnom odeljku pretpostavljeno je da na trenutni izlaz mreže utiče samo trenutni uzorak na ulazu mreže, tj. mreža ni na koji način ne pamti eventualne prethodne ulaze u mrežu. Neke pojave, kao što je i govor, mogu se predstaviti sekvencom opservacija, u kojoj na trenutnu vrednost opservacija utiču i vrednosti prethodnih opservacija. Za predstavu takvih pojava mogu da se iskoriste rekurentne neuronske mreže (engl. *Recurrent Neural Networks,* RNN). To su mreže kod kojih postoje petlje čime se omogućava da se određene informacije zadrže neko vreme u mreži.



Slika 4.9 Primer rekurentne neuronske mreže

Primer rekurentne mreže dat je na slici 4.9. U ovakvoj strukturi trenutni izlaz iz jednog neurona istovremeno se dovodi i na njegov ulaz. Ova činjenica može se iskoristiti da se struktura sa slike 4.9 predstavi u tzv. „razmotanom" (engl. *unrolled*) obliku prikazanom na slici 4.10 koji omogućava da se uspešno vrši procedura propagacije unazad. Jedan rekurentni neuron u stvari je predstavljen sekvencom standardnih neurona. Broj neurona u „razmotanom" obliku određuje koliko izlaza iz prethodnih vremenskih trenutaka će biti zapamćeno. Ovaj broj ne sme biti preveliki jer može da utiče na trajanje obuke. Jedan od problema koji se javlja u „razmotanom" obliku mreže tokom treninga jesu „nestajući" gradijenti. Naime, dešava se da gradijenti postepeno počinju da opadaju i njihova vrednost počinje da teži nuli čime mreža u stvari prestaje da uči. Ovaj problem doveo je do uvođenja



Slika 4.10 „Razmotani" oblik rekurentne mreže

LSTM (engl. *Long Short-term Memory*) neurona [96], [97]. LSTM neuron je rekurentna struktura koja se takođe može predstaviti u „razmotanom“ obliku prikazanom na slici 4.10. Gradivni blokovi ne sastoje se samo od blokova sa aktivacionom funkcijom nego od složenih struktura koje su predstavljene na slici 4.11.

Izlaz LSTM neurona određen je nizom jednačina:

$$f_t = f(\boldsymbol{W_f}[\boldsymbol{y}_{t-1}; \boldsymbol{x_t}] + \boldsymbol{b}_f), \tag{4.13}$$

$$i_t = f(\boldsymbol{W_i}[\boldsymbol{y}_{t-1}; \boldsymbol{x_t}] + \boldsymbol{b}_i), \tag{4.14}$$

$$\tilde{C}_t = tanh(\boldsymbol{W_c}[\boldsymbol{y}_{t-1}; \boldsymbol{x_t}] + \boldsymbol{b}_c), \tag{4.15}$$

$$C_t = C_{t-1} f_t + \tilde{C}_t i_t, \tag{4.16}$$

$$o_t = f(\boldsymbol{W_o}[\boldsymbol{y}_{t-1}; \boldsymbol{x_t}] + \boldsymbol{b}_o), \tag{4.17}$$

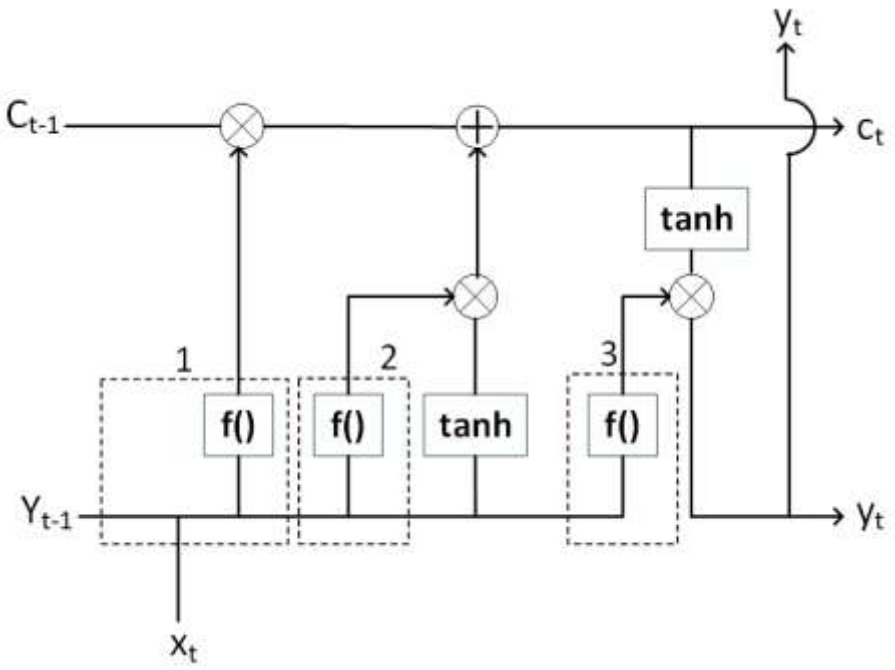
$$y_t = o_t \tanh(C_t). \tag{4.18}$$

U prethodnim jednačinama *C* označava memoriju prethodnog stanja LSTM neurona, $\boldsymbol{W_f}$, $\boldsymbol{W_i}$, $\boldsymbol{W_c}$ i $\boldsymbol{W_o}$ matrice težinskih koeficijenata, $\boldsymbol{x_t}$ trenutne ulaze, $\boldsymbol{y}_{t-1}$ izlaze iz prethodnog vremenskog trenutka, *f(·)* sigmoidalnu aktivacionu funkciju, dok je sa *tanh(·)* označena tangens-hiperbolična funkcija

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{4.19}$$

Parametri *C* i *y* u početnom trenutku, tj. C_0 i y_0, imaju vrednost nula.

Jednakost (4.13) definiše koji deo prethodnog stanja se pamti. Ukoliko je vrednost jednakosti (4.13) jedan to u stvari znači da se u potpunosti pamti prethodno stanje, a ako je vrednost nula ono se u potpunosti odbacuje. Ova vrednost predstavljena je oznakom 1 na slici 4.11. Blok 1 naziva se još i kapija zaboravljanja (engl. *forget gate*). Računanje novog stanja dato je kroz jednakosti (4.14)-(4.16). Blok koji je na slici predstavljen oznakom 2, a koji odgovara jednakosti (4.14) naziva se ulazna kapija (engl. *input gate*). Međuvrednost $\tilde{C}_t$, data jednačinom (4.15), predstavlja tzv. kandidata za trenutno stanje. Računanje izlaza



Slika 4.11 Blok šema LSTM neurona

LSTM neurona dato je u jednačinama (4.17) i (4.18). Jednačina (4.18) na slici 4.1 odgovara bloku sa oznakom 3 koji se još naziva i izlazna kapija (engl. *output gate*).

4.4. Primena DNN u sintezi govora

Prvi primeri upotrebe neuronskih mreža u sintezi govora potiču iz 90-ih godina XX veka. U [98] predstavljen je sistem za predikciju akustičkih obeležja koji se sastoji od tri odvojene neuronske mreže. Prva od njih vrši predikciju fonetske transkripcije na osnovu teksta. Druga mreža vrši predikciju trajanja fonema, a treća predikciju određenih formantnih parametara na osnovu kojih je moguće rekonstruisati govor. Sve tri mreže se sastoje od samo jednog skrivenog sloja. Sistem je obučavan korišćenjem 300 rečenica, ali u samom radu nije data evaluacija kvaliteta sintetizovanog govora. U [99] takođe su korišćene mreže sa jednim skrivenim slojem za predikciju LSP (engl. *Linear Spectral Pairs*) [100] parametara alofona na osnovu ulaznih lingvističkih parametara. Trening skup sastojao se od 10 rečenica, ali takođe nije data analiza kvaliteta generisanog govora, nego poređenje performansi različitih konfiguracija neuronskih mreža. U [101] predlaže se korišćenje zasebnih mreža za predikciju

kepstralnih obeležja koja odgovaraju svakom odvojenom fonemu. Mreže koje modeluju obeležja predstavljene su sa 3 skrivena sloja. U ovom radu data je i detaljnija analiza kvaliteta dobijenog govora koja u potpunosti podržava korišćenje neuronskih mreža u sintezi govora.

I pored dobijenih rezultata primena neuronskih mreža u sintezi govora u tom periodu nije stekla naročitu popularnost. Ovo je pre svega bilo uzrokovano njihovom kompleksnošću. Dostupni računarski resursi, u poređenju sa savremenim, bili su skromnih mogućnosti i skupi kako bi bili naširoko korišćeni. Međutim, u XXI veku računarski resursi su postali znatno moćniji i jeftiniji i neuronske mreže su se masovno počele primenjivati u različitim oblastima što je dovelo i do ponovne primene neuronskih mreža u sintezi govora. U [102], [103] opisano je korišćenje određenih tipova neuronskih mreža za modelovanje raspodela kojima su opisana akustička obeležja govora. Korišćenje dubokih neuronskih mreža za preslikavanje lingvističkih obeležja u akustička opisano je u [33]. Pristup sintezi korišćenjem neuronskih mreža baziran na arhitekturi opisanoj u [33] najčešće je korišćen u literaturi, kao i tokom eksperimenata opisanih u disertaciji i biće detaljnije opisan u nastavku.

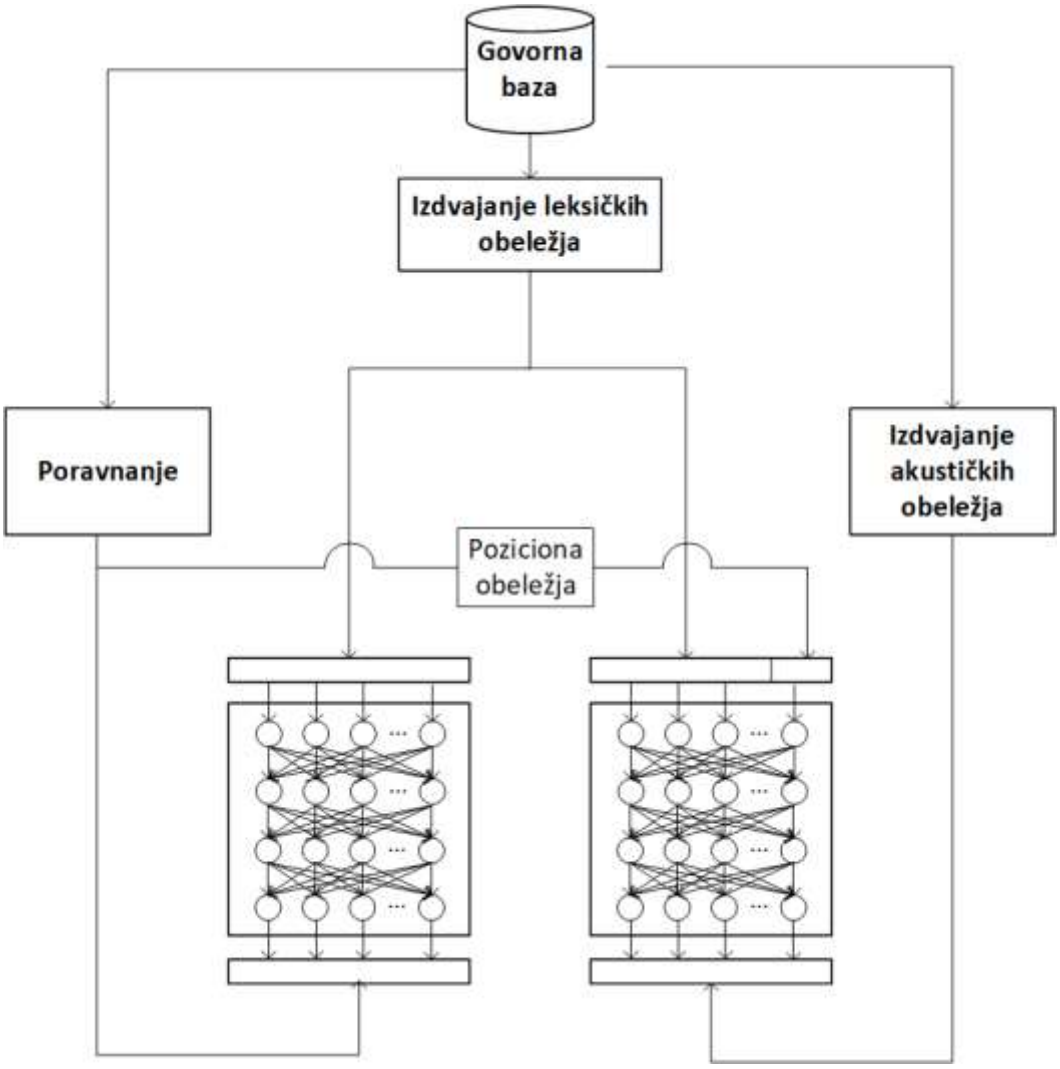
4.4.1. Standardna primena DNN u sintezi govora

Standardan DNN sintetizator govora sastoji se od dve neuronske mreže. Prva neuronska mreža koristi se za predviđanje trajanja fonema, a druga za predikciju akustičkih obeležja. Proces obuke ovih neuronskih mreža prikazan je na slici 4.12.

Ulaze u obe neuronske mreže predstavljaju leksička obeležja. Leksička obeležja su predstavljena u vidu odgovora na binarna pitanja, tj.imaju vrednost nula ili jedan. Primeri korišćenih pitanja su:

- „Da li je trenutni fonem vokal?”,
- „Da li je trenutni fonem A? ”,
- „Da li je trenutni fonem naglašen?”.

Izlazi mreže za procenu trajanja sastoje se od trajanja pojedinih HMM stanja odgovarajućeg fonema izraženih u broju frejmova, te je jedan od koraka u pripremi podataka za obuku neuronske mreže procedura poravnanja, odnosno određivanje granica između



Slika 4.12 Obuka u DNN sintezi

fonema i pojedinačnih stanja unutar fonema. Za ove potrebe uglavnom se koriste metode razvijene u okviru automatskog prepoznavanja govora, a njihovo detaljno objašnjenje prevazilazi obim ove disertacije. Više detalja može se pronaći u [104], [105].

Akustička obeležja koja treba da predvidi druga neuronska mreža sastoje se od obeležja koja izdvaja odgovarajući vokoder. Obeležja izdvojena vokoderom obično se dele u tri grupe:

- spektralna obeležja,
- osnovna učestanost,

- obeležja aperiodičnosti.

Pošto je govor u velikoj meri vremenski varijabilan, fonem se ne može opisati samo jednim vektorom akustičkih obeležja. Akustička obeležja se, kao i u slučaju HMM obuke, izdvajaju za prozore, tj. frejmove, koji su obično međusobno pomereni za pet milisekundi. Za razliku od mreže za predviđanje trajanja kod kojih jedan par ulaz-izlaz predstavlja odgovarajuća obeležja jednog fonema, jedan uzorak u obuci mreže za predikciju akustičkih obeležja odgovara jednom frejmu. Ukoliko bi se kao ulaz mreže koristila ranije pomenuta leksička obeležja, tada mreža ne bi mogla npr. da razlikuje početni frejm od krajnjeg frejma jednog fonema, a koji međusobno ne moraju biti slični. Da bi se ovaj problem prevazišao, ulazima mreže za predviđanje akustičkih obeležja pored leksičkih obeležja korišćenih kao ulaz mreže za predikciju trajanja, dodaju se i tzv. poziciona obeležja, koja opisuju položaj trenutnog frejma u okviru odgovarajućeg fonema. Ova obeležja formiraju se na osnovu ranije određenih poravnanja i uključuju sledeće informacije:

- redni broj HMM stanja kojem trenutni frejm pripada,
- relativnu poziciju frejma unutar trenutnog stanja (određenu kao količnik broja frejmova pre, odnosno posle posmatranog frejma i ukupnog broja frejmova u stanju),
- relativnu poziciju frejma unutar fonema (određenu kao količnik broja frejmova pre, odnosno posle posmatranog frejma i ukupnog broja frejmova u fonemu).

U [106] pokazano je da se i u slučaju primene neuronskih mreža bolji rezultati postižu ako se kao izlaz mreže za predviđanje akustičkih obeležja pored statičkih koriste i dinamička (prvi i drugi izvodi) obeležja. Pored pomenutih obeležja koristi se i još jedno dodatno obeležje koje definiše da li je trenutni frejm zvučan ili bezvučan. Ovo obeležje se još naziva i VUV (engl. *Voiced UnVoiced*) obeležje. U HMM pristupu nije potrebno korišćenje ove informacije pošto je ona modelovana implicitno uvođenjem verovatnoće na uniji prostora verovatnoća.

Normalizacija ulaznih podataka jedan je od najčešće korišćenih alata u procesu mašinskog učenja [107]. Cilj normalizacije pri korišćenju neuronskih mreža jeste da sva obeležja budu u istom opsegu kako bi se smanjila pristrasnost mreže na neko od ulaznih

obeležja. Proces normalizacije primenjuje se i u obuci neuronskih mreža u sintezi govora. Ulazna obeležja normalizovana su primenom *min-max* normalizacije

$$x_i' = (max_t - min_t)\frac{x_i - min}{max - min} + min_t, \tag{4.20}$$

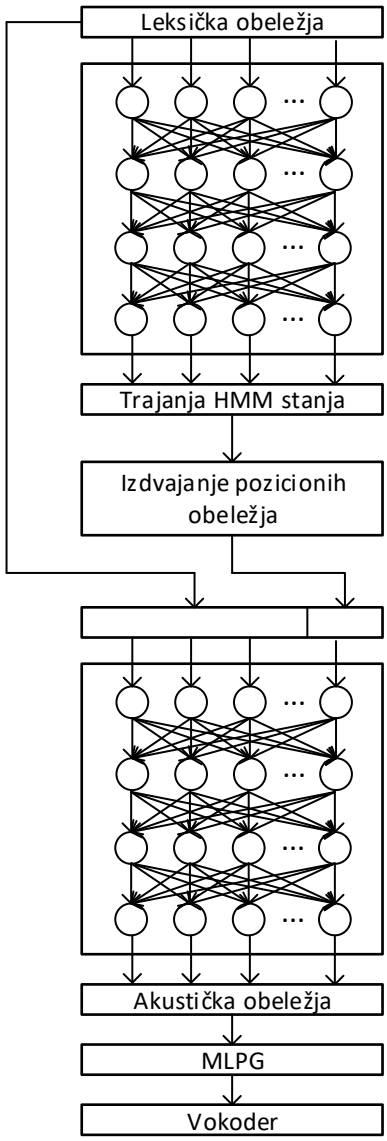
gde su *min* i *max* minimalna i maksimalna vrednost datog obeležja određena nad podacima skupa za obuku, a min_t i max_t željene minimalne i maksimalne vrednosti u normalizovanom skupu. U primeni za sintezu govora željena minimalna vrednost se postavlja na 0.01, a željena maksimalna vrednost na 0.99.

Izlazni podaci normalizuju se primenom *z*-normalizacije

$$x_i' = \frac{x_i - \mu_i}{\sigma_i}, \tag{4.21}$$

gde su μ_i i σ_i srednja vrednost i standardna devijacija koje su izračunate za dato obeležje, redom.

Proces generisanja govora na bazi primene DNN prikazan je na slici 4.13. Prvo se na osnovu leksičkih obeležja izdvojenih iz teksta vrši predikcija trajanja HMM stanja datog fonema. Ovi podaci koriste se da bi se izračunala poziciona obeležja, koja se zajedno sa leksičkim obeležjima koriste kao ulaz u mrežu za predikciju akustičkih obeležja. Pošto su izlaz mreže i statička i dinamička obeležja, nakon primene MLPG algoritma dobijaju se finalna obeležja koja se koriste kao ulaz vokodera. Izlaz vokodera je sintetizovan govor.



Slika 4.13 Postupak sinteze govora korišćenjem DNN pristupa

4.5. Poređenje HMM i DNN pristupa u sintezi neutralnog govora

Poređenje kvaliteta sintetizovanog govora korišćenjem HMM i DNN pristupa biće dato na primeru sintetizatora govora za srpski jezik. Prvi HMM sintetizator za srpski jezik predstavljen je u [108], dok je prva verzija DNN sinteze za srpski jezik opisana u [109]. Međutim, ova dva sintetizatora nisu u potpunosti uporediva pošto su u njihovom kreiranju

korišćena različita obeležja. Naime, HMM sintetizator baziran je na generisanju govora korišćenjem MGC koeficijenata, izdvojenih direktno na osnovu spektra signala i MLSA (engl. *Mel Log Spectrum Approximation*) filtra [110], dok je DNN sintetizator kreiran korišćenjem WORLD vokodera.

Za potrebe poređenja razvijen je HMM sintetizator koji takođe koristi WORLD vokoder. Parametri koje generiše HMM sistem opisan u [108] su MGC koeficijenti i osnovna učestanost. HMM sistem zasnovan na upotrebi WORLD vokodera, pored ovih obeležja[5] takođe generiše i koeficijente aperiodičnosti čiji broj zavisi od učestanosti odabiranja (npr. za učestanost odabiranja od 16 kHz koristi se samo jedan koeficijent, a za učestanost odabiranja od 22 kHz koriste se dva koeficijenta). Pošto koeficijenti aperiodičnosti nisu definisani za bezvučne regione, modelovani su na isti način kao i osnovna učestanost – korišćenjem HMM modelovanja za slučaj kada su raspodele opisane verovatnoćom nad unijom prostora verovatnoća (kao što je opisano u odeljku 3.3).

Baza korišćena u eksperimentima sastoji se od 3 sata i 20 minuta govora ženskog govornika i snimljena je u profesionalnom studiju učestanošću odabiranja od 44.1 kHz, da bi za potrebe eksperimenata učestanost odabiranja bila smanjena na 22.05 kHz. Od navedene količine materijala oko 40 minuta odnosi se na tišine.

HMM sintetizator kreiran je upotrebom HTS alata [111]. HMM modeli opisani su sa pet emitujućih stanja. Parametri za kontrolisanje veličine stabala u procesu klasterizacije postavljeni su na standardne vrednosti definisane u okviru HTS alata. Za razvoj DNN sistema korišćen je alat Merlin [112]. Neuronske mreže sastoje se od 4 sloja sa 512 neurona, kako one za predikciju akustičkih obeležja, tako i one za predikciju trajanja. Neuroni u prva tri skrivena sloja koriste tangens-hiperboličnu funkciju kao aktivacionu funkciju, dok su u četvrtom sloju korišćeni LSTM neuroni. Izlazni sloj je linearan. Više detalja oko odabira optimalnih parametara neuronske mreže može se pronaći u [113].

Akustička obeležja koja koriste oba sistema sastoje se od 40 MGC koeficijenata, osnovne učestanosti i dva koeficijenta aperiodičnosti (kao što je ranije napomenuto, broj ovih

[5] Kao što je napomenuto u odeljku 4.3 WORLD vokoder generiše spektralne obvojnice, ali su zbog velike dimenzionalnosti one predstavljene korišćenjem MGC koeficijenata.

koeficijenata određen je učestanošću odabiranja), uz napomenu da neuronske mreže koriste i dodatno obeležje koje definiše da li je trenutni frejm zvučan ili bezvučan.

Sve rečenice koje su korišćene za potrebe poređenja datih sistema bile su isključene iz procesa njihove obuke. U obuci DNN sistema sav materijal za obuku nekoliko puta prolazi kroz mrežu. Jedna iteracija korišćenja celog materijala za obuku naziva se epoha. Takođe, u procesu obuke DNN sistema 10% svih rečenica za obuku koristi se u procesu validacije. Rečenice koje se koriste u procesu validacije ne koriste se u procesu računanja gradijenata, tj. tokom algoritma propagacije unazad. One imaju ulogu u računanju vrednosti momentuma i koeficijenta učenja, čija vrednost može da se menja od epohe do epohe. Ovakva procedura standardna je za sve prikazane DNN eksperimente, stoga neće biti posebno naglašavana u nastavku teksta.

Rezultati objektivnog poređenja ova dva pristupa dati su u tabeli 4.1. Rezultati su generisani na osnovu deset test rečenica koje nisu korišćene u toku obuke. Prilikom generisanja parametara korišćeni su isti podaci o trajanjima pojedinačnih stanja koji su izdvojeni na osnovu poravnanja koja su kreirana u toku DNN obuke.

Na osnovu vrednosti prikazanih u tabeli 4.1 lako se zaključuje da se bolji rezultati dobijaju korišćenjem DNN pristupa. Naročito veliko odstupanje dobija se za BAP vrednosti (koeficijente aperiodičnosti). Ovakvo ponašanje može se opravdati suboptimalnim tretmanom koeficijenata aperiodičnosti prilikom HMM sinteze. Naime, vrednosti koeficijenata aperiodičnosti koji se dobijaju kao izlaz WORLD vokodera za zvučne frejmove nikad nisu negativne. Takvo ograničenje ne postoji prilikom generisanja parametara od strane HMM sistema, pa se dešava da neki od njih dobijaju minimalne negativne vrednosti, koje dovode dovode do nemogućnosti WORLD vokodera da ispravno rekonstruiše govor na osnovu dobijenih koeficijenata. Zbog dobijanja negativnih vrednosti koeficijenata

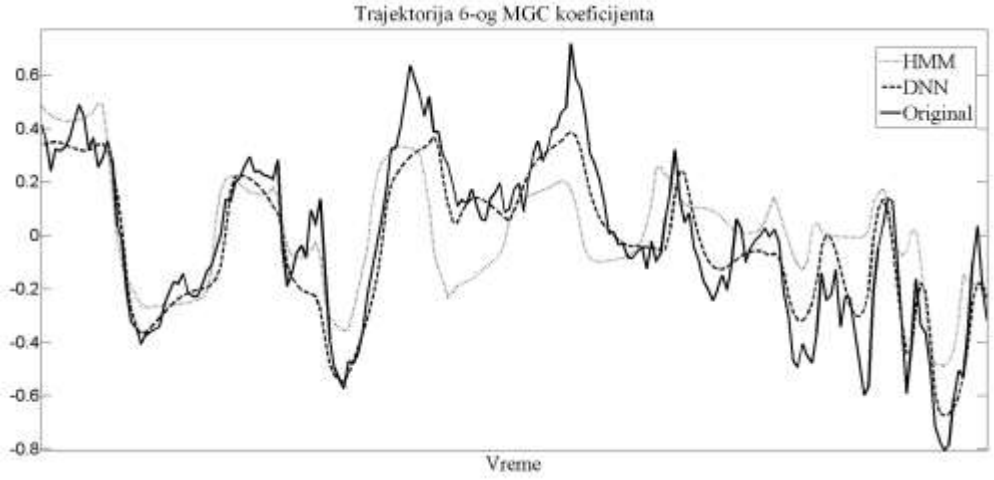
Tabela 4.1 Poređenje objektivnih mera za HMM i DNN sistem pri sintezi neutralnog govora

	MCD(dB)	BAP(dB)	F0-RMSE (Hz)	F0 CORR	VUV (%)
DNN	4.25	0.25	15.02	0.9	5.84
HMM	6.56	0.88	19.40	0.83	9.33

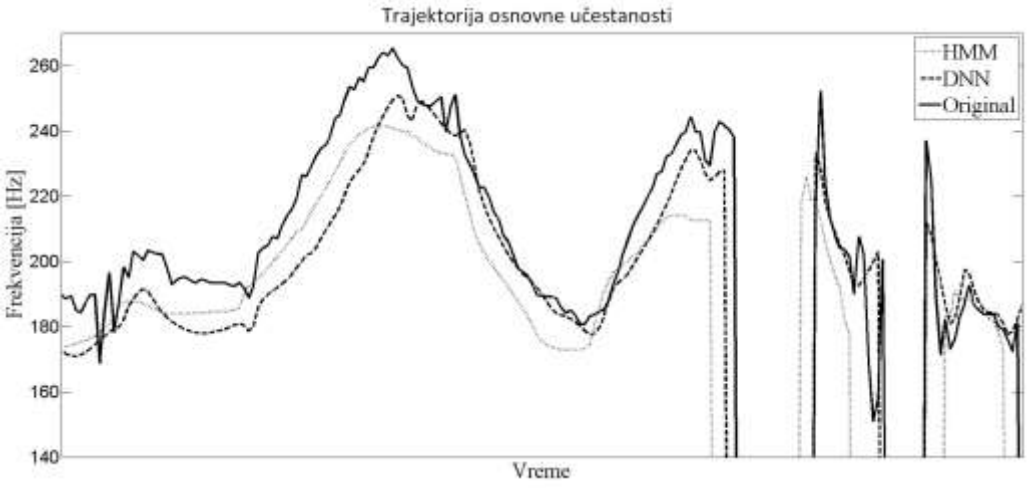
aperiodičnosti u procesu HMM sinteze vrši se dodatni korak u kom se negativni koeficijenti aperiodičnosti postavljaju na vrednost nula. Ove vrednosti se dosta razlikuju od vrednosti u originalnim rečenicama i značajno utiču na veliko odstupanje objektivne mere za aperiodičnost.

Na slici 4.14 dato je još jedno poređenje generisanih parametara. Na slici 4.14a prikazane su trajektorije šestog MGC koeficijenta generisane pomoću HMM i DNN sistema, kao i trajektorije iz originalne rečenice. Originalne i generisane trajektorije učestanosti prikazane su na slici 4.14b. Prilikom generisanja parametara pomoću oba sintetizatora korišćena su trajanja koja su dobijena na osnovu originalnih poravnanja dobijenih u toku

a



b

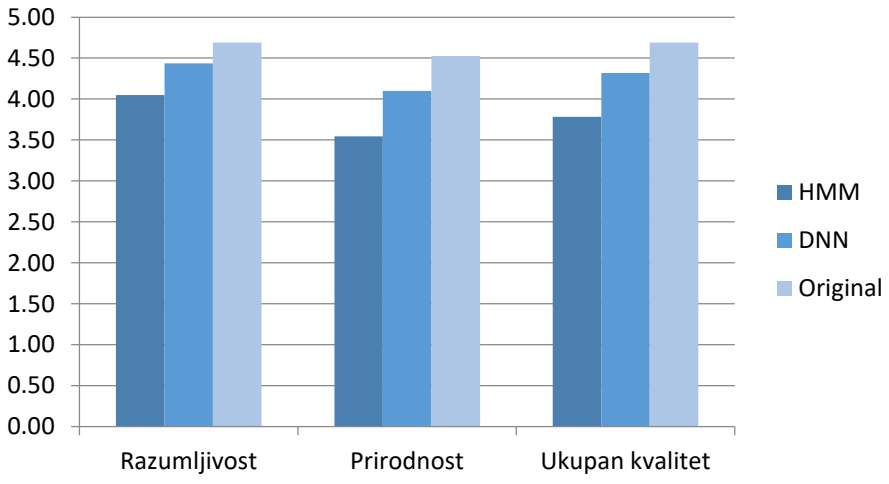


Slika 4.14 Poređenje trajektorija MGC koeficijenata (a) i osnovne učestanosti (b) generisanih od strane HMM i DNN sistema sa originalnim trajektorijama

pripreme podataka za obuku DNN sistema. Na osnovu date slike može se zaključiti da je trajektorija generisana upotrebom neuronskih mreža sličnija originalnoj trajektoriji nego trajektorija dobijena korišćenjem HMM sintetizatora. HMM trajektorija nije uspela da isprati nagle promene koje postoje u originalnoj trajektoriji.

Sistemi su međusobno poređeni i testom slušanja baziranom na pristupu opisanom u [114]. U testu je učestovalo četrdeset slušalaca. Slušaoci je trebalo da preslušaju tri audio fajla. Svaki audio fajl sadržao je četiri iste rečenice. Rečenice u prvom fajlu generisane su korišćenjem HMM sintetizatora, u drugom korišćenjem DNN sintetizatora, dok se treći fajl sastojao od originalnih rečenica iz baze. Za svaki preslušani fajl zadatak slušalaca je bio da odgovore na deset pitanja dajući ocene od 1 do 5. Od datih pitanja, pet se odnosilo na razumljivost govora, četiri na prirodnost, a u jednom pitanju je trebalo da ocene ukupan kvalitet govora. Na slici 4.15 prikazani su uprosečeni rezultati dobijeni računanjem srednjih vrednosti svih pitanja koja pripadaju jednoj kategoriji. Sa slike je vidljivo da je DNN sistem prema sva tri testirana kriterijuma nadmašio HMM sistem. Analizom pojedinačnih pitanja dolazi se do zaključka da je u svakom pitanju prosečna ocena dobijena za DNN bila veća nego za HMM sintetizator, dok se najveća ocena dobijala za originalne rečenice.

Zanimljivo je primetiti da čak ni originalne rečenice iz baze nisu ocenjene najvećom ocenom.



Slika 4.15 Rezultati subjektivnog poređenja HMM i DNN sintetizatora

5. Sinteza ekspresivnog govora

Na početku istraživanja sa ciljem sinteze ekspresivnog govora u dostupnoj literaturi nisu pronađeni pristupi predloženi za DNN sintezu. Znatno više pažnje bilo je posvećeno pokušajima da se u jednom sistemu omogući istovremena sinteza govora više govornika (engl. *multi-speaker synthesis*). Upravo pristupi koji su bili predloženi za ove potrebe poslužili su kao inspiracija za sintezu govora u više različitih stilova.

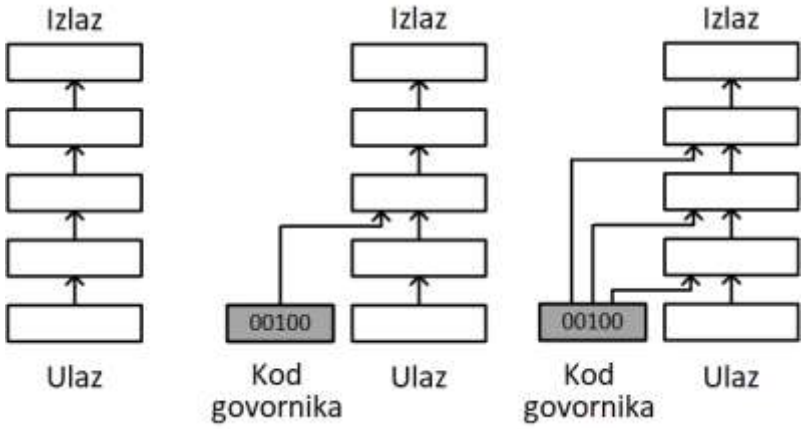
5.1. Osnovni predloženi pristup za sintezu ekspresivnog govora

U [115] predloženo je korišćenje kodova govornika kako bi se omogućila sinteza govora u više različitih glasova. Ova metoda zasniva se na proširenju ulaznih lingvističkih podataka podacima o identitetu govornika kojem pripada trenutni uzorak u obuci. Ukoliko u sistemu postoji *N* govornika, govornik sa rednim brojem *i* opisan je vektorom $S^i = [s_1^i, s_2^i, \dots, s_N^i]$ pri čemu elementi vektora S^i zadovoljavaju jednakost

$$s_k^i = \begin{cases} 1, & k = i \\ 0, & k \neq i \end{cases}. \tag{5.1}$$

U originalnom radu navodi se da se vektor koji opisuje govornika može dodati ulaznom vektoru (lingvističkih obeležja), jednom skrivenom sloju ili svim skrivenim slojevima, kao što je prikazano na slici 5.1. Detaljna analiza performansi ovog pristupa za modelovanje više govornika može se pronaći u [116].

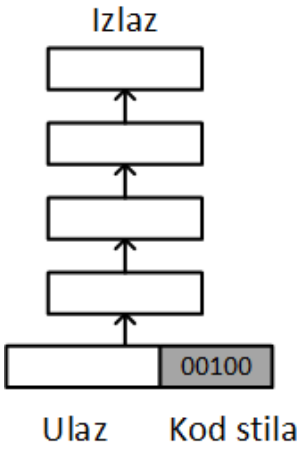
Slično kao u pristupu baziranom na kodovima govornika predloženi metod za istovremeno modelovanje više stilova ulazna lingvistička obeležja proširuje vektorom koji se formira analogno jednakosti (5.1), s tim da dodatni vektor obeležja definiše da li trenutni frejm pripada jednom od postojećih stilova u bazi za obuku.



Slika 5.1 Simultano modelovanje više govornika na bazi kodova stila

Predložena metoda, nazvana metoda kodova stila prezentovana je u [117], a grafički prikaz dat je na slici 5.2. Slične ideje su prezentovane u [118] i [119], u relativno kratkom periodu nakon objavljivanja rada [117].

Predloženi pristup može da se primeni i u HMM sintezi. Kontekst koji opisuje neki fonem proširuje se informacijom o pripadnosti odgovarajućem stilu. Ovakav pristup u stvari odgovara metodi koja je već korišćena u HMM sintezi i opisana u [70].



Slika 5.2 Predložena metoda kodova stila

5.2. Poređenje HMM i DNN pristupa u sintezi ekspresivnog govora

Za potrebe poređenja HMM i DNN pristupa u sintezi ekspresivnog govora iskorišćene su slične karakteristike odgovarajućih sistema koje su opisane u poglavlju 4.5. HMM modeli opisani su sa pet emitujućih stanja, a i svi ostali parametri u obuci HMM modela opisani u odeljku 4.5 takođe su nepromenjeni. Neuronske mreže za predikciju trajanja, kao i za predikciju akustičkih obeležja, sastoje se od četiri skrivena sloja sa po 1024 neurona. Neuroni u prva tri sloja koriste tangens-hiperboličnu funkciju kao aktivacionu funkciju, dok je četvrti sloj sastavljen od LSTM neurona. Vektor akustičkih obeležja sastoji se od 40 MGC koeficijenata, jednog koeficijenta aperiodičnosti, osnovne učestanosti kao i njihovih prvih i drugih izvoda. Vektor akustičkih obeležja je u slučaju DNN sistema proširen informacijom o zvučnosti trenutnog frejma.

Iako je poređenje performansi dva sistema u slučaju generisanja neutralnog govora dato na primeru srpskog jezika, to nije bilo moguće za potrebe poređenja ovih sistema u generisanju ekspresivnog govora, pošto u trenutku kada je istraživanje rađeno nije postojala odgovarajuća baza ekspresivnog govora za srpski jezik. Zbog nedostatka baze korišćena je baza engleskog govornika koja se sastoji od 4 sata i 20 minuta neutralnog govora i 20 minuta govora u ljutitom stilu. Učestanost odabiranja korišćenog govora je 16 kHz. Oznaka pripadnosti odgovarajućem stilu data je na nivou cele rečenice. Dakle, prilikom označavanja ekspresivnih delova baze nije se vodilo računa da li se u svakoj pojedinačnoj reči mogu prepoznati karakteristike određenog stila, nego se ta odluka donosila na nivou cele rečenice i ta informacija potom prosleđivala na pojedinačne reči i foneme.

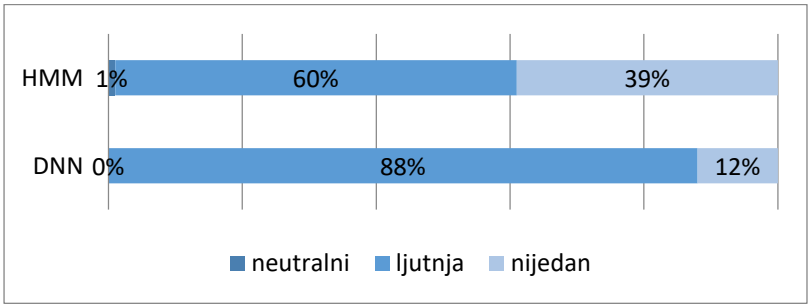
Kvalitet sintetizovanih rečenica prvo je poređen objektivno, a rezultati su dati u tabeli 5.1. I za neutralni i za ljutiti stil sve objektivne mere su bolje u slučaju DNN sistema. Rezultat koji je pomalo iznenađujući dobijen je u slučaju mel-kepstralne distorzije. Ova vrednost je manja za govor koji pripada ljutitom stilu, nego za govor koji pripada neutralnom stilu i pored značajno veće količine govora koji odgovara neutralnom stilu.

Tabela 5.1 Poređenje objektivnih mera za HMM i DNN sistem pri sintezi ekspresivnog govora

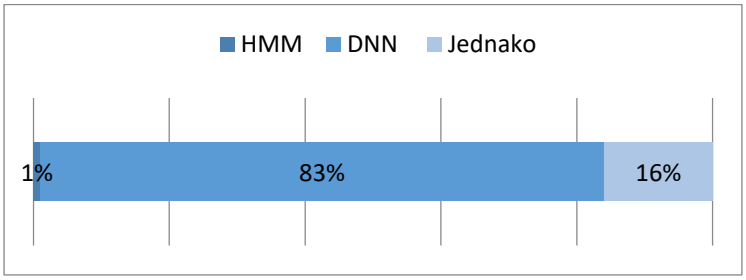
Stil	Sistem	MCD(dB)	BAP(dB)	F0-RMSE (Hz)	F0 CORR	VUV (%)
Neutralni	DNN	4.46	0.13	19.97	0.63	6.46
	HMM	6.34	0.31	20.41	0.59	9.36
Ljutiti	DNN	4.29	0.15	20.84	0.63	5.52
	HMM	6.73	0.18	23.60	0.5	8.64

Performanse datih sistema takođe su poređene i pomoću nekoliko testova slušanja. U prvom testu slušanja učestovala su 24 slušaoca. U okviru jednog zadatka slušaoci je trebalo da preslušaju dve rečenice, jednu sintetizovanu u neutralnom stilu i istu rečenicu sintetizovanu u ljutitom stilu. Posle preslušavanja trebalo je da odluče u kojoj od rečenica je više bila naglašena emocija ljutnje. Moguć je bio i odgovor da se emocija ljutnje podjednako čuje u obe rečenice. Ovaj test sastojao se od 20 parova rečenica, od čega je 10 sintetizovano korišćenjem HMM sintetizatora, a 10 korišćenjem DNN sintetizatora. Rezultati ovog testa dati su na slici 5.3. Zaključuje se da su mnogo bolji rezultati dobijeni korišćenjem DNN pristupa. U slučaju rečenica generisanih HMM metodom neki ispitanici su deo neutralnih rečenica prepoznali kao ljutiti govor, dok u slučaju DNN pristupa takvih grešaka nije bilo.

U drugom testu slušanja ispitanici su imali zadatak da preslušaju dve iste rečenice sintetizovane u ljutitom stilu. Jedna od ponuđenih rečenica dobijena je korišćenjem DNN sintetizatora, a druga korišćenjem HMM sintetizatora. Zadatak ispitanika je bio da ocene u kojoj od ponuđenih rečenica je emocija ljutnje izražena na prirodniji način. Bio je ponuđen i odgovor da je izraženost emocije u ponuđenim rečenicama jednaka. Rezultati drugog testa



Slika 5.3 Prepoznavanje emocije sintetizovane HMM i DNN pristupom



Slika 5.4 Prirodnost emocije sintetizovane HMM i DNN pristupom

slušanja prikazani su na slici 5.4 koja takođe potvrđuje nadmoćnost DNN pristupa nad HMM pristupom u sintezi ekspresivnog govora.

Dva prethodno opisana eksperimenta pokazala su da se primenom DNN pristupa postiže bolja izraženost emocije nego u slučaju korišćenja HMM sinteze. Međutim, da bi se stekla potpuna slika o performansama predložene metode neophodno je proveriti i kvalitet dobijenog emotivnog govora u poređenju sa neutralnim, uzimajući u obzir činjenicu da je količina dostupnog materijala za obuku u neutralnom stilu daleko veća nego količina emotivnog materijala. Treba imati na umu da je prema rezultatima prikazanim u tabeli 5.1 vrednost mel-kesptralne distorzije za govor u ljutitom stilu čak bila nešto veća nego vrednost dobijena za neutralni stil. Međutim, objektivne mere ne moraju uvek da budu potpuno korelisane sa rezultatima subjektivnih testova. Zbog ovog neslaganja sproveden je još jedan test slušanja u kojem je zadatak slušalaca bio da ocene ukupni kvalitet rečenica sintetizovanih korišćenjem DNN sintetizatora u standardnom MOS testu. Ispitanici su preslušavali ukupno 20 rečenica, od čega je 10 sintetizovano u neutralnom, a 10 u ljutitom stilu. Kvalitet svake rečenice je trebalo da bude ocenjen na skali od 1 do 5. Najveća ocena 5 označava najbolji mogući kvalitet, dok je najniža ocena 1. Prosečna ocena za rečenice sintetizovane u neutralnom stilu iznosila je 3.9, dok je govor sintetizovan u ljutitom stilu dobio ocenu 3.8.

Prethodni rezultati pokazuju da metoda kodova stila može uspešno da se primeni za generisanje govora u datom stilu (emociji) bez obzira na to što je količina dostupnog materijala u datom stilu nekoliko puta manja.

Pošto je nadmoćnost DNN pristupa sintezi govora nad HMM pristupom potvrđena i u slučaju sinteze ekspresivnog govora ostatak istraživanja baziran je na primeni DNN sinteze.

5.3. Detaljna analiza performansi kodova stila u DNN sintezi

U odeljku 5.2 metoda kodova stila je primenjena u sintezi samo jednog stila. Međutim, broj stilova koji se može koristiti nije ograničen. Da bi se proverilo da li uvođenje dodatnih stilova utiče na kvalitet sinteze sprovedeni su eksperimenti u kojima su uključena dva dodatna stila – srećni i pomirljivi. Osnovne karakteristike korišćenih stilova prikazane su u tabeli 5.2. Iz date tabele vidi se da srećni stil ima najveću prosečnu učestanost, kao i da je standardna devijacija osnovne učestanosti ovog stila takođe najveća. Pomirljivi i ljutiti stil slični su prema svojim karakteristikama, iako je standardna devijacija osnovne učestanosti nešto manja za ljutiti stil. Neutralni stil odlikuje najveća brzina govora, kao i najmanja prosečna vrednost osnovne učestanosti.

Tabela 5.2 Karakteristike govornih stilova korišćenih za detaljniju analizu performansi kodova stila

Stil	Brzina govora (fonema/s)	Prosečna *f0* (Hz)	Standardna devijacija *f0* (Hz)
Neutralni	12.7	98.7	34.1
Pomirljivi	10.8	101.9	25.1
Srećni	11.4	170.2	71.4
Ljutiti	10.9	103.9	30.3

Za razliku od eksperimenata opisanih u odeljku 5.2, u eksperimentima koji će biti predstavljeni u ovom odeljku korišćeno je dva sata neutralnog govora i po 10 minuta govora za svaki od stilova. U svim eksperimentima arhitektura mreže identična je arhitekturi mreže korišćene u eksperimentima opisanim u odeljku 5.2.

Eksperimenti sa dodatnim stilovima mogu se podeliti u dve grupe. U prvoj grupi eksperimenata konstruisani su odvojeni sintetizatori za svaki stil korišćenjem samo neutralnog govora i govornog materijala koji odgovara željenom stilu. Objektivne mere za svaki od pojedinačnih sintetizatora, koje su izračunate uprosečavanjem rezultata za 30 test

Tabela 5.3 Objektivne mere za sintetizatore sa samo jednim stilom

	MCD(dB)	BAP(dB)	F0-RMSE (Hz)	F0 CORR	VUV (%)
Srećni	5.50	0.19	41.48	0.79	5.59
Pomirljivi	4.70	0.13	16.85	0.73	4.88
Ljutiti	4.79	0.17	18.67	0.62	5.68

rečenica, prikazane su u tabeli 5.3. Na osnovu datih rezultata može se zaključiti da postoje određene razlike u modelovanju pojedinačnih stilova, ali da se one delimično mogu objasniti i karakteristikama datih stilova.

Najbolji rezultati dobijeni su za pomirljivi stil koji se odlikuje najmanjom brzinom govora, kao i najmanjom osnovnom učestanošću i odgovarajućom standardnom devijacijom. Najlošiji rezultati dobijeni su za srećni still, kod kojeg je dobijeno ubedljivo najveće odstupanje srednje kvadratne greške za osnovnu učestanost, ali je već ranije napomenuto da je ovo stil sa najvećom standardnom devijacijom osnovne učestanosti.

Druga grupa eksperimenata sastojala se od konstruisanja samo jednog sintetizatora upotrebom neutralnog govora i materijala koji odgovara svim dostupnim stilovima. Objektivne mere za iste test rečenice koje su korišćene za izračunavanje rezultata prikazanih u tabeli 5.3 date su u tabeli 5.4. Isti zaključci o kvalitetu stilova koji su navedeni prilikom opisa rezultata dobijenih sa samo jednim stilom mogu se navesti i u ovom slučaju – najbolji rezultati dobijaju se za pomirljivi stil, a najlošiji za srećni stil.

Direktno poređenje rezultata u modelovanju jednog stila sa simultanim modelovanjem više stilova dato je u tabeli 5.5. Predstavljeni rezultati dobijeni su uprosečavanjem rezultata

Tabela 5.4 Objektivne mere za sintetizator sa više stilova

	MCD(dB)	BAP(dB)	F0-RMSE (Hz)	F0 CORR	VUV (%)
Srećni	5.46	0.19	42.85	0.77	5.64
Pomirljivi	4.67	0.13	17.11	0.72	4.92
Ljutiti	4.75	0.17	18.46	0.63	5.70

Tabela 5.5 Poređenje objektivnih mera za modelovanje jednog stila i modelovanje više stilova

	MCD(dB)	BAP(dB)	F0-RMSE (Hz)	F0 CORR	VUV (%)
1 stil	5.00	0.16	25.55	0.71	5.38
3 stila	4.96	0.16	26.14	0.71	5.42

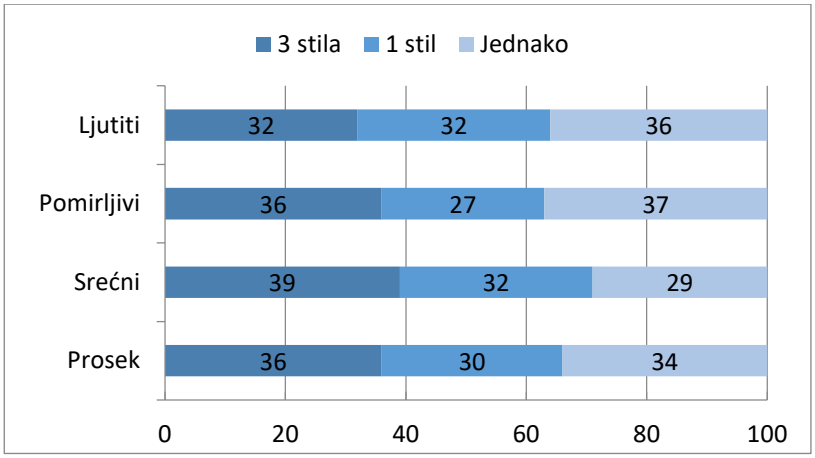
po svim dostupnim stilovima. Analizom ovih rezultata može se zaključiti da uvođenje dodatnih stilova skoro da i nema uticaja na dobijene vrednosti objektivnih mera.

Uspešnost modelovanja pojedinačnih stilova i simultanog modelovanja više stilova proverena je i testovima slušanja u kojima je učestovalo 15 slušalaca. Test slušanja sastojao se od 30 zadataka. U okviru jednog zadatka slušaocima je zadato da preslušaju dve rečenice, jednu sintetizovanu sintetizatorom koji modeluje samo jedan stil i drugu sintetizovanu korišćenjem sintetizatora koji modeluje više stilova istovremeno. Slušaoci su zatim određivali u kojoj od dve rečenice je traženi stil (odnosno emocija) bolje izražena. Bio je moguć i odgovor da je stil jednako izražen u obe rečenice. Slušaocima je jasno bilo naznačeno koji zadatak odgovara kojem stilu. Svakom od 3 korišćena stila odgovaralo je po 10 zadataka. Rezultati ovog testa prikazani su na slici 5.5.

Izraženost stila u rečenicama koje odgovaraju pomirljivom i srećnom stilu, a koje su generisane korišćenjem sintetizatora koji modeluje više stilova istovremeno, ocenjena je nešto bolje nego u slučaju rečenica generisanih sistemima koji modeluju pojedinačne stilove. Međutim, i procenat rečenica koje su slušaoci ocenili kao jednako izražajne takođe je visok. U slučaju rečenica koje pripadaju ljutitom stilu ne postoji prednost ni jednog pristupa po pitanju izraženosti.

Na slici 5.5 prikazani su i rezultati koji su dobijeni uprosečavanjem ocena za sva tri stila. I oni pokazuju da su slušaoci kao izražajnije ocenili rečenice sintetizovane sintetizatorom koji simultano modeluje više stilova.

Svi prethodni testovi pokazali su da uvođenje dodatnih stilova ne utiče na degradaciju dobijenih rezultata. Čak se i u testovima slušanja dobijaju rezultati koji blago favorizuju ovakav pristup. Po pitanju primenjivosti, sintetizatori koji modeluju više stilova istovremeno



Slika 5.5 Rezultati subjektivnog testa poređenja modelovanja jednog stila sa simultanim modelovanjem više stilova

su u prednosti, jer ne zahtevaju istovremeno rukovanje sa većim brojem neuronskih mreža. Ovo može biti naročito korisno ukoliko postoji potreba da se u istoj rečenici upotrebi više stilova (iako ovakvi eksperimenti nisu bili obuhvaćeni u istraživanju).

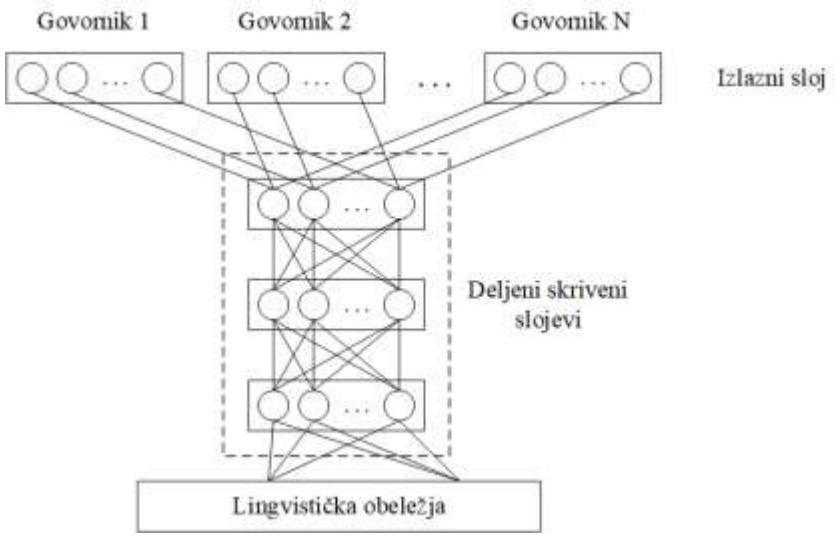
6. DNN pristupi za sintezu ekspresivnog govora korišćenjem male količine ekspresivnog materijala

Kao što je već ranije spomenuto, jedan od velikih problema prilikom sinteze ekspresivnog govora predstavlja nedostatak ili jako ograničena količina ekspresivnog materijala. Jedan od ciljeva disertacije jeste i ispitivanje mogućnosti sintetizovanja govora u različitim stilovima na osnovu ograničene količine govornog materijala za date stilove. Pored već opisanih kodova stila, kao osnovnog pristupa, u ovom poglavlju biće opisane i dve nove metode koje se mogu koristiti u ekspresivnoj sintezi, a koje su originalno predstavljene u [120].

6.1. Arhitektura sa deljenim skrivenim slojevima

Kao i u slučaju kodova stila, arhitektura sa deljenim skrivenim slojevima (ADSS) inicijalno je predložena u okviru simultanog modelovanja više govornika [121] i prikazana je na slici 6.2. Svaki od govornika u bazi predstavljen je zasebnim izlaznim slojem, dok su skriveni slojevi deljeni između svih govornika. Ovaj pristup zasniva se na pretpostavci da su zajednički slojevi uglavnom nezavisni od govornika i modeluju jezički nezavisne informacije, dok izlazni slojevi modeluju informacije koje su specifične za svakog govornika. Ovakva pretpostavka može se smatrati ispravnom uzimajući u obzir način primene neuronskih mreža u sintezi govora. Naime, neuronska mreža transformiše lingvistička obeležja u akustička. Lingvistička obeležja zavise samo od ulaznog teksta i ne uzimaju u obzir karakteristike govornika, dok akustička obeležja zavise od svakog govornika ponaosob i stoga se svaki izlazni sloj može smatrati zasebnom reprezentacijom akustičkog prostora.

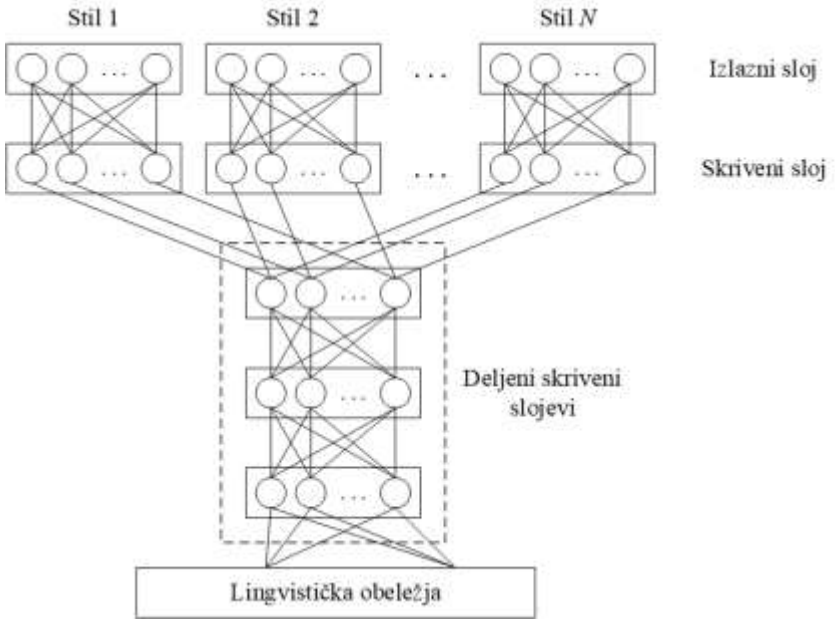
Koristeći prethodno opisane pretpostavke, u [121] predloženo je da se obuka mreže vrši tako što se propagacija unazad vrši samo kroz odgovarajuće izlazne slojeve, a potom kroz



Slika 6.2 Arhitektura sa deljenim skrivenim slojevima u modelovanju više govornika

deljene skrivene slojeve. Količina podataka za obuku koja odgovara svakom govorniku mora biti dovoljno velika, da bi ovaj pristup bio uspešan.

Slična arhitektura predložena je i za sintezu ekspresivnog govora, pri čemu izlazni slojevi ne predstavljaju govornike nego odgovarajuće stilove. Za razliku od arhitekture korišćene u



Slika 6.1 Predložena arhitektura sa deljenim skrivenim slojevima u modelovanju više stilova

modelovanju više govornika, u predloženom pristupu pored izlaznih slojeva razdvojen je i jedan skriveni sloj kao što je prikazano na slici 6.1.

Budući da je početna pretpostavka da je količina materijala po stilu ograničena izlazni, slojevi ne bi mogli dovoljno dobro da se obuče, ako bi se obučavali samo na materijalu u odgovarajućem stilu. Iz tog razloga predloženo je da se prvo izvrši obuka cele mreže (i svih izlaznih slojeva) koristeći samo neutralni materijal, a da se potom izvrši dopunska obuka neurona u razdvojenim slojevima, pri čemu vrednosti parametara u neuronima u deljenim slojevima mreže ostaju nepromenjene.

6.2. Dodatna obuka neuronske mreže

Ova ideja bazirana je na pristupu opisanom u [122]. Ovaj pristup koristi se za kreiranje novog govornika sa ograničenom količinom materijala. Polazi se od neuronske mreže koja je prethodno već kreirana za govornika sa dovoljnom količinom materijala, a potom se samo vrši dopunska obuka mreže sa materijalom koji odgovara novom govorniku. U originalnom radu pokazano je da se korišćenjem ovog pristupa i samo 10 minuta govora može postići isti kvalitet sintetizovanog govora kao i kada se mreža obučavala na sat vremena materijala za obuku. Čak i korišćenjem samo pet minuta govora željenog govornika dobijali su se prihvatljivi rezultati.

Prethodni pristup se u sintezi emocija koristi tako što se prvo obučava mreža koja odgovara samo neutralnom govoru, a potom se za svaku emociju/stil vrši dodatna obuka mreže. Ukoliko postoji N emocija kreira se N novih modela. Za razliku od pristupa opisanog u odeljku 6.1 gde se vrši samo obuka poslednjih slojeva mreže, u ovom pristupu svi delovi mreže dodatno se obučavaju.

6.3. Eksperimentalni rezultati

Evaluacija prethodno opisanih pristupa izvršena je korišćenjem baze engleskog govornika koja se sastoji od 270 minuta neutralnog govora i po pet minuta svakog od sledećih stilova: ljutnja, sreća, odrešitost. Radi se o istom govorniku čija baza je korišćena u

eksperimentima opisanim u poglavlju 5. Korišćena su ista ulazna i izlazna obeležja, kao i arhitektura mreže opisana u eksperimentima u poglavlju 5. U slučaju ADSS pristupa postoje tri zajednička skrivena sloja sa po 1024 neurona koji koriste tangens-hiperboličnu aktivacionu funkciju. Skriveni sloj koji odgovara svakom stilu sastoji se od po 1024 LSTM neurona.

Tabela 6.1 Simboličke oznake korišćenih sistema

Arhitektura	**Simbolička predstava**
Kod stila	Sis1
ADSS	Sis2
Doobuka mreže	Sis3

Radi preglednosti u svim narednim slikama i tabelama opisani sistemi predstavljeni su simboličkim oznakama koje su prikazane u tabeli 6.1.

Rezultati objektivnih testova prikazani su u tabeli 6.2. Najbolje vrednosti mel-kepstralne distorzije, bez obzira na sintetizovani stil, dobijene su korišćenjem sistema baziranog na kodovima stila (Sis1). Među ostalim objektivnim merama ne postoje drastična odstupanja u zavisnosti od korišćenog sistema. Rečenice koje pripadaju stilu ljutnje imaju najbolje vrednosti objektivnih mera, dok su one najlošije za srećni stil.

Subjektivni kvalitet sintetizovanih stilova testiran je pomoću dva testa slušanja. Prvi test

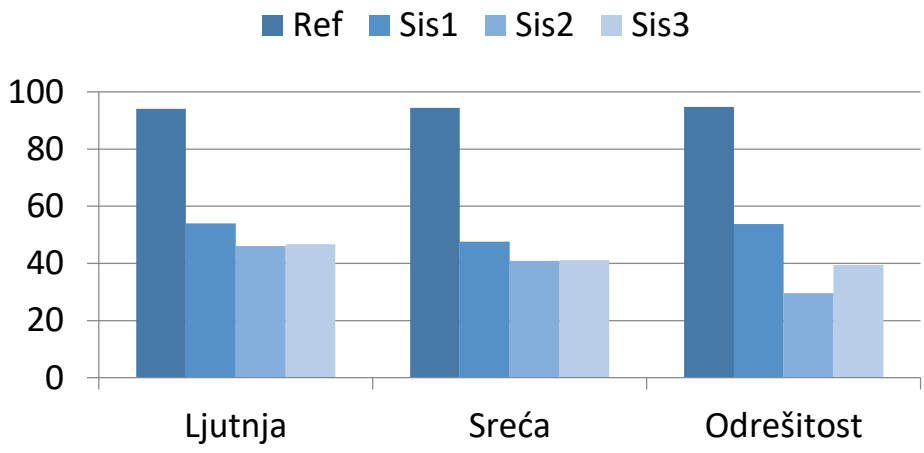
Tabela 6.2 Objektivne mere za predložene pristupe

		MCD(dB)	BAP(dB)	F0-RMSE (Hz)	F0 CORR	VUV (%)
Ljutnja	Sis1	4.82	0.18	18.62	0.61	6.23
	Sis2	5.20	0.19	18.97	0.60	6.46
	Sis3	4.96	0.18	18.61	0.61	6.23
Sreća	Sis1	5.48	0.20	40.43	0.76	5.75
	Sis2	5.86	0.21	41.05	0.76	6.26
	Sis3	5.63	0.20	38.53	0.79	5.73
Odrešitost	Sis1	5.28	0.19	34.58	0.72	4.56
	Sis2	5.64	0.21	33.34	0.73	5.15
	Sis3	5.50	0.20	33.78	0.71	5.74

predstavlja MUSHRA test i sastojao se od pet zadataka pridruženih svakom stilu. U okviru svakog zadatka, pored referentne rečenice koja je jasno bila naznačena, slušaoci je trebalo da preslušaju još četiri dodatne rečenice. Jedna od tih rečenica bila je identična referentnoj, dok ostale tri predstavljaju istu rečenicu koja je sintetizovana upotrebom jednog od tri sistema opisanih u prethodnim odeljcima. Slušaoci su ocenjivali prirodnost svake od četiri rečenice u poređenju sa referentnom na skali od 0 do 100. Učesnicima je predočeno da treba da obrate pažnju na intonaciju rečenice i smislenost stila, a da zanemare moguće artefakte koji postoje u govoru.

U originalnim MUSHRA testovima (korišćenim za testiranje kvaliteta vokodera), podrazumeva se da pored skrivenog referentnog fajla postoji i jedan fajl izrazito lošeg kvaliteta koji nije generisan ni pomoću jednog od testiranih sistema. Ovakav fajl naziva se sidro (engl. *anchor*). Međutim, u primerima MUSHRA testova korišćenih za ocenjivanje kvaliteta sinteze govora dostupnih u literaturi, nije pronađeno da se ovakvi fajlovi uključuju u proces testiranja. Stoga ni u korišćenom MUSHRA testu nisu korišćeni fajlovi koji predstavljaju sidro.

U testu je učestovalo 20 slušalaca, a rezultati su prikazani na slici 6.3. Referentni „sistem” (prirodni govor) dobio je prosečnu ocenu 94, što je i očekivano. Od predloženih pristupa najbolju ocenu dobio je sistem baziran na kodovima stila, što su pokazale i

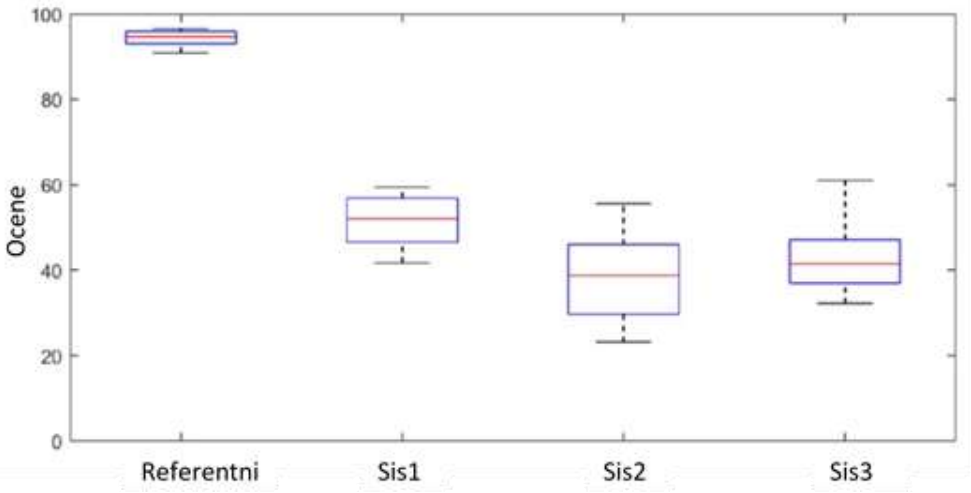


Slika 6.3 Rezultati MUSHRA testa za ocenu prirodnosti ekspresivnog govora dobijenog predloženim pristupima

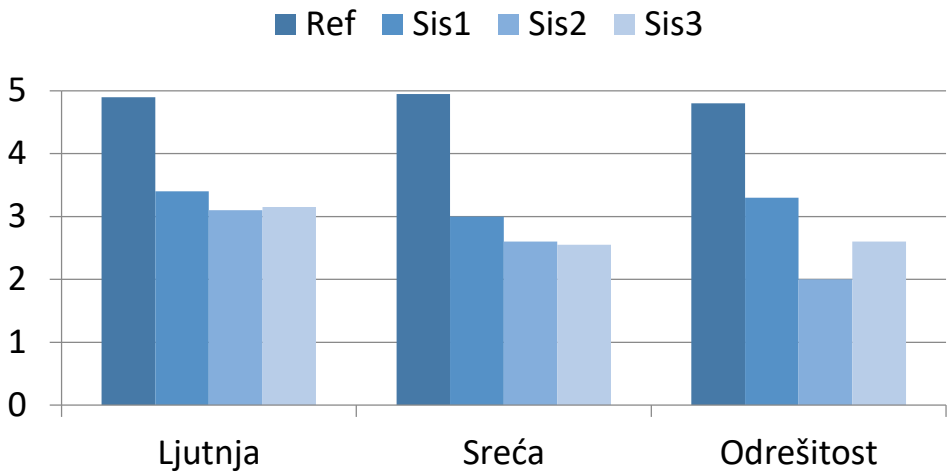
objektivne mere. Takođe, emocija ljutnje bila je najbolje ocenjena bez obzira na to koji sistem je korišćen za njenu sintezu. Prirodnost ljutitog i srećnog stila skoro podjednako je ocenjena u slučaju kada su test rečenice sintetizovane sistemima baziranim na ADSS pristupu i dodatnoj obuci mreže, dok je odrešiti stil bolje ocenjen u sintezi koja koristi dodatnu obuku mreže nego ADSS pristup.

Detaljniji prikaz dobijenih rezultata primenom *boxplot* pristupa dat je na slici 6.4. Na datoj slici crvenom bojom su predstavljene srednje vrednosti, dok plavi kvadrati opisuju opseg vrednosti koje se prostiru od prvog do trećeg kvartila. Sa prikazanih grafikona može se zaključiti da sistem baziran na kodovima stila nije samo dobio najbolju prosečnu ocenu nego se takođe pokazao i najstabilnijim. Prirodnost sistema baziranog na dodatnoj obuci je ocenjena nešto bolje nego ADSS pristup.

Drugi test slušanja predstavljao je MOS test koji je bio fokusiran na ocenjivanje ukupnog kvaliteta sistema. U okviru ovog testa slušaoci su preslušavali tri grupe rečenica. Svaka grupa predstavljala je jedan stil i sastojala se od četiri podgrupe. U okviru jedne podgrupe nalazilo se pet rečenica iz originalne baze. U preostalim podgrupama nalazile su se rečenice istog tekstualnog sadržaja kao i u prvoj podgrupi, ali su sve rečenice iz jedne podgrupe sintetizovane jednim od sistema prikazanih u tabeli 6.1. Zadatak slušalaca bio je da rečenice iz jedne podgrupe ocene od 1 (jako loš kvalitet) do 5 (najbolji mogući kvalitet – prirodni



Slika 6.4 *Boxplot* analiza MUSHRA testa



Slika 6.5 Rezultati MOS testa za ocenu kvaliteta ekspresivnog govora dobijenog predloženim pristupima

govor).

U ovom testu takođe je učestovalo 20 slušalaca, a rezultati su prikazani na slici 6.5. Rezultati ovog testa pokazuju da se govor najboljeg kvaliteta dobija korišćenjem metode kodova stila, kao i da je najbolje ocenjen govor sintetizovan u ljutitom stilu.

Upoređivanjem rezultata prikazanih na slikama 6.3 i 6.5 može se zaključiti da su dobijeni grafikoni veoma slični. Ovakvi rezultati pomalo su neočekivani, jer je u prvom testu učesnicima naglašeno da treba da zanemare eventualne artefakte u govoru i pažnju obrate na prirodnost, pre svega na intonaciju rečenica u skladu sa datim stilom, dok im je u drugom testu sugerisano da ocenjuju ukupan kvalitet rečenica. Ovakvi rezultati dovode do zaključka da je učesnicima koji nisu iskusni u polju govornih tehnologija ponekad teško objasniti razliku između pojedinih aspekata govora koje treba ocenjivati.

I objektivni i subjektivni testovi pokazali su da su najbolji rezultati, i po pitanju izraženosti stila, kao i po kriterijumu ukupnog kvaliteta sintetizovanog govora, dobijeni korišćenjem pristupa kodova stila. Pristupi bazirani na dodatnoj obuci i ADSS arhitekturi postižu približno iste rezultate, iako govor sintetizovan ADSS arhitekturom sadrži nešto više artefakata.

Među predloženim metodama samo je metoda kodova stila informaciju o stilu/emociji koristila kao novo ulazno obeležje i samo se kod nje istovremeno obučavaju svi delovi mreže bez obzira na ulazni stil. Kod druga dva pristupa parametri neuronske mreže menjaju se na osnovu ograničene količine govornog materijala po stilu, odnosno kod dodatne obuke menjaju se vrednosti svih neurona u mreži, dok se kod ADSS pristupa menjaju samo vrednosti parametara neurona koji pripadaju poslednjem skrivenom i izlaznom sloju u mreži. Na osnovu ovoga može se zaključiti da je bolje uraditi obuku cele mreže sa svim dostupnim govornim materijalom, nego raditi na izmeni samo nekih vrednosti na osnovu ograničene količine podataka po stilu.

Metoda kodova stila nije fleksibilna za dodavanje novih stilova u sistemu. Ukoliko se želi dodati novi stil u sistemu baziranom na kodovima stila neophodno je uraditi ponovnu obuku cele mreže sa svim dostupnim materijalom, dok je kod druga dva pristupa dovoljno uraditi samo dodatnu obuku sa materijalom koji pripada novom stilu.

7. Transplantacija stilova (emocija)

U uvodnom poglavlju pomenuto je da je snimanje govorne baze zahtevno i da potreba za postojanjem delova baze koji su snimljeni u različitim stilovima dodatno komplikuje proces snimanja nove baze. Ovaj problem doveo je do razvoja metoda koje omogućavaju da se glas nekog govornika sintetizuje u određenom stilu, iako govorni materijal koji odgovara tom stilu ne postoji u bazi tog govornika, već u bazi nekog drugog govornika. Ovakvi postupci se u literaturi nazivaju transplantacijom stilova ili emocija.

Prvi primeri ovakvih metoda pojavljuju se u okviru sinteze bazirane na HMM modelima. U [123] predlaže se postupak koji se sastoji od dve faze. U prvoj fazi vrši se kreiranje modela prosečnog govornika korišćenjem samo neutralnog materijala više različitih govornika. Potom se izračunavaju transformacione matrice koji model prosečnog govornika preslikavaju u model koji odgovara ekspresivnom govoru. Pretpostavka je da baza svakog govornika koji je korišćen u kreiranju modela prosečnog govornika sadrži i ekspresivne delove. U drugoj fazi kreirane adaptacione matrice primenjuju se na modelu neutralnog govora nekog novog govornika kako bi se dobio njegov ekspresivni model.

Pristup baziran na adaptaciji opisan je i u [124] i sastoji se od 3 faze. Prvo se formira prosečni model korišćenjem svog dostupnog materijala, i neutralnog i ekspresivnog. Taj prosečni model potom se adaptira korišćenjem samo neutralnog modela i dobija se neutralni prosečni model. Neutralni prosečni model služi kao polazni model za formiranje dve grupe transformacionih matrica. Prva grupa transformiše neutralni prosečni model u neutralni model nekog novog govornika, dok druga grupa transformacionih matrica služi da se neutralni prosečni model transformiše u ekspresivni model (korišćenjem polaznog ekspresivnog skupa za obuku). Ekspresivni model novog govornika dobija se sekvencijalnom primenom prethodno pomenutih grupa transformacionih matrica.

U [125] predlaže se transplantacija emocija korišćenjem emotivnog aditivnog modela (EAM). EAM predstavlja razliku između emotivnog i neutralnog govora i obučava se

korišćenjem neutralnog i emotivnog govora nekoliko različitih govornika. Emotivni model novog govornika dobija se dodavanjem EAM modela njegovom neutralnom modelu.

Metode za transplantaciju stilova razvijaju se i u okviru DNN sinteze. U [126] predložene su tri arhitekture koje omogućavaju transplantaciju stila. Paralelna arhitektura zasniva se na postojanju odvojenih izlaznih slojeva za svakog govornika, kao i odvojenih izlaznih slojeva koji odgovaraju korišćenim emocijama (stilovima). Izlazna obeležja dobijaju se kombinacijom obeležja iz izlaznog sloja odgovarajućeg govornika i odgovarajućeg stila. Izlazni deo serijske arhitekture sastoji se od dva dela. U prvom delu nalaze se slojevi koji odgovaraju govornicima, a u drugom delu slojevi koji odgovaraju emocijama. Prilikom sinteze aktivira se prvo sloj odgovarajućeg govornika, a potom se njegovi izlazi koriste kao ulaz sloja koji predstavlja stil. Treća arhitektura zasnovana je na ranije pomenutoj arhitekturi kodova stila. Informacije o trenutnom govorniku i trenutnom stilu koriste se kao dodatna ulazna obeležja mreže. Govornik sa rednim brojem i predstavljen je vektorom $S^i = [s_1^i, s_2^i, \dots, s_N^i]$, pri čemu elementi vektora zadovoljavaju jednakost

$$s_k^i = \begin{cases} 1, & k = i \\ 0, & k \neq i \end{cases}. \tag{7.1}$$

Na sličan način je i emocija sa rednim brojem j opisana vektorom $E^j = [e_1^j, e_2^j, \dots, e_N^j]$, pri čemu važi

$$e_k^j = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}. \tag{7.2}$$

Autori su treću arhitekturu nazvali arhitektura sa dodatnim ulazima. Eksperimentalni rezultati su pokazali da je po pitanju transplantacije stilova serijski model najlošiji, dok su rezultati dobijeni korišćenjem paralelnog modela neznatno bolji u poređenju sa arhitekturom sa dodatnim ulazima.

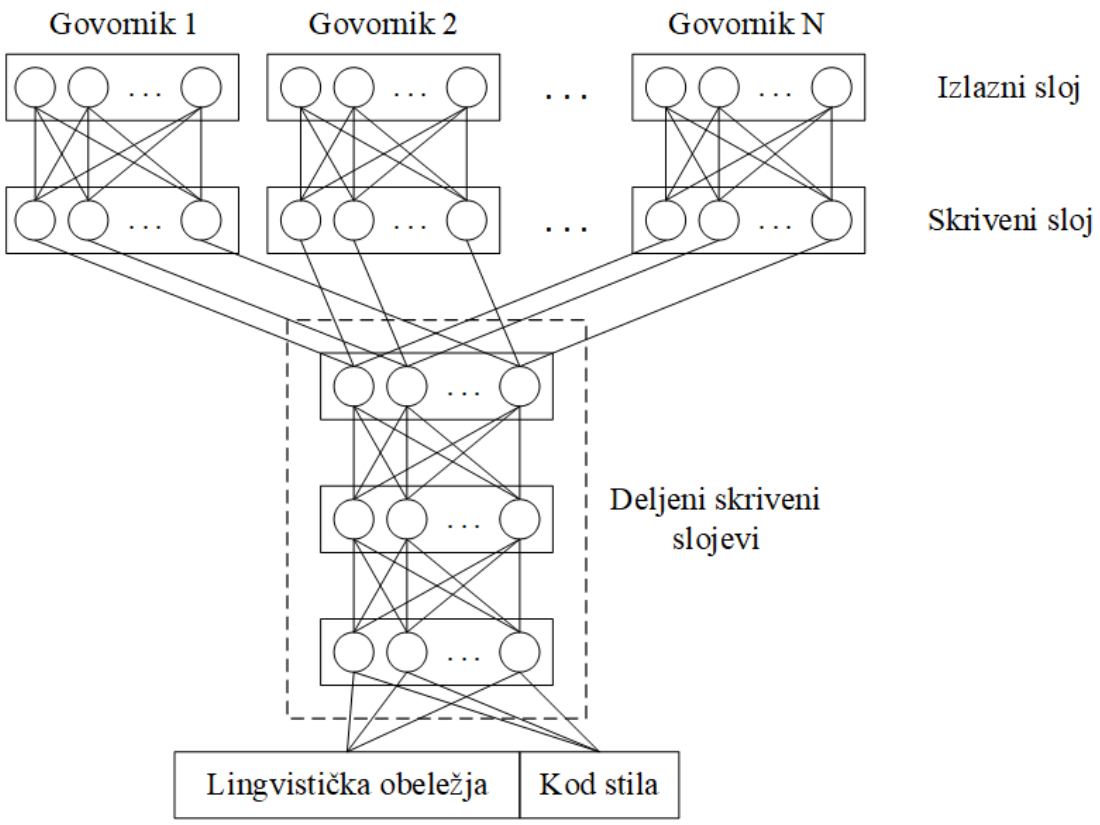
Svi prethodno opisani pristupi bazirani su na korišćenju materijala velikog broja govornika pri čemu postoji i dosta materijala ekspresivnog govora. Tako se npr. u [126] koristi baza koja sadrži 16 govornika sa samo neutralnim govorom, od čega su svi govornici snimili jedan deo baze sa istim rečenicama. Pored tih 16 govornika bazu čini i pet dodatnih govornika za koje je snimljeno oko sat vremena svakog od tri korišćena stila: neutralnog, veselog i tužnog.

Za razliku od već opisanih pristupa, u [127] predstavljena je metoda koja sistem za sintezu ekspresivnog govora, baziran na neuronskim mrežama i materijalu samo jednog govornika, adaptira korišćenjem samo neutralnog materijala nekog novog govornika. Adaptacija se vrši korišćenjem modifikovanog LHUC pristupa (engl. *Learning Hidden Layer Contribution*) [128]. Autori navode da predloženi pristup postiže dobre rezultate po pitanju sličnosti govora generisanog adaptiranim sistemom i originalnog govora, ali nešto lošije po pitanju izraženosti emocije u sintetizovanom govoru.

7.1. Predloženi pristup za transplantaciju stila

Istraživanje u vezi sa transplantacijom stilova bazirano je na pretpostavci da polazna baza sadrži mali broj različitih govornika, kao da je i količina dostupnog materijala za svaki stil ograničena.

Predložena arhitektura prikazana je na slici 7.1 i inspirisana je metodom kodova stila i ADSS arhitekturom. Informacija o stilu koristi se kao dodatna ulazna informacija u mreži, dok svaki od govornika ima odvojeni izlazni deo. Kao i u slučaju ADSS arhitekture za modelovanje stilova kod jednog govornika, opisane u odeljku 6.1, predložena arhitektura razlikuje se od originalne ADSS arhitekture u činjenici da deo koji odgovara razdvojenim izlazima ne uključuje samo linearni izlazni sloj, nego i jedan skriveni sloj mreže. Inicijalni eksperimenti su pokazali da se bolji rezultati sa predloženom arhitekturom postižu primenom tzv. postupka „uskog grla” (engl. *bottleneck*). Ovaj postupak podrazumeva da se u jednom od skrivenih slojeva koristi manji broj neurona nego u ostalim skrivenim slojevima. Primeri primene „uskog grla” u DNN sintezi mogu se pronaći u [129]. Uloga deljenih skrivenih slojeva u ADSS arhitekturi jeste formiranje globalnih transformacija između ulaznih i izlaznih obeležja nezavisnih od govornika, a smanjenje broja neurona u nekom od skrivenih slojeva pomaže kreiranju kompaktnijih transformacija.



Slika 7.1 Predložena arhitektura za transplantaciju stilova

7.2. Eksperimentalni rezultati

U ovom odeljku biće predstavljeni rezultati transplantacije emocija korišćenjem predloženog pristupa. Predloženi pristup biće poređen sa ranije pomenutom arhitekturom sa dodatnim ulazima (ADU). U [126] pored ADU arhitekture predložene su još dve dodatne arhitekture: serijska i paralelna. Serijska arhitektura je pokazala najlošije rezultate od predloženih i zbog toga je isključena iz poređenja. Paralelna arhitektura, zasnovana na postojanju razdvojenih izlaznih slojeva za govornike i stilove, ne može biti uspešno obučena zbog ograničene količine ekspresivnog materijala koji je korišćen u eksperimentima, i iz tog razloga takođe nije uključena u poređenje.

U nastavku će se nova arhitektura, opisana u poglavlju 7.1, označavati skraćenicom ADSST (arhitektura sa deljenim skrivenim slojevima za potrebe transplantacije).

7.2.1. Baza za transplantaciju

Za potrebe testiranja performansi predložene arhitekture korišćena je govorna baza na engleskom jeziku koju čini materijal jednog muškog i jednog ženskog govornika. Delovi govornog materijala koji odgovara muškom govorniku korišćeni su u eksperimentima opisanim u prethodnim poglavljima. U eksperimentima čiji rezultati su prikazani u ovom odeljku korišćeno je 2 sata neutralnog govora za svakog govornika i po 10 minuta govora svakog od tri stila: srećni, izvinjavajući i odrešiti. U procesu snimanja stilovi su govornicima opisani na sledeći način:

- **srećni** – stil kojim operater u pozivnom centru pozivaocu saopštava neku veoma važnu vest, kao npr: „Upravo ste osvojili 1000 dolara”,
- **izvinjavajući** – stil kojim operater u pozivnom centru pozivaocu saopštava postojanje određenog problema, kao što je npr. situacija da je njegov račun blokiran zbog greške u sistemu,
- **odrešiti** – stil kojim operater u pozivnom centru razgovara sa zahtevnim mušterijom koji ima problem sa razumevanjem nekih jednostavnih instrukcija zbog čega kod operatera postoji i mala doza nervoze.

Govorna baza snimljena je sa ciljem kreiranja TTS sistema koji može da se primenjuje u pozivnim centrima i u skladu sa tim potrebama su dati stilovi i definisani. Sadržaj rečenica koje odgovaraju ekspresivnom delu baze nije stilski zavisan. Učestanost odabiranja u korišćenom materijalu je 16 kHz.

Osnovne karakteristike baze prikazane su u tabeli 7.1. Može se primetiti da muški govornik govori brže u odnosu na ženskog. Jedino je u slučaju odrešitog stila prosečna brzina govora ista. Oba govornika najbrže govore u neutralnom stilu, a najsporije u odrešitom. Prosečne vrednosti osnovne učestanosti znatno su veće kod ženskog govornika što je i očekivano. Najveća prosečna učestanost kod oba govornika izražena je u srećnom stilu, dok je je najniža u neutralnom stilu. Standardna devijacija osnovne učestanosti za oba govornika skoro je identična u neutralnom i odrešitom stilu, dok bitna odstupanja postoje u preostala dva stila. Standardna devijacija je dosta veća u srećnom stilu muškog govornika u poređenju sa istim stilom ženskog govornika, dok za izvinjavajući stil važi obrnuto. Ovakva situacija je

Tabela 7.1 Karakteristika govornih stilova korišćenih za analizu performansi transplantacije stilova

	Muški govornik			Ženski govornik		
	Brzina govora (fonema/s)	Prosečna *f0* (Hz)	Standardna devijacija *f0* (Hz)	Brzina govora (fonema/s)	Prosečna *f0* (Hz)	Standardna devijacija *f0* (Hz)
Neutralni	12.7	98.7	34.1	11.5	188.3	34.1
Srećni	11.4	170.2	71.4	11.0	239.7	53.3
Izvinjavajući	10.8	101.9	25.1	9.7	215.7	38.4
Odrešiti	9.5	131.0	50.4	9.5	216.3	50.6

i očekivana pošto različiti govornici mogu na različite načine iskazati iste emocije u govoru. Tako se u [130] navodi da je možda i glavni uzrok degradacije performansi sistema za prepoznavanje emocija iz govora upravo varijabilnost između različitih govornika, iako i neki drugi faktori mogu imati uticaja na ponašanje sistema.

7.2.2. Opis poređenih sistema

Akustička obeležja koja su korišćena odgovaraju obeležjima pomenutim u ranijim odeljcima i izdvojena su pomoću WORLD vokodera. Vektor statičkih obeležja sastoji se od 40 MGC koeficijenata, osnovne učestanosti, jednog BAP obeležja i jednog VUV obeležja. Za sva statička obeležja, osim za VUV obeležje, koriste se i delta i delta-delta obeležja. Ukupna dimenzionalnost vektora akustičkih obeležja je 127.

Vektor lingvističkih obeležja sastoji se od 540 obeležja (uz podsećanje da se na ulazu u mrežu za predikciju akustičkih obeležja koristi 9 dodatnih pozicionih obeležja). U pomenutih 540 obeležja uključen je i kod stila dužine 4. ADU sistem pored ovih obeležja uključuje i kod govornika dužine 2 (s obzirom da se koriste samo dva govornika).

Obe arhitekture sastoje se od 4 skrivena sloja, od čega prva tri skrivena sloja koriste tangens-hiperboličnu aktivacionu funkciju. U inicijalnim eksperimentima broj neurona u svim skrivenim slojevima postavljen je na 1024. Pokazalo se da se upotrebom tehnike „uskog grla" u ADSST pristupu mogu postići bolji rezultati. Isprobano je nekoliko različitih vrednosti broja neurona u skrivenim slojevima, a najbolji rezultati su dobijeni za sledeću

kombinaciju veličine skrivenih slojeva: 1024-512-64-512. Tehnika „uskog grla" takođe je isprobana i za ADU pristup, međutim nisu uočena bitna poboljšanja i zbog toga broj neurona u svim skrivenim slojevima za ovu arhitekturu nije menjan u odnosu na početnih 1024 neurona po sloju.

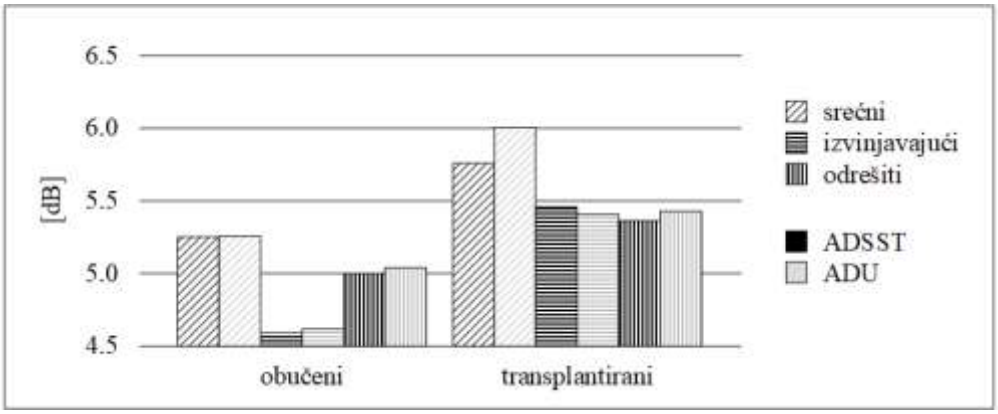
Oba pristupa su poređena po pitanju kvaliteta obučenih i transplantiranih stilova. Pojam obučenog stila odnosi se na govor generisan u određenom stilu pomoću sistema koji je obučavan na svom dostupnom materijalu, a koji uključuje materijal oba govornika u sva tri stila. Pojmom transplantirani stil označava se govor sintetizovan u nekom stilu, a pri čemu za tog govornika dati stil nije bio uključen u postupak obuke, ali je postojao u materijalu koji odgovara drugom govorniku.

7.2.3. Objektivne mere

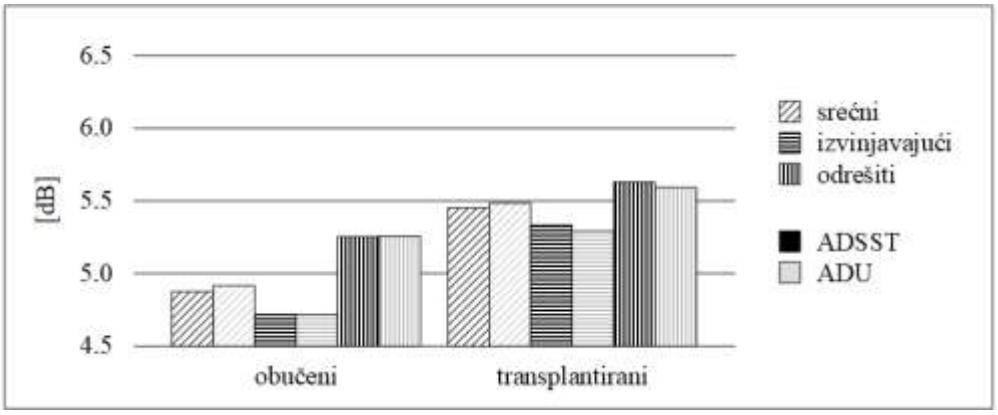
Zbog jednostavnosti prikaza rezultata analiza objektivnih mera biće bazirana na analizi mel-spektralne distorzije (MCD) i srednje kvadratne greške (RMSE) osnovne učestanosti. Dobijene vrednosti za MCD za muškog govornika prikazane su na slici 7.2a, a za ženskog govornika na slici 7.2b. Primećuje se da su vrednosti očekivano lošije u slučaju transplantiranog stila nego u obučenom stilu, bez obzira na govornika i stil. Obe testirane arhitekture postižu približno iste rezultate za govor koji pripada obučenim stilovima muškog govornika. Izvinjavajućem i odrešitom transplantiranom stilu muškog govornika takođe odgovaraju približno iste vrednosti MCD bez obzira na korišćeni pristup, sa razlikama manjim od 0.04 dB i 0.07 dB. Za srećni transplantirani stil vrednost MCD je za 0.24 dB bolja kod govora generisanog pomoću predloženog ADSST pristupa. U slučaju ženskog govornika obe arhitekture postižu približne vrednosti za svaki od stilova, sa međusobnim razlikama manjim od 0.05 dB.

Iz gornje analize zaključuje se da pri transplantaciji stilova oba pristupa generišu govor čije su vrednosti mel-kepstralne distorzije približne u većini primera, ali da je ADSST nadmašio ADU u slučaju koji je po pitanju transplantacije možda i najzahtevniji, a to je transplantacija srećnog stila sa ženskog na muškog govornika. Naime, iz tabele 7.1 vidi se da je srećni stil muškog govornika stil sa najvećom standardnom devijacijom osnovne

a



b



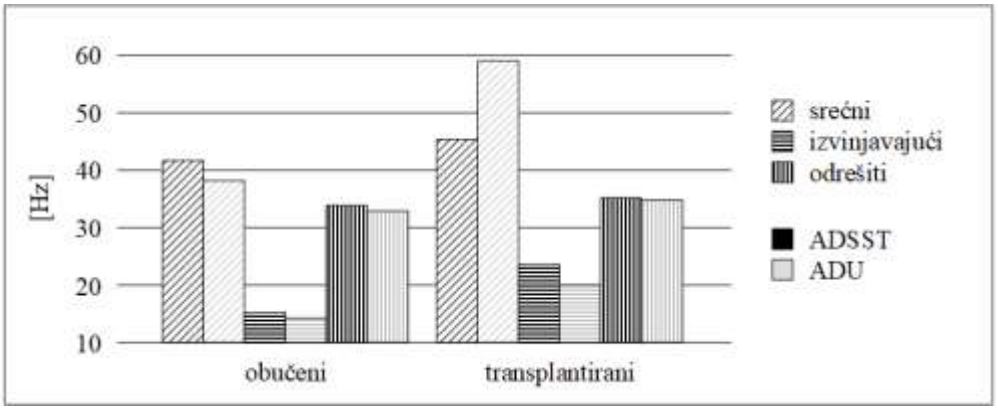
Slika 7.2 Poređenje MCD vrednosti muškog (a) i ženskog (b) govornika za obučeni i transplantirani stil

učestanosti, kao i stil sa najvećim odstupanjem prosečne učestanosti u odnosu na neutralni stil.

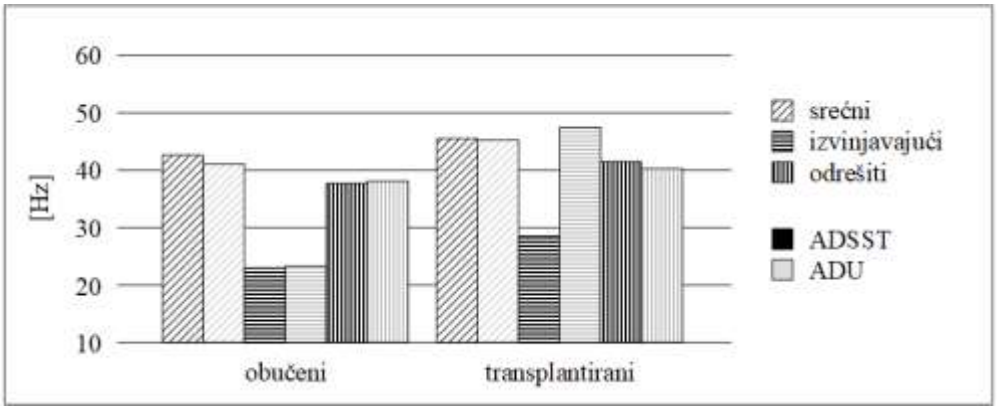
Poređenje RMSE vrednosti osnovne učestanosti dato je na slici 7.3a za muškog govornika, a za ženskog na slici 7.3b. Analizom rezultata može se doći do sličnih zaključaka kao i u slučaju analize MCD mera. Oba pristupa postižu podjednaku grešku kada se generiše govor koji odgovara obučenim stilovima. Transplantacija srećnog stila sa ženskog govornika na muškog ponovo je bila problematična, a ADSST pristup je i u ovom slučaju nadmašio ADU i to za 13.6 Hz.

Pored srećnog stila ADSST pristup je postigao i bolji rezultat u slučaju transplantacije izvinjavajućeg stila sa muškog na ženskog govornika. I ova situacija se može objasniti analizom osnovnih osobina stilova prikazanih u tabeli 7.1. Naime, izvinjavajući stil ženskog govornika ima veću standardnu devijaciju nego isti stil muškog govornika. Iz ovog se može

a



b



Slika 7.3 Poređenje RMSE vrednosti osnovne učestanosti muškog (a) i ženskog (b) govornika za obučeni i transplantirani stil

zaključiti da se ADSST u opštem slučaju ponaša bolje od ADU pristupa u slučaju transplantacije govora sa govornika sa manjom varijabilnošću osnovne učestanosti nekog stila ka govorniku sa većom varijabilnošću osnovne učestanosti govora koji pripada istom stilu.

Takođe se može uočiti da je razlika objektivnih mera između obučenog i transplantiranog stila najmanja za odrešiti stil. Iz tabele 7.1 vidi se da je standardna devijacija osnovne učestanosti govora koji odgovara odrešitom stilu oba govornika skoro identična. Iz ovog se može zaključiti da je sličnost nivoa ekspresivnosti nekog stila među govornicima bitan faktor za uspešnost transplantacije stila.

7.2.4. Subjektivna evaluacija

Kvalitet sintetizovanog govora generisanog pomoću oba pristupa takođe je proveren i pomoću dva odvojena testa slušanja.

U prvom testu slušanja zadatak učesnika je bio da rečenice koje čuju klasifikuju prema pripadnosti jednom od tri stila. Pored govora generisanog poređenim sistemima ovaj test uključivao je i klasifikaciju rečenica resintetizovanih korišćenjem originalnih akustičkih obeležja izdvojenih iz test rečenica. Korišćenjem resintetizovanih, a ne originalnih test rečenica, želeo se izbeći eventualni uticaj artefakata vokodera na rezultat klasifikacije.

U testu je učestovalo 30 slušalaca. Svaki slušalac trebalo je da klasifikuje ukupno 60 rečenica, odnosno 20 rečenica po stilu, koje su reprodukovane na slučajan način. Svaka grupa od 20 rečenica sastoji se od pet podgrupa od po 4 rečenice. Podgrupe odgovaraju različitim pristupima za generisanje govora – svaka od dve arhitekture može da generiše govor za obučeni ili transplantirani stil, a peta podgrupa predstavlja resintetizovane fajlove. Na početku testa učesnicima je dat opis svakog od stilova koje je trebalo da prepoznavaju, a opis stilova bio im je dostupan i tokom samog testa.

Matrica konfuzije koja odgovara prepoznavanju stilova prikazana je u tabeli 7.2, a grafički prikaz tačnosti klasifikacije dat je na slici 7.4. Prvo treba primetiti da je rezultat prepoznavanja stilova za resintetizovane rečenice ženskog govornika zadovoljavajući i iznosi 90%, dok je kod muškog govornika lošiji i iznosi 63%. Lošijoj tačnosti prepoznavanja najviše je doprineo odrešiti stil, čija tačnost prepoznavanja iznosi svega 33% i koji je najčešće bio zamenjen sa izvinjavajućim stilom.

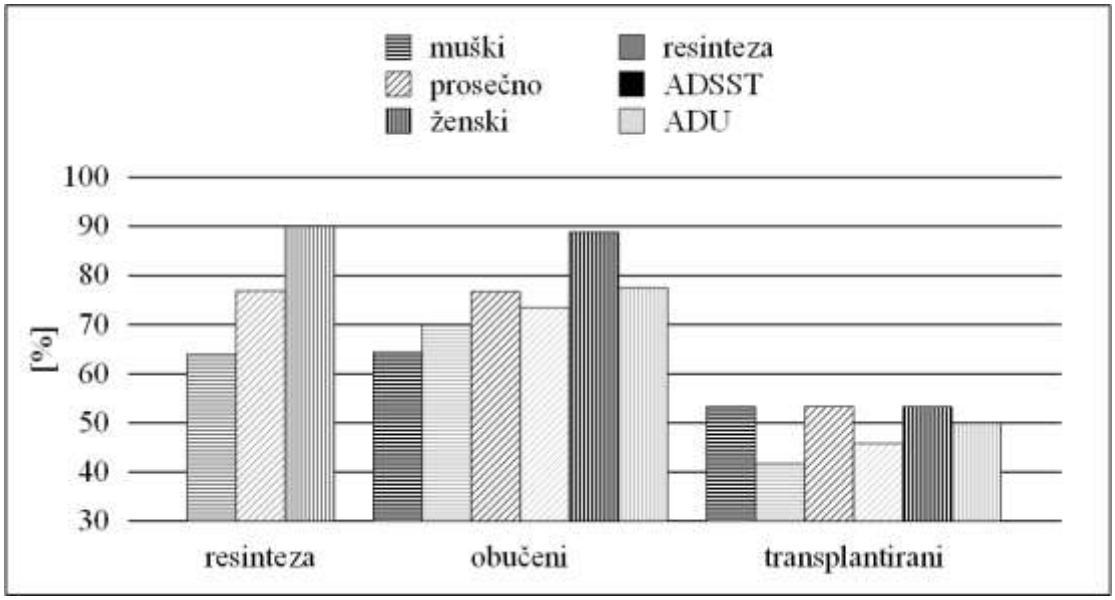
Rečenice koje odgovaraju obučenim stilovima su tačnije klasifikovane za sve stilove oba govornika u odnosu na rečenice dobijene transplantacijom što odgovara rezultatima prikazanim u [126]. Tačnost klasifikacije rečenica koje pripadaju obučenim stilovima ženskog govornika veća je nego kod muškog govornika, što se takođe poklapa sa rezultatima dobijenim za resintetizovane rečenice.

Tabela 7.2 Matrica konfuzije za prepoznavanje stilova

[%]		Resinteza			Obučeni						Transplantirani					
					ADSST			ADU			ADSST			ADU		
		S	I	O	S	I	O	S	I	O	S	I	O	S	I	O
Muški	Srećni	**80**	0	20	**83**	0	17	**87**	0	13	**80**	10	10	**50**	17	33
	Izvinjavajući	8	**78**	13	0	**87**	13	7	**73**	20	10	**37**	53	23	**32**	45
	Odrešiti	23	43	**33**	23	53	**23**	33	17	**50**	37	20	**43**	40	17	**43**
	Tačnost	63.9			64.4			70.0			53.3			41.7		
Ženski	Srećni	**98**	2	0	**93**	0	7	**63**	3	33	**52**	10	38	**30**	5	65
	Izvinjavajući	3	**83**	13	2	**92**	7	3	**93**	3	5	**60**	35	5	**62**	33
	Odrešiti	12	0	**88**	17	2	**82**	20	5	**75**	35	17	**48**	35	7	**58**
	Tačnost	90.0			88.9			77.2			53.3			50.0		
Prosek	Srećni	**89**	1	10	**88**	0	12	**75**	2	23	**66**	10	24	**40**	11	49
	Izvinjavajući	6	**81**	13	1	**89**	10	5	**83**	12	8	**48**	44	14	**47**	39
	Odrešiti	18	22	**61**	20	28	**53**	27	11	**63**	36	18	**46**	38	12	**51**
	Tačnost	76.9			76.7			73.6			53.3			45.8		

Tačnost prepoznavanja i za obučene i za transplantirane stilove veća je za rečenice koje su generisane korišćenjem predložene ADSST arhitekture. Ova arhitektura postiže tačnost od 76.7% za obučene stilove, što se poklapa sa tačnošću klasifikacije rečenica dobijenih resintezom, dok za iste test rečenice ADU pristup ima tačnost od 73.6%. U slučaju transplantiranih stilova tačnost klasifikacije za ADSST pristup je 53.3%, dok je za ADU pristup tačnost 45.8%. Iako se ovi rezultati poklapaju sa rezultatima objektivnih mera i pokazuju prednost ADSST arhitekture nad ADU arhitekturom, ipak postoje određene razlike za različite stilove.

Prednost ADSST pristupa nad ADU pristupom u transplantaciji srećnog stila sa ženskog na muškog govornika, potvrđena je i objektivnim merama, kao i rezultatima klasifikacije stilova. Međutim, drugi slučaj u kojem je ADSST nadmašio ADU u pogledu objektivnih mera za transplantirani stil, a to je transplantacija izvinjavajućeg stila ka ženskom govorniku,



Slika 7.4 Tačnost prepoznavanja stilova

nije potvrđen rezultatima klasifikacije. Naime, tačnost klasifikacije, u ovom slučaju približno je jednaka za oba pristupa.

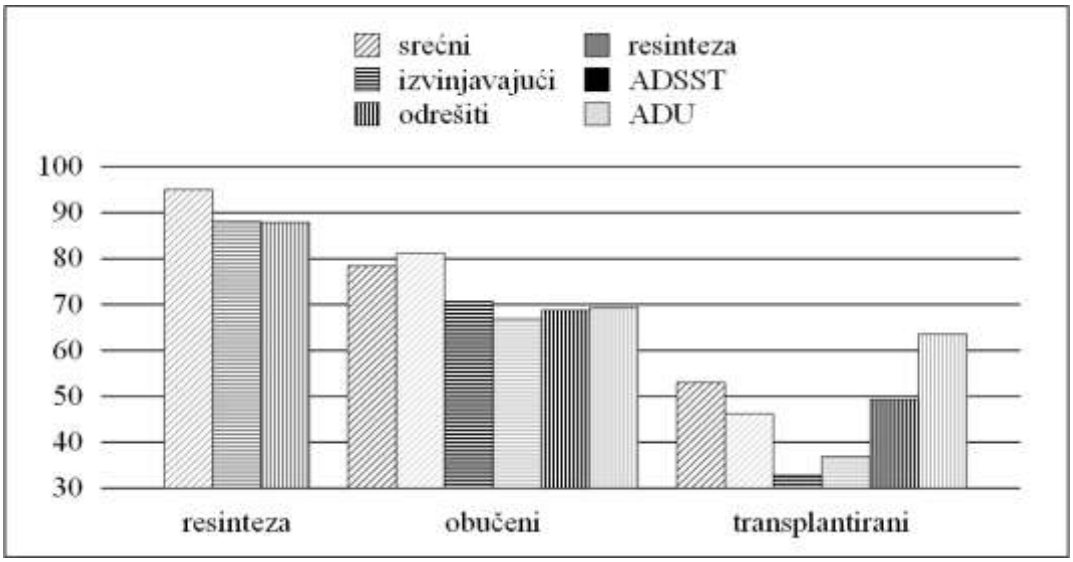
Semantički sadržaj rečenica koje su korišćene u testu je neutralan, tj. sadržaj rečenice nije odgovarao stilu u kojem je rečenica izgovorena. U [131], [132] pokazano je da se zadovoljavajuća tačnost prepoznavanja emocija može postići i u semantički neutralnim rečenicama. Međutim, određena istraživanja pokazala su da je rezultat prepoznavanja emocija ipak bolji ukoliko rečenice sadrže i semantičke karakteristike stila, a ne samo prozodijske [133]. O uticaju semantičkog sadržaja na klasifikaciju sintetizovanog ekspresivnog govora nisu pronađeni podaci u literaturi, tako da ovo može biti jedan novi zanimljiv pravac u daljem istraživanju.

Da bi se dodatno ispitala sličnost sintetizovanih rečenica, ali i ocenio ukupni kvalitet sinteze, sproveden je drugi subjektivni test, baziran na MUSHRA testovima. Korišćeno je ukupno 18 rečenica – tri rečenice po svakom stilu za oba govornika. Referentna rečenica dobijena je resintezom, tj. korišćenjem originalnih akustičkih obeležja (ponovo se želeo ukloniti eventualni uticaj vokodera na rezultate). Predložene rečenice sastoje se od rečenica generisanih pomoću oba sistema za slučaj obučenog i transplantiranog stila, rečenice sintetizovane u neutralnom stilu i ponovljene referentne rečenice.

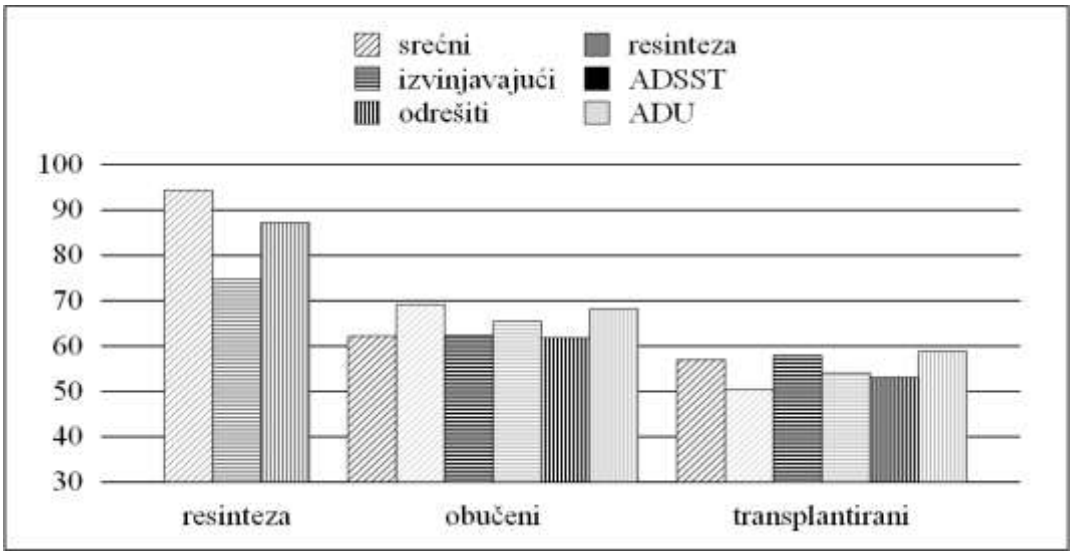
Ovaj test sastojao se od dva dela. U prvom delu svaki od učesnika je trebalo da oceni da li rečenice odgovaraju željenom stilu. Učesnicima je skrenuta pažnja da prilikom ovog ocenjivanja odbace eventualne artefakte koje čuju u govoru. U drugom delu testa, koji je takođe uključivao 18 rečenica, trebalo je da ocene ukupni kvalitet sintetizovanog govora. U oba testa minimalna ocena je iznosila 0, a maksimalna 100.

Rezultati ocenjivanja sličnosti prikazani su na slici 7.5. Iako rezultati zavise od govornika i od stila, mogu se doneti određeni zaključci. Najbolje su ocenjene resintetizovane rečenice,

a



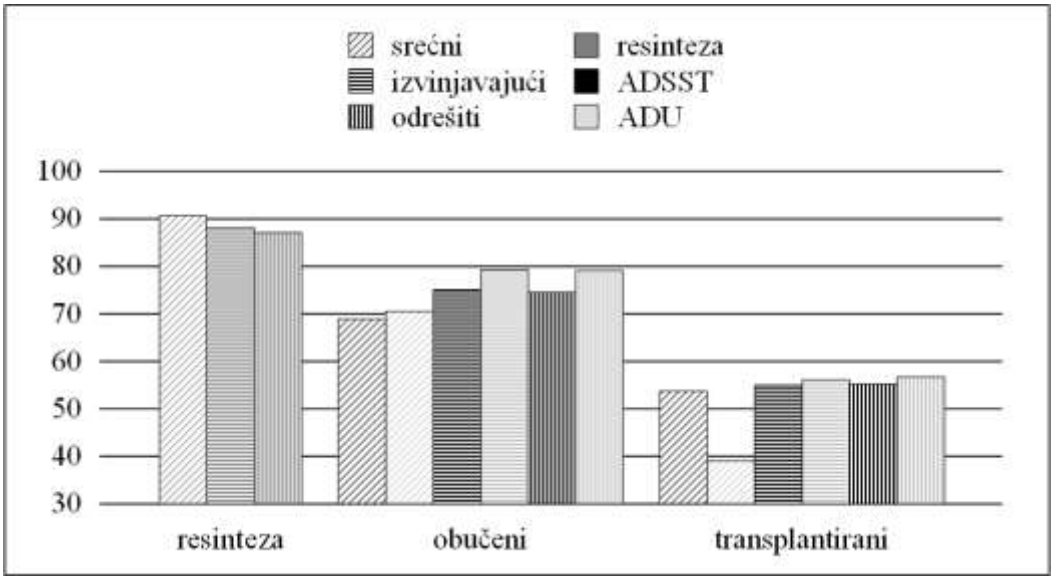
b



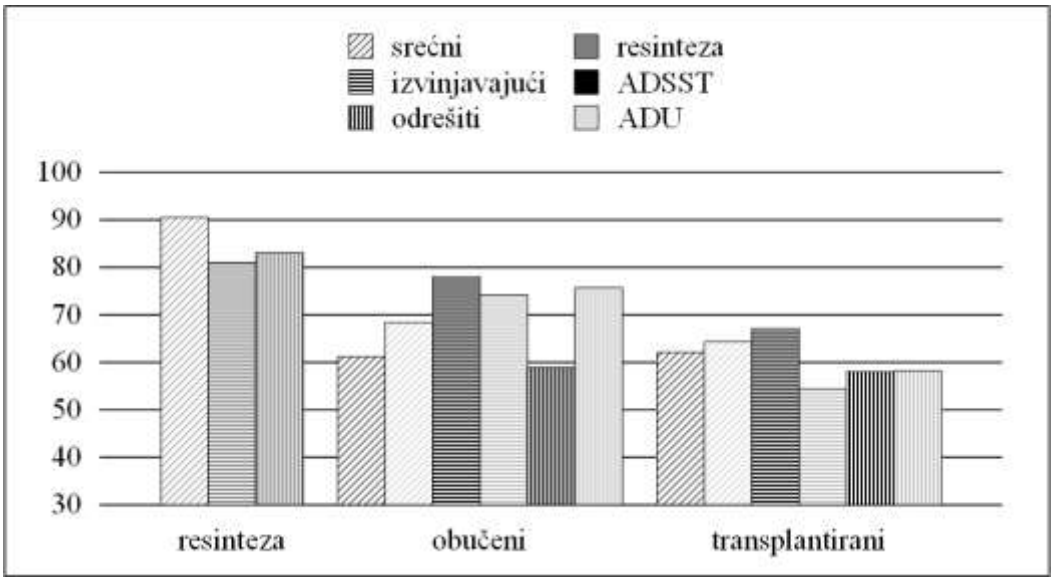
Slika 7.5 Rezultati MUSHRA testa ocene sličnosti sintetizovane emocije sa originalnom za muškog (a) i ženskog (b) govornika

što je i bilo očekivano. Prosečna ocena sličnosti resintetizovanog govora iznosi 90.3 za muškog govornika, a za ženskog 85.4. Iako su rezultati klasifikacije stila u slučaju muškog govornika bili lošiji nego za ženskog govornika, u situaciji u kojem je slušaocima poznata referentna rečenica postižu se dobri rezultati. Vrednosti dobijene za obučene stilove su veće nego vrednosti dobijene za transplantirane stilove, a ADU arhitektura nešto je bolje ocenjena nego ADSST. Prosečna ocena sličnosti obučenih stilova za ADU arhitekturu je 70, a za

a



b



Slika 7.6 Rezultati MUSHRA testa ocene ukupnog kvaliteta sintetizovanog govora za muškog (a) i ženskog (b) govornika

ADSST 67. Za transplantirane stilove je sličnost govora generisanog ADU pristupom ocenjena sa 51.7, a ADSST pristupom 50.5.

Prikaz rezultata ocenjivanja ukupnog kvaliteta sintetizovanog govora dat je na slici 7.6. Resinteza je ponovo najbolje ocenjena, a govor koji pripada obučenim stilovima je bolje ocenjen nego govor transplantiranih stilova. Kvalitet govora obučenih stilova sintetizovanog pomoću ADU arhitekture ocenjen je sa 71.2, a govora generisanog pomoću ADSST 66.1. Za transplantirane stilove je kvalitet govora generisanog ADU pristupom ocenjen sa 54.8, a ADSST pristupom 58.5.

ADU pristup je po pitanju kvaliteta govora obučenih stilova ocenjen bolje nego ADSST, međutim ADSST je nadmašio ADU u slučaju transplantiranih stilova.

8. Zaključak

Od sistema za konverziju teksta u govor se na početku njihove primene očekivalo da generisani govor bude razumljiv i sa što je moguće manje artefakata. Smatra se da su savremeni TTS sistemi ispunili ove zahteve, ali se sada moraju suočiti sa novim izazovima. Naime, više nije dovoljno da govor bude sintetizovan samo jednim stilom. Potrebno je da generisani govor svojim karakteristikama odgovara kontekstu u kojem se i koristi. U okviru ove disertacije analizirani su postupci koji omogućavaju sintezu ekspresivnog govora korišćenjem parametarskih pristupa sintezi govora.

Kao osnovna metoda za sintezu ekspresivnog govora predložena je metoda kodova stila koja je uspešno primenjena kako u okviru HMM sinteze, tako i u DNN sintezi. Pokazano je da se DNN pristupom postiže bolja izraženost emocija. Ukupni kvalitet ekspresivnog govora nije se značajno smanjio u odnosu na kvalitet neutralnog govora, iako je količina ekspresivnog materijala nekoliko puta manja nego količina neutralnog govornog materijala. Pošto je takođe potvrđeno da DNN pristup nadmašuje HMM pristup po pitanju ukupnog kvaliteta govora, ostatak istraživanja urađen je samo za sintezu zasnovanu na korišćenju dubokih neuronskih mreža.

Pored osnovnog pristupa za sintezu ekspresivnog govora upotrebom kodova stila, predložena su još dva dodatna pristupa: dodatna obuka neuronske mreže i sinteza zasnovana na arhitekturi sa deljenim skrivenim slojevima. Uporedna analiza sva tri pristupa na ograničenoj količini materijala nekoliko dostupnih stilova pokazala je da se najbolja prirodnost i kvalitet govora postižu upotrebom metode kodova stila, dok su performanse ostala dva pristupa podjednake.

Zahtev za postojanje ekspresivnih delova usložnjava proces snimanja govorne baze. Ovaj zahtev doveo do razvoja metoda koje omogućavaju transplantaciju emocije ili govornog stila prisutne u bazi jednog govornika u sintetizovani govor nekog drugog govornika. U okviru disertacije predložen je jedan postupak za transplantaciju zasnovan na arhitekturi sa deljenim skrivenim slojevima i kodovima stila. Predloženi pristup poređen je sa referentnim pristupom

iz literature. Pokazano je da predloženi pristup u slučaju transplantacije, kada u sistemu postoje dva govornika, nadmašuje referentni pristup i po objektivnim i po subjektivnim merama. Sprovedeni testovi prepoznavanja emocija pokazuju da je tačnost klasifikacije emocija u sintetizovanom govoru približno ista rezultatima koji se dobijaju sa rečenicama koje odgovaraju prirodnom govoru.

8.1. Dalji pravci istraživanja

Istraživanja prikazana u disertaciji bazirana su na ukupnoj oceni kvaliteta sintetizovanog govora, odnosno na objektivnim merama koje se odnose na akustičke parametre, što je u skladu sa trenutnim tendencijama u istraživačkoj zajednici. Međutim, u ekspresiji stilova svakako su značajna i obeležja trajanja [65], te bi se pažnja u proceni kvaliteta sintetizovanog govora mogla usmeriti na otkrivanje uticaja kvaliteta predikcije trajanja na ukupan kvalitet.

U odeljku 7.2.4 dati su rezultati klasifikacije stilova u sintetizovanim rečenicama sa neutralnim semantičkim sadržajem. Istraživanja koja su rađena sa originalnim govorom pokazuju da uključivanje odgovarajućeg semantičkog sadržaja povećava tačnost prepoznavanja emocija. Svakako bi trebalo proveriti da li, i u kolikoj meri, ovakav zaključak važi i za sintetizovani govor, pošto slični podaci nisu pronađeni u literaturi.

Analiza predloženog pristupa za transplantaciju stilova data je za slučaj dva govornika i tri stila. Pretpostavka je da bi predložena arhitektura mogla da postigne i veći kvalitet transplantiranih stilova u scenariju sa većim brojem govornika, više stilova i više materijala po svakom stilu. Glavnu prepreku ovakvoj analizi predstavlja dostupnost odgovarajućih baza.

Nedavno su u okviru DNN sinteze na bazi neuronskih mreža predloženi pristupi koji omogućavaju sintezu govora glasom jednog govornika, ali u više različitih jezika [134]. Interesantno bi bilo istražiti mogućnost međujezičke transplantacije emocija/stilova.

Sva analiza sinteze ekspresivnog govora data je za engleski jezik, zbog postojanja odgovarajuće baze. Autor bi svakako voleo da se predloženi algoritmi primene i za sintezu govora na srpskom jeziku. Pošto je proces pripreme ekspresivne govorne baze na srpskom jeziku u toku [135], uskoro bi mogla biti dostupna i ekspresivna sinteza na srpskom jeziku.

Literatura

[1] J. T. Wood, *Communication mosaics: An introduction to the field of communication*. Cengage Learning, 2013.

[2] M. R. Schroeder, "A brief history of synthetic speech," *Speech Commun.*, vol. 13, no. 1–2, pp. 231–237, 1993.

[3] S. King, "A beginners' guide to statistical parametric speech synthesis," Edinburgh, 2010.

[4] M. Friestad and E. Thorson, "The Role of Emotion in Memory for Television Commercials.," 1985.

[5] J. A. Jacko, *Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications*. CRC press, 2012.

[6] M. Abe, "Speaking styles: statistical analysis and synthesis by a text-to-speech system," in *Progress in speech synthesis*, Springer, 1997, pp. 495–510.

[7] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press, 1996.

[8] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, "Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-Robot Interaction," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, 2016, pp. 493–500.

[9] M. Rusko, S. Darjaa, M. Trnka, R. Sabo, and M. Ritomský, "Expressive Speech Synthesis for Critical Situations," *Comput. Informatics*, vol. 33, no. 6, pp. 1312–1332, 2015.

[10] T. Dutoit, *An introduction to text-to-speech synthesis*, vol. 3. Springer Science & Business Media, 1997.

[11] I. Jauk, "Unsupervised Learning for Expressive Speech Synthesis," Universitat Politècnica de Catalunya BarcelonaTech (UPC).

[12] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," *Psychol. Monogr. Gen. Appl.*, vol. 68, no. 8, p. 1, 1954.

[13] D. O'Shaughnessy, L. Barbeau, D. Bernardi, and D. Archambault, "Diphone speech

synthesis," *Speech Commun.*, vol. 7, no. 1, pp. 55–65, 1988.

[14] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–467, 1990.

[15] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[16] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, vol. 1, pp. 373–376.

[17] N. Campbell and A. W. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in speech synthesis*, Springer, 1997, pp. 279–292.

[18] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[19] A. W. Black and P. A. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," 1997.

[20] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[21] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-gen TTS system," in *Joint meeting of ASA, EAA, and DAGA*, 1999, vol. 1, pp. 18–24.

[22] M. Beutnagel, A. Conkie, and A. K. Syrdal, "Diphone synthesis using unit selection," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[23] H. Segi, T. Takagi, and T. Ito, "A concatenative speech synthesis method using context dependent phoneme sequences with variable length as search units," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[24] L. Latacz, Y. O. Kong, and W. Verhelst, "Unit selection synthesis using long non-uniform units and phonemic identity matching," *target*, vol. 1, no. 2, pp. 1–2, 2007.

[25] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 2001, vol. 2, pp. 837–840.

[26] J. Přibil, A. Přibilová, and J. Matoušek, "Automatic Classification of Types of Artefacts Arising During the Unit Selection Speech Synthesis," in *International Conference on Text, Speech, and Dialogue*, 2017, pp. 38–46.

[27] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective information processing*, Springer, 2009, pp. 111–126.

[28] H. Dudley, "The carrier nature of speech," *Bell Labs Tech. J.*, vol. 19, no. 4, pp. 495–515, 1940.

[29] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[30] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[31] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 21, no. 3, pp. 587–597, 2013.

[32] V. Tsiaras, R. Maia, V. Diakoloukas, Y. Stylianou, and V. Digalakis, "Linear dynamical models in speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 300–304.

[33] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7962–7966.

[34] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, vol. 2. Walter de Gruyter, 2012.

[35] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971–995, 1980.

[36] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: Improving production quality," *IEEE Trans. Acoust.*, vol. 37, no. 12, pp. 1870–1887, 1989.

[37] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *J. Acoust. Soc. Am.*, vol. 70, no. 2, pp. 321–328, 1981.

[38] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Commun.*, vol. 1, no. 3–4, pp. 199–229, 1982.

[39] P. Birkholz, "VocalTractLab 2.1 User Manual." Technische Universität Dresden, 2013.

[40] A. Toutios and S. Narayanan, "Articulatory synthesis of French connected speech from EMA data," in *INTERSPEECH*, 2013, pp. 2738–2742.

[41] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, and S. S. Narayanan, "Articulatory Synthesis Based on Real-Time Magnetic Resonance Imaging Data," in

INTERSPEECH, 2016, pp. 1492–1496.

[42] H. Cryer and S. Home, "Review of methods for evaluating synthetic speech," *RNIB Cent. Access. Inf. (CAI), Tech. Rep.*, vol. 8, 2010.

[43] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.

[44] N. Campbell, "Evaluation of speech synthesis," in *Evaluation of text and speech systems*, Springer, 2007, pp. 29–64.

[45] I. Recommendation, "Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra)," *Int. Telecommun. Union, Geneva*, 2001.

[46] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[47] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, 2009.

[48] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[49] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016.

[50] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in *Third International Conference on Spoken Language Processing*, 1994.

[51] M. Morise and Y. Watanabe, "Sound quality comparison among high-quality vocoders by using re-synthesized speech," *Acoust. Sci. Technol.*, vol. 39, no. 3, pp. 263–265, 2018.

[52] P. Ekman and W. V Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, p. 124, 1971.

[53] P. Ekman, "Facial expression and emotion," *Am. Psychol.*, vol. 48, no. 4, p. 384, 1993.

[54] R. Kehrein, "The prosody of authentic emotions," in *Speech Prosody 2002, International Conference*, 2002.

[55] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[56] S. Brognaux, "Expressive speech synthesis: Research and system design with hidden Markov models," Université catholique de Louvain, 2015.

[57] I. Stravinsky, "Expresso : Transformation of Expressivity in Speech," pp. 4–7, 2010.

[58] H. Fujisaki, "Information, Prosody, and Modeling with Emphasis on Tonal Features of Speech," *Speech Prosody*, no. January 2004, pp. 1–10, 2004.

[59] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1145–1153, 2006.

[60] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality transformation," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[61] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[62] J. E. Cahn, "The generation of a ect in synthesized speech," *J. Am. Voice I/O Soc.*, vol. 8, pp. 1–19, 1990.

[63] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Commun.*, vol. 16, no. 4, pp. 369–390, 1995.

[64] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Tutorial and Research Workshop (ITRW) on speech and emotion*, 2000.

[65] J. Vroomen, R. Collier, and S. Mozziconacci, "Duration and intonation in emotional speech," in *Third European Conference on Speech Communication and Technology*, 1993.

[66] D. Govind and S. R. M. Prasanna, "Expressive speech synthesis: a review," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 237–260, 2013.

[67] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Commun.*, vol. 40, no. 1–2, pp. 161–187, 2003.

[68] G. O. Hofer, K. Richmond, and R. A. J. Clark, "Informed blending of databases for emotional speech synthesis," 2005.

[69] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for american english," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1099–1108, 2006.

[70] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[71] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Commun.*, vol. 52, no. 5, pp. 394–404, 2010.

[72] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, p. IV--1233.

[73] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," *IEICE Trans. Inf. Syst.*, vol. 92, no. 3, pp. 489–497, 2009.

[74] M. Schroeder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," University of the Saarland, 2004.

[75] D. Ververidis and C. Kotropoulos, "A review of emotional speech databases," in *Proc. Panhellenic Conference on Informatics (PCI)*, 2003, pp. 560–574.

[76] N. Campbell, "Databases of expressive speech," *J. Chinese Lang. Comput.*, vol. 14, no. 4, pp. 295–304, 2004.

[77] N. Campbell, "The recording of emotional speech: JST/CREST database research," in *Proc LREC*, 2002.

[78] R. Barra Chicote *et al.*, "Spanish expressive voices: Corpus for emotion research in spanish," 2008.

[79] Y. Zhao *et al.*, "Constructing stylistic synthesis databases from audio books," in *Ninth International Conference on Spoken Language Processing*, 2006.

[80] B. Hans-Hermann, "Origins and extensions of the k-means algorithm in cluster analysis," *J. Electron. d'Histoire des Probab. la Stat. Electron. J. Hist. Probab. Stat.*, vol. 4, no. 2, 2008.

[81] E. Székely, T. G. Csapó, B. Tóth, P. Mihajlik, and J. Carson-Berndsen, "Synthesizing expressive speech from amateur audiobook recordings," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012, pp. 297–302.

[82] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[83] J. Yamagishi, "An introduction to hmm-based speech synthesis," *Tech. Rep.*, 2006.

[84] D. Jurafsky and J. H. Martin, *Speech and language processing*, vol. 3. Pearson London, 2014.

[85] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[86] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. 85, no. 3, pp. 455–464, 2002.

[87] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 3, pp. 1315–1318.

[88] J. Gonzalvo Fructuoso, "Síntesi basada en models ocults de Markov aplicada a l'espanyol ia l'anglès, les seves aplicacions i una proposta híbrida," Universitat Ramon Llull, 2010.

[89] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[90] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Acoust. Sci. Technol.*, vol. 21, no. 2, pp. 79–86, 2001.

[91] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.

[92] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychol. Rev.*, vol. 65, no. 6, p. 386, 1958.

[93] B. Widrow and M. A. Lehr, "30 years of adaptive neural networks: perceptron, madaline, and backpropagation," *Proc. IEEE*, vol. 78, no. 9, pp. 1415–1442, 1990.

[94] M. Minsky and S. A. Papert, *Perceptrons: An introduction to computational geometry*. MIT press, 2017.

[95] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," 1985.

[96] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[97] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[98] T. Weijters and J. Thole, "Speech synthesis with artificial neural networks," in *Neural*

Networks, 1993., IEEE International Conference on, 1993, pp. 1764–1769.

[99] G. C. Cawley and P. D. Noakes, "LSP speech synthesis using backpropagation networks," in *Artificial Neural Networks, 1993., Third International Conference on*, 1993, pp. 291–294.

[100] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT—From LPC to LSP—," *Speech Commun.*, vol. 5, no. 2, pp. 199–215, 1986.

[101] C. Tuerk and T. Robinson, "Speech synthesis using artificial neural networks trained on cepstral coefficients," in *Third European Conference on Speech Communication and Technology*, 1993.

[102] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 21, no. 10, pp. 2129–2139, 2013.

[103] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8012–8016.

[104] M. Li, Z. Wu, and L. Xie, "On the impact of phoneme alignment in DNN-based speech synthesis," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 196–201.

[105] S. Suzić, T. Delić, D. Pekar, and V. Ostojić, "Novel alignment method for DNN TTS training using HMM synthesis models," in *Intelligent Systems and Informatics (SISY), 2017 IEEE 15th International Symposium on*, 2017, pp. 271–276.

[106] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4455–4459.

[107] K. L. Priddy and P. E. Keller, *Artificial neural networks: an introduction*, vol. 68. SPIE press, 2005.

[108] E. Pakoci and R. Mak, "HMM-based speech synthesis for the Serbian language," in *Proceedings of the 56th ETRAN Conference*, 2012, pp. 1–4.

[109] T. Delić and M. Sečujski, "Speech synthesis in Serbian based on artificial neural networks," in *Telecommunications Forum (TELFOR), 2016 24th*, 2016, pp. 1–4.

[110] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Japan (Part I Commun.*, vol. 66, no. 2, pp. 10–18, 1983.

[111] H. Zen *et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," in *SSW*,

2007, pp. 294–299.

[112] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[113] T. Delić, M. Sečujski, and S. Suzić, "A review of Serbian parametric speech synthesis based on deep neural networks," *Telfor J.*, vol. 9, no. 1, pp. 32–37, 2017.

[114] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Comput. Speech Lang.*, vol. 19, no. 1, pp. 55–83, 2005.

[115] N. Hojo, Y. Ijima, and H. Mizuno, "An Investigation of DNN-Based Speech Synthesis Using Speaker Codes.," in *INTERSPEECH*, 2016, pp. 2278–2282.

[116] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-Based Speech Synthesis Using Speaker Codes," *IEICE Trans. Inf. Syst.*, vol. 101, no. 2, pp. 462–472, 2018.

[117] T. Delić, S. Suzić, S. Ostrogonac, S. Đurić, and D. Pekar, "Multi- style Statistical Parametric TTS," in *Zbornik radova konferencije Digitalna obrada govora i slike (DOGS 2017)*, 2018, pp. 5–8.

[118] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*, 2017, pp. 1613–1616.

[119] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Commun.*, vol. 99, pp. 135–143, 2018.

[120] S. Suzić, T. Delić, V. Jovanović, M. Sečujski, D. Pekar, and V. Delić, "A comparison of multi-style DNN-based TTS approaches using small datasets," in *MATEC Web of Conferences*, 2018, vol. 161, p. 3005.

[121] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 3829–3833.

[122] T. Delić, S. Suzić, M. Sečujski, and D. Pekar, "Rapid Development of New TTS Voices by Neural Network Adaptation."

[123] H. Kanagawa, T. Nose, and T. Kobayashi, "Speaker-independent style conversion for HMM-based expressive speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7864–7868.

[124] J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi, and J. M. Montero, "Emotion transplantation through adaptation in HMM-based speech

synthesis," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 292–307, 2015.

[125] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[127] J. Parker, Y. Stylianou, and R. Cipolla, "Adaptation of an Expressive Single Speaker Deep Neural Network Speech Synthesis System," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5309–5313.

[128] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1450–1463, 2016.

[129] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4460–4464.

[130] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models-Analysis and normalisation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7522–7526.

[131] J. R. Davitz and L. J. Davitz, "The communication of feelings by content-free speech," *J. Commun.*, vol. 9, no. 1, pp. 6–13, 1959.

[132] W. F. Johnson, R. N. Emde, K. R. Scherer, and M. D. Klinnert, "Recognition of emotion from vocal cues," *Arch. Gen. Psychiatry*, vol. 43, no. 3, pp. 280–283, 1986.

[133] M. D. Pell, A. Jaywant, L. Monetta, and S. A. Kotz, "Emotional speech processing: Disentangling the effects of prosody and semantic cues," *Cogn. Emot.*, vol. 25, no. 5, pp. 834–853, 2011.

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

[135] M. Secujski, S. Ostrogonac, S. Suzic, and D. Pekar, "Speech database production and tagset design aimed at expressive text-to-speech in Serbian," in *Proceedings of the 10th DOGS Conference*, 2014, pp. 51–54.

[136] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete

data via the EM algorithm," *J. R. Stat. Soc. Ser. B*, pp. 1–38, 1977.