UNIVERZITET U BEOGRADU
ELEKTROTEHNIČKI FAKULTET

Milana M. Milošević

IDENTIFIKACIJA GOVORNIKA U USLOVIMA EMOTIVNOG GOVORA

doktorska disertacija

Beograd, 2020.

UNIVERSITY OF BELGRADE
SCHOOL OF ELECTRICAL ENGINEERING

Milana M. Milošević

SPEAKER IDENTIFICATION IN CONDITIONS OF EMOTIONAL SPEECH

Doctoral Dissertation

Belgrade, 2020.

Mentor

dr Željko Đurović, redovni profesor
Univerzitet u Beogradu, Elektrotehnički fakultet

Članovi komisije

dr Zoran Perić, redovni profesor
Univerzitet u Nišu, Elektronski fakultet

dr Goran Kvaščev, vanredni profesor
Univerzitet u Beogradu, Elektrotehnički fakultet

_____

_____

_____

_____

Datum usmene odbrane

_____

Zahvalnica i posveta

Hvala prof. dr Željku Đuroviću, mom mentoru, na vođenju ovog istraživanja i mog razvoja kao inženjera, stručnjaka, predavača. Hvala kolegi sa doktorskih studija, Željku Nedeljkoviću, na gotovo deceniji dugoj saradnji i prijateljstvu. Hvala porodici na podršci i ljubavi. Hvala prijateljima na moralnoj podršci, razumevanju i ohrabrenju. Hvala Elektrotehničkom fakultetu i EMPA - Švajcarskim Federalnim laboratorijama za materijale, nauku i tehnologije i Ulrike Glavitsch na saradnji, prijateljstvu i podršci.

Podaci o doktorskoj disertaciji

Naslov disertacije: Identifikacija govornika u uslovima emotivnog govora

Rezime: Fenomen emotivnog govora retko je modelovan u dosadašnjem istraživanju prepoznavanja govornika. Ključni izazov u sistemima za prepoznavanje govornika je razlika u emotivnom stanju govornika prilikom obuke sistema i trenutka kada je potrebno izvršiti prepoznavanje.

U našem istraživanju, izvršena je analiza algoritama za modeliranje i identifikaciju govornika koja je obuhvatila dosada poznate metode. Eksperimentalno su evaluirane metoda standardnog modela Gausovih mešavina, i tehnika i-vektora. Varijacijama sadržaja govora za obuku modela govornika u našim eksperimentima koje se odnosi na korišćenje i emotivnog govora za obuku modela govornika - različit broj rečenica izgovorenih u određenom emotivnom stanju i različit ukupan broj rečenica pokazali smo da sistem postaje robusniji i postiže bolji procenat prepoznavanja govornika i na osnovu neutralnog i na osnovu emotivnog govora. Nakon toga testirane su varijacije konfiguracije modela govornika u smislu modeliranja govornika sa više od jednog modela, na osnovu grupisanja emotivnog govora i kasnije prepoznavnja na osnovu ovako distribuiranog modela.

Na kraju, izvedena je varijacija strukture modela govornika. U našem istraživanju, primenom subtractive klasterizacije, određen je broj komponenti Gausove mešavina govornika na osnovu uzoraka njegovog glasa i to na automatski način. Teorisjki je izvedena zavisnosti parametra subtractive klasterizacije od broja vektora obučavajućeg uzorka, njihove dimenzionalnosti, kao i raspodele. Dobijeni rezultati pokazali su da se sa manjim brojem komponenti govornik može modelovati podjednako uspešno, što daje prostor za dalja istraživanja mogućnosti primene subtractive klasterizacije za generisanje modela govornika.

Ključne reči: klasifikacija, klasterizacija, subtractive klasterizacija, obrada govora, identifikacija govornika, Gausove mešavine, i-vektori, karakteristike govora, emocije u govoru

Naučna oblast: Elektrotehnika i računarstvo

Uža naučna oblast: Obrada signala

UDK: 621.3

About

Title: Speaker identification in conditions of emotional speech

Abstract: The phenomenon of emotional speech is rarely modeled in up to date speaker recognition research. The key challenge for a speaker recognition system is the difference in emotional state of a speaker in the phase of system training and in the phase of the system usage.

In our research, different speaker modeling and classification algorithms were analyzed. In our experiments, Gaussian mixture models, as fundamental technique in speaker recognition, and i-vectors, as state-of-the-art technique for commercial use were evaluated. It was showed that by variating emotional content of speech and the number of sentences for speaker model training improves system robustness in recognition of speakers from both neutral and emotion impacted speech. Also, variation in model configuration by creating three models for each speaker was examined. Each of these models was trained by appropriate emotional speech grouped by emotion valence.

The last part of research is dedicated to variation of speaker model structure by determining the number of components in Gaussian mixture based on subtractive clustering. The number of components was determined automatically from the training utterances for each speaker by using subtractive clustering. Theoretical dependencies of subtractive clustering parameters and the number of feature vectors in training sample, feature vector dimensionality and their distribution were induced. The results obtained showed that speaker can be modeled with fewer components in Gaussian mixture successfully, which opens space for further research of subtractive clustering application for speaker model training.

Key words: classification, clasterisation, subtractive clustering, Gaussian mixture models, speech processing, speaker recognition, i-vectors, speech features, emotions in speech

Scientific field: Electrical engineering and computer science

Scientific subfield: Signal processing

UDC: 621.3

Sadržaj

I Uvod u prepoznavanje govornika, model i karakterizacija govora 1

1 Uvod 3
1.1 Motivacija . . . 3
1.2 Pregled literature . . . 4
1.3 Predmet i cilj istraživanja . . . 5
1.4 Naučni doprinos . . . 6
1.5 Struktura teze . . . 6

2 Prepoznavanje govornika 9
2.1 Glas kao deo identiteta osobe . . . 9
2.2 Sistem za prepoznavanje govornika . . . 10
2.3 Specifični zadaci prepoznavanja govornika i primene . . . 11
2.4 Pristupi u prepoznavanju govornika . . . 12

3 Model generisanja signala govora 13
3.1 Psiholingvistički model govornog signala . . . 14
3.2 Fiziološki model govornog signala . . . 15
3.3 Izazovi modeliranja govora . . . 20

4 Karakterizacija signala govora 23
4.1 Grupe karakteristika . . . 23
4.2 Izračunavanje vektora karakteristika . . . 24
4.3 Redukcija dimenzija vektora karakteristika . . . 28
4.4 Uticaj emocija na karakteristike signala govora . . . 30

II Istraživanje, rezultati i naučni doprinos 33

5 Klasifikacije i modeliranje govornika u uslovima emotivnog govora 35
5.1 Opis eksperimenata i analize . . . 36
5.2 Algoritmi klasifikacije i modeliranja govornika . . . 39

5.3 Rezultati eksperimenata i diskusija . . . 42
5.4 Ostale tehnike prepoznavanja govornika . . . 55
5.5 Rezime i zaključci . . . 63

6 Određivanje broja Gausovih mešavina subtractive klasterizacijom 65
6.1 Subtractive klasterizacija . . . 65
6.2 Analiza Gausovih slučajnih promenljivih . . . 66
6.3 Parametri klasterizacije . . . 84
6.4 Opis eksperimenta . . . 85
6.5 Rezultati eksperimenta . . . 86
6.6 Rezime i zaključci . . . 89

III Zaključak, buduće istraživanje i korišćena literatura 91

7 Zaključak 93

Literatura 97

IV Skraćenice, liste slika, tabela i dodaci 109

A Lista skraćenica 111

B Lista slika 113

C Lista tabela 115

D Baze govora 117
D.1 Baze za prepoznavanje govora i govronika . . . 117
D.2 Baze emotivnog govora . . . 118

E Hibridna klasifikacija 123
E.1 Segmentalne karakteristike . . . 123
E.2 Rezultati eksperimenata . . . 124
E.3 Rezime i zaključci . . . 125

Deo I

Uvod u prepoznavanje govornika, model i karakterizacija govora

1. Uvod

Istraživanje ljudskog glasa je multidisciplinarno polje - presek istraživanja inženejrstva sa medicinom, psihologijom, lingvistikom i umetnošću (Slika 1.1). Obrada govora, prepoznavanje govora, prepoznavanje jezika, prepoznavanje govornika itd, imaju široku primenu u ostalim naučnim disciplinama.

Slika 1.1: Multidisciplinarnost istraživanja o glasu.

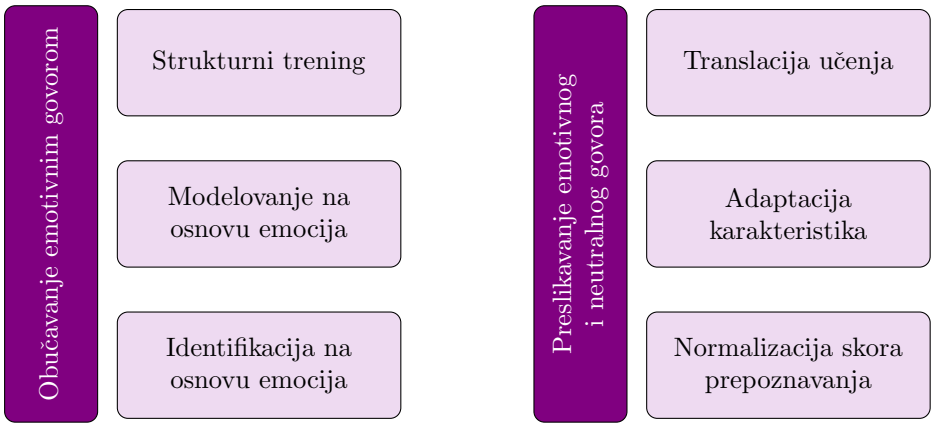
1.1 Motivacija

Prepoznavanje govornika je jedna od klasičnih grana automatske obrade govora. Osnovni cilj izučavanja glasa u ovoj oblasti je odgovor na pitanje "Ko je to izgovorio?". Poslednjih godina, ova oblast ponovo je dobila na značaju zahvaljujući razvoju pametnih, personalizovanih sistema [1, 2]. Identifikacija govornika jedan je od načina autentifikacije korisnika za korišćhenje servisa u svakodnevnom životu kao što su telefonsko bankarstvo, pretraga interneta i preuzimanje zaštićenih informacija [3].
Glas, kao biometrijski podatak na osnovu koga se osoba može identifikovati, smatra se kombinacijom bihevioralnih i fizioloških biometrija [4]. Osim primene u svakodnevnom životu, identifikacija govornika je od interesa u forenzičkim istraživanjima, telefonskim servisnim centrima, centrima za hitne slučajeve, u transferu poverljivih podataka itd. U tim slučajevima, govornici

uglavnom ne koriste uobičajeni ton glasa. U muzici, jedna od tema automatske analize snimaka je i prepoznavanje pevača [5, 6].
Kada je govornik uznemiren, pod stresom, plače, smeje se, šapuće, viče ili peva, karakteristike njegovog glasa su izmenjene [7]. U sistemima za identifikaciju govornika, to može dovesti do greške koja dalje izaziva frustraciju korisnika, a u hitnim slučajevima, naročito kada su vojne i bezbednosne primene u pitanju, može dovesti do ozbiljnih posledica.
Ovi primeri ilustruju važnost zadatka prepoznavanja govornika i razvoj pouzdanih sistema koji na njega mogu odgovoriti. Cilj u dizajnu ovih sistema je robusnost, a da se pri tome koristi što manje podataka za obuku. U dosadašnjem istraživanju, analizirani su i razvijeni mnogobrojni algoritmi klasifikacije i klasterizacije primenjeni u ovoj oblasti.

1.2 Pregled literature

Prvi radovi u oblasti prepoznavanja govornika dolaze iz Belovih laboratorija, šezdesetih godina, [8, 9], kada je razvijen prvi sistem za prepoznavanje govornika na bazi banaka filtara za poređenje spektra [10–12]. Preteča današnjih sistema za prepoznavanje govornika ipak je primena Skrivenih Markovljevih modela [13], koje su Rose i Reynolds ubrzo uprostili [14, 15] na Gausove mešavine. Ovaj pristup čini osnovu današnjih modernih sistema za prepoznavanje govornika u kombinaciji sa mašinama potpornih vektora, dubokim neuralnim mrežama [16], i-vektorima [17–19] i alternativnim tehnikama akustičkog modelovanja [20].
Fenomen emotivnog govora retko je modelovan u dosadašnjem istraživanju prepoznavanja govornika. Gledano iz perspektive emotivnog stanja govornika, ključni izazov u sistemima za prepoznavanje govornika je razlika u emotivnom stanju govornika prilikom obuke sistema i trenutka kada je potrebno izvršiti prepoznavanje. Karakteristike glasa menjaju se pod uticajem emocija, što sistem lako može dovesti u zabludu. Jednan od razloga za malu pažnju od strane istraživača je delom i zahvaljujući nedostatku snimaka emotivnog govora za različite govornike. Možemo ih grupisati u dva pristupa: obučavanje sistema emotivnim govorom i preslikavanje emotivnog i neutralnog govora (Slika 1.2).



Slika 1.2: Emotivni govor i pristupi prepoznavanju govornika

Obučavanje sistema emotivnim govorom podrazumeva da se za obuku jednog ili više modela govornika upotrebljava njegov emotivni govor. Scherer [21] i Klasmeyer [22] predložili su prikupljanje snimaka emotivnog govora u fazi obuke sistema korišćenjem softvera za prikupljanje uzoraka govora koji govorniku zadaje različite zadatke sa ciljem da izazove reakciju govornika ili emotivno stanje. Ovaj pristup primenjuje se i za istraživanje prepoznavanja emocija u govoru. U nekoliko drugih studija, modeli govora obučavani su korišćenjem odglumljenog

emotivnog govora iz neke od dostupnih baza za proučavanje emocija u govoru. Ghiurcau i ostali obučavali su model Gausovih mešavina birajući nasumično rečenice emotivnog i neutralnog govora za obuku [23]. Mansour i Lachiri trenirali su i testirali Skrivene Markoviljeve modele i dva različita tipa Mašina potpornih vektora [24, 25].

Modelovanje na osnovu emocija [26,27] sastoji se u kreiranju i obuci više modela, pri čemu su obučavajuće rečenice grupisane u tri grupe na različite načine: (1) grupisanjem po sličnosti emotivnih stanja i (2) nasumičnim grupisanjem. Za pojedinačnog govornika obučena su tri modela - po jedan za svaku od ove tri grupe. Prilikom testiranja, najpre je vršena detekcije emocije na osnovu prozodijskih obeležja [26, 27], a nakon toga je biran model u odnosu na koji se vrši finalna evaluacija. Shahin je u svojim istraživanjima [28–30] predložio **identifikaciju na osnovu emocija** i dva načina na koja se ona može sprovesti. Prvi pristup sastoji se u kreiranju modela za svakog govornika za svako posebno emotivno stanje. Izlaz svakog od modela je verovatnoća da je generisao zadatu test rečenicu. Prepoznati govornik je onaj čiji model daje najveću verovatnoću. Drugi pristup bio je da se najpre prepozna emocija iz test rečenice, a da se potom donese i odluka o govorniku uzimajući u obzir samo datu emociju. Oba eksperimenta sprovedena su tako da su zavisni od izgovorenog teksta. Govornike je u oba slučaja modelirao korišćenjem Skrivenih Markovljevih modela i kepstralnih koeficijenata.

Adaptacija modela govora pojavljuje se u nekoliko studija u vidu prilagođavanjem modela neutralnog govora na emotivni govor ,,bojenjem"[1]modela emotivnim govorom, ili ,,beljenjem"[1] u fazi testiranja. Chen i Yang [31] tretirali su emocije na isti način kao i distorzije koje se javljaju usled uticaja kanala prenosa glasa i varijabilnosti usled razlike u sesijama [32] u kreiranju modela govornika. Li i ostali [33, 34] predlažu **konverziju emotivnog govora** na način da se karakteristike neutralnog govora modifikuju na osnovu seta pravila nastalih uparivanjem neutralnog i malog uzorka emotivnog govora [33]. Tako modifikovanim govorom uspešno obučavaju standardni model Gausovih mešavina, a sličan pristup primenili su i Krothapalli i ostali [35] koji su za alat preslikavanja odabrali neuralnu mrežu. U svom daljem radu, Li i ostali istražuju polazeći od drugačije pretpostavke - nisu svi delovi signala govora podjednako modifikovani usled emocija [36] predlažući metodu izdvajanja ovih delova signala na osnovu visine glasa govornika, kao i **normalizaciju skora prepoznavanja govornika** favorizovanjem baš ovih delova signala koji su manje izmenjeni emocijama. Ideja je da se otklone pomeraji i skaliranje modela nastali zbog emocija u govoru. Ovaj pristup kasnije su obogatili superviziranim učenjem [37]. Kada je istraživanje na temu broja mešavina u GMM u pitanju, pristupi se sastoje ili u optimizaciji određenog kriterijuma ili u ispitvanju rezultata dobijenim sa brojem mešavina iz određenog opsega. Lee i ostali primenili su inkrementalni K-means algoritam za određivanje optimalnog broja GMM [38]. Kombinacijom kriterijuma Wang i ostali [39] pokazali su da se broj komponenti raspodele u teoriji može precizno odrediti, međutim eksperimente su sproveli samo na generisanim podacima.

1.3 Predmet i cilj istraživanja

Predmet ovog istraživanja je modeliranje govornika u uslovima emotivnog govora. Glas govornika se menja pod uticajem emocija, što za poslednicu ima izmene performansi sistema u smislu uspešnosti prepoznavanja govornika. Sa druge strane, ako se ovakve varijacije mogu modelirati kroz sistem, makar odglumljenim govorom, moguće je uraditi adaptaciju sistema. Takođe, ispitivano je koliko na uspešnost modela utiče svaka pojedinačna emocija, pol govornika, jezik snimaka i količina iskorišćenih podataka za obuku modela. Dalje, osim modeliranja emocija, pokušano je i adekvatno modeliranje svakog od govornika brojem komponenti Gausovih mešavina.

[1]Asocijacija na pojam bojenja i beljenja signala u stohastičkim sistemima.

Cilj istraživanja je ostvarivanje povećane preciznosti sistema za prepoznavanje govornika upotrebom emotivnog govora i modela prilagođenog svakom od govornika. Na osnovu analize uticaja emotivnog stanja na izmene karakteristika govora, konstruisan je sistem za čiju se obuku, osim neutralnog govora koristi i emotivni govor. Takođe, pokušan je i test sa različitim brojem Gausovih mešavina na osnovu subtractive klasterizacije. Cilj je dakle da se sistem za prepoznavanje govornika učini robusnijim za promene emotivnog stanja govornika.

1.4 Naučni doprinos

Rezulatati ove doktorske disertacije su:

- Procena uspešnosti upotrebe glasa kao biometrisjskog podatka i sistema za identifikaciju govornika poznatih u literaturi u uslovima emotivnog govora.
- Ustanovljen je uticaj svake od pojedinačnih emocija radosti, besa, tuge i straha u odnosu na neutralno emotivno stanje kroz procenat uspešnosti sistema prepoznavanja.
- Ustanovljen je uticaj pola, kvaliteta snimka i veličina baze na rezultate sistema za prepoznavanje govornika u uslovima emotivnog govora.
- Eksperimentalno su upoređeni različiti modeli govornika: Gausove mešavine i mel kepstralni koeficijenti, i-vektori i mel kepstralni koeficijenti, kao i Gausove mešavine sa brojem gausovih funkcija određenim za svakog od govornika pojedinačno korišćenjem subtractive klasterizacije.
- Eksperimentalno su upoređeni modeli različite konfiguracije obučeni emotivnim govorom: miks model i tri model.
- Modeli miks model i tri model upoređeni su i u zavisnosti od broja trening rečenica u istim uslovima testiranja.
- Razvijen sistem evaluiran je na nekoliko baza emotivnog govora, različite veličine i jezika.
- Izvedeno je očekivanje minimalnog i maksimalnog rastojanja vektora karakteristika obučavajućeg skupa za primenu u subtractive klasterizaciji, na osnovu čega je automatski određen broj komponenti GMM za svakog od govornika.
- Eksperimentalno su evaluirani rezultati sa modelima GMM čiji je broj komponenti određen subtractive klasterizacijom.

1.5 Struktura teze

Teza se sastoji od uvodnog poglavlja, tri poglavlja preglednog tipa, zatim dva istraživačka poglavlja, praćena poglavljem zaključka, pregledom literature i dva dodatka.
Drugo poglavlje posvećeno je zadacima prepoznavanja govornika, izazovima u ovoj oblasti sa posebnim osvrtom na varijabilnosti u glasu.
U trećem poglavlju opisani su fiziološki i psiho-lingvistički model generisanja signala govora. Opisani su mehanizmi generisanja reči od misaonih formi, kao i govorni organi, njihova fizionomija i aksutika.
Četvrto poglavlje daje pregled karakteristika govora, njihovog značaja za prepoznavanje govornika, načina za izračunavanje datih karakterisstika, kao i uticaja emocija na njih.

U petom poglavlju data je analiza algoritama primenjivanih u identifikaciji govornika sa posebnim osvrtom na tehnike korišćene za adaptaciju na emotivna stanja govornika. U ovom poglavlju opisani su eksperimenti kao i korišćeni podaci. Takođe prikazani su i diskutovani rezultati eksperimenata sprovedeni sa pojedinim tehnikama iz aspekta emocija, pola, količine obučavajućih podataka.
Šesto poglavlje opisuje primenu subtractive klasterizacije za određivanje broja komponenti Gausovih mešavina i početnih klastera u modelu govornika, kao i rezultate dobijene na osnovu primene ove tehnike. Rezultati su analizirani i diskutovani iz različitih aspekata.
Rezime teze, kao i sublimirani zaključci dati su u sedmom poglavlju. Na kraju, dat je pregled korišćene literature.
Prvi dodatak sadrži detalje o bazama emotivnog govora i bazama za prepoznavanje govornika.
Drugi dodatak sadrži nezavisno istraživanje o prepoznavanju govornika na osnovu suprasegmentalnih karakteristika.

2. Prepoznavanje govornika

2.1 Glas kao deo identiteta osobe

Osobu karakterišu različiti biometrijski podaci kao što su slika lica, potpis, glas, linije šake, način hoda. Glas, kao biometrijski podatak na osnovu koga se osoba može identifikovati, smatra se kombinacijom fizioloških biometrija i biometrija ponašanja [4]. Biometrijski podaci mogu se porediti po više različitih kriterijuma:

- Koliko dobro biometrijski podatak identifikuje osobu?
- Da li je jednostavno prikupiti biometrijski podatak?
- Da li su potrebni specijalizovani uređaji?

Poređenje glasa sa ostalim biometrijskim podacima po različitim kriterijumima, prikazano je u Tabeli 2.1 [40].

Tabela 2.1: Poređenje parametara različitih vrsta biometrijskih podataka [40].

Tip biometrijskog podatka	Preciznost	Jednostavnost korišćenja	Prihvatljivost za korisnike	Jednostavnost implementacije	Cena
Otisak prsta	Visoka	Srednja	Niska	Viskoka	Srednja
Geometrija šake	Srednja	Viskoa	Srednja	Srednja	Visoka
Glas	**Srednja**	**Visoka**	**Visoka**	**Visoka**	**Niska**
Retina	Visoka	Niska	Niska	Niska	Srednja
Iris	Srednja	Srednja	Srednja	Srednja	Visoka
Potpis	Srednja	Srednja	Visoka	Niska	Srednja
Lice	Niska	Visoka	Visoka	Srednja	Niska

Iako glas nije najpreciznija biometrija, to se delimično kompenzuje jednostavnošću implementacije, činjenicom da ne zahteva nikakvu specijalnu opremu i lakom dostupnošću. Ponekad je glas i jedina dostupna biometrija. Prepoznavanje osoba na osnovu glasa je od interesa u forenzičkim istraživanjima, telefonsikm centralama za hitne slučajeve, kao i u prenosu podataka visokog stepena poverljivosti [2], sistemima za kontrolu pristupa itd [4]. U takvim situacijama, govornik obično neće upotrebiti neutralan ton glasa. Karakteristike glasa uznemirene osobe, osobe pod stresom, one koja plače, šapuće, smeje se ili viče razlikuju se u odnosu na karakteristike njenog glasa kada je u neutralnom stanju [7]. Razlika u karakteristikama glasa može

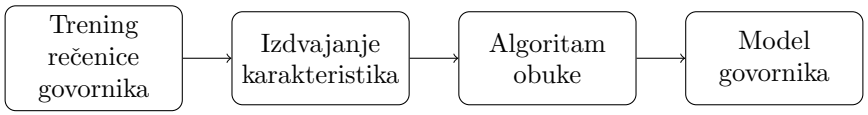
dovesti do pogrešne procena sistema za prepoznavanje govornika, što za posledicu može imati da se osobi ne dozvoli pristup određenom servisu ili prostoriji ili pogrešno identifikovanje osobe u veštačenjima dokaza. Jedna od posledica može biti frustracija korisnika, ali može izazvati i ozbiljnije posledice u vojnim ili policijskim misijama ili pravne posledice. U tom smislu, cilj prepoznavanja govornika je kreiranje što robusnijeg sistema korišćenjem što manje podataka.

2.2 Sistem za prepoznavanje govornika

Životni vek sistema za prepoznavanje govornika sastoji se iz više faza:

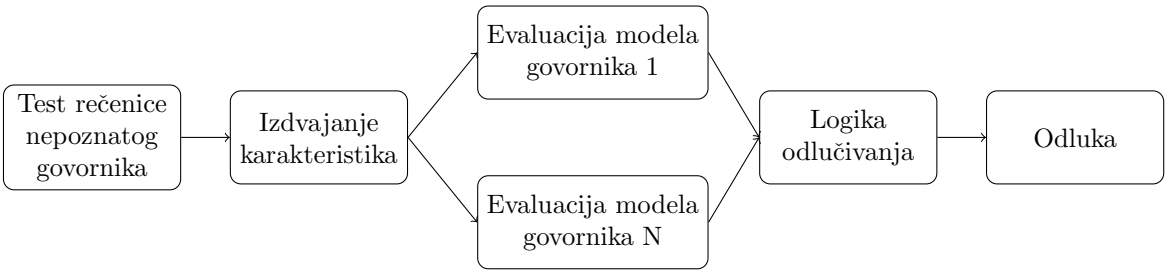
- obuka sistema,
- testiranje sistema,
- instalacija sistema u realnom okruženju i
- upotreba sistema.

U zavisnosti od prirode sistema, može se vršiti i dodatno obučavanje tokom faze upotrebe sistema. U našim razmatranjima fokusirali smo se na faze obuke i testiranja sistema.
Obuka sistema podrazumeva kreiranje modela glasova željenih govornika. Opšta šema obuke sistema za prepoznavanje govornika prikazana je na Slici 2.1. Najpre se vrši upis glasova novih



Slika 2.1: Opšta šema kreiranja modela govornika u sistemu za prepoznavanje govornika

govornika u bazu glasova sistema. Iz ovih snimaka izračunavaju se karakteristike glasa koje su od interesa kada je u pitanju identitet govornika. Na osnovu ovih karakteristika, korišćenjem odgovorajućeg algoritma obuke, formira se model govrnika.
Druga faza, faza testiranja sistema, podrazumeva određivanje kome pripada određeni uzorak glasa. Kao i u slučaju obuke sistema, vrši se izdvajanje karakteristika glasa. Zatim, u zavisnosti od specifičnog zadatka prepoznavanja govornika, na osnovu ovih karakteristika vrši se evaluacija test rečenice u odnosu na postojeće modele govornika. Algoritam zaključivanja zatim donosi odluku koji govornik iz sistema je izgovorio zadatu rečencu (Slika 2.2).



Slika 2.2: Opšta šema testiranja sistema za prepoznavanje govornika

2.3 Specifični zadaci prepoznavanja govornika i primene

Prepoznavanje govornika ima široku oblast primene pod različitim uslovima. U okviru prepoznavanja govornika možemo prepoznati nekoliko različitih zadataka [41]:

- verifikacija govornika,
- identifikacija govornika,
- segmentacija i klasifikacija događaja i govornika,

Svi specifični zadaci imaju za cilj projektovanje sistema za prepoznavanje koji će raditi što preciznije tj koji će imati mali broj grešaka. Zahvaljujući sve većoj komercijalnoj primeni sistema za prepoznavanje govornika, pojavila se i njihova zloupotreba - pokušaji da se sistem prevari snimkom glasa određenog govornika ili imitacijom [42–44]. Tako se razvio još jedan zadatak u okviru prepoznavanja govornika:

- otkrivanje podmetanja u sistemu.

Zadatak **verifikacije govornika** je da odgovori na pitanje da li određeni glas odgovara predloženom identitetu govornika. Odluka da li je govornik onaj koji se tvrdi da jeste ili nije treba da bude doneta sa dovoljnom verovatnoćom, odnosno da sistem bude siguran sa dovoljnom pouzdanošću da li je osoba koja govori zaista i ona koja se tvrdi da jeste. U ovakvim sistemima, uzorak glasa osobe čiji se identitet potvrđuje poredi se na osnovu izabranog modela sa modelom glasa koji se čuva uz identitet u sistemu.
Identifikacija govornika je nešto zahtevniji zadatak od prethodnog - potrebno je odrediti kome od govornka pripada zadati glas. Cilj ovog sistema je da odredi kome pripada određeni uzorak glasa u odnosu na modele zabeležene u sistemu ili da donese odluku da glas ne pripada nikome. Ujedno, ovo su i dva pristupa u identifikaciji govornika:

- identifikacija na zatvorenom skupu - kada se očekuje da će se pred sistem u fazi korićenja predstavljati samo osobe čiji su glasovi bili dostupni i za obuku sistema, i
- identifikacija na otvorenom skupu - sistem je obučen za identifikaciju određenog skupa glasova, i očekuje se da će se u fazi upotrebe sistema pojaviti i glasovi govornika koji nisu u tom skupu.

U slučaju zatvorenog skupa govornika, za zadati glas potrebno je odrediti koji model govornika ima najveću verovatnoću da mu taj glas odgovara. U slučaju otvorenog skupa, osim određivanja modela sa najvećom verovatnoćom potrebno je da i vrednost bude dovoljno velika - inače sistem zaključuje da glas ne odgovara ni jednom od postojećih modela govornika.
U realnim uslovima, audio snimci sadrže mnogo više od samog glasa koji želimo da identifikujemo ili čiji sadržaj želimo da prepoznamo - uličnu buku, sirene, muziku, dečiji govor, bebin plač, više glasova istovremeno, hrkanje, kijavicu itd [41, 45]. Prepoznavnanje i izdvajanje ili uklanjanje ovih događaja je često od značaja za dalju obradu govora, kao na primer, delovi snimka koji odgovaraju samo jednom govorniku ili detektovanje određenih emotivnih stanja [46], [47]. Klasifikacija ovakve vrste elemenata audio snimaka zadatak je **segmentacije i klasifikacije događaja i govornika** [41].
Na kraju, sa razvojem sistema za sintezu govora i kvalitetnih i dostupnih mikrofona za snimanje glasa, paralelno se razvila još jedna grana u prepoznavanju govornika: **detekcija napada na sistem** (*anti-spoofing*). Osnovni tipovi napada na sisteme za prepoznavanje govornika su [48]:

- imitacija - napadač imitira glas originalnog govornika,

- odgovor - napadač koristi snimak glasa originalnog govornika,
- sinteza glasa - od zadatog teksta vrši se sinteza govora adaptirana karakteristikama originalnog govornika,
- konverzija glasa - glas napadača je izmenjen na način da deluje kao glas originalnog govornika

Ova oblast ima za cilj da detektuje napad u slučaju da postoji i time poveća pouzdanost sistema za prepoznavanje govornika.

2.4 Pristupi u prepoznavanju govornika

U zavisnosti od načina interakcije sistema za prepoznavanje govornika i sadržaja govora koji se uzima u obzir, suštinski se razlikuju prepoznavanje govornika nezavisno od teksta i zavisno od teksta. Prepoznavanje govornika **nezavisno od teksta** ne uzima u obzir leksički sadržaj govora tj zanemaruje se šta je rečeno, već se identitet govornika određuje na osnovu univerzalnih karakteristika koje opisuju vokalni trakt govornika.
U slučaju prepoznavanja govornika **zavisno od teksta** uzima se u obzir sadržaj govora - ili je u pitanju unapred definisan tekst, ili govornik ima asistenciju šta treba da izgovori ili pitanje na koje treba da da odgovor vezano za prethodno definisani PIN kod i slično [48]. Prepoznavanje se tada može izvršiti na osnovu analize određenih fonema ili karakteristika izračunatih na samo određenim delovima govora [20].

3. Model generisanja signala govora

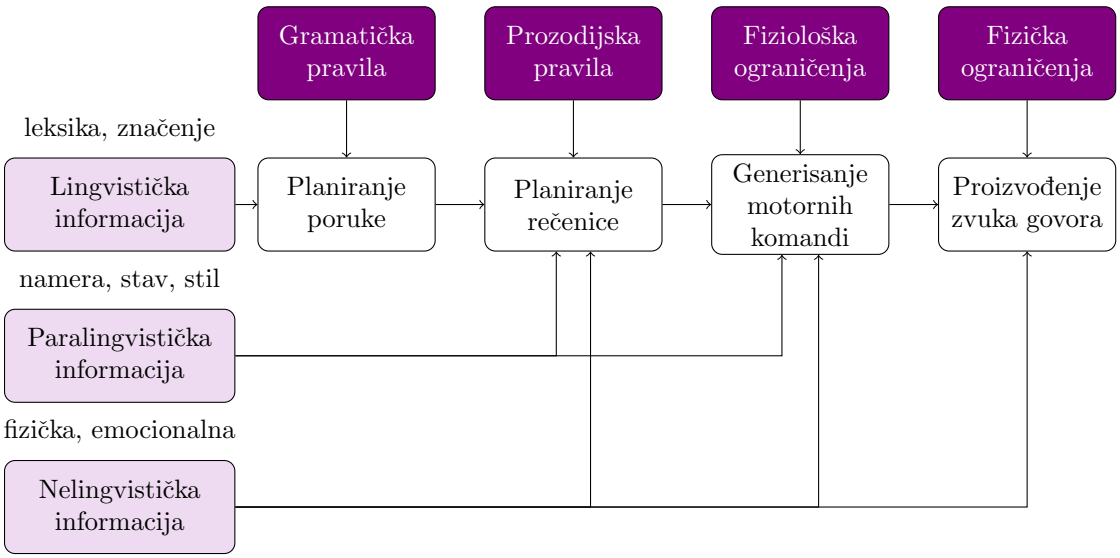
Govor je izraz misli i osećanja artikulacijom zvukova [49]. Strožije definisano generisanje govornog signala odnosi se na kognitivne procese koji transformišu misli i koncepte u lingvistički formiranu rečenicu [50, 51]. Kao takav, govor je kompleksna pojava koja se može posmatrati u različitim kontekstima i iz više aspekata. Informacije koje se izražavaju kroz govor mogu se podeliti u tri grupe [52]

- **Lingvističke informacije** - simboli i pravila za njhovo kombinovanje u smislene rečenice.
- **Paralingvističke informacije** - informacije koje se ne mogu naći u pisanom jeziku, kao što je prozodija. Govornik dodaje ove informacije kako bi dopunio ili modifikovao lingvistički sadržaj govora.
- **Nelingvističke informacije** - uključuju faktore kao što su godine, pol, emotivno stanje goornika, model ponašanja, način razmišljanja itd.

Ove informacije učestvuju u generisanju govora u različitim stadijumima. Dva osnovna aspekta modeliranja procesa generisanja govora su psiholingvistički i fiziološki model. Šema procesa kreiranja govora po fazama prikazana je na Slici 3.1 [52].

Psiholingvistički model govora podrazumeva modeliranje procesa stvaranja signala govora od namere do impulsa šta i kako izgovoriti. To podrazumeva osmišljavanje koncepta, sastavljanje rečenice i njene intonacije. Uticaji su brojni: jezik kojim se govori, kultura i region odakle potiče govornik, profesija, obrazovanje i socijalni status, ali i trenutno raspoloženje i emotivno stanje govornika.

Fiziološki model govora podrazumeva sam proces artikulacije osmišljene rečenice. Elementi ovog modela su svi organi koji učestvuju u artikulaciji glasa, njihova međusobna interakcija i akustika. Uticaj na ovaj model prevashodno imaju biološka svojstva govornih organa govornika. Ona zavise od starosti govornika, pola, ali i emocija i trenutnog zdravstvenog stanja govornika. Takođe, postoje i integrativni modeli govora koji predviđaju interakcije između psiholingvističkog stadijuma i fiziološkog stadijuma generisanja govornog signala [53].



Slika 3.1: Šema generisanja govora na osnovu tri tipa informacije [52].

3.1 Psiholingvistički model govornog signala

Psiholingvistički model signala govora opisuje mehanizme za generisanje govora od komunikativne namere do konačne artikulacije govora. Većina ovih modela razvijeni su na osnovu posmatranja grešaka u govoru, a nedoumice oko načina interakcije i dalje postoje. Dva pristupa u psiholingvističkom modelovanju su serijski (modularni) i paralelni(interaktivni). Serijskim pristupom sistem za generisanje govora modelira se fazama koje se odvijaju jedna za drugom, a u svakoj od ovih faza generiše se različit tip informacije [54–56]. Paralelni pristup podrazumeva da je sistem mnogo bolje povezan i da omogućava direktnu obradu i neposredne interakcije više izvora informacija [57]. Radi potpunosti modela govora, u nastavku su opisane faze pripreme signala govora na osnovu uprošćenog Leveltovog serijskog modela [56]. Na Slici 3.2 prikazani su koraci pripreme signala govora do same artikulacije, kao i međurezultati ovih koraka.

3.1.1 Konceptualna priprema

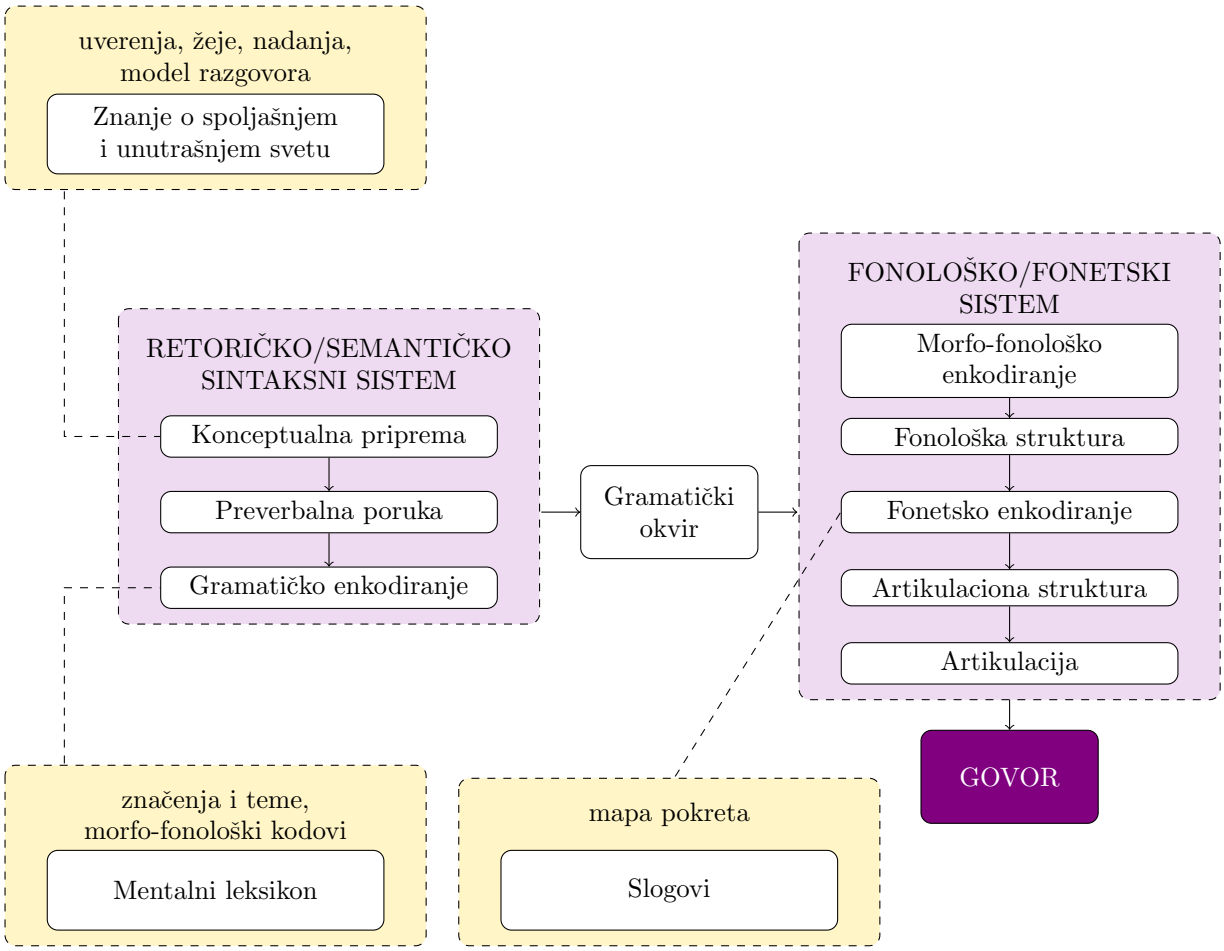
Poruka nastaje kao misao govornika ili kao reakcija u odnosu sa sagovornikom, koja je struktura koja nosi komunikativnu nameru govornika [56]. U ovom koraku, vrši se izbor i redosled dostupne konceptualne informacije koje će biti deo poruke, i to tako da su pogodne za leksičko izražavanje [51]. Konačna poruka je struktura koja se sastoji od leksičkih koncepta, koji se mogu izražavati kroz reči i paralingvističke informacije [56].

3.1.2 Gramatičko enkodiranje

U ovom koraku, na osnovu leksičkog koncepta, aktivira se izbor odgovarajućih rečeničnih struktura u mentalnom leksikonu reči, što za rezultat daje gramatički okvir rečenice [56].

3.1.3 Morfo-fonološko enkodiranje

Sledeći korak procesa je kompozicija rečenice, odnosno postupno građenje reči iz slogova, fraza i intonacije, prateći gramatička pravila [56].



Slika 3.2: Koraci u psiholingvističkoj fazi generisanja govora (serijski model) [56]

3.1.4 Fonetsko enkodiranje

Fontesko enkodiranje je proces generisanja pokreta odgovarajućih mišića na osnovu slogova definisanim u morfo-fonološkom enkodiranju i u ovom koraku se u rečenicu dodaje i prozodijski sadržaj [51]. Veze slog-pokret mišića izgrađuju se do kraja prve godine života [56].

3.1.5 Artikulacija

Artikulacija je izvršavanje fonetskog enkodiranja kroz govorni aparat, a njen rezultat je konačno govor [56].

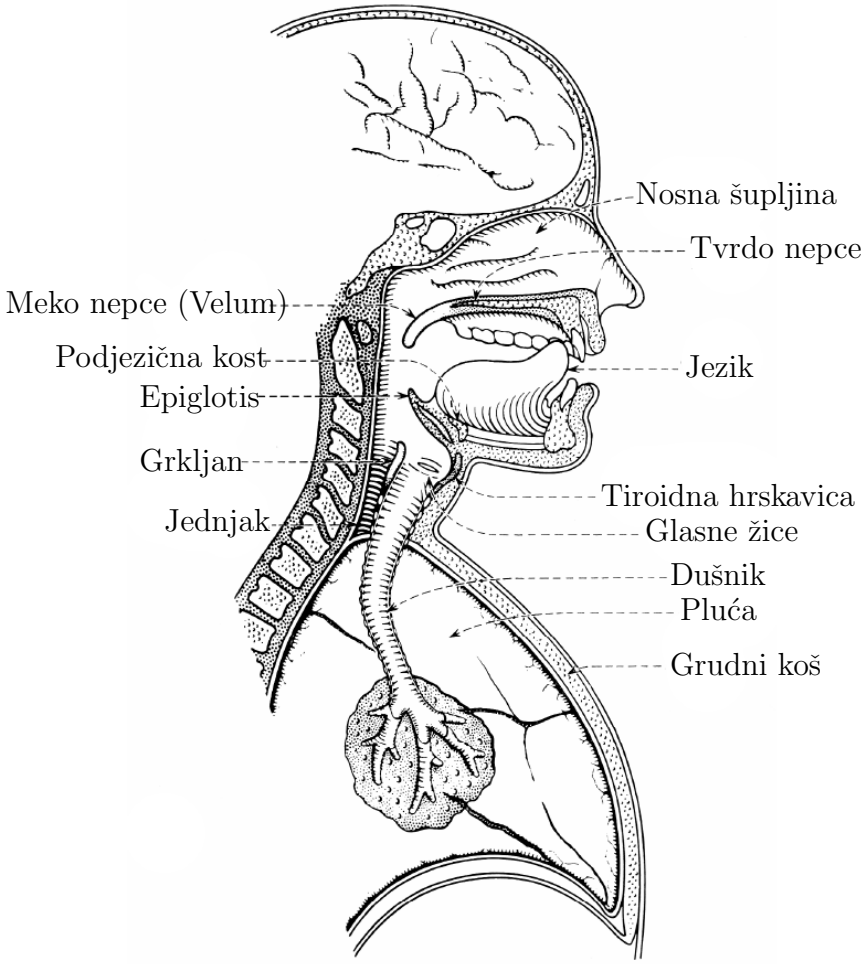
3.1.6 Automonitoring

Automonitoring je proces posmatranja sopstvenog govora, unutrašnjeg i izgovorenog, koji omogućava samo-korekcije pri govoru, i uključuje iste elemente kao i da slušamo nekog drugog [51]. Potreba za samokorekcijama javlja se kada primećujemo smetnje prilikom govora, naročito one koje imaju posledice na sadržaj i značenje onoga što smo želeli da izgovorimo [56].

3.2 Fiziološki model govornog signala

Govor nastaje izvršavanjem motornih komandi koje su generisane psiho-lingvističkim procesom opisanim u prethodnoj sekciji. Konačan oblik signala govora, pored psiho-lingvističkih faktora,

formira se na osnovu uticaja fizioloških i fizičkih svojstva govornih organa (Slika 3.3).



Slika 3.3: Šematski prikaz govornih organa [58]

.

3.2.1 Govorni organi

Ljudski govorni aparat (Slika 3.3) sastoji se iz tri celine: respiratornih organa, fonatornog sistema i vokalnog trakta [59].

Respiratorni organi

Respiratorni organi koji učestvuju u generisanju signala govora su pluća, bronhije i dušnik. Uloga **pluća** je da obezbede izvor energije govora - prilikom udisaja i izdisaja, pluća stvaraju struju vazduha koja pomoću **bronhija** i **dušnika** ulazi u fonatorni sistem i vokalni trakt.

Fonatorni sistem

Fonatorni sistem i vokalni trakt menjaju ovu struju vazduha pretvarajući je u različite zvučne talase. Osnova fonatornog sistema je **grkljan** - kompleksan sistem lakopokretljivih hrskavica, veznog i mišićnog tkiva, čiju osnovnu funkciju realizuju dva bočna mišića u unutrašnjem delu grkljana, zategnuta između prednjeg i zadnjeg dela koja se nazivaju **glasne žice tj glasnice** [59]. Glasnice se primiču i odmiču menjajući promer otvora grkljana koji se naziva **glotis** i trouglastog je oblika kada su glasnice potpuno razmaknute [59]. Promena promera glotisa

menja pritisak struje vazduha - promena promera sa određenom učestanošću stvara zvučne talase određene frekvencije [58]. Osnovna učestanost oscilacija glasnica naziva se fundamentalna frekvencija i određuje boju glasa. Uobičajene vrednosti fundamentalne frekvencije za muškarce, žene i decu date su u Tabeli 3.1 [60].

Tabela 3.1: Prosečna, minimalna i maksimalna vrednost fundamentalne frekvencije [60].

Fundamentalna frekvencija[Hz]			
grupa	prosečna	minimalna	maksimalna
Muškarci	125	80	200
Žene	225	150	350
Deca	300	200	500

Pored akustičke, glasnice imaju i biološku funkciju da zaštite respiratorni trakt prilikom gutanja hrane. Tada su u potpunosti primaknute. Grkljan se dalje nastavlja u vokalni trakt.

Vokalni trakt

Vokalni trakt je govorni organ od grkljana do usana koji se sastoji od ždrela (farinksa) koje je membranozna cev koja se račva u usnu i nosnu šupljinu i u akustičkom smislu, predstavljaju sistem rezonatora koji modifikuje vazdušne talase [61]. Između delova ždrela koji su povezani na grkljan i usnu šupljinu nalazi se **jednjak** koji štiti grkljan od hrane pri gutanju. Usna i nosna šupljina neprestano menjaju oblik i veličinu prilikom govora. **Usnu šupljinu** čine usne, zubi, tvrdo nepce, meko nepce koje se završava resicom, alveolarni rub i jezik. **Meko nepce** usmerava vazdušnoj struji ka usnoj šupljini prilikom izgvora vokalnih glasova, ili ka nosnoj šupljini prilikom izgovora nazalnih glasova. Zahvaljujući pokretima **jezika** zvučni talasi se emituju u okolinu u finalnom obliku [59].

3.2.2 Mehaničko-akustički model govornog signala

Mehanički model govornog signala daje opis fizike interakcije govornih organa i zvučnog talasa. Efekti koje je potrebno uzeti u obzir prilikom modelovanja voklanog trakta su [61]:

- geometrija vokalnog trakta,
- toplotna provodnost i viskozno trenje,
- konačna čvrstina i vibracije zidova vokalnog trakta,
- zračenje na usanama,
- uticaj nosne šupljine
- uticaj pobude glasnih žica.

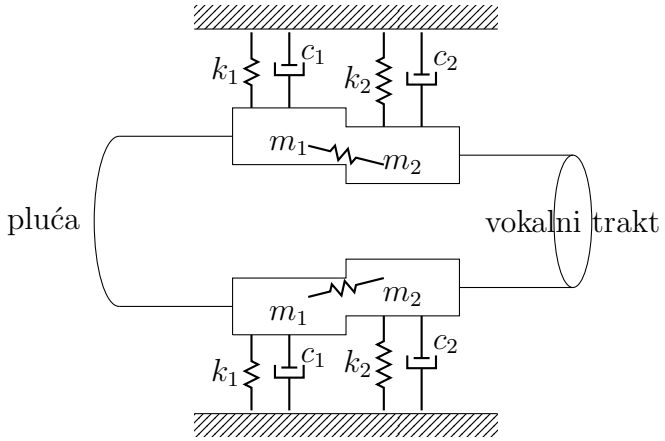
Početni, krajnje uprošćeni model je uniformna cev i jednačine koje opisuju protok vazduha. Model se generalizuje i čini približnijim realnom slučaju uopštenjem na neuniformnu segmentnu cev. Zatim se uzimaju u obzir nesavršenosti sistema: efekat gubitka u vokalnom traktu, zračenja na usnama i efekat nosne šupljine. Drugi deo modeliranja posvećen je modelu glotalnog sistema i njegovom interakcijom sa vokalnim traktom. Krajnji rezultat je digitalni model celokupnog mehanizma.
U slučaju analize glasa za potrebe modeliranja govornika, od interesa je model koji opisuje promenu karakteristika glasa koje odvajaju jednog govornika od drugog kao što je fundamentalna frekvencija. Tri osnovna paramtera odgovorna za promenu fundamentalne frekvencije su [58, 62, 63]:

- a_{g0} - površina između glasnih žica tzv. neutralna glotalna površina,
- p - subglotalni pritisak
- κ - parametar tenzije koji kontroliše fundamentalnu frekvenciju oscilacija glasnih žica.

Dodatno pojašnjenje poslednjeg paramtera - odstupanja glasnih žica i tenzija su glavni faktori koje govornik koristi kako bi kontrolisao izgovaranje glasova.

Mehanički model govornog signala

Pojednostavljeni, deterministički model prikazan je na Slici 3.4, a prvi put su ga predložili Ishizaka i Flanagan [64]. Ovim modelom obuhvata se osnova mehanike i akustike govornih organa.



Slika 3.4: Mehanički model generisanja govornog signala [63].

Svaka od glasnih žica prikazana je kao sistem sa dve mase. Prepostavlja se da su glasne žice idealno simetrične i da na taj način i osciluju, te se dinamika ovog sistema može opisati jednačinama [63]:

$$\xi_1(\boldsymbol{\eta})\dot{\nu}_g + \xi_2(\boldsymbol{\eta})|\nu_g|\nu_g + \xi_3(\boldsymbol{\eta})\nu_g + \frac{1}{\tilde{c}_1}\int_0^t \big(\nu_g(\tau) - \nu_r(\tau)\big)d\tau - u = 0 \tag{3.1}$$

$$[M]\ddot{\boldsymbol{\eta}} + [C]\dot{\boldsymbol{\eta}} + [K]\boldsymbol{\eta} + h(\boldsymbol{\eta}, \dot{\boldsymbol{\eta}}, \nu_g, \dot{\nu}_g) = 0 \tag{3.2}$$

gde je $\boldsymbol{\eta}(t) = (x_1(t), x_2(t), \nu_1(t), \nu_2(t), \nu_r(t))^T$, x_1 i x_2 su pomeraji masa, ν_1 i u_2 opisuju zapreminski protok vazduha kroz dve cevi koje predstavljaju vokalni trakt i ν_r je zapreminski protok vazduha kroz usta. p je oznaka za subglotalni pritisak, a ν_g je funkcija koja predstavlja signale glotalnog pulsa. Pritisak koji se emituje na izlazu dat je funkcijom:

$$p_r(t) = \nu_r(t) \cdot r_r, \quad r_r = \frac{128\rho v_c}{9\pi^3 R_2^2}, \tag{3.3}$$

gde je ρ gustina vazduha, v_c brzina zvuka, R_2 prečnik druge cevi. Od ostalih promenljivih, $\tilde{c}_1 = l_1\pi R_1^2/\rho v_c^2$, gde je l_1 dužina prve cevi, R_1 prečnik prve cevi, i μ koeficijent viskoznog trenja. Jednačina 3.2 opisuje problem vibracije u svakom od dva podsistema pojedinačno, dok jednačina 3.1 spaja dva podsistema.
Parametri $\xi_1, \xi_2, \xi_3, \boldsymbol{h}$, kao i matrice $[\boldsymbol{M}], [\boldsymbol{C}], [\boldsymbol{K}]$ dobijeni su na osnovu modela predloženog u [64] nakon algebarskih manipulacija. U ovim izrazima figurišu parametri koji su relevantni

za fundamentalnu frekvenciju:

$$\xi_1(\boldsymbol{\eta}) = \left(\frac{0.19\rho}{a_{g0}+2l_g x_1} + 2l_g x_1\right) + \frac{\rho}{(a_{g0}+2l_g x_2)}\left[0.5 - \frac{a_{g0}+2l_g x_2}{a_1}\left(1 - \frac{a_{g0}+2l_g x_2}{a_1}\right)\right] \quad (3.4)$$

$$\xi_2(\boldsymbol{\eta}) = 12\mu l_g \frac{d_1}{(a_{g0}+2l_g x_1)^3} + 12l_g^2 \frac{d_2}{(a_{g0}+2l_g x_2)^3} + \tilde{r}_1 \quad (3.5)$$

$$\xi_3(\boldsymbol{\eta}) = \left(\frac{\rho d_1}{a_{g0}+2l_g x_1} + \frac{\rho d_2}{a_{g0}+2l_g x_2} + \tilde{l}_1\right), \quad (3.6)$$

gde je:

$$\tilde{l}_1 = \frac{\rho l_1}{2\pi R_1^2}, \quad \tilde{l}_r = \frac{8\rho}{3\pi^2 R_n}, \quad r_1 = \frac{2}{R_1}\sqrt{\rho\mu\tfrac{\omega}{2}}, \quad \omega = \sqrt{\frac{k_1}{m_1}}, \quad a_1 = \pi y_1^2. \quad (3.7)$$

Modeliranje kontrole izgovora određenih glasova i samim tim fundamentalne frekvencije, ostvareno je putem paramtera tenzije κ na sledeći način:

$$m_1 = \frac{\hat{m}_1}{\kappa}, \quad k_1 = \hat{k}_1\kappa, \quad (3.8)$$

$$m_2 = \frac{\hat{m}_2}{\kappa}, \quad k_2 = \hat{k}_2\kappa, \quad (3.9)$$

$$k_c = \hat{k}_c\kappa, \quad (3.10)$$

gde su paramteri mase i elastičnosti u modelu zapisani preko fiksnog dela vrednosti obeleženog sa ˆ i parametra tenzije.

3.2.3 Modeliranje stohastičkih procesa govornog signala

Stohastički pogled na signal govora odnosi se upravo na parametre koji utiču na fundamentalnu frekvenciju govora: parametar tenzije (κ), subglotalni pritisak (p) i neutralna glotalna površina (a_{g0}) [62]. Ove veličine opisane su odgovarajućim slučajnim promenljivima: K, Π i A_{g0}.

Apriori funkcije gustine verovatnoće

Prema [62], apriori funkcija gustine verovatnoće parametra K izabrana je da bude uniformna raspodela na prilagođenom intervalu $[\alpha, \beta]$ tj $K \sim U[\alpha, \beta]$:

$$f_K(x) = \frac{1}{\beta - \alpha}, \quad \alpha \leqslant x \leqslant \beta. \quad (3.11)$$

U slučaju promenljivih A_{g0} i Π, apriori funkcija gustine verovatnoće izvedena je u obliku:

$$f_{A_{g0}}(x) = e^{-\lambda_0 - \lambda_1 x - \lambda_2 (x)^2}, \quad x \geqslant 0, \quad (3.12)$$

$$f_\Pi(x) = \frac{1}{E\{\Pi\}}\left(\frac{1}{\delta_\Pi^2}\right)^{-\frac{1}{\delta_\Pi^2}} \times \frac{1}{\Gamma(1/\delta_\Pi^2)}\left(\frac{x}{E\{\Pi\}}\right)^{\frac{1}{\delta_\Pi^2}-1} e^{-\frac{x}{\delta_\Pi^2}} \quad (3.13)$$

Iz prethodnih jednačina, λ_0, λ_1 i λ_2 su rešenja jednačine:

$$\int_{-\infty}^{+\infty} f_{A_{g0}}(x)dx = 1, \quad (3.14)$$

$$\int_{-\infty}^{+\infty} x f_{A_{g0}}(x)dx = E\{A_{g0}\}, \quad (3.15)$$

$$\int_{-\infty}^{+\infty} x^2 f_{A_{g0}}(x)dx = c. \quad (3.16)$$

Osim toga, korišćena je notacija $E\{A_{g0}\}$ označava očekivanje slučajne promenljive, a gde je drugi momenat $c = A_{g0}^2(1 + \delta_{A_{g0}}^2)$ slučajne promenljive A_{g0}, a $\delta_{A_{g0}}$ njen koeficijent varijacije. Takođe, $\delta_\Pi = \sigma_\Pi / E\{\Pi\}$ koeficijent varijacije slučajne promenljive Π, gde je σ_Π standardna devijacija. Gama funkcija iskorišćena u ovim jednačinama data je sa:

$$\Gamma(\alpha) = \int_0^{+\infty} \xi^{\alpha-1} e^{-\xi} d\xi, \tag{3.17}$$

Aposteriori funkcije gustine verovatnoće

Od tri parametra od uticaja na fundamentalnu frekvenciju, naveći uticaj ima parametar tenzije glasnih žica koji nije moguće izmeriti [63]. Na osnovu eksperimentalnih podataka, i Bajesove formule [63], aposteriori funkcije gustine verovatnoće za parametar K i fundamentalnu frekvenciju f_0, date su funkcijama:

$$f_K^p(x) = L^{Bayes}(x) f_K(x), \tag{3.18}$$

$$f_{F_0}^p(x) = E\{L^{Bayes}(K) f_{F_0|K}(x|K)\}, \tag{3.19}$$

gde se Bajesova formula izračunava na osnovu vrednosti dobijenih u simulacijama:

$$L^{Bayes}(x) = \frac{\prod_{l=1}^{\nu_{exp}} f_{F_0|K}(f_0^{exp,l}|x)}{E\{\prod_{l=1}^{\nu_{exp}} f_{F_0|Q}(f_0^{exp,l}|K)\}}. \tag{3.20}$$

3.3 Izazovi modeliranja govora

Zadatak prepoznavanja govornika je utoliko lakši ukoliko se modeli govornika koje je potrebno identifikovati u sistemu više međusobno razlikuju. U idealnom slučaju karakteristike govornika na osnovu kojih se formiraju modeli govornika imale bi velike varijacije kada su različiti govornici u pitanju i veoma male kada je jedan govornik u pitanju [65]. Prema podeli u radu Hansena i Hasana [66] uzroci varijabilnosti glasa jednog istog govornika mogu dolaziti od govornika lično, razgovora ili tehnlogije koja obrađuje signal govora [66, 67].
U kategoriju uzroka varijabilnosti glasa koje potiču od **govornika lično** svrstavaju se trenutna stanja govornika koja utiču na njegovo psiho-fizičko stanje:

(1) **stres ili tenzija** aktivnosti koji govornik trenutno obavlja, kao što je na primer vožnja;

(2) **način govora**, kada govornik govori nežnim glasom, obraća se detetu ili peva, šapuće ili viče;

(3) **emocije** - govornik izražava svoje emotivno stanje krzoz način govora;

(4) **fiziološko stanje** - govornik je prehlađen ili pati od neke druge bolesti, pod uticajem je lekova ili alkohola;

(5) **imitacija** - govornik namerno menja glas kako bi izbegao detekciju ili imitira drugog govornika;

(6) **starenje** se može svrstati u kategoriju ličnih varijabilnosti u glasu.

Način govora nije isti kad osoba govori spontano, u prijateljskom razgovoru, poslovnom okruženju, prezentuje, obraća se detetu, ljubimcu ili mašini. Spontani govor kao izvor varijabilnosti uključuje jezik koji se govori, dijalekt, socijalni kontekst, prisnost sa sagovornikom, da li je govornik nešto čita, pita, da li je u pitanju monolog, snimanje glasovne poruke, dijalog, obraćanje auditorijumu itd.

Tehnologija i eksterni izvori varijabilnosti uključuju gde i kako se zvuk snima: elektromehanička svojstva kanala prenosa, mikrofona i prijemnog uređaja, pozadinska buka, akustika prostorije ili udaljen mikrofon, kvalitet podataka, trajanje uzorka, frekvencija odabiranja, kvalitet snimka, kompresija itd.

4. Karakterizacija signala govora

Signal govora može se okarakterisati i na druge načine na osnovu fiziološkog modela: frekvencijskom domenu, kepstralnom domenu, prediktivnim koeficijentima, kvalitativno itd. Karakteristike govora koje ćemo upotrebiti zavise od zadataka obrade govora koje rešavamo, a u ovom poglavlju fokus je stavljen na karakteristike koje kvalitetno opsiju govornike, kao i emocije izražene u govoru. Osobine idealnih karakteristika za zadatak prepoznavanja govornika definisane su na sledeći način [65, 68]:

(1) pojavljuju se prirodno i često u normalnom govoru,

(2) lako su merljive,

(3) razlikuju se što je više moguće od govornika do govornika, a konstantne su za istog govornika,

(4) ne menjaju se sa vremenom i ne zavise od zdravstvenog stanja govornika,

(5) ne pogađa ih razuman nivo pozadinske buke, niti na njih ima uticaj kanal prenosa,

(6) ne menjaju se i pored napora govornika da izmeni glas, ili bar pokušaj maskiranja glasa i imitacije na njih nema uticaj.

4.1 Grupe karakteristika

Na osnovu fizičke interpretacije, karakteristike signala govora mogu se podeliti u nekoliko kategorija [69, 70]: karakteristike izvora govora, karakteristike kvaliteta glasa, spektralne karakteristike, prozodijske karakteritike i rečnik govornika (tj karakteristike visokog nivoa [70]).
Karakteristike izvora govora opisuju signal glotalne pobude kod zvučnih glasova. Takvi su na primer oblik glotalnog pulsa i fundamentalna frekvencija. Razumno je pretpostaviti da nose informacije koje su karakteristične za govornika [70].
Spektralne karakteristike ekstrahuju se iz signala govora na kratkim vremenskim intervalima, frejmovima trajanja od $20 - 30ms$ [71], jer na toj dužini trajanja signal možemo smatrati stacionarnim. Na osnovu signala u zadatom frejmu izračunava se vektor spektralnih karakteristika. U istraživanjima je korišćeno još nekoliko vrsta spektralnih karakteristika. Estimacija spektra može biti zasnovana na diskretnoj Furijeovoj transformaciji ili alternativno na linearnoj predikciji. Na osnovu diskretne Furijeove transformacije dalje se izračunavaju karakteristike kao

što su mel-frekvencijski kepstralni koeficjenti (MFCC), a korišćenjem linearne predikcije mogu se dalje izračunati linearni prediktivni kepstralni koeficijenti (LPCC), formantne učestanosti, njihovi opsezi [71] itd.

Prozodijske karakteristike odnose se na aspekte govora kao što su akcenti, intonacija, brzina i ritam govora itd. Prozodijske karakteristike, za razliku od spektralnih, mogu da se protežu duž više slogova, reči ili čak cele rečenice [72]. Prozodija reflektuje razlike u stilu govora, jezičko poreklo, vrstu rečenice i emocije. Izazov za prepoznavanje govornika nezavisno od teksta je modelovanje različitih nivoa prozodijskih informacija da bi se zabeležila razlika među govornicima. Najvažniji prozodijski parametar je fundamentalna frekvencija. Kombinacijom fundamentalne frekvencije sa drugim spektralnim karakteristikama može biti jako efektna, naročito u okruženju sa visokim nivoom buke [70]. Druge prozodijske karakteristike su na primer brzina govora, trajanje fonema ili pauza, raspodela energije.

Kvalitet glasa opisuje se kao [73]: **normalan, zadihan, hrapav, grub, napet**. Kvalitet glasa ili njegova promena mogu okarakterisati govornika [74] ili njegovo emotivno stanje [75]. Međutim, najveći nedostatak ove vrste karakteristika je to što nisu direktno merljive [70].

Govornici se razlikuju ne samo na osnovu boje glasa i načina izgovora, nego i po **jeziku i rečniku koji upotrebljavaju** [74]. Ovakav pristup prepoznavanju govornika zahteva analizu sadržaja govora, kao i znanje o karakterističnom rečniku nekog govornika. Rezultujuće karakteristike u ovom slučaju nisu numeričkog karaktera [70].

4.2 Izračunavanje vektora karakteristika

U nastavku opisani su koraci za izračunavanje karakteristika glasa koje se najčešće primenjuju u obradi govora.

4.2.1 Početna obrada signala.

Izdvajanje kratkovremenskih karakteristika iz signala govora počinje podelom signala na frejmove, obično trajanja između 15 i 25 ms. Frejmovi se uzimaju sa pomerajem od 10 do 15 ms. Na svaki od frejmova primenjuje se prozorska funkcija. Najčešće se koristi funkcija Hammingovog prozora [76]:

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right), \quad n = 0, ..., N-1, \tag{4.1}$$

gde je N broj odbiraka u jednom frejmu/prozoru, mada postoje i druge prozorske funkcije, kao što su funkcija pravougaonog ili Hannovog prozora [77]:

$$w(n) = \sin^2\left(\frac{\pi n}{N}\right), \quad n = 0, ..., N. \tag{4.2}$$

Nakon ovog početnog koraka, dolazi dalja obrada koja zavisi od konkretnog tipa karakteristika koji se izdvaja.

4.2.2 Linearni prediktivni kepstralni koeficijenti (LPCC)

Linearni prediktivni kepstralni koeficijenti (LPCC) [71, 78] zasnovani su na pretpostavci da se jedan odbirak signala govora u datom trenutku može predvideti kao linearna kombinacija prethodnih odbiraka. Koriste se u različitim zadacima obrade govora, i kao osnova za izračunavanje

kompleksnijih karakteristika. Ovi koeficijenti izračunavaju se na osnovu **autokorelacione funkcije**. Primenjuje se modifikovana kovarijaciona metoda:

$$R_{i,j} = \frac{1}{2(N-p)}\left(\sum_{n=p}^{N-1} x(n-i)x(n-j) + \sum_{n=0}^{N-1-p} x(n+i)x(n+j)\right), \quad 0 \leqslant i,j \leqslant p \tag{4.3}$$

gde je p broj autokorelacionih koeficijenata, a N dužina frejma u odbircima, kako je ranije definisano. Uobičajena vrednost je $p = 8$. LPC koeficijenti a_m, $m = 1, .., p$, na osnovu kojih se izvode LPCC koeficijenti, dobijaju se rešavanjem sledeće jednačine:

$$- \begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,p} \\ R_{2,1} & R_{2,2} & \cdots & R_{2,p} \\ \vdots & & \ddots & \vdots \\ R_{p,1} & & & R_{p,p} \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_{1,0} \\ R_{2,0} \\ \vdots \\ R_{p,0} \end{bmatrix} \tag{4.4}$$

Na kraju LPCC kepstralni koeficijenti izvode se korišćenjem sledeće referentne formule:

$$c_0 = R_{0,0} \tag{4.5}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leqslant m \leqslant p \tag{4.6}$$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad p < m \leqslant M-1 \tag{4.7}$$

gde je M broj LPCC koeficijenata [71].

4.2.3 Log-frekvencijski koeficijenti spektra snage (LFPC)

Log-frekvencijski koeficijenti spektra snage (LFPC) [79] pružaju informaciju o distribuciji energije spektra i korisni su u detekciji emocija u govoru. Ovi koeficijenti izračunavaju se na sledeći način: signal podeljen na frejmove, na koje je primenjena prozorska funkcija transformiše se u frekvencijski domen korišćenjem Furijeove transformacije. Spektar se zatim deli na M opsega korišćenjem seta pravougaonih filtara. Centralne frekvencije i širine filtara izračunavaju se na sledeći način:

$$b_1 = C \tag{4.8}$$

$$b_m = \alpha b_{m-1}, \quad 2 \leqslant m \leqslant M \tag{4.9}$$

$$f_m = f_1 + \sum_{j=1}^{m-1} b_j + \frac{b_m + b_1}{2} \tag{4.10}$$

gde su b_m and f_m širina opsega i centralna frekvencija m-tog filra. Uobičajene vrednosti konstanti su $C = 54Hz$, $f_1 = 127Hz$ i $\alpha = 1.4$. Pravougaoni prozor definisan je formulom:

$$w_m = 1, \quad \frac{f_m - b_m}{2} \leqslant f \leqslant \frac{f_m + b_m}{2} \tag{4.11}$$

gde je $m = 1, \ldots, M$, $f \in \frac{nF_S}{N}, n = 0, \ldots, \frac{N}{2}$ i F_S frekvencija odabiranja. Energija je izračunata kao kvadrat sume izlaza svakog od filtara:

$$S(m) = \sum_{f=\frac{f_m - b_m}{2}}^{\frac{f_m + b_m}{2}} \Big(X(f)w_m(f)\Big)^2 \tag{4.12}$$

gde je $X(f)$ Furijeova spektralna komponenta na frekvenciji f. Konačna mera energije frekvencijskog opsega izračunata je logaritmovanjem i skaliranjem filtra:

$$SE(m) = \frac{10 \log_{10} \Big(S(m) \Big)}{N_m} \tag{4.13}$$

gde je N_m broj spektralnih komponenti m-tog filtra. To je i finalna formula za izračunavanje M spektralnih koeficijenata, gde se uobičajeno postavlja $M = 12$.

4.2.4 Mel-frekvencijski kepstralni koeficijenti (MFCC)

Najpopularnija vrsta koeficijenata za prepoznavanje govornika i emotivnih stanja su mel-frekvencijski kepstralni koeficijenti (Mel Frequency Cepstral Coefficients - MFCC) [80]. Smatra se da MFCC predstavljaju signal govora baziran na ljudskom sluhu, a da ovi koeficijenti sadrže informacije o identitetu govornika i njegovom emotivnom stanju, i druge informacije kao što su glasnost i sadržaj govora [81]. MFCC koeficijenti izračunavaju se prema sledečoj proceduri: Signal na koji je primenjena početna obrada u vidu podele na frejmove i primene Hammingovog prozora (jednačina 4.1), transformiše se u frekvencijski domen primenom Furijeove ransformacije. Potom se izračunava spektar snage, kao kvadrat magnitude spektra. Sledeći korak je primena banke od M trougaonih filtara, koji su ekvidistantni na mel skali:

$$H_m(\varphi) = \begin{cases} \frac{\varphi - \varphi_{b_{m-1}}}{\varphi_{b_m} - \varphi_{b_{m-1}}} & , \varphi_{b_{m-1}} \leqslant \varphi \leqslant \varphi_{b_m} \\ \frac{\varphi_{b_{m+1}} - \varphi}{\varphi_{b_{m+1}} - \varphi_{b_m}} & , \varphi_{b_m} \leqslant \varphi \leqslant \varphi_{b_{m+1}} \\ 0 & , \text{inače} \end{cases} \tag{4.14}$$

gde je $m = 1, \ldots, M$ indeks filtra, a φ predstavlja diskretnu frekvenciju na mel-skali. Granične frekvencije $\varphi_{b_0}, \ldots, \varphi_{b_{M+1}}$ dele mel-skalu na $M + 1$ ekvidistantan opseg. Maksimum mel-skale odgovara vrednosti $\frac{F_S}{2}$ na linearnoj (Hz) skali. Filtri se dalje transformišu u lineranu skalu zahvaljujući relaciji mel i linearne skale:

$$\varphi = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \tag{4.15}$$

Banka filtara je normalizovana na takav način da je suma koeficijenata za svaki filtar jednaka jedinici. Ovim korakom, banka filtara dobila je konačan oblik. Njenom primenom na spektar snage dobija se mel-spektar snage. Na kraju, mel-kepstrali koeficijenti generišu se primenom diskretne kosinusne transformacije na logaritam spektra snage. Rezultat je M kepstralnih koeficijenata, gde se za vrednost uobičajeno bira $M = 13$ ili $M = 21$ koeficijent.

4.2.5 Perceptualni linearni predikcijski keprstralni koeficijenti (PLP)

Ova vrsta karakteristika bazirana je na sličnim pretpostavkama kao i LPC koeficijenti, stim da su PLP koeficijenti (Perceptual Linear Prediction - PLP) konzistentni sa ljudskim sluhom [82]. Postupak izračunavanja PLP koeficijenta počinje na isti način kao i u slučaju MFCC koefcijenata: na signal podeljen na frejmove primenjuje se funkcija Hammingovog prozora, pa Furijeova transformacija odakle se izračunava spektar snage. Za razliku od MFCC, sada se preslikavanje vrši na Barkovu frekvencijsku skalu:

$$\Omega(f) = 6 \ln \left(\frac{f}{1200\pi} + \sqrt{\left(\frac{f}{1200\pi} \right)^2 + 1} \right), \tag{4.16}$$

gde je f linearna frekvencija, a Ω frekvencija na Barkovoj skali [83]. Na spektar snage koji se dobija kao rezultat primenjuje se konvolucija sa krivom $\Psi(\Omega)$ datom sledećom formulom:

$$\Psi(\Omega) = \begin{cases} 10^{2.5(\Omega+0.5)} & , -1.3 < \Omega < -0.5 \\ 1 & , -0.5 < \Omega < 0.5 \\ 10^{-(\Omega-0.5)} & , 0.5 < \Omega < 2.5 \\ 0 & , \Omega < -1.3 \Omega > 2.5 \end{cases} \tag{4.17}$$

4.2.6 Kratkovremenska energija

Amplituda signala govora ima različite varijacije tokom vremena i ove varijacije mogu se izraziti kroz energiju signala govora:

$$E(n) = \sum_{m=-\inf}^{+\inf} \Big(x(m) w(n-m) \Big)^2, \tag{4.18}$$

gde je $x(m)$ odbirak signala govora m, a $w(m)$ funkcija Hammingovog prozora (jednačina 4.1). Generalno, bezglasni segmenti govora imaju mnogo manje enegrgije od segmenata koji sadrže reči [84], pa je ova karakteristika pogodna za izdvajanje reči od tišine i pauza, kao i za karakterizaciju različitih samoglasnika i suglasnika za prepoznavanje govornika [20].

4.2.7 Nulti prolazi (ZCR)

Broj nultih prolaza (Zero-crossing rate(ZCR)) [85] je koliko puta signal govora promeni znak u toku trajanja jednog frame-a. [84]. Promena znaka signala dešava se učestalije tokom pauza između reči i tokom bezglasnih delova govora [86], tako da se može iskoristiti za detekciju pauza i reči u rečenici. Ova informacija može se iskoristiti za dalju obradu signala govora ili direktno, kao opis dinamike govora [87]. ZCR izračunava se kao:

$$ZCR(n) = \sum_{k=n-\frac{N}{2}}^{n+\frac{N}{2}-1} x_0(k), \quad x_0(k) = \begin{cases} 0, & sgn(x(k)) = sgn(x(k-1)) \\ 1, & sgn(x(k)) \neq sgn(x(k-1)) \end{cases} \tag{4.19}$$

4.2.8 Formanti i njihova širina

Formanti tj formantne učestanosti predstavljaju rezonantne učestanosti funkcije prenosa vokalnog trakta [72]. Vrednosti prvog i drugog formanta dominantno su određene izgovorenim sadržajem, dok su formanti višeg reda vezani za karakteristike govornika [20]. Osim same vrednosti formantne učestanosti, u obzir se uzimaju i širina formanata, amplituda. Formanti se uglavnom izračunavaju korišćenjem LPC kepstra [88].

4.2.9 Fundamentalna frekvencija (F_0)

Fundamentala frekvencija glasa, F_0 definisana je brzinom oscilacija glasnica [89]. Izračunavanje fundamentalne frekvencije izazov je već pola veka. Razvijene su brojne tehnike zasnovane na reprezentaciji signala u vremenskom domenu, frekvencijskom domenu ili hibridno [90]. Poteškoće za tehnike estimacije su [91]:

(1) govor nije idealno periodičan usled promena fundamentalne frekvencije i pokreta vokalnog trakta;

(2) teško je proceniti fundamentalnu frekvenciju delova govora koji nije dovoljno vokalizovan i na njegovim počecima/krajevima;

(3) spoljašnji faktori, poput šuma, degradiraće performanse estimacije.

Algoritam za izračunavanje fundamentalne frekvencije na osnovu LPC analize [92], počinje podelom signala na frejmove dužine N, sa pomerajem S, potom se primenjuje procedura za poboljšanje periodične strukture signala korišćenjem Hannovog prozora (jednačina 4.2). Kratkovremenska energija signala, za ove potrebe, dobijena konvolucijom prozora usrednjavanja i kvadrata amplitude signala pomnožena je sa originalnim signalom [92]:

$$\begin{aligned} w_S &= w * w, \\ E &= w_S * x^2, \\ x_2 &= x \cdot e, \end{aligned} \tag{4.20}$$

Dalje, na svaki frejm signala primenjena je prozorska funkcija 4.2. Zatim se vrši linearna prediktivna kepstralna (LPC) analiza. Izračunavanje LPC koeficijenata a_k dato je u prethodnoj sekciji (jednačina 4.4), dok su sada potrebni reziduali r(n,m):

$$r(n,m) = x_2(n,m) + \sum_{k=1}^{p} a_k \cdot x_2(n-k,m), \tag{4.21}$$

gde je p red LPC analize. Sledeći korak je izračunavanje Hilbertove transformacije reziduala $r_h(n,m)$, koja je definisana funkcijom prenosa $H(\omega) = -j \cdot \mathrm{sgn}(\omega)$. Hilbertova envelopa signala izračunava se kao:

$$h(n,m) = \sqrt{r_h^2(n,m) + r^2(n,m)}. \tag{4.22}$$

Fundamentalna perioda m-tog frejma, $T_0(m)$, je tada prvi pik autokorelacione funkcije Hilbertove envelope. Fundamentalna frekvencija izračunava se kao:

$$F_0(m) = \frac{F_S}{T_0(m)}, \tag{4.23}$$

gde je F_S - frekvencija odabiranja signala. Relevantno je spomenuti i pojam *pitch*-a. U litaraturi se često pojavljuje i koristi kao sinonim za fundamentalnu frekvenciju, mada je definisan kao ljudska percepcija fundamentalne frekvencije [89].

4.2.10 Kvalitet glasa

Ova karakteristika glasa ima opisne kategorije, koje se reflektuju numerički u neke druge karakteristike govora. Zadihan glas okarakterisan je velikim protokom vazduha, što je suprotno od napetog glasa koji ima mali protok vazduha [73]. Hrapav glas odlikuju niska frekvencija i različite nepravilnosti koje se pojavljuju tokom govora [73]. Grub glas je za razliku od hrapavog okarakterisan normalnom fundamentalnom frekvencijom, ali sadrži aperiodičnosti ili šum u spektru [73].

4.3 Redukcija dimenzija vektora karakteristika

Izdvajanjem različitih karakteristika govora i njihovih statistika, vektori karakteristika na osnovu kojih se vrši dalja obrada mogu biti veoma veliki. Vektor karakteristika u sebi nosi mnogobrojne informacije od kojih nisu sve od podjednake važnosti za zadatak obrade govora. Osim toga, ne nose svi elementi vektora karakteristika potpuno novu informaciju, nego postoji

određena korelacija između komponenti. Različiti algoritmi za redukciju dimenzija razvijeni da bi od početnog, sirovog vektora karakteristika algebarskim transformacijama došli do novog vektora, koji je manjih dimenzija i sadrži samo relevantne informacije. U nastavku su opisane dve najčešće korišćene metode za redukciju dimenzija: analiza glavnih komponenata i linarna disckriminaciona analiza.

4.3.1 Analiza glavnih komponeneata (PCA)

Analiza glavnih komponenata (Principle Component Analysis - PCA) [93] je standardna tehnika analize podataka za transformaciju originalnog seta podataka u niže dimenzioni nekorelisani, ortogonalni prostor korišćenjem dekompozicije sopstvenih vektora. Proces transformacije podataka započinje od izračunavanja srednje vrednosti svih vektora, koja se potom oduzima od svih vektora. Time se postiže da je ukupna srednja vrednost nula vektor:

$$\boldsymbol{\Psi} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\Gamma_i}, \tag{4.24}$$

$$\boldsymbol{\Phi_i} = \boldsymbol{\Gamma_i} - \boldsymbol{\Psi}. \tag{4.25}$$

gde je M dimenzija originalnih vektora karakteristika $\boldsymbol{\Gamma_i}$, $i = 1, 2, \cdots, N$, gde je N ukupan broj vektora. Sada imamo novoformiranu matricu podataka dimenzija $M \times N$:

$$\boldsymbol{X} = [\boldsymbol{\Phi_1}, \boldsymbol{\Phi_2}, \cdots, \boldsymbol{\Phi_N}]. \tag{4.26}$$

Njena kovarijaciona matrica određuje se kao:

$$\boldsymbol{C} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\Phi_i}\boldsymbol{\Phi_i}^T \tag{4.27}$$

Od interesa su sopstvene vrednosti matrice C, $\boldsymbol{\Lambda} = [\boldsymbol{\lambda_1}, \boldsymbol{\lambda_1}, \cdots, \boldsymbol{\lambda_K}]$, gde je $K = min(M, N)$, i spostveni vektori $\boldsymbol{U_K} = [\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_K}]$. Komponente koje sadrže najznačajnije informacije su one koje odgovaraju najvećim sopstvenim vrednostima, pa se redukcija u prostor u k dimenzija, $k < K$ vrši izborom k najvećih sopstvenih vrednosti i odgovarajućih sopstvenih vektora $\boldsymbol{U_k} = [\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_k}]$. Projektovanjem matrice podataka $\boldsymbol{X}$ na ovaj prostor $\boldsymbol{U_k}$ dimenzija $(M \times k)$ dobija se transformisani set podataka:

$$\boldsymbol{\Omega} = \boldsymbol{U_k}^T\boldsymbol{X}\ (k \times N), \tag{4.28}$$

$$\boldsymbol{\omega_i} = \boldsymbol{U_k}^T\boldsymbol{\Phi_i}, \quad i \in [1, N]. \tag{4.29}$$

4.3.2 Linarna disckriminaciona analiza (LDA)

Linearna diskriminaciona analiza (Linear discriminant analysis - LDA) [94] preslikava M-dimenzione vektore karakteristika koji pripadaju različitim klasama (npr različitim govornicima) u niže dimenzioni prostor tako da je rastojanje klasa maksimalno u tom novom prostoru. Neka je skup klasa $s \in [1, S]$, od kojih svaka ima N_S elemenata. Definišu se međuklasno rasejanje S_B i unutarklasno rasejanje S_W:

$$\boldsymbol{S_B} = \frac{1}{S}\sum_{s=1}^{S}(\boldsymbol{\Psi_s} - \boldsymbol{\Psi})(\boldsymbol{\Psi_s} - \boldsymbol{\Psi})^T, \tag{4.30}$$

$$\boldsymbol{S_W} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N_S}\sum_{i=1}^{N_S}(\boldsymbol{\Gamma_i} - \boldsymbol{\Psi_s})(\boldsymbol{\Gamma_i} - \boldsymbol{\Psi_s})^T, \tag{4.31}$$

gde je $\boldsymbol{\Psi}$ vektor sredanje vrednosti svih odbiraka, a $\boldsymbol{\Psi_S}$ vektor srednje vrednosti klase s. $\boldsymbol{\Gamma_i}$ su orginalni vektori karakteristika, $i = 1, 2, \cdots, N$, gde je N ukupan broj vektora. Sada, cilj da odbirci iz iste klase budu što bliže, a odbirci iz različitih klasa što dalje može se opisati uslovom [95]:

$$\max \frac{|\boldsymbol{S_B}|}{|\boldsymbol{S_W}|} \sim \max |\boldsymbol{S_W^{-1}} \boldsymbol{S_B}|, \tag{4.32}$$

gde je sa $|\cdot|$ označena determinanta matrice. Zadatak je da izaberemo redukovan broj komponenata vektora karakteristika K, $K < M$ primenom matrice transformacija A na originalni vektor, tako da uslov 4.32 bude zadovoljen i u transformisanom prostoru. Pokazuje se da se to postiže izborom najdominantnijih K sopstvenih vrednosti $\boldsymbol{\Lambda} = [\boldsymbol{\lambda_1}, \boldsymbol{\lambda_1}, \cdots, \boldsymbol{\lambda_K}]$ matrice $\boldsymbol{S_W^{-1}} \boldsymbol{S_B}$, kojima odgovaraju sopstveni vektori $\boldsymbol{U_K} = [\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_K}]$ [95]. Transforimisani vektori karakteristika dobijaju se u obliku:

$$y = \boldsymbol{U_K^T} \cdot x. \tag{4.33}$$

4.4 Uticaj emocija na karakteristike signala govora

Emocije, koje su svakodnevna pojava u govoru u značajnoj meri pogađaju neke od krakteristika signala govora. U istraživanjima prepoznavanja emocija u govoru, zabeležen je i način na koji emocije utiču na određene karakteristike signala govora. Primeri uticaja na vrednost i varijansu fundamentalne frekvencije F_0, konture slogova, ritam govora, dužinu fonema i naglašavanje reči, emocija besa, sreće, tuge i straha u odnosu na neutralno stanje prikazani su u Tabeli 4.1. Tabela je popunjena na osnovu prethodnih istraživanja [96–98].

Tabela 4.1: Promene karakteristika govora za emotivna stanja u odnosu na neutralno stanje.

Radost	
F_0	povišena srednja vrednost i varijansa
Ritam	brz
Dužina fonema	male promene
Akcentuacija	nekoliko slogova je akcentovano, kao i poslednja reč
Konture slogova	veliki broj rastućih i malo opadajućih kontura slogova
Bes	
F_0	povišena srednja vrednost i varijansa
Ritam	brz
Dužina fonema	male promene
Akcentuacija	najmanje neakcentovanih i najviše jako akcentovanih reči; poslednja reč nije akcentovana
Konture slogova	Veliki broj rastućih i opadajućih kontura
Tuga	
F_0	smanjena srednja vrednost i varijansa
Ritam	spor
Dužina fonema	velika varijansa trajanja fonema
Akcentuacija	slična kao i za neutralno stanje
Konture slogova	sličan broj slogova sa konstantnom F_0 kao i neutralno stanje, veoma mali broj slogova sa rastućom konturom
Strah	
F_0	povišena srednja vrednost i varijansa
Ritam	brži od neutralnog stanja, sporiji od radosti i besa
Dužina fonema	skraćena, isprekidan govor, pauze između reči
Akcentuacija	slična kao i za radost
Konture slogova	sličan broj slogova sa konstantnom F_0 kao i neutralno stanje, najmanji broj slogova sa opadajućom F_0

Deo II

Istraživanje, rezultati i naučni doprinos

5. Klasifikacije i modeliranje govornika u uslovima emotivnog govora

Osnovni zadatak mašinskog prepoznavanja govornika, je kako da na što precizniji, i što jednostavniji način opišemo govornika na osnovu karakteristika izdvojenih iz obučavajućih uzoraka govora i to na način da se što lakše može utvrditi ko je izgovorio neki novi uzorak govora (Slika 2.1 i 2.2). Dodatno, u slučaju prisustva emotivnog govora, želeli smo da postignemo da sa što manje ulaznih podataka možemo da prepoznamo govornike čak i kada su ljuti, uplašeni, uzbuđeni ili tužni. Primena bilo koje od tehnika prepoznavanja govornika kompromis je neophodne količine ulaznih podataka, robusnosti i kompleksnosti sistema. U našim istraživanjima, koristili smo Gausove mešavine i i-vektore za model i klasifikaciju, i odgovorili smo na postavljena pitanja na četiri nivoa:

(1) analizom algoritama za modeliranje i klasifikaciju govornika,

(2) varijacija sadržaja govora za obuku modela,

(3) varijacija konfiguracije obuke modela govornika,

(4) varijacija strukture modela govornika.

Analiza algoritama za modeliranje i klasifikaciju govornika obuhvatila je dosada poznate metode klasifikacije i njihov način u prepoznavanju govornika. Teorijski su obrađene tehnike Gausovih mešavina, skrivenih Markovljevih modela, mašina potpornih vektora, i-vektora, dubokih neuralnih mreža i x-vektora. Eksperimentalno su evaluirane metoda standardnog modela Gausovih mešavina, kao tehnika koja je osnov modernog prepoznavanja govornika, i tehnika i-vektora koja se smatra savremenom tehnikom rasprostranjenom u komercijalnoj primeni.
Varijacija sadržaja govora za obuku modela govornika, odnosi se na korišćenje i emotivnog govora za obuku modela govornika - različit broj rečenica izgovorenih u određenom emotivnom stanju i različit ukupan broj rečenica. Jedan model govornika obučavali smo sa rečenicama neutralnog govora, ali i sa rečenicama emotivnog govora - radosti, besa, tuge i straha.
Varijacija konfiguracije modela govornika je modeliranje govornika sa više od jednog modela, takođe na osnovu grupisanja emotivnog govora i kasnije prepoznavnja ne osnovu ovako distribuiranog modela.
Varijacija strukture modela govornika odnosi se na određivanje broja mešavina za svakog od govornika na osnovu inicijalne klasterizacije, dakle da broj nije isti i nije zadat unapred nego se određuje automatski. Tome je posvećeno sledeće poglavlje.

Pregled rezultata dosadašnjih istraživanja koja su slična našem prikazan je u Tabeli 5.1.

Tabela 5.1: Rezultati sistema za prepoznavanje govornika u uslovima emotivnog govora.

Baza	BRG	Trening podaci	RN	RE	Ref
Berlin + IITKGP-SESC	30	8 neutralnih rečenica	98.33	50.33	[99]
IEMOCAP	10	160 rečenica iz svih stanja	80.77	75.47	[100]
			/	91.34	[25]
Berlin	10	9 neutralnih rečenica	99	57	[23]
		34 nasumične rečenice	/	98.57	
EPS	8	0.5-1 minuta neutralnog govora	/	67.91	[26]
		2 minuta govora iz svih stanja klasterovanih na 3 modela	/	90.52	
		2 minuta govora iz svih stanja nasumično grupisanih na 3 modela	/	68.46	
MASC	50	2 neutralna pasusa	95.63	51.28	[37]
	25	100+ rečenica	95.37	56.17	[31]
	50	100+ rečenica	93.07	61.64	[32]

RN - rezultati testiranja neutralnim govorom, RE - rezultati testiranja emotivnim govorom, BRG - broj govornika.

5.1 Opis eksperimenata i analize

Ključni izazov u sistemima za prepoznavanje govornika je razlika u njegovom emotivnom stanju prilikom obuke sistema i prepoznavanja, jer se karakteristike glasa menjaju pod uticajem emocija. U našim istraživanjima, konkretan zadatak je prepoznavanje govornika u uslovima emotivnog govora na zatvorenom skupu govornika, a nezavisno od teksta. Analizirali smo efekte pojedinačnih emocija na odstupanje od modela govornika, kao i različiti jezici i baze različite veličine. Korišćene karakteristike u svim eksperimentima su MFCC koeficijenti i to 13 koeficijenata na prozorima od $20ms$ i to sa pomerajem od $10ms$. Ukupno je dizajnirano pet modela govornika, dok su upotrebljeni algoritmi obučavanja bili Gausove mešavnie (GMM) sa 30 mešavina i i-vektori [2]. Rezultati su osim na ruskoj bazi podataka [2], dopunjeni ispitivanjima na bazama srpskog, italijanskog i engleskog jezika.

5.1.1 Korišćeni podaci

Podaci korišćeni za eksperimente su rečenice odglumljenog govora iz baza emotivnog govora ruskog (RUSLANA [101]), srpskog (GEES [102, 103]), engleskog (SAVEE [104]), italijanskog (EMOVO [105]). Osim toga korišćena je i baza neutralnog govora švajcarskog nemačkog TEVOID [106]. Pregled ovih baza dat je u Tabeli 5.2. Emocije koje su bile od interesa su strah, tuga, bes i radost, kao i neutralno emotivno stanje. Ukupan broj govornika u bazama emotivnog govora je $61 + 6 + 6 + 4 = 77$, dok TEVOID baza ima ukupno 50 govornika. Više o bazama govora dato je u Apendixu D.

Tabela 5.2: Baze emotivnog govora

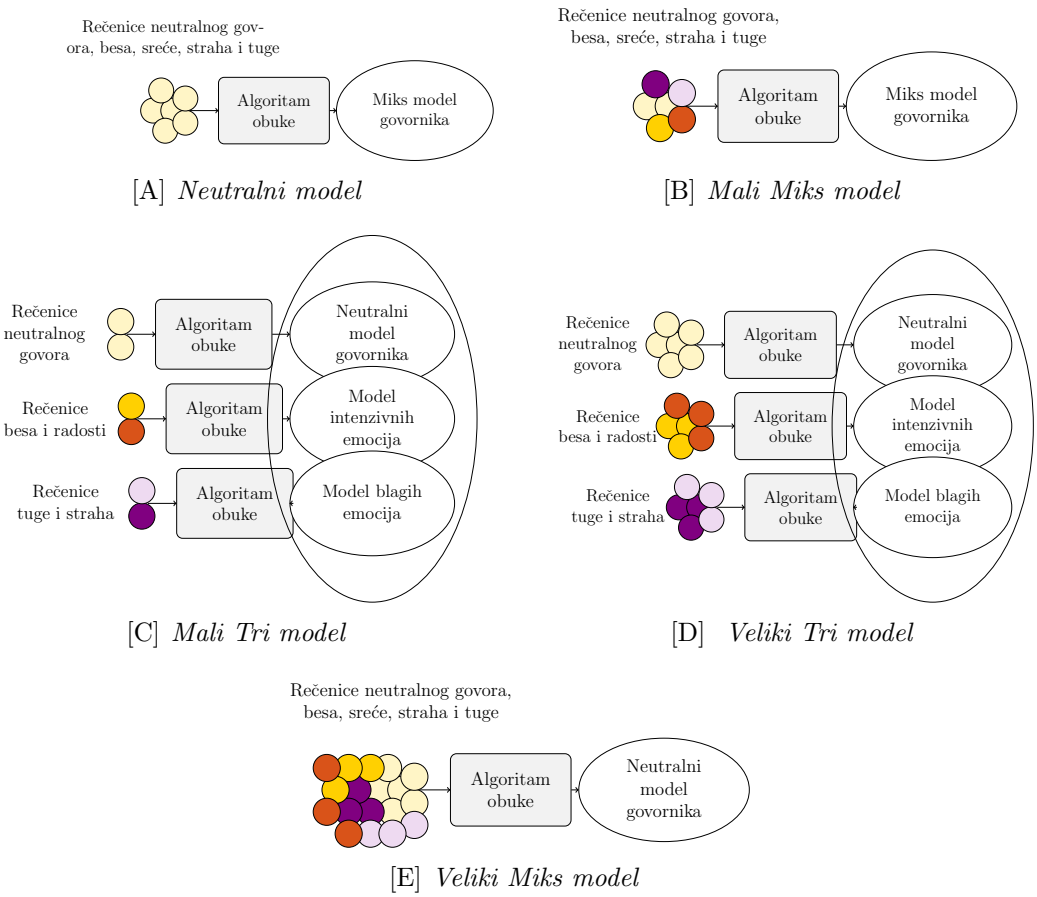
Baza	Jezik	Emocije	BRG	Tekst po govorniku	Ref.
RUSLANA	Ruski		61	10 predefinisanih rečenica	[101]
GEES	Srpski		6	30 reči, 30 kratkih, 30 dugih rečenica, 1 praragraf	[102]
EMOVO	Italijnski		6	14 rečenica	[105]
SAVEE	Engleski		4	4rečenice	[104]

Neutralno Radost Bes Tuga Strah Iznenađenje Gađenje

Evaluacija je rađena u okviru svake od baza zasebno.

5.1.2 Način obuke modela govornika

Obučavano je pet modela govornika: [A] *neutralni model*, [B] **mali miks model**, [C] *mali tri model*, [D] *veliki tri model* i [E] *veliki miks model*. Svaki od ovih modela obučavan je sa ograničenim setom podataka - šest ili osmanaest rečenica. Šema obuke prikazana je sa Slici 5.1, a podaci korišćeni u obuci i opis modela dat je u Tabeli 5.3.



Slika 5.1: Šema obuke modela govornika u eksperimentima [A]-[E] [2].

Neutralni model obučavan je sa šest rečenica neutralnog govora za svakog od govornika (Slika 5.1[A]) [2].
Mali miks model je prvi eksperiment u kome je emotivni govor uključen u trening na način da je umesto šest rečenica od kojih su sve neutralni govor, za obuku modela koristimo dve rečenice emotivnog govora i po jednu rečenicu od svakog emotivnog stanja kao što je dato u Tabeli 5.3 [2].
Mali tri model je sledeći model sa kojim smo eksperimentisali. Podaci koji su upotrebljeni u obuci *malog misks modela* sada su iskorišćeni na drugačiji način - umesto jednog obučena su tri modela za jednog govornika. Šema obuke predastavljena je na Slici 5.1[C]. Neutralni model govornika obučen je sa dve neutralne rečenice, *model intenzivnih emocija* obučen je sa jednom rečenicpom besa i jednom rečenicom radosti, a *model blagih emocija* obučen je sa jednom rečenicom tuge i jednom rečenicom straha. Ovakva podela u dva različita modela odgovara podeli emocija prema nivou uzbuđenosti [107].
Veliki tri model je takođe pristup u kome se za svakog govornika kreiraju tri modela govornika kao i kod *malog tri modela* govornika, stim da je broj rečenica korišćenih za obuku povećan tako da se za svaki od modela (*model blagih emocija*, *model intenzivnih emocija* i *neutralni model*) koristi po šest rečenica (Slika 5.1[C]). *Neutralni model* je, dakle, isti kao i u prvom

eksperimentu. *Model intenzivnih emocija* obučen je upotrebom tri rečenice govora besa i tri rečenice govora radosti. *Model blagih emocija* obučen je sa tri rečenice govora tuge i tri rečenice govora straha. Obuka ovog modela sprovedena je sveukupno sa 18 različitih rečenica [2].

Na kraju, isti broj rečenica upotrebljen je za obuku samo jednog modela - *velikog miks modela*. Šest neutralnih i po tri rečenice svakog od emotivnih stanja upotrebljene su kao ulaz u algoritam obuke jedinstvenog modela govorinika [2].

Tabela 5.3: Broj i emocije rečenica za obuku svakog od pet modela [2].

Model govornika	BRM	BRR	UBRR
[A] Neutralni	1	6 neutralnih	6
[B] Mali miks model	1	2 neutralne, 1 radost, 1 bes, 1 strah, 1 tuga	6
[C] Mali tri model	3	2 neutralne	6
		1 radost, 1 bes	
		1 strah, 1 tuga	
[D] Veliki tri model	3	6 neutralnih	18
		3 radost, 3 bes	
		3 strah, 3 tuga	
[E] Veliki miks model	1	6 neutralnih, 3 radost, 3 bes, 3 strah, 3 tuga	18

BRG - broj govornika, BRM - broj modela po govorniku, BRR - broj rečenica za obuku jednog modela, UBRR - ukupan broj rečenica iskorišćenih za obuku.

5.1.3 Način evaluacije modela

U svim eksperimentima, obučeni modeli testirani su upotrebom četiri rečenice od svakog emotivnog stanja (uključujući i neutralno). To je ukupno 20 rečenica po govorniku. Rečenice koje su upotrebljene u evaluaciji modela različite su od rečenica upotrebljenih u obučavanju modela. Osim toga, skup rečenica koji je odvojen za evaluaciju isti je u svim eksperimentima. Na taj način postignuta je ekvivalencija među eksperimentima, radi lakešeg upoređivanja [2].

Tabela 5.4: Struktura skupa rečenica za testiranje. BRG je broj govornika.

Baza	BRG	neutralno	radost	bes	tuga	strah	ukupno
RUSLANA	61	4x61 = 244	4x61 = 244	4x61 = 244	4x61 = 244	4x61 = 244	1220
GEES	6	4x6=24	4x6=24	4x6=24	4x6=24	4x6=24	120
EMOVO	6	4x6=24	4x6=24	4x6=24	4x6=24	4x6=24	120
SAAVE	4	4x4=16	4x4=16	4x4=16	4x4=16	4x4=16	80

Testiranje je izvršeno tako što su MFCC koeficijenti izračunati za svaku od test rečenica. U slučaju *neutralnog* i *miks modela* test rečenice evaluirane su u odnosu na model govornika. Izlazne verovatnoće evaluacije modela sortirane su i model sa najvećom verovatnoćom proglašavan je kao model govornika koji je izgovorio zadatu test rečenicu. U slučaju tri modela, test rečenice evaluirali smo u odnosu na model koji odgovara emotivnom stanju test rečenice, sortirali i onda određivali koji je od govornika onaj koji je prepoznat. Specifična implementacija zavisi od konkretnog algoritma obuke, o čemu će biti reči u narednim poglavljima.

5.1.4 Mera uspešnosti prepoznavanja

Uspešnost predloženih modela u svim eksperimentima merena je procentom prepoznavanja (RR - recognition rate). Njegovo izračunavanje dato je formulom:

$$RR = \frac{N_{corr}}{N_{total}}, \tag{5.1}$$

gde je N_{corr} broj test rečenica za koje je govornik uspešno identifikovan, a N_{total} ukupan broj test rečenica.

5.2 Algoritmi klasifikacije i modeliranja govornika

5.2.1 Gaussove mešavine (GMM)

GMM predstavljaju osnovni algoritam za moderno prepoznavanje govornika, na osnovu koga su razvijene naprednije tehnike prepoznavanja govornika. Za evaluacije sistema sa emocijama u govoru uglavnom je korišćena osnovna implementacija ove tehnike. Upotrebu Gausovih mešavina za zadatak prepoznavanja govornika uveo je Reynolds [108]. Model Gausove mešavine je zbir Gausovih raspodela od kojih je svakoj dodeljen težinski koeficijent [109]. Zbir težinskih koeficijenata jednak je jedinici. Smisao svake od komponenata je da odgovara određenom fonetskom događaju kao što su samoglasnici, nazali, frikativi itd. [108]. Gausova raspodela data je jednačinom:

$$g(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T\right), \tag{5.2}$$

A jednačina Gausove mešavine data je sa:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} \alpha_i g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}), \tag{5.3}$$

gde je M broj Gausovih raspodela u mešavini, $\boldsymbol{x}$ je D-dimenzioni vektor karakteristika, α_i su težinski koeficijenti, $\boldsymbol{\mu_i}$ su vektori srednjih vrednosti svake od Gausovih komponenti, a $\boldsymbol{\Sigma_i}$ kovarijacione matrice. Pri tome važi da je:

$$\sum_{i=1}^{M} \alpha_i = 1. \tag{5.4}$$

Ovo praktično znači da je kompletan model poznat kada se odrede srednje vrednosti $\boldsymbol{\mu_i}$, kovarijacione matrice $\boldsymbol{\Sigma_i}$ i α_i težinski koeficijenti za svaku od M komponenti. Postupak izračunavanja ovih parametara dat je u nastavku. Korišćena iterativna procedura Expectation Maximisation (EM). Početni parametri inicijalizuju se postavljanjem vektora srednjih vrednosti i kovarijacionih matrica na nulu, dok su težinski koeficijenti postavljeni na istu vrednost:

$$\begin{aligned} \boldsymbol{\alpha_i} &= \frac{1}{M}, \\ \boldsymbol{\mu_i} &= \mathbf{0}, \\ \boldsymbol{\Sigma_i} &= \mathbf{0}. \end{aligned} \tag{5.5}$$

Unutar iterativne procedure parametri se ažuriraju na sledeći način:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}} &= \frac{1}{T}\sum_{t=1}^{T} Pr(i|\boldsymbol{x_t}, \lambda),\\ \hat{\boldsymbol{\mu}}_{\boldsymbol{i}} &= \frac{\sum_{t=1}^{T} Pr(i|\boldsymbol{x_t}, \lambda)\boldsymbol{x_t}}{\sum_{t=1}^{T} Pr(i|\boldsymbol{x_t}, \lambda)},\\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{i}} &= \frac{\sum_{t=1}^{T} Pr(i|\boldsymbol{x_t}, \lambda)\boldsymbol{x_t^2}}{\sum_{t=1}^{T} Pr(i|\boldsymbol{x_t}, \lambda)} - \boldsymbol{\mu_i^2},\end{aligned} \tag{5.6}$$

gde se *a posteriori* verovatnoća za komponentu i, modela λ izračunava kao:

$$Pr(i|\boldsymbol{x_t}, \lambda) = \frac{\alpha_i g(\boldsymbol{x_t}, |\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})}{\sum_{k=1}^{M} \alpha_k g(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}. \tag{5.7}$$

Obuka modela kada kriterijum postigne vrednost ispod određeni praga. Kriterijum se izračunava kao razlika sume logaritama verovatnoća modela za trenutnu i prethodnu iteraciju.

5.2.2 Univerzalni pozadinski model (UBM)

Univerzalni pozadinski model (Universal Background Model - UBM) [110] je obuka Gausove mešavine za prosečnog govornika na osnovu uzoraka rečenica za obuku svakog od govornika od interesa. Obično se sastoji od velikog broja mešavina - 512, 1024 ili više. Ovako dobijeni model zatim se adaptira na mali uzorak govora pojedinačnih govornika. Na ovaj način, kreirani pozadinski model je polazni model za dalju obuku - ovaj model ima dovoljno podataka da bi se obučio i predstavlja stabilan inicijalni model za pojedinačne modele. Osnovna primena mu je za verifikaciju govornika, a predstavlja i jedan od koraka u obuci i-vektora.

5.2.3 Združena analiza faktora i i-vektori

Upotreba i-vektora (identity vector) prvi put predstavili su Dehak i ostali [17]. Ovaj pristup nadograđuje ideju iza združene analize faktora [18] pretpostavljajući da supervektor $\boldsymbol{x}$ karakteristika koji zavisi od sesije i govornika može biti modelovan na sledeći način:

$$\boldsymbol{x} = \boldsymbol{q} + \boldsymbol{T}\boldsymbol{w}, \tag{5.8}$$

gde je $\boldsymbol{q}$ komponenta koja ne zavisi od govornika i kanala, $\boldsymbol{T}$ je matrica totalne varijabilnosti, a $\boldsymbol{w}$ i-vektor [17]. Modelovanje govornika na ovaj način vrši se kroz korake opisane u nastavku. Počinjemo modelovanje određivanjem univerzalnog pozadinskog modela (Universal Backgrund Model - UBM). Ovo je model ,,prosečnog govornika" koji se svodi na GMM koji se sastoji od K Gausovih funkcija, pri čemu je uobičajeno $K = 512$ ili 1024 ili 2048. Za obučavanje ovog modela koriste se pozadinski podaci, koji ne moraju biti od podataka koji su vezani za govornike od interesa. Nakon toga prelazi se na korak procene Baum-Welch-ove statistike. Ove statistike neophodne su za izračunavanje i-vektora za govornu sekvencu, a date su sa:

$$N_k = \sum_{t=1}^{T} P(k|\boldsymbol{y_t}, \Omega), \tag{5.9}$$

$$F_k = \sum_{t=1}^{T} P(k|\boldsymbol{y_t}, \Omega) \tag{5.10}$$

gde je $k = 1, 2, \cdots, K$ indeks Gausove raspodele u Gausovoj mešavini za UBM modela Ω, a $P(k|\boldsymbol{y_t}, \Omega)$ *posteriori* verovatnoća komponente k da generiše vektor $\boldsymbol{y_k}$. Za procenu i-vektora neophodno je da se izračuna Baum-Welchova statistika prvog reda:

$$F_k^b(\boldsymbol{y_t}) = \sum_{t=1}^{T} P(k|\boldsymbol{y_t}, \Omega)(\boldsymbol{y_t} - \boldsymbol{\mu_k}), \tag{5.11}$$

gde je $\boldsymbol{\mu_k}$ srednja vrednost k-te komponentete UBM modela. Sledeći korak je ekstrakcija i-vektora:

$$\boldsymbol{w} = (\boldsymbol{I} + \boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{N}(u)\boldsymbol{T})^{-1}\boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{F}^b(u), \tag{5.12}$$

gde je $\boldsymbol{N(u)}$ dijagonalna matrica dimenzija $KD \times KD$, čiji su blokovi $\boldsymbol{N_k}I$, $k = 1, 2, \cdot, K$. $\boldsymbol{F_b(u)}$ je supervektor dimenzija $KD \times 1$ konstruisan povezivanjem svih Baum-Welch-ovih statistika $F_k^b(\boldsymbol{y_t})$ za datu govornu sekvencu u. $\boldsymbol{\Sigma}$ je dijagonalna kovarijaciona matrica koja modeluje ostatak verovatnoće koji nije obuhvaćen modelom i matricom $\boldsymbol{T}$ [17]. Na kraju, vrši se redukcija dimenzija korišćenjem LDA [94] algoritma za dodatno smanjenje dimezija i-vektora.
Ideja za primenu i-vektora u uslovima emotivnog govora došla je od Chena i ostalih [32]. Emocije posmatraju ekvivalentno distorzijama koje se javljaju usled uticaja kanala prenosa glasa i varijabilnosti usled razlike u sesijama u kreiranju modela govornika. Modifikacije koje su predložili odnose se na kriterijum za prepoznavanje rečenice u fazi testa, kao i redukciju dimenzija u poslednjem koraku izračunavanja i-vektora. Kriterijum koji su iskoristili za određivanje udaljenosti test rečenice od modela govornika je:

$$\|w_{target}, w_{test}\| = \frac{w_{target}^t \cdot w_{test}}{\|w_{target}\|\|w_{test}\|} \tag{5.13}$$

Druga modifikacija odnosi se na primenu unutarklasne kovarijacione normalizacije, koja dolazi iz pristupa *jedan i svi* (5.4.2) za klasifikaciju u više klasa upotrebom mašina potpornih vektora. Ovaj pristup poredili su sa klasičnin LDA pristupom [32], standardnim GMM-UBM modelom i analizom faktora na osnovu emocija. Eksperimente su sprovodili na MASC bazi [111], tako da je govor 18 govornika iskorišćen kao razvojni uzorak, dok je za ostalih 50 govornika rađena obuka i kasnije test. Rezultati koje su dobili pokazali su da je ovaj pristup bolji u proseku u odnosu na klasičan LDA i GMM-UBM, međutim i dalje ne bolji od analize faktora na osnovu emocija.

Analiza faktora na osnovu emocija

Osim i-vektora kao takvih, Chen i ostali [32] razvili su ideju i-vektora u Analizu faktora na osnovu emocija (EFA - Emotional Factor Analysis). Supervektor $\boldsymbol{x}$ koji opisuje uzorak govora može se razložiti na dva supervektora - onaj koji zavisi od govornika $\boldsymbol{s}$ i onaj koji zavisi od emocije $\boldsymbol{e}$:

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{e}. \tag{5.14}$$

Supervektor govornika dalje se razlaže na supervektor nezavisan i od govornika i od emocije, na prostor govornika i na rezidualni prostor:

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{z}, \tag{5.15}$$

gde je $\boldsymbol{V}$ matrica sopstvenih vrednosti, $\boldsymbol{M}$ dijagonalna matrica koja označava rezidualni prostor, a $\boldsymbol{y}$ i $\boldsymbol{z}$ faktori govornika i rezidualnog prostora [32]. Raspodela supervektora koji zavisi od emocija opisana je jednačinom:

$$\boldsymbol{c} = \boldsymbol{U}\boldsymbol{\xi}, \tag{5.16}$$

gde je $\boldsymbol{U}$ sopstvena matrica za emocije, a $\boldsymbol{\xi}$ faktor za emocije. Primena analize faktora na osnovu emocija zahteva određivanje podprostora $\boldsymbol{U}, \boldsymbol{V}$ i $\boldsymbol{x}$ na obeleženoj razvojnoj bazi govora, a model govornika $(\boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{z})$ na zadatim trening rečenicama. Skor prepoznavanja izračunava se na osnovu verovatnoće test rečenice na modelu sa kompenzacijom sesije $(\boldsymbol{x} - \boldsymbol{U\xi})$. Estimacija podprostora sastoji se iz tri dela: estimacija matrice sopstvenih vrednosti, estimacije matrice za emocije i estimacije rezidualne matrice. Pri tome, procena svake od ovih matrica koristi algoritam maksimizacije očekivanja na sličan način.

5.3 Rezultati eksperimenata i diskusija

Pet eksperimenata opisanih u sekciji 5.1 sproveli smo sa Gausovim mešavinama (GMM) u osnovnoj implementaciji i sa i-vektorima.

5.3.1 Evaluacija test rečenica

GMM model govornika i klasifikacija

Odabrano je da se obuka vrši sa 30 mešavina za svaki pojedinačni model. To znači da je za *miks modele* kreiran jedan model po govorniku od 30 mešavina, a za *tri modele* po tri modela od po 30 mešavina po govorniku. Testiranje je izvršeno na način da su MFCC koeficijenti, izračunati iz test rečenica evaluirani u odnosu na model svakog govornika. Skor je za jednu rečenicu izračunavat kao suma logaritama vrednosti Gausove mešavine zadatog govornika za svaki od vektora date test rečenice:

$$S_g(u_{test}) = \sum_{k=1}^{N_{test}} \log\left(\sum_{i=1}^{M} \alpha_{g,i} g(\boldsymbol{x}_k|\boldsymbol{\mu_{g,i}}, \boldsymbol{\Sigma_{g,i}}),\right), \tag{5.17}$$

gde je u_{test} test rečenica za koju određujemo ko ju je izgovorio, N_{test} broj MFCC vektora izdvojenih iz te rečenice, a g oznaka za govornika za čiji model određujemo skor $S_g(u_{test})$ za u_{test}. Govornik čiji je model dao **najveći skor** proglašavan je prepoznatim govornikom. Test rečenica je ukupno bilo 20 po govorniku, i te rečenice nisu bile korišćene u obuci modela (Više u Sekciji 5.1). U slučaju kada je model testiran u odnosu na određenu emociju, za test su iskorišćene samo rečenice zadate emocije iz ukupnog test skupa tj četiri po govorniku (Tabela 5.4).

i-vektor model govornika i PLDA klasifikacija

Iskoristili smo MSR toolbox [112] i voicebox [113] alate za sprovođenje i evaluaciju opisanog algoritma i-vektora. Korišćene karakteristike bile su kao i u slučaju Gausovih mešavina 13 MFCC koeficijenata. GMM-UBM model sa 32 Gausove mešavine na osnovu razvojnih podataka, koji su se sastojali od 10000 rečenica (200 rečenica $\times$ 50 govornika) iz TEVOID baze [106].

Eksperiment opisan u Sekciji 5.1 ponovili smo sa i-vektorima kao klasifikatorom i uz upotrebu LDA za radukciju dimenzija. Rezultati su prikazani u Tabelama 5.5 - 5.9.

U odnosu na prethodno istraživanje samo na ruskoj bazi [2], napravljena je modifikacija u odabriu dimenzije vektora i finalne dimenzije, kako bi bilo tehnički moguće da se sprovedu eksperimenti sa malim brojem govornika i rečenica za obuku.

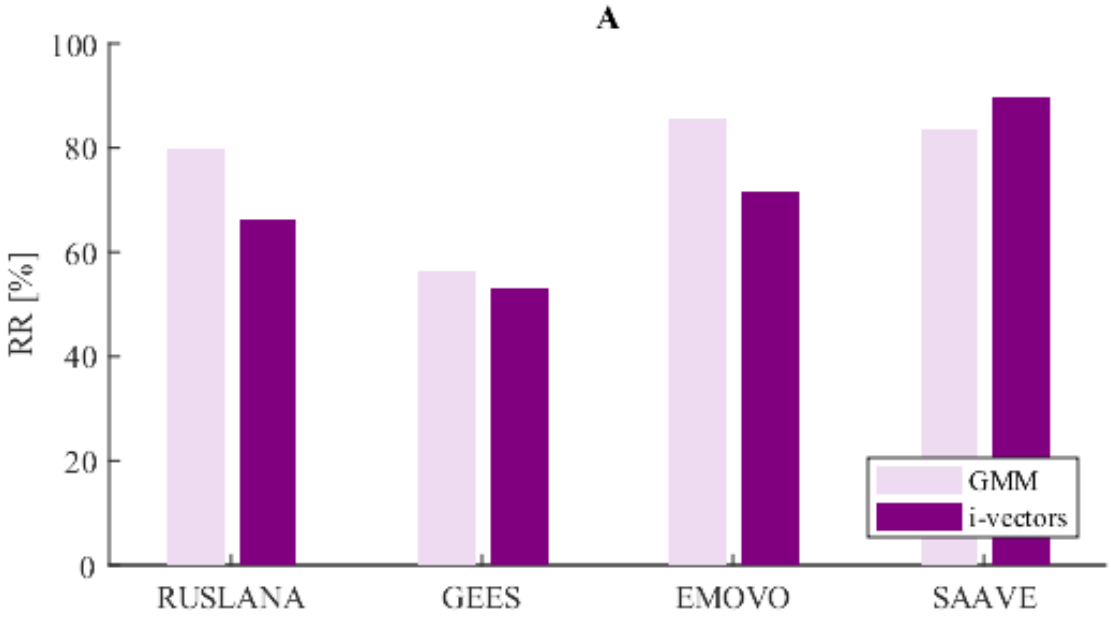
5.3.2 Rezultati eksperimenata

[A] Neutralni model

Procenat prepoznavanja za GMM neutralnog govora testiran sa neutralnim govorom bio je skoro 100.0%. Testiranje ovog modela sa svim rečenicama daje rezultat je od 56.67% za GEES bazu do 83.75% za SAAVE bazu. U slučaju modela i-vektora rezultati testiranja sa neutralnim govorom su između 93.85%, koliko je za RUSLANA bazu i 100.0%, koliko je za SAAVE bazu. Upotrebom svih rečenica za testiranje ovog modela dobijaju se rezultati prepoznavanja između 53.33%, koliko je za GEES bazu i 90.00%, što je rezultat u slučaju SAAVE baze. Kompletni rezultati prikazani su u Tabeli 5.5, a rezultati testiranja svim rečenicama prikazani su i grafički na Slici 5.2.

Tabela 5.5: Rezultati testiranja *Neutralnog modela*.

[A] *Neutralni model*						
Baza	Neutralno	Radost	Bes	Tuga	Strah	Sve
GMM						
RUSLANA	**99.59**	71.73	67.21	88.52	74.18	**80.16**
GEES	**100.0**	45.83	45.83	33.33	58.33	**56.67**
EMOVO	**100.0**	95.83	62.50	95.83	75.00	**85.83**
SAVEE	**100.0**	62.50	75.00	93.75	87.50	**83.75**
i-vektori						
RUSLANA	**93.85**	52.87	52.87	77.05	54.51	**66.48**
GEES	**95.83**	29.17	16.67	62.50	62.50	**53.33**
EMOVO	**95.83**	79.17	37.50	79.17	66.67	**71.67**
SAVEE	**100.0**	81.25	75.00	100.0	93.75	**90.00**



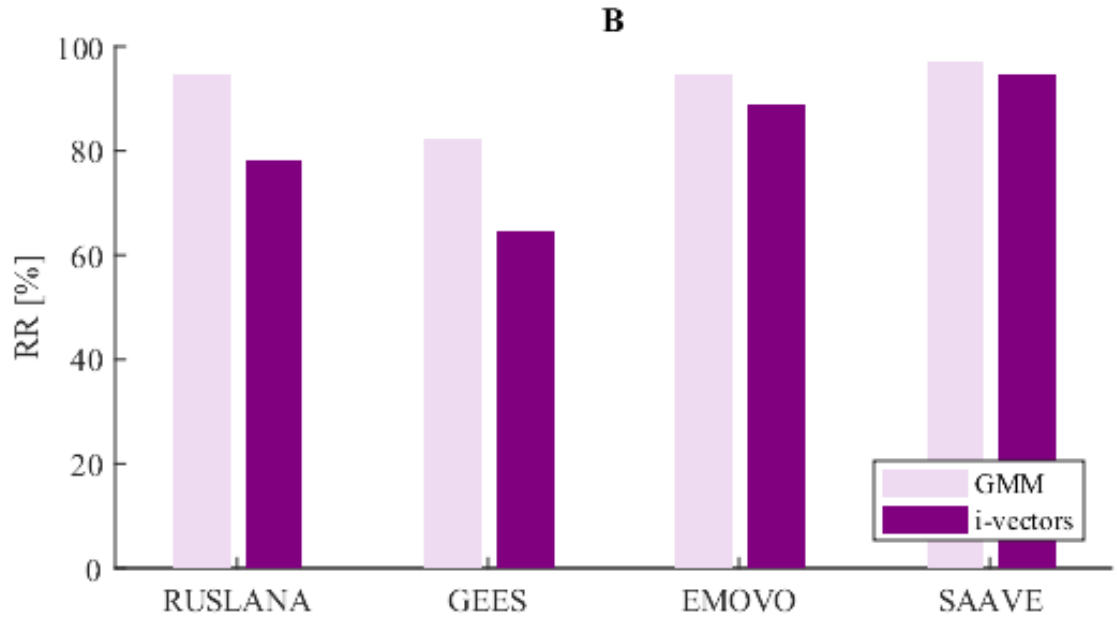
Slika 5.2: Rezultati *Neutralnog modela* za GMM i i-vektore za sve emocije.

[B] Mali miks model

Testiranje malog miks modela GMM sa neutralnim govorom dalo je procenat prepoznavanja 97.54% – 100.0%. Ovaj model testiran sa svim test rečenicama dao je rezultat između 82.50% i 97.50%. Eksperiment sproveden sa i-vektorima kao modelom dao je prepoznavanje 82.38% – 100.0% kada je za test korišćen samo neutralni govor, odnosno 65.00% – 95.00% za ceo test skup. Kompletni rezultati prikazani su u Tabeli 5.6, a rezultati testiranja svim rečenicama prikazani su i grafički na Slici 5.3.

Tabela 5.6: Rezultati testiranja *Malog miks modela*.

[B] *Mali miks model*						
Baza	Neutralno	Radost	Bes	Tuga	Strah	Sve
GMM						
RUSLANA	**97.54**	91.80	93.03	95.90	95.90	**94.83**
GEES	**100.0**	83.33	66.67	75.00	95.83	**82.50**
EMOVO	**100.0**	100.0	87.50	100.0	87.50	**95.00**
SAVEE	**100.0**	93.75	93.75	100.0	100.0	**97.50**
i-vektori						
RUSLANA	**82.38**	77.87	77.87	81.15	72.95	**78.20**
GEES	**87.50**	54.17	45.83	62.50	75.00	**65.00**
EMOVO	**100.0**	95.83	66.67	91.67	91.67	**89.17**
SAVEE	**100.0**	87.50	100.0	93.75	93.75	**95.00**



Slika 5.3: Rezultati *Malog miks modela* za GMM i i-vektore za sve emocije.

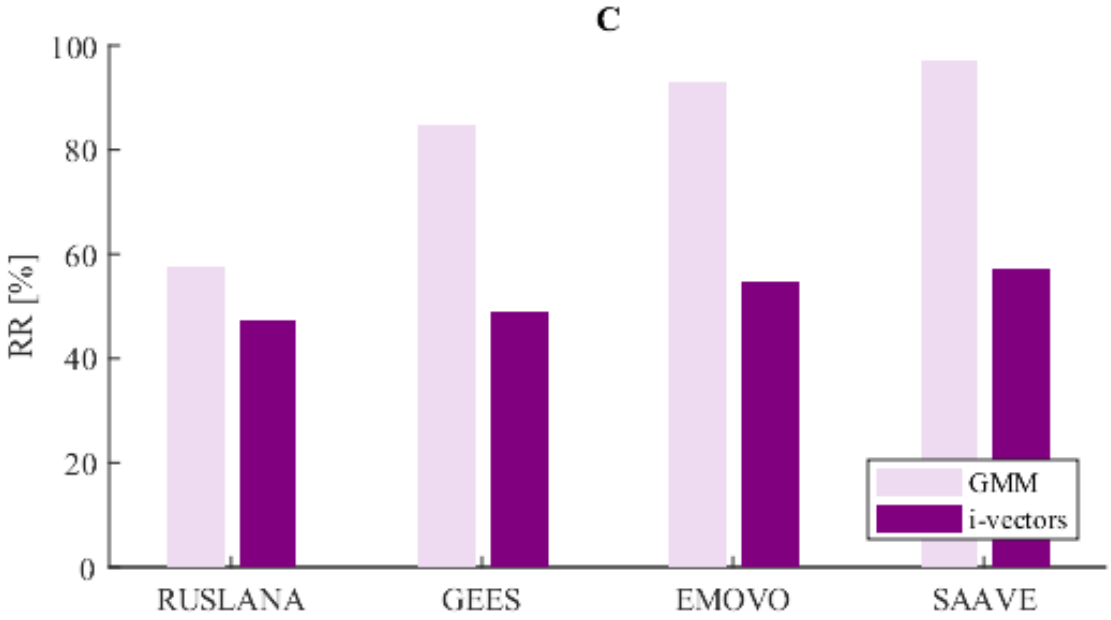
[C] Mali tri model

Mali tri model sa GMM metodom modeliranja govornika u slučaju testiranja neutralnim govorom daje tačnost prepoznavanja od 80.33% – 100.0%, dok je za testiranje svim rečenicama taj rezultati između 57.95% i 97.50%. Modeliranje i-vektorima u ovom eksperimentu daje uspešnost prepoznavanja 62.50% – 71.72% za test sa neutralnim govorom i 47.54% – 57.50%

kada se za testiranje iskoriste sve rečenice iz test skupa. Kompletni rezultati prikazani su u Tabeli 5.7, a rezultati testiranja svim rečenicama prikazani su i grafički na Slici 5.4.

Tabela 5.7: Rezultati testiranja *Malog tri modela*.

[C] *Mali tri model*						
Baza	Neutralno	Radost	Bes	Tuga	Strah	Sve
GMM						
RUSLANA	**80.33**	45.90	52.46	48.77	62.3	**57.95**
GEES	**100.0**	83.33	70.83	91.67	79.17	**85.00**
EMOVO	**100.0**	95.83	79.17	100.0	91.67	**93.33**
SAVEE	**100.0**	93.75	93.75	100.0	100.0	**97.50**
i-vektori						
RUSLANA	**71.72**	38.11	43.85	41.80	42.21	**47.54**
GEES	**66.67**	41.67	33.33	50.00	54.17	**49.17**
EMOVO	**66.67**	70.83	37.50	45.83	54.17	**55.00**
SAVEE	**62.50**	81.25	50.00	43.75	50.00	**57.50**



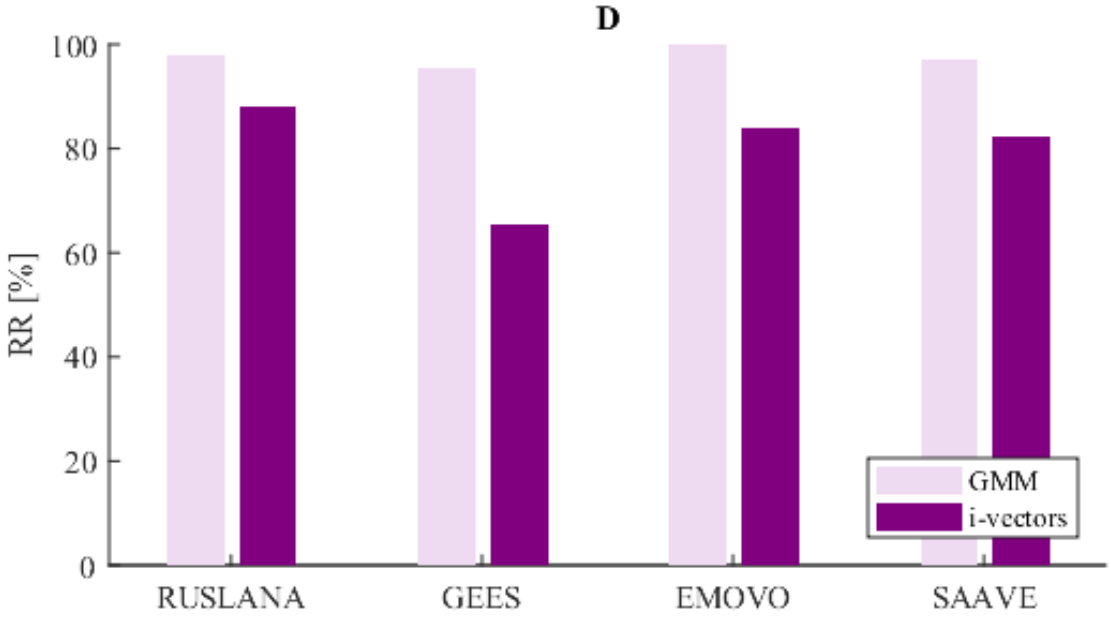
Slika 5.4: Rezultati *Malog tri modela* za GMM i i-vektore za sve emocije.

[D] Veliki tri model

Modeliranjem GMM u slučaju eksperimenta sa *Velikim tri modelom* postignut je procenat prepoznavanja od gotovo 100% u testu sa neutralnim govorom i 99.59% – 100%. Modeliranje i-vektorima rezultovalo je prepoznavanjem u opsegu od 91.67% do 94.67% kada je testirano neutralnim govorom i od 65.83% – 88.28% kada je testirano sa celim test skupom.
Kompletni rezultati prikazani su u Tabeli 5.8, a rezultati testiranja svim rečenicama prikazani su i grafički na Slici 5.5.

Tabela 5.8: Rezultati testiranja *Velikog tri modela*.

[D] *Veliki tri model*						
Baza	Neutralno	Radost	Bes	Tuga	Strah	Sve
GMM						
RUSLANA	**99.59**	97.54	98.77	95.49	99.18	**98.11**
GEES	**100.0**	100.0	87.5	95.83	95.83	**95.83**
EMOVO	**100.0**	100.0	100.0	100.0	100.0	**100.0**
SAVEE	**100.0**	93.75	93.75	100.0	100.0	**97.50**
i-vektori						
RUSLANA	**94.67**	84.02	86.48	86.89	89.34	**88.28**
GEES	**91.67**	58.33	45.83	66.67	66.67	**65.83**
EMOVO	**91.67**	79.17	75.00	83.33	91.67	**84.17**
SAVEE	**93.75**	68.75	62.50	93.75	93.75	**82.50**



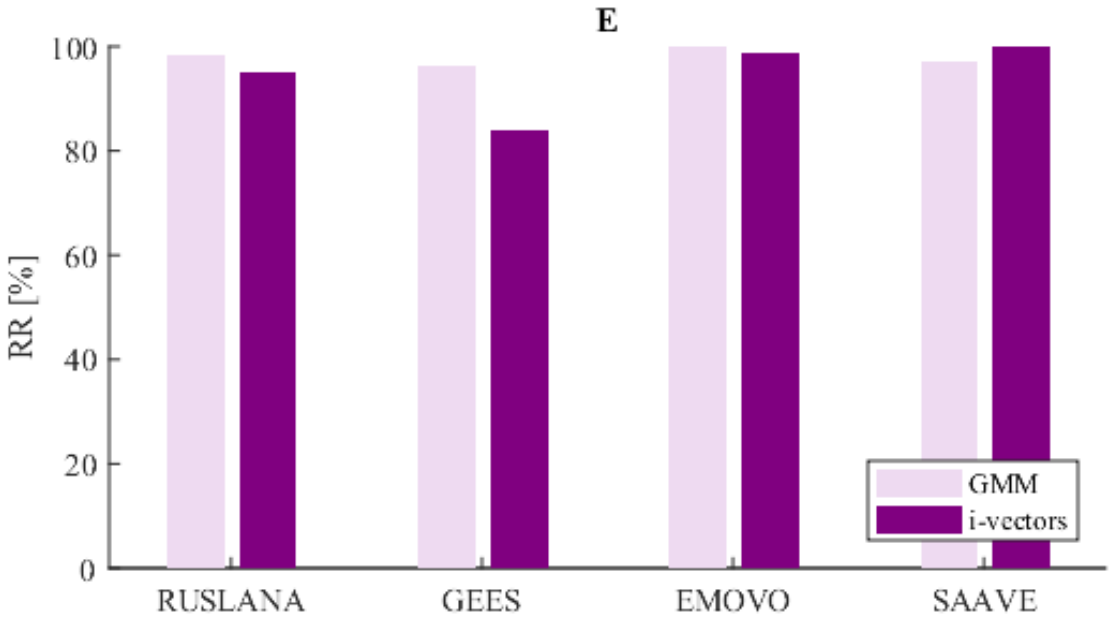
Slika 5.5: Rezultati *Velikog tri modela* za GMM i i-vektore za sve emocije.

[E] Veliki miks model

Poslednje sprovedeni eksperiment, kada je GMM iskorišćen kao model govornika, postigao je gotovo apsolutnu tačnost kada je testiran sa neutralnim govorom, a između 96.67% i 100.0% kada je testiran sa ukupnim test govorom. U slučaju i-vektora, rezultati su bili sledeći 96.72% – 100.0% za test neutralnim govorom i 84.17% – 100.0% za test sa svim rečenicama. Kompletni rezultati prikazani su u Tabeli 5.9, a rezultati testiranja svim rečenicama prikazani su i grafički na Slici 5.6.

Tabela 5.9: Rezultati testiranja *Velikog miks modela.*

[E] *Veliki miks model*						
Baza	Neutralno	Radost	Bes	Tuga	Strah	Sve
GMM						
RUSLANA	**99.59**	97.54	98.36	98.77	98.77	**98.61**
GEES	**100.0**	95.83	91.67	91.67	100.0	**96.67**
EMOVO	**100.0**	100.0	100.0	100.0	100.0	**100.0**
SAVEE	**100.0**	93.75	93.75	100.0	100.0	**97.50**
i-vektori						
RUSLANA	**96.72**	95.08	95.08	96.31	93.03	**95.41**
GEES	**100.0**	83.33	79.17	79.17	79.17	**84.17**
EMOVO	**100.0**	100.0	100.0	95.83	100.0	**99.17**
SAVEE	**100.0**	100.0	100.0	100.0	100.0	**100.0**



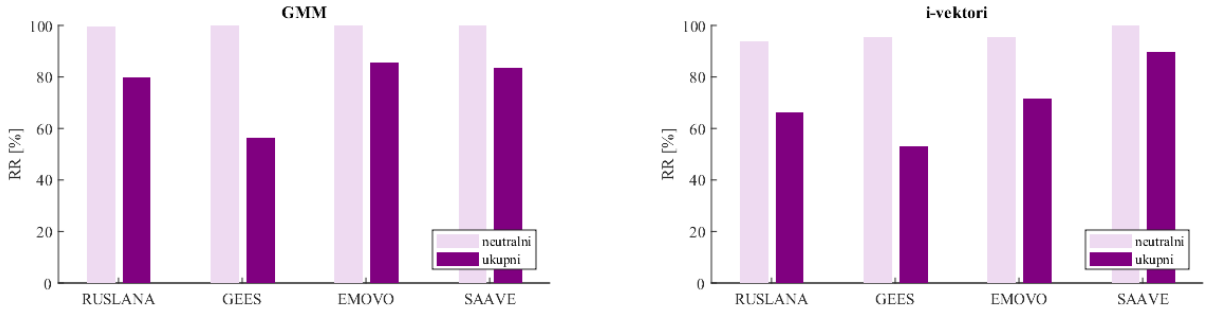
Slika 5.6: Rezultati *Velikog miks modela* za GMM i i-vektore za sve emocije.

5.3.3 Analiza uspešnosti i robusnosti modela

Dobijeni rezultati eksperimenata A-E (Tabele 5.5 - 5.9) pokazali su da se uspešnost prepoznavanja razlikuje pri poređenju testiranja neutralnim i emotivnim govorom. Očekivano, svi modeli su uspešniji u prepoznavanju govornika iz neutralnog govora u odnosu na test uzorke emotivnog govora. U slučaju korišćenja GMM kao tehnike modeliranja i prepoznavanje govornika, modeli govora *Neutralni, Mali miks, Veliki tri* i *Veliki miks model* testirani neutralnim govorom daju skoro idealno prepoznavanje od 97.54% – 100.0%. Jedino odstupanje postoji kod *Malog tri modela* i to u slučaju RUSLANA baze - prepoznavanje je oko 80.33%, dok je za ostale baze i za ovaj model prepoznavanje 100%. Tehnika i-vektora daje nešto slabije rezultate, međutim i dalje testiranje neutralnim govorom ima značajno bolje rezultate nego u slučaju kada se testiranje vrši emotivnim govorom.

Neutralni model govornika, sa upotrebom i-vektora za klasifikaciju, testiran neutralnim govorom ostvaruje rezultate između 93.85% – 100.0% u testu neutralnim govorom i između 53.33% i 90.00% u testovima sa emotivnim govorom (Slika 5.7). Rezultati za *Mali miks model* sa ovo tehnikom sa testom neutralnog govora su od 82.38.00% za bazu RUSLANA, a potpuna tačnost

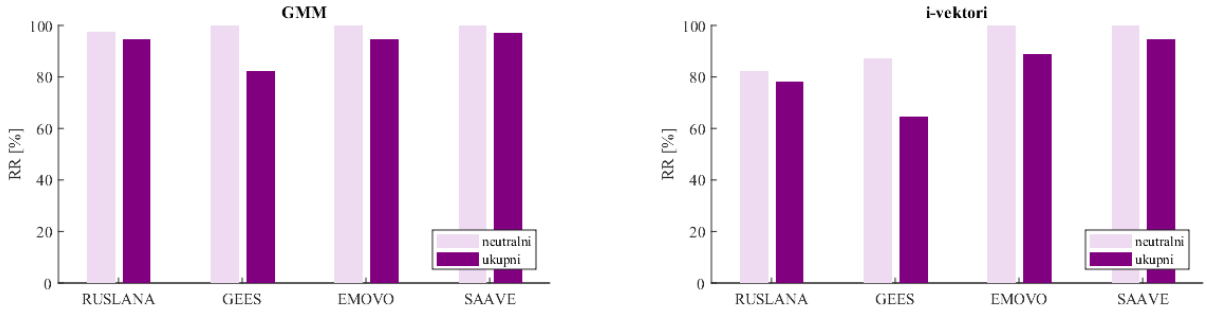
postiže se za EMOVO i SAAVE bazu. Kada se testiranje vrši svim test rečenica rezultat je 65.00% − 95.00%. *Mali tri model*, čija je uspešnost za testiranje neutralnim govorom ispod 80%, a emotivnim ispod 60%, najmanje je uspešna konfiguracija za i-vektore.



[A] GMM [B] i-vektori

Slika 5.7: Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Neutralni model* govornika.

Mali miks model posebno je interesantan da se po uspešnosti uporedi sa *Neutralnim modelom*. Ova dva modela su iste strukture i obučavaju se istom količinom podataka, stim da je za obuku *Malog miks modela* jedan deo rečenica neutralnog govora zamenjen emotivnim govorom. Uspešnost prepoznavanja govornika na osnovu neutralnog govora za ovaj model je neznatno slabija u odnosu na *Neutralni model* na svim bazama podataka. Sa druge strane, u testu sa celokupnim test skupom *Mali miks model* postiže poboljšanje između 5.00% za SAAVE bazu i i-vektore do 25.83% za GEES bazu i GMM.

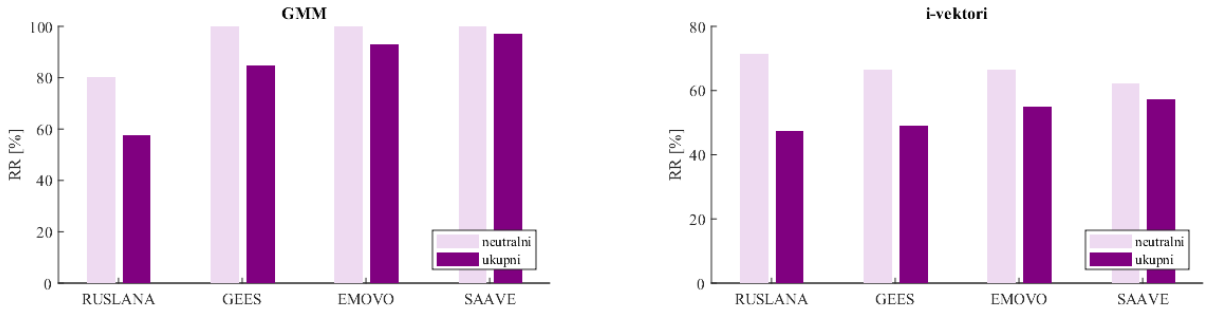


[A] *Mali miks model* - GMM [B] *Mali miks model* - i-vektori

Slika 5.8: Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Mali miks model* govornika.

Mali tri model koristi iste obučavajuće podatke kao i *Mali miks model*, stim da se umesto jedne mešavine tj modela i-vektora, obučavaju tri modela/klasifikatora. Kao i ostali modeli, i ovaj model sa većim uspehom prepoznaje neutralni govor u odnosu na emotivni govor. Kada se GMM koristi za modeliranje govornika, ovaj model, kao i *Neutralni model* daje 100% uspešnost za sve baze osim za RUSLANA za koju je manje uspešan od *Neutralnog modela*. Rezultat testiranja na svim test rečenicama ima slično ponašanje - *Mali tri model* uspešniji je u prepoznavanju u odnosu na *Neutralni model* za sve baze osim za RUSLANA, za koju se performansa značajno degradira. U slučaju i-vektora kao klasifikatora, *Mali tri model* pokazuje lošije performanse u odnosu na *Neutralni model* i kada je testiranje izvršeno sa neutralnim i sa emotivnim govorom.

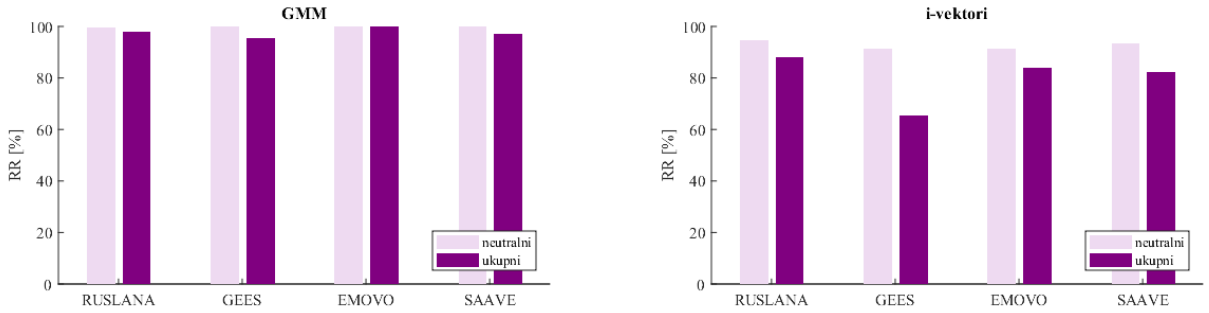
Ovakvo ponašanje može se razumeti u kontekstu obuke pojedinačnog modela. Za obuku svakog od tri modela/klasifikatora koji sačinjavaju *Mali tri model* govornika koristi se svega dve rečenice. Iako se koriste isti podaci kao i za obuku *Malog miks modela*, po pojedinačnom modelu ima tri puta manje podataka, što očigledno nije dovoljno za kvalitetnu obuku.



[A] *Mali tri model* - GMM [B] *Mali tri model* - i-vektori

Slika 5.9: Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Mali tri model* govornika.

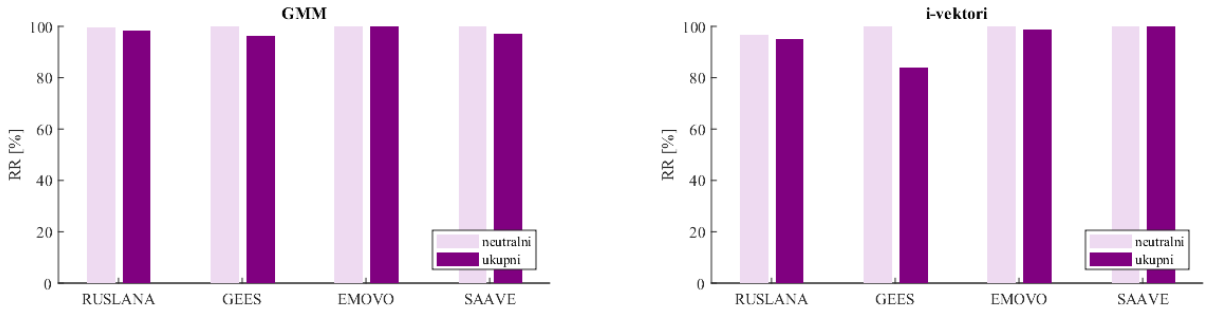
Rezultati koji ovo potvrđuju pokazani su u eksperimentima sa *Velikim tri modelom*. Za obuku ovog modela upotrebljeno je ukupno šest rečenica po modelu/klasifikatoru kojih je ukupno tri za svakog govornika. Rezultati pokazuju značajno poboljšanje u odnosu na *Mali miks model*. Rezultat postignut na neutralnom govoru identičan je kao i za *Neutralni model*, isti je model u pitanju, a prepoznavanje emotivnog govora je značajno bolje sa procentom prepoznavanja za različite baze između 96.67% i 100.0%. U slučaju i-vektora, poboljšanje postoji i u prepoznavanju neutralnog govora i u prepoznavanju emotivnog govora u odnosu na *Mali tri model*, međutim, *Mali miks model* daje bolje rezultate u slučaju EMOVO i SAAVE baze.



[A] *Veliki tri model* - GMM [B] *Veliki tri model* - i-vektori

Slika 5.10: Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Veliki tri model* govornika.

Poslednji eksperiment sa *Velikim miks modelom* dao je i najbolje rezultate i u slučaju korišćenja GMM i u slučaju korišćenja i-vektora. Ovaj eksperiment pokazao je da je korišćenje svih raspoloživih podataka za obuku jednog modela značajno bolje u odnosu na deljenje na *neutralni model*, *model intenzivnih emocija* i *model blagih emocija*.



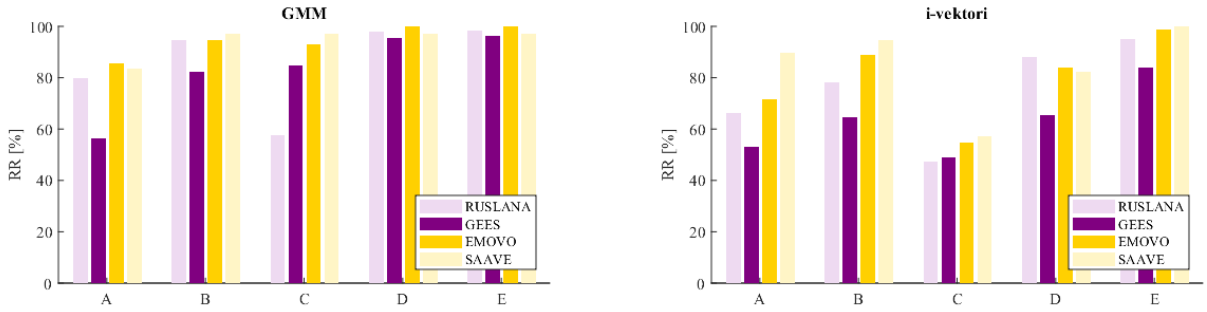
[A] *Veliki miks model* - GMM

[B] *Veliki miks model* - i-vektori

Slika 5.11: Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Veliki miks model* govornika.

5.3.4 Poređenje algoritma modeliranja/klasifikacije govornika

Interesantno je videti uporedne rezultate za tehnike GMM i i-vektora u različitim eksperimentima. Rezultati testiranja svim test rečenicama, dobijeni modeliranjem govornika sa GMM za svih pet eksperimenata A-E, na sve četiri baze prikazan je i grafički na Slici 5.12[A], dok su za i-vektore rezultati prikazani na Slici 5.12[B].



[A]

[B]

Slika 5.12: Rezultati testiranja po eksperimentima za različite baze za [A] GMM i [B] i-vektore.

Bez obzira što su i-vektori današnji standard za prepoznavanje govornika, dobijeni rezultati pokazali su da je GMM pogodnija tehnika za modelovanje od i-vektora u slučaju baza sa relativno malim brojem govornika. Malim brojem govornika smatramo broj manji od 100 govornika. Ovo zapažanje poklapa se sa rezultatima istraživanja Nayana i ostali [114].
GMM pokazao je bolje rezultate u svim eksperimentima na bazama RUSLANA, GEES i EMOVO, i u eksperimentima C i D na SAAVE bazi.

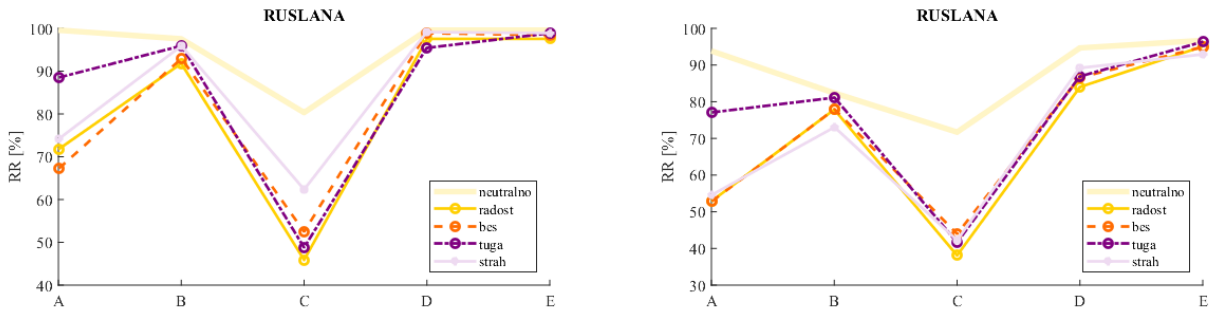
5.3.5 Analiza rezultata na osnovu korišćene baze

Interesantno je uporediti rezultate postignute na bazama EMOVO i GEES koje obe imaju po šest govornika - U svim eksperimentima i za GMM i za i-vektore rezultati su bolji na italijanskoj bazi. Rezultate možemo obrazložiti samim subjektivnim doživljajem glasova govornika u jednoj i drugoj bazi - na slušanje se značajnije razlikuju italijanski govornici. To naročito važi kada je emotivni govor u pitanju - može se reći da su emocije čak i prenaglašene kod ovih govornika.

Najbolji rezultat za tehniku i-vektora postignut je, očekivano, na SAAVE bazi u svim eksperimentima osim u eksperimentu D sa *Velikim tri modelom* - u ovom ekperimentu najbolji rezultat postignut je na bazi RUSLANA koja je ujedno i najveća. Ovu naizgled neobičnu pojavu možemo protumačiti u kontekstu konstrukcije modela i samih i-vektora. Algoritam se pokazao stabilniji u obuci na većem skupu podataka za pojedinačni model. Naime u obuci modela i-vektora koriste se podaci za obuku svih govornika. U tom smislu, za RUSLANU u obuci jednog od tri modela ima najviše podataka jer ima i najviše govornika. Ipak, ovi rezultati su ispod rezultata dobijenih za eksperiment E i *Veliki miks model*.

5.3.6 Analiza uticaja emotivnog stanja

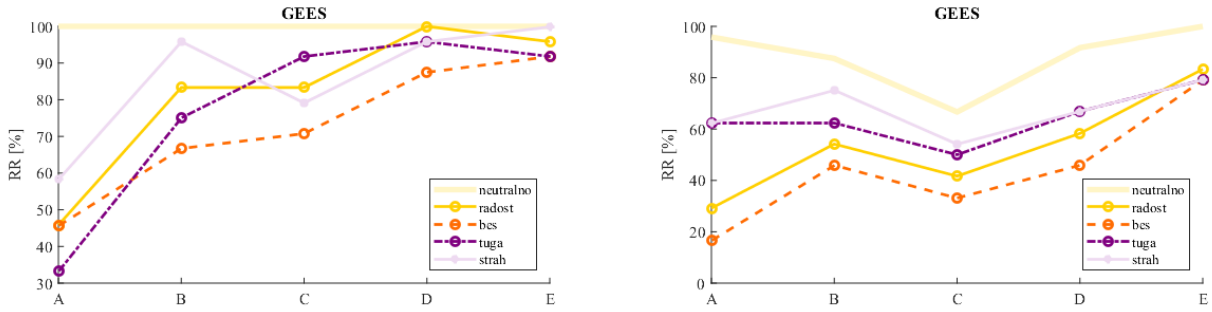
Važan aspekt analize rezultata je i uspešnost prepoznavanja svake od pojedinačnih emocija. Uticaj emocija razlikuje se po eksperimentima, tehnikama i bazama.
Za bazu RUSLANA (Slika 5.13), u eksperimentima sa GMM, rezultati za sve emocije su praktično isti kada su u pitanju eksperimenti B, D i E. Rezultati eksperimenata A i C za ovu bazu pokazuju da se govornici najlakše prepoznaju kada je njihov govor neutralan. U eksperimentu A, posle neutralnog govora najmanji uticaj ima tuga, dok sve ostale emocije imaju sličan uticaj. U eksperimentu C, osim neutralnog stanja, strah nešto manje utiče na prepoznavanje od svih ostalih emocija. Kada je u pitanju tehnika i-vektora, važe slična opažanja, stim da je u ekperimentu C strah sada zajedno sa ostalim emocijama.



[A] GMM [B] i-vektori

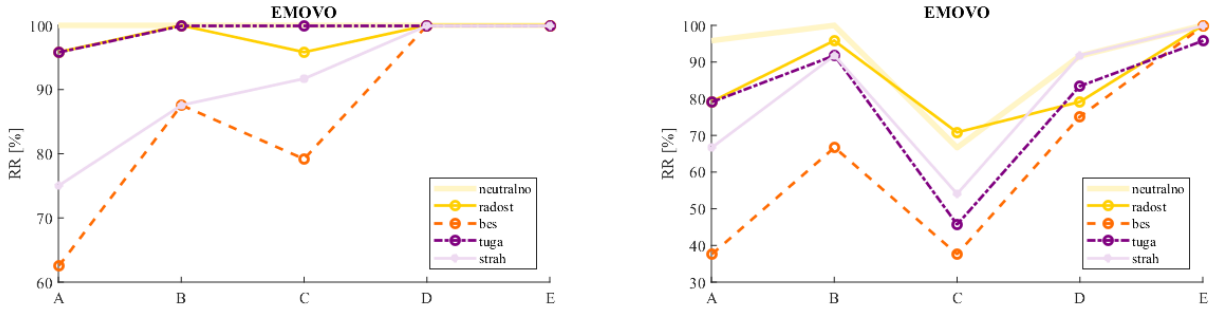
Slika 5.13: Rezultati po emocijama za bazu RUSLANA.

Eksperimenti sprovedeni na GEES bazi potvrđuju da se govornici koji govore neutralnim govorom najlakše prepoznaju i kada se koristi GMM i kada se koriste i-vektori (Slika 5.14). *Neutralni model* govornika sa GMM najmanje je pogođen emocijom straha, zatim radosti i besa, a najviše emocijom tuge. Takvo ponašanje razlikuje se od i-vektora gde tuga i strah omogućavaju podjednaku uspešnost prepoznavanja, a bolje od radosti i besa. Osim ovog modela, ostali eksperimenti sa GMM i svi eksperimenti sa i-vektorima najteže savladavaju govor besa. Za i-vektore odmah posle besa je i radost - dakle intenzivne emocije, dok blage emocije tuge i straha imaju manji uticaj na prepoznavanje govornika. U eksperimentima sa GMM, u zavisnosti od eksperimenta, emocije imaju različit uticaj.
Posmatranjem rezultata na EMOVO bazi (Slika 5.15), zaključujemo da je takođe bes emocija od najvećeg negativnog uticaja na uspešnost prepoznavanja. U slučaju GMM, prati ga strah. U slučaju i-vektora, osim besa koji se izdvaja, ostale emocije imaju manji uticaj koji se razlikuje od eksperimenta do eksperimenta. Zanimljivo je primetiti da su *Veliki tri model* i *Veliki miks model* GMM u potpunosti savladali sve emocije. Slična je situcija kod i-vektora za *Veliki miks model*, dok je *Veliki tri model* sa nešto manjim uspehom.



[A] GMM [B] i-vektori

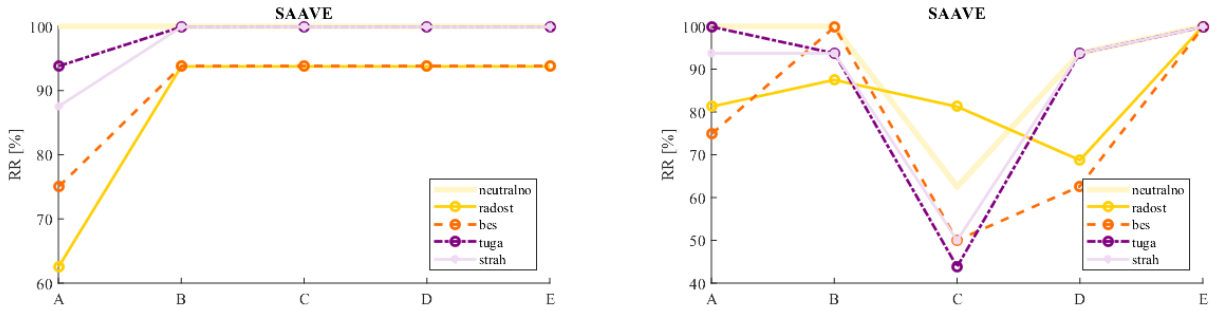
Slika 5.14: Rezultati po bazama, emocijama i tehnikama.



[A] GMM [B] i-vektori

Slika 5.15: Rezultati po bazama, emocijama i tehnikama.

Poslednja u nizu je SAAVE baza koja ima samo četiri muška govornika (Slika 5.16). U eksperimentima A-E sa GMM, dobro prepoznavanje govornika ostvaruje se, osim za neutralni govor i za tugu i strah, dok značajniju degradaciju performansi proizvode bes i radost. Kada je u pitanju tehnika i-vektoraRezultati su raznoliki. Može se primetiti da se *Mali miks model* i *Veliki miks model* sa uspehom izlaze na kraj sa svim emocijama.



[A] GMM [B] i-vektori

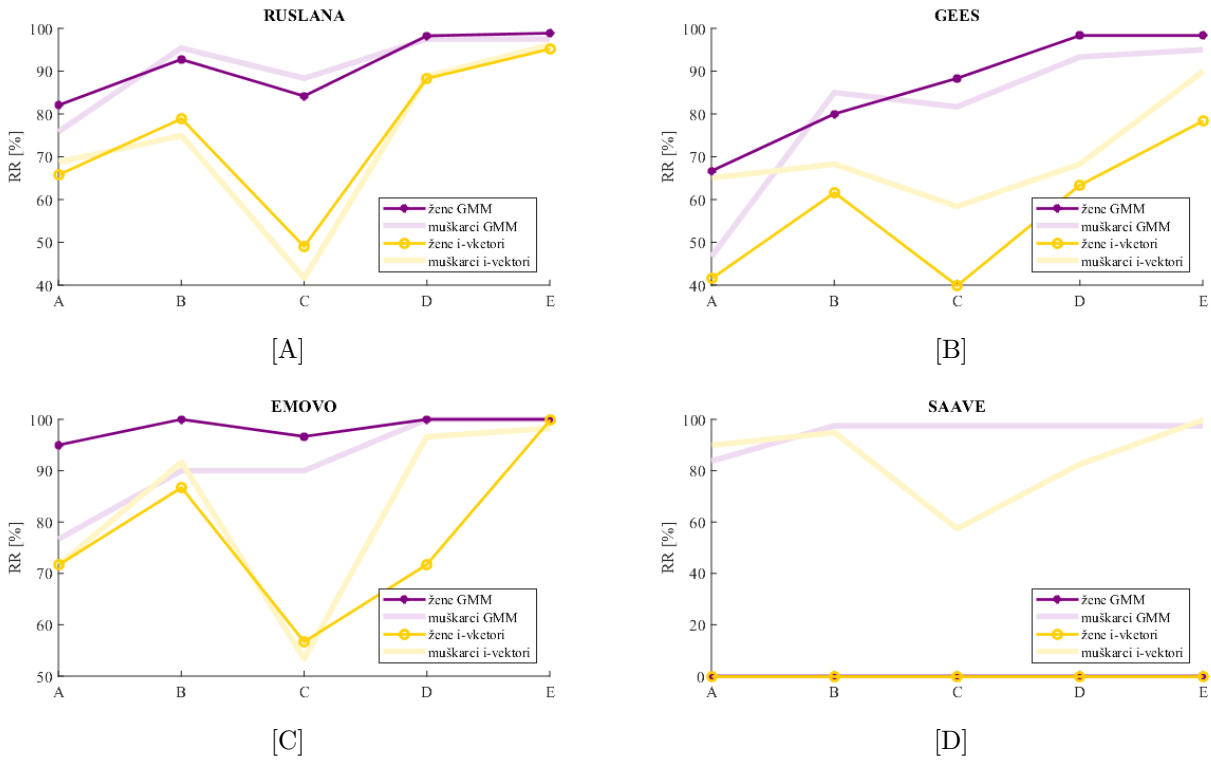
Slika 5.16: Rezultati po bazama, emocijama i tehnikama.

Sveukupno, možemo zaključiti da se govornici najlakše prepoznaju kada su u neutralnom emotivnom stanju, dok najteži zadatak sistemu za prepoznavanje govornika zadaje bes.
Emotivna stanja menjaju raspored energije spektra govora, a bes i radost pomeraju energiju

spektra govora ka višim učestanostima [115]. Prema evaluaciji energije za GEES bazu, bes ima naveću maksimalnu energiju govora, kao i maksimalnu standardnu devijaciju [102].

5.3.7 Analiza uticaja pola

Rezultate smo prikazali i po uspešnosti prepoznavanja za muške i ženske govornike (Slika 5.17). U eksperimentu sa *Neutralnim modelom*, GMM klasifikacija lakše prepoznaje žene nego muškarce, a najveća razlika postoji kod EMOVO baze. Klasifikacija i-vektorima za RUSLANA i EMOVO bazu daje praktično iste rezultate, dok je za GEES bazu slučaj da se muškarci značajno lakše prepoznaju nego žene.
Uspešnost prepoznavanja govornika i govornica na RUSLANA bazi daje slične rezultate i za GMM i za i-vektore kao klasifikator, stim da je u nekim eksperimentima bolje prepoznavanje muškaraca, a u drugima žena. Zanimljivo je ovu informaciju postaviti u kontekst broja jedne i druge vrste govornika - naime muških govornika je svega 12 naspram 49 koliko je ženskih. Manji broj omogućava da se muški govornici izdvoje kao grupa, a i samo razlikovanje da je manje zahtevan problem. To nas može dovesti do zaključka da je jednostavnije raspoznavanje ženskih glasova. Kada su u pitanju GMM i EMOVO baze, GMM klasifikator bolje prepoznaje ženske govornike. Za i-vektore na GEES bazi lakši zadatak predstavljaju muškarci. Na EMOVO bazi je slična situacija, stim da je razlika značajna jedino u eksperimentu D. SAAVE baza ima samo muške govornike, pa ova vrsta analize nije primenljiva, rezultate smo prikazali zbog kompletnosti.



Slika 5.17: Rezultati po bazama i polu.

5.3.8 Diskusija

U dosadašnjim istraživanjima prepoznavanja govornika u uslovima emotivnog govora, GMM je najčešće korišćen model. Ghiurcau i ostali [23] obučavali su model Gausovih mešavina bi-

rajući nasumično rečenice za obuku iz Berlin baze [116]. Sproveli su ukupno četiri eksperimenta da ispitaju uticaj emocija na kvalitet modela i uticaj emocija. **Prvi eksperiment** podrazumeva devet rečenica za obuku modela govornika i jednu rečenicu za test, pri čemu su sve rečenice u okviru iste emocije. **Drugi eksperiment** podrazumevao je obuku modela neutralnim rečenicama, a testiranje rečenicama različitih emocija. U okviru ovog eksperimenta varirali su broj MFCC i GMM, najbolji rezultat ostvarili su sa 50 GMM, međutim nije prelazio 57%. **Treći eksperiment** suštinski je isti kao i drugi eksperiment, stim da je zasebno realizovano testiranje za svaku od emocija pojedinačno. **Četvrti eksperiment** sproveden je sa ciljem poboljšanja ovako dobijenih rezultata, te su za obuku modela, ovaj put, upotrebili po 5-6 nasumično odabranih rečenica iz svakog emotivnog stanja po govorniku. Ukupno, za obuku modela iskoristili su 34 rečenice dok je testiranje izvršeno sa jedno rečenicom. Postupak je ponovljen 35 puta. Rezultati dobijeni u ovom eksperimentu dostižu 98.57% Autori [23] su zaključili da je poboljšanje dobijeno u ovom slučaju značajno, ali nedovoljno i da je potrebno nastaviti istraživanja u ovom smeru.
Modelovanje na osnovu emocija koje su predložili Wu i Yang [26] izvršeno je UBM-GMM sa 32 mešavine i tri modela po govorniku. Karakteristike govora koje su iskoristili kao ulaz u ovaj sistem je 32 MFCC i vrednost fundamentalne frekvencije izračunatih iz EPS [117] baze. Ova baza ima osam govornika i za svakog snimke u 14 emotivnih stanja (više u Apendixu D). Obuku **modela govornika na osnovu emocija** izvršili obučavanjem tri modela - svaki model za jednu grupu emocija. Emocije su najpre grupisali po sličnosti statistika fundamentalne frekvencije:

- nervoza, dosada, neutralno i stid,
- prezir, hladni bes, interesovanje, ponos i tuga,
- uzbuđenje, bes, panika, očaj i gađenje.

Testiranje ovog modela izvršeno je lokalno i globalno. **Lokalno testiranje** je u dva koraka - prvo je vršena detekcija emocije, a zatim je tražen model te grupe emocija koji daje najbolje performanse. U slučaju **globalnog testiranja**, test rečenica je evaluirana u odnosu na sve modele pa je biran najbolji model. Kako bi potvrdili performanse, konstruisali su još dve vrste modela sa kojima su poredili rezultate. Prva vrsta model je **model govornika obučen samo neutralnim govorom**. Druga vrsta modela je bazirana je takođe na grupama emocija, stim da je ovaj put **gupisanje izvršeno nausmično**:

- hladni bes, gađenje, intersovanje, panika i tuga,
- nervoza, prezir, bes, ponos i stid,
- uzbuđenje, dosada, očaj i neutralno.

Rezultati koje su dobili govore da je model na osnovu emocija ostvario procenat prepoznavanja od 90.52% u lokalnom i 76.43% u globalnom testu u poređenju sa 67.91% koje je ostvario neutralni model i 68.46% koje ostvaruje model sa nasumično grupisanim emocijama.
Chen i Yang [31] predložili su pristup koji spada u grupu preslikavanja emotivnog i neutralnog govora - **translacije učenja**. Njihova polazna hipoteza je da svaka od komponenti Gausove mešavine predstavlja određeni akustički događaj i da taj događaj postoji odgovarajuća komponenta u emotivnom modelu [31]. U fazi obuke sistema, koriste se neutralni i emotivni govor iz MASC baze [111]. Konstruiše se neutralni model govornika i set pravila koja preslikavaju komponente iz neutralnog modela u emotivni. Za preslikavanje koriste Kullback-Leibler divergenciju. Ovim algoritmom uspeli su da ostvare poboljšanje prepoznavanja u uslovima emotivnog govora od 2.81%.

Konverzija emotivnog govora koju predlažu Li i ostali [33, 34] uključuje sintezu emotivnog govora na osnovu neutralnog. U fazi testiranja, određuje se emocija test rečenice, njen prozodijski sadržaj i LPC analiza. Na osnovu ovih informacija, konstruiše se obučavajući skup za model govornika on neutralniog govora koji se konvertuje u emotivni govor. Za svakog od govornika, formira se emotivni model govornika od konvertovanog govora u odnosu na koje se onda vrši ispitivanje test rečenice. Samo modeliranje govornika vrši se MFCC-GMM modelom na podacima iz EPS baze [117]. Rezultati koje su postigli su 70.22% uspešnog prepoznavanja u odnosu na 62.81% koliko daje model obučen neutralnim govorom. Sličan pristup primenjuju Shan i Yang [118].
Krothapalli i ostali [35] fokusirali su se na konverziju karakteristika emotivnog govora u neutralni govor korišćenjem neuralne mreže dok je model govornika MFCC-GMM. Baze govora koje su koristili su Berlin [116], Hindi i Telugu baze [119, 120].
Li i ostali [36, 37] polazeći od pretpostavke da nisu svi delovi signala govora podjednako modifikovani usled emocija, predlažu metodu izdvajanja ovih delova signala na osnovu visine glasa govornika i normalizaciju skora prepoznavanja favorizovanjem baš ovih delova signala koji su manje izmenjeni emocijama.

5.4 Ostale tehnike prepoznavanja govornika

5.4.1 Skriveni Markovljevi modeli (HMM)

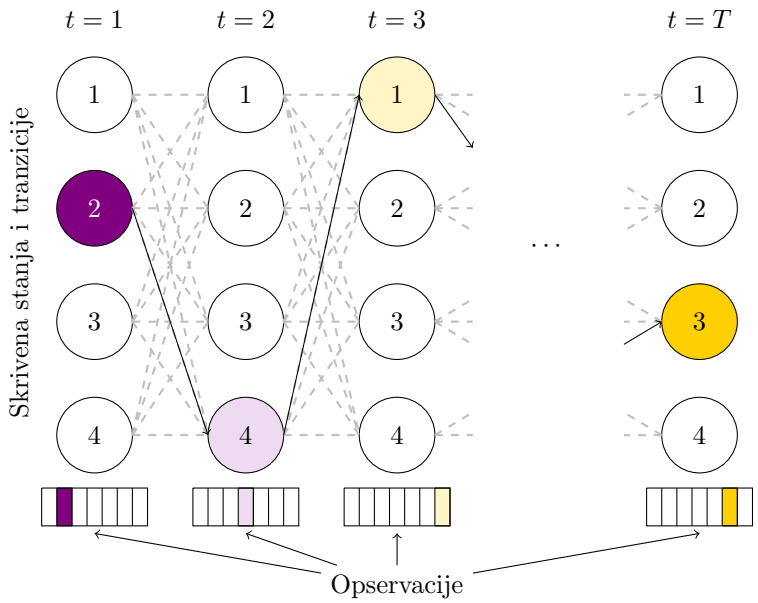
Skriveni Markovljevi modelli (Hidden Markov Models - HMM) [71] široko su primenjeni u oblastima prepoznavanja govora i prepoznavanja emocija u govoru, pa ne čudi što se pojavljuju i u pokušaju prepoznavanja govornika u uslovima emotivnog govora. Ovaj pristup modeliranja zasnovan je na stohastičkom procesu u dva nivoa [71]. Prvi proces opisuje tranziciju između stanja modela koja nisu merljiva - nazivaju se skrivenim stanjima. Drugi proces generiše opservacije na osnovu trenutnog stanja u kome se nalazi prvi proces. U zadacima prepoznavanja govora, tranzicije između skrivenih stanja odgovaraju tranizicijama između izgovorenih fonema u reči. Iako nema tačne fizičke interpretacije, ovaj model često se primenjuje i u zadacima prepoznavanja emocija u govoru [69, 79, 121]. Istorijski gledano, na osnovu HMM sa jednim skrivenim stanjem nastao je model GMM, te je korišćenje kompleksnijeg HMM modela napušteno. Ipak kada su emocije u govoru prisutne, ima smisla i primena ove vrste klasifikacije. Osnovna karakteristika HMM je sposobnost da modeluju dinamičke promene karakteristika govora. Postoje različiti načini implementacije HMM [69], a ovde će biti opisan model ergodičke strukture.
Prvi, skriveni proces koji je sadržan u HMM prelazi iz stanja u stanje sa određenom verovatnoćom promene. Da bi proces bio Markovljev, buduće stanje procesa zavisi isključivo od trenutnog stanja, a ne zavisi od prošlih stanja. Primer jedne sekvence promene stanja ilustrovan je na slici 5.18.
Moguća stanja modela data su skupom S, a skup opservacija skupom V:

$$\begin{aligned} S &= (s_1, s_2, \ldots, s_N), \\ V &= (v_1, v_2, \ldots, v_M), \end{aligned} \tag{5.18}$$

gde je N broj skrivenih stanja, a M broj mogućih opservacija. Definišimo Q kao fiksan niz stanja dužine T, a odgovarajući niz opservacija O:

$$\begin{aligned} Q &= (q_1, q_2, \ldots, q_T), \\ O &= (o_1, o_2, \ldots, o_T). \end{aligned} \tag{5.19}$$

Jedan skriveni Markovljev model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ opisan je sa tri parametra: matricom verovatnoće promene stanja $\boldsymbol{A}$, matircom verovatnoće opservacija $\boldsymbol{B}$ i vektorom početnih verovatnoća stanja



Slika 5.18: Tranzicije skrivenih stanja kod ergodičkih skrivenih Markovljevih modela [122].

$\boldsymbol{\pi}$.
Matrica verovatnoće promene stanja $\boldsymbol{A}$ definiše verovatnoću da proces posle stanja i uđe u stanje j, nezavisno od trenutka t:

$$\boldsymbol{A} = |a_{i,j}|,\ a_{i,j} = P(q_t = s_j | q_{t-1} = s_i). \tag{5.20}$$

Matrica verovatnoće opservacija $\boldsymbol{B}$ definiše verovatnoću da je opservacija k generisana iz skrivenog stanja i, nezavisno od trenutka t:

$$\boldsymbol{B} = |b_i(k)|,\ b_i(k) = P(x_t = v_k | q_t = s_i). \tag{5.21}$$

Vektor početnih verovatnoća stanja $\boldsymbol{\pi}$ definiše verovatnoću da određeno stanje bude početno stanje:

$$\boldsymbol{\pi} = [\pi_i],\ \pi_i = P(q_1 = s_i). \tag{5.22}$$

Prva pretpostavka, da buduće stanje zavisi isključivo od trenutnog stanja opisano je sa:

$$P(q_t | q_1, \ldots, q_{t-1}) = P(q_t | q_{t-1})). \tag{5.23}$$

Druga pretpostavka koja se koristi u HMM je da opservacija u trenutku t zavisi isključivo od trenutnog stanja modela tj da je nezavisna od prethodnih opservacija i stanja:

$$P(o_t | o_1, \ldots, o_{t-1}), q_1, \ldots, q_{t-1})) = P(o_t | q_t). \tag{5.24}$$

Određivanje parametara HMM modela λ vrši se Baum-Welchovim algoritmom [71].
Shahin je predložio nekoliko različitih modela HMM [28–30] sa primenom na prepoznavanje govornika u uslovima emotivnog govora. Eksperimenti su sprovedeni na konstruisanoj bazi od 40 govornika - deset različitih rečenica svaki od govornika izgovarao je četiri puta za neutralno i po jednom za svako od emotivnih stanja straha, besa, radosti, tuge i gađemnja. Standardni HMM primenjen je na dva različita načina [28]. U prvom eksperimentu, modeliranje je izvršeno u odnosu na rečenicu - modelirana je svaka rečenica na način kako je izgovorena od strane svakog od govornika i za svako od pet načina izgovaranja. U fazi testiranja određen je model koji je najverovatniji da generiše test rečenicu i kao prepoznat govornik određen je onaj koji tu rečenicu

izgovara. Veliki broj modela koji se kreira u ovom pristupu daje vrlo sporu fazu evaluacije. Modifikovani pristup primenjen je u drugom eksperimentu podrazumeva obuku modela emocija i prepoznavanje emocije test rečenice, a zatim i testiranje koji od modela govornika u okviru date emocije daje najveću verovatnoću da je generisao upravo tu rečenicu. Prosečan rezultat prepoznavanja je 68.3%.
Mansour i Lachiri koristili su i skrivene Markoviljeve modele da uporede efikasnost različitih vrsta kepstralnih koeficijenata za prepoznavanje govornika u uslovima emotivnog govora i u prisustvu šuma [24]. Koristili su podatke iz Berlin baze [116].
Cirkularni skriveni Markovljev model drugog reda [123] predstavljaju modifikaciju standardnog HMM tako što je skriveni stohastički proces modelovan 3D matricom. Verovatnoće tranzicija iz stanja u stanje opisane su:

$$\boldsymbol{A} = |a_{i,j,k}|,\ a_{i,j,k} = P(q_t = s_k | q_{t-1} = s_j, q_{t-2} = s_i). \tag{5.25}$$

Dalja nadogradnja ovog modela predstavlja konfiguraciju modela koja je umesto sleva na desno cirkularna. To znači da se u svako stanje modela može vratiti, da nema terminalnog stanja, kao i da je ispunjen uslov: $a_{i,j} = a_{j,i}$. Pokazalo se da obe modifikacije daju bolje performanse od standardnog modela u uslovima vikanja [29] - procenat prepoznavanja je 75% [29, 30].
Suprasegmentalne pojave u govoru protežu se duž više glasova i fonema, kao što su na primer fundamentalna frekvencija i akcenat [29]. **Suprasegmentalni skriveni Markovljevi modeli** imaju osobinu da više skrivenih stanja HMM agregiraju u jedno, suprasegmentalno stanje i kao takvi pogodni su za modeliranje prozodije, a koja je pogodna za detekciju emocija u govoru [124]. Integracija akustičkih i prozodijskih informacija korišćenjem standardnog HMM modela govornika λ i suprasegmentalnog modela govornika ψ opisana je sledećom formulom:

$$\log P(\lambda, \Psi | O) = (1 - \alpha) \cdot \log P(\lambda | O) + \alpha \cdot \log P(\Psi | O). \tag{5.26}$$

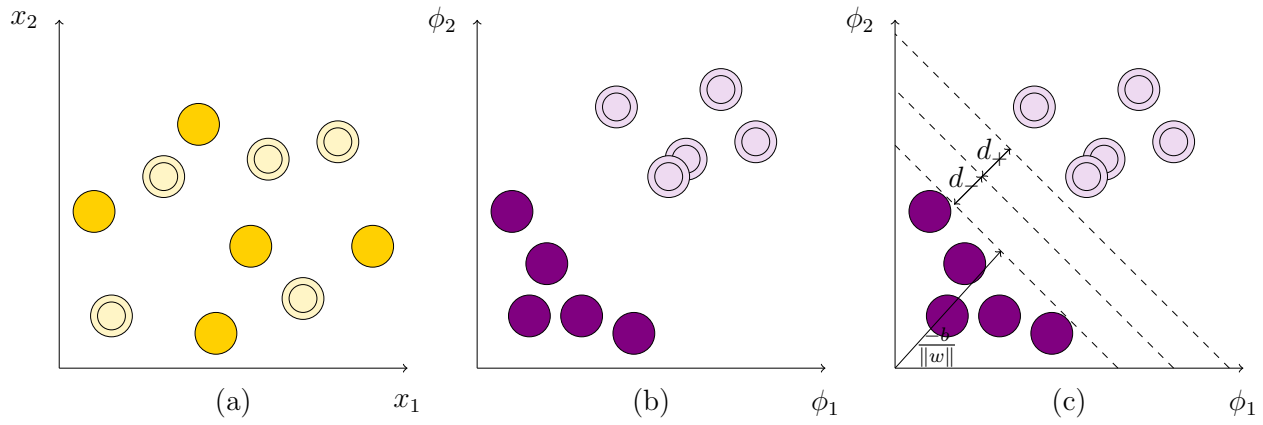
Ova vrsta modela dala je uspešnost prepoznavanja od 68% u uslovima vikanja, dok je u slučaju besa vrednost 71% [29, 30].

5.4.2 Mašine potpornih vektora (SVM)

Osnovna ideja Mašina potpornih vektora (Support vector machines - SVM) [125] je definisanje funkcije $f(x)$ takve da je moguće razdvajanje prostora karakteristika u dva podprostora (koji odgovaraju dvema klasama) nekom hiperravni tako da je rastojanje između tako podeljenih klasa maksimalno. Obuku klasifikatora u slučaju dve linearnoo separabilne klase počinje definisanjem ulaznih podataka kao seta parova:

$$X = \{(x_1, y_1), \ldots, (x_N, y_N)\},\ x_i \in \mathbb{R}^d,\ y_i \in \{-1, +1\},\ i = 1, \cdots, N, \tag{5.27}$$

gde je N broj ulaznih vektora. U nastavku je opisan postupak određivanja parametara klasifikacije za tri slučaja koja su od interesa: dve linearno separabilne klase, dve linearno neseparabilne klase i slučaj više klasa.



Slika 5.19: Primer (a) linearno neseparabilnih klasa, (b) linearno separabilnih klasa (c) optimalne hipperravni koja razdvaja dve linearno separabilne klase.

Linearno separabilne klase

Najjednostavniji slučaj svakako jesu dve linearno separabilne klase. Hiperravani koje ih razdvajaju definisane su jednakošću $\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b = 0$, a funkcija odlučivanja definiše se sa:

$$f(x) = \operatorname{sgn}\Big(\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b\Big) = \begin{cases} +1 & \langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b > 0, \\ -1 & \langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b \leqslant 0. \end{cases} \tag{5.28}$$

gde su b i $\boldsymbol{w}$ parametri hiperravni, a $\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle$ skalarni proizvod $\boldsymbol{w}$ i $\boldsymbol{x}$ [126]. Procesom optimizacije dolazi se do hiperravni koja maksimizuje rastojanje od obe klase. Nakon rešavanja problema optimizacije [126], parametar $\boldsymbol{w}$ dat je u obliku:

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i x_i. \tag{5.29}$$

Svaka od tačaka x_i opisan Lagranžovim multiplikatorima u toku rešavanja problema optimizacije. Tačke za koje je ispunjeno da je $\alpha_i[y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b) - 1]$ i da su Lagranžovi multiplikatori veći od nula $\alpha_i > 0$ nazivaju se "potpornim vektorima". Tada je optimalna funkcija odlučivanja (koja maksimizuje rastojanje klasa i hiperravni) data formulom:

$$f(\boldsymbol{x}) = \operatorname{sgn}\Bigg(\sum_{i=1}^{N_{SV}} \alpha_i y_i \langle \boldsymbol{x_i^{SV}} \cdot \boldsymbol{x} \rangle + b\Bigg), \tag{5.30}$$

gde je N_{SV} broj potpornih vektora.

Linearno neseparabilne klase

U slučaju da klase nisu linearno separabilne, često je moguće mapirati originalne vrednosti u visokodimenzioni prostor gde ove vrednosti postaju linearno separabilne. To se postiže kernel funkcijom:

$$\begin{aligned} \Phi : \mathbb{R}^d &\to \mathbb{R}^D, \ (D >> d) \\ \boldsymbol{x} &\to \Phi(\boldsymbol{x}), \end{aligned} \tag{5.31}$$

gde je Φ kernel funkcija, a D dimenzija novog prostora. Najčešće korišćene kernel funkcije su Gausova funkcija sa radijalnom bazom (RBF), polinomijalna, sigmoidalna itd. U ovom

visokodimenzionom prostoru, koeficijenti funkcije odlučivanja treba da budu izabrani na način da je margina između klasa maksimalna [127]. Funkcija odlučivanja tada je data u obliku:

$$f(x) = \mathrm{sgn}\Big(\langle \boldsymbol{w} \cdot \Phi(\boldsymbol{x})\rangle + b\Big) = \tag{5.32}$$

Gde su b i $\Phi(\boldsymbol{x})$ parametri hiperravni. Na kraju, pod uslovima koji su dati u [126], funkcija odlučivanja predstavljena je u formi:

$$f(x) = \mathrm{sgn}\Big(\sum_{i=1}^{n_{SV}} \alpha_i y_i \langle \Phi(\boldsymbol{x_i}) \cdot \Phi(\boldsymbol{x})\rangle + b\Big),\ 0 \leqslant \alpha_i \leqslant C, \tag{5.33}$$

gde je n_{SV} broj potpornih vektora, a C regularizacioni parametar odnosa greške. Neka je odnos dva vektora u prostoru karakteristika definisan sa:

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \Phi(\boldsymbol{x_i}) \cdot \Phi(\boldsymbol{x_j})\rangle. \tag{5.34}$$

U slučaju korišćenja RBF, kernel funkcija data je sa:

$$k(\boldsymbol{x}, \boldsymbol{x_i}) = \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{x_i}\|^2}{2\sigma^2}\Big), \tag{5.35}$$

gde parametar σ definiše širinu Gausove funkcije i može se koristiti kao parametar kojim se podešava nivo generalizacije. Tada je razdvajajuća hiperravan definisana sa:

$$f(\boldsymbol{x}) = \mathrm{sgn}\Big(\sum_{i=1}^{N_{SV}} \alpha_i y_i k(\boldsymbol{x_i}, \boldsymbol{x}) + b\Big). \tag{5.36}$$

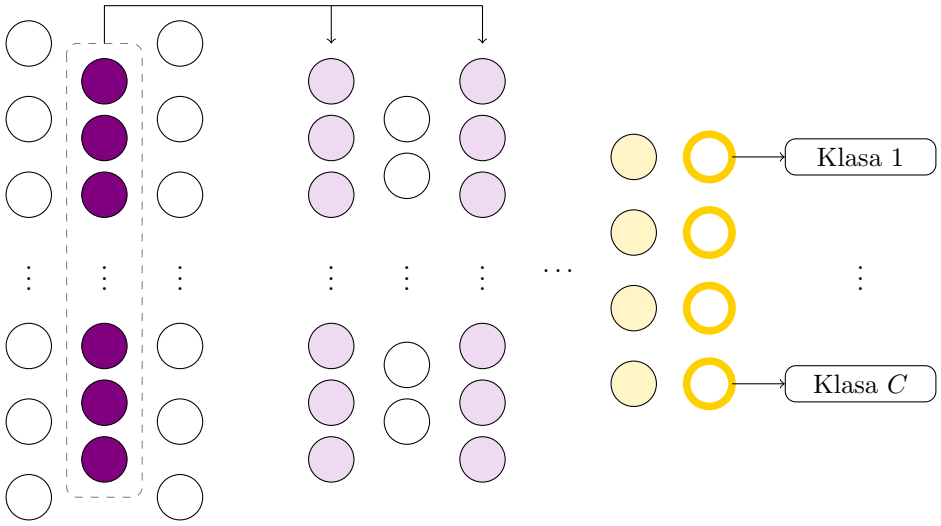
Slučaj više klasa

Praktična primena obično zahteva klasifikaciju u više od dve klase. U nastavku je dato tri načina na koji se SVM može primeniti na zadatak klasifikacije više klasa [128]. **Jedan i jedan** - Ovaj pristup podrazumeva da se klasifikacija primenjuje na svakom od parova klasa. Tada se za problem m klasa obučava $\frac{m(m-1)}{2}$ SVM klasifikatora. **Usmereni aciklični graf** Nadogradnja prethodnog pristupa na način da se umesto jednostavnog odlučivanja gradi graf odluke. **Jedan i svi** - Ideja ovog pristupa je da se konstruiše po jedan klasifikator za svaku od klasa. Klasifikatori pojedinačnih klasa tada diskriminišu vektore karakteristika jedne klase u odnosu na sve ostale klase. Novi vektor klasifikuje se u neku od klasa na osnovu onog klasifikatora koji daje maksimalnu vrednost izlaza.
Ova vrsta klasifikatora ima rasprostranjenu primenu u različitim oblastima. Jedan je od najuspešnijih klasifikatora kada je u pitanju prepoznavanje emocija. U prepoznavanju govornika, Campbell i ostali [129] predložili su kombinaciju GMM supervektora sa SVM i različitim kernelima. U prepoznavanju govornika u uslovima emotivnog govora, Mansour i Lachiri iskoristili su SVM kao klasifikator da uporede efikasnost MFCC i SDC-MFCC karakteristika [25] na IEMOCAP bazi [130]. Kernel funkcije u ovim SVM su i polinomijalna i gausova, a konfiguracije koje su primenjene su i *jedan i jedan* i *jedan i svi*. Rezultati koje su dobili pokazuju da *jedan i svi* SVM sa gausovskim kernelom i SDC-MFCC karakteristikama daju najbolje rezultate 91.34%.

5.4.3 Duboke neuralne mreže (DNN)

Duboka neuralna mreža (Deep Neural Network - DNN) [131,132] je višeslojna veštačka neuralna mreža koja ima nekoliko skrivenih slojeva između ulaznog i izlaznog sloja. Višeslojne neuralne

mreže pokazale su se veoma korisnim za klasifikaciju kompleksnih podataka [133] jer svaki od slojeva može da "nauči" sa različitim stepenom generalizacije. Konkretna konfiguracija DNN zavisi od zadatka klasifikacije na koji se odgovara [133]. Efikasan način obuke ovakvih neuralnih mreža leži u obuci svakog od slojeva zasebno, nakon čega se obučava mreža kao celina (Slika 5.20). Proces formiranja M skrivenih slojeva DNN počinje od jedne *feed-foreward* neuralne mreže sa ulaznim, jednim skrivenim i izlaznim slojem. Pri tome je broj čvorova u ulaznom i izlaznom sloju ove mreže D jednak dimenziji vektora karakteristika, dok je broj čvorova u skrivenom sloju d ispunjava uslov $d < D$. Trening ove mreže se sprovodi algoritmom propagacije unazad (*back-propagation*) sa ciljem da se ulazni vektor reflektuje na izlaz mreže. Po završetku obuke ove mreže, izlazni sloj se uklanja, a skriveni sloj posmatra se kao novi izlazni sloj. Na taj način stvara se redukovana slika ulaznog vektora karakteristika. Sledeći skriveni slojevi formiraju se ponavljanjem ovog postupka: za obuku skrivenog sloja i, $i = 1, \cdots, M$, sa brojem čvorova d_i koristi se sloj d_{i-1} kao ulazni i izlazni sloj i na ovu mrežu koraka i primenjuje se algoritam propagacije unazad. Finalni izlazni sloj mreže obučava se superviziranim treningom, tako što se za ulaz koristi izlaz poslednjeg skrivenog sloja, a izlaz su labele C klasa u koje se vrši klasifikacija. U poslednjem koraku, povezuju se slojevi mreže obučavani nezavisno i vrši se fino podešavanje mreže kao celine upotrebom algoritma propagacije unazad.

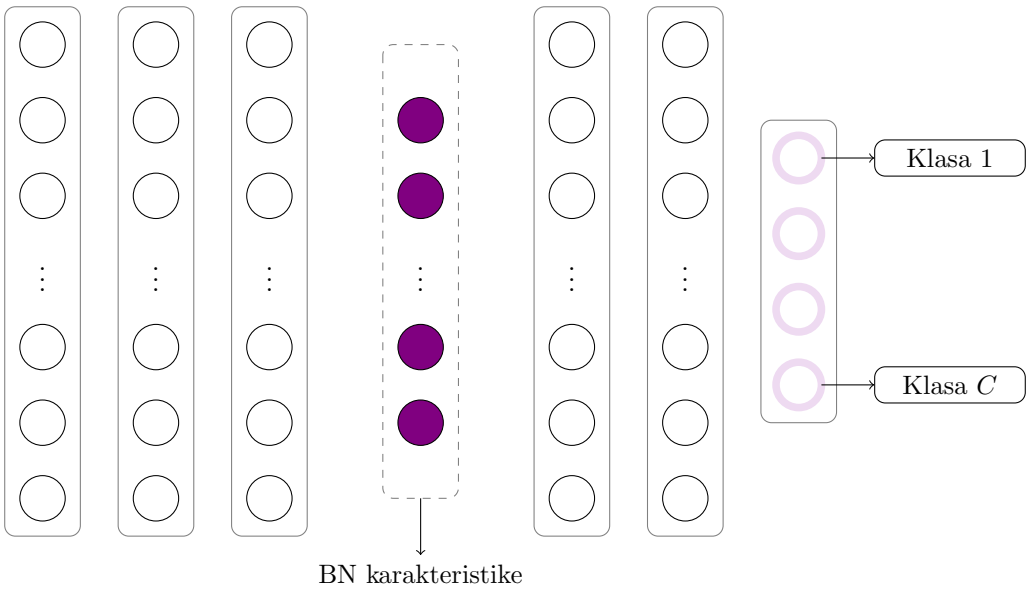


Slika 5.20: Proces obučavanja duboke neuralne mreže [122]

Ovakav način obuke neuralne mreže omogućio je njihovu efikasniju obuku, što je zajedno sa porastom procesorske snage omogućilo da ova metoda dođe u centar naučne pažnje u poslednjih nekoliko godina. Duboke neuralne mreže primenjene su uspešno i u prepoznavanju govornika. Pokazale su se kao veoma efikasan ekstraktor karakteristika koje su dalje ulaz za druge sisteme klasifikacije, dok kao klasifikator uspevaju da dostižu performanse standardnih sistema i-vektora u slučaju kratkih test rečenica. Dve su značajne primene DNN. Prva je topologija za izdvajanje karakteristika uskog grla - Bottleneck (BN) [134]. Druga je topologija DNN za izdvajanje takozvanih x-vektora [135], a takođe se ispituju i kao klasifikatori govornika. U nastavku su obe tehnike analizirane sa više detalja.

Karakteristike uskog grla i duboke neuralne mreže

Karakteristike uskog grla (bottleneck - BN) generišu se kao izlaz skrivengog sloja DNN koji ima mali broj čvorova u odnosu na ostale slojeve [136]. Grafički prikaz topologije mreže dat je na slici 5.21. Na taj način praktično se vrši redukcija dimenzija ulaznog vektora i ekstrakcija informacije koja je relevantna za klasifikaciju.



Slika 5.21: Izdvajanje suženog (BN) seta karakteristika korišćenjem DNN [137].

Ovaj pristup prvo je uspešno primenjen u prepoznavanju jezika [138], a zatim je primenjen sa uspehom i u prepoznavanju govornika [134, 137, 139]. Ričardson i ostali [137] u svom radu prezentovali su uniformnu neuralnu mrežu koja ima dvojaku namenu: prepoznavanje jezika i prepoznavanje govornika. Korišćenjem BN karakteristika, postigli su značajno poboljšanje u odnosu na sisteme koji koriste spektralne karakteristike i i-vektore.
Struktura duboke neuralne mreže za izdvajanje BN karakteristika sastoji se od slojeva koji su potpuno međusobno povezani i imaju fiksan broj čvorova: ulaznog sloja, nekoliko skrivenih slojeva i izlaznog sloja [137]. Transformacija izlaza određenog sloja je aktivacija čvorova sledećeg sloja i izračunava se matricom transformacije, a izlaz sloja se zatim izračunava primenom aktivacione funkcije:

$$a^{(i)} = M^{(i)} \cdot x^{(i-1)}. \tag{5.37}$$

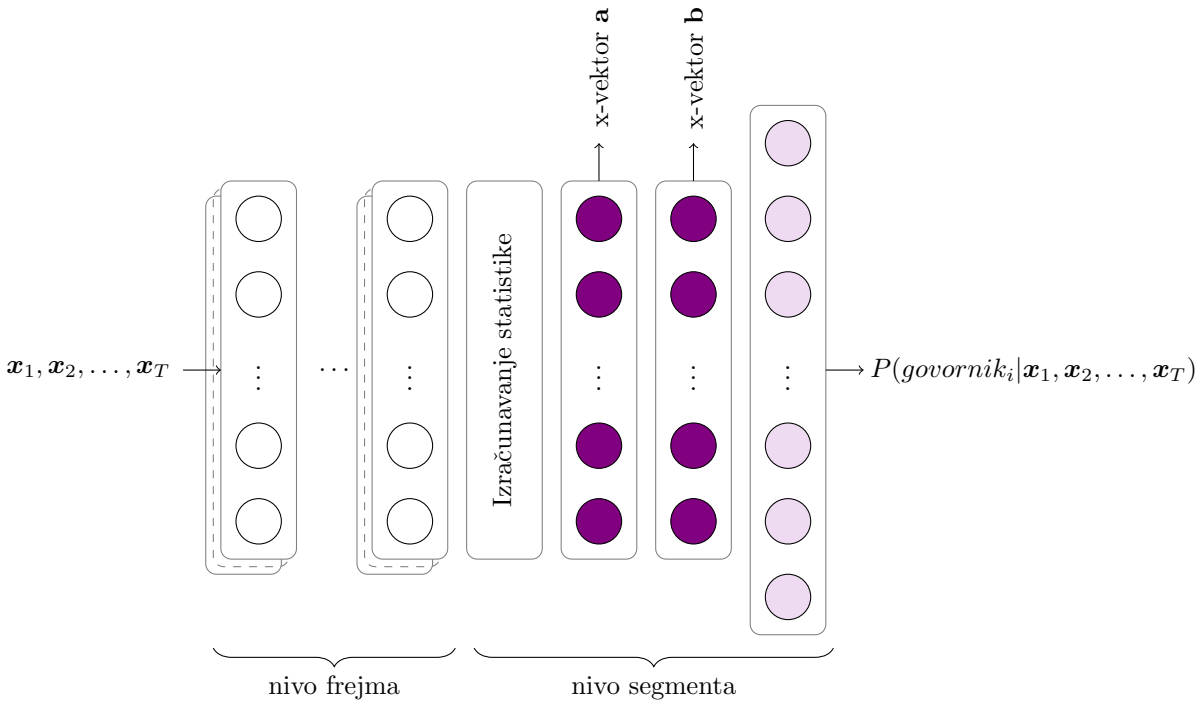
$$x^{(i)} = h^{(i)} a^{(i)}. \tag{5.38}$$

Namena DNN diktira tip aktivacione funkcije poslednjeg sloja mreže. Sprovedeno je nekoliko istraživanja na ovu temu [137, 139, 140], gde je DNN obučavana sa ciljem prepoznavanja govornika, a korišćenjem BN karakteristika. U ovim radovima [137, 139], aposteriori verovatnoće koje su izlaz neuralne mreže, korišćene su kao zamena za posteriori verovatnoće GMM u ekvivalentnom sistemu i-vektora. Ričardson i ostali [137] uporedili su sva četiri sistema i-vektora. Varijacije ulaznih karakteristika bile su da li se koriste MFCC ili BN karakteristike, a varijacije aposteriori verovatnoće bile su da li se upotrebljavaju GMM ili DNN aposteriori verovatnoće. Rezultati su pokazali da BN karakteristike daju bolje rezultate u odnosu na obične MFCC karakteristike u klasičnom sistemu i-vektora. Međutim, dolazi do degradacije performansi kada BN karakteristike kombinuju sa DNN aposteriori verovatnoćama.

X-vektori i duboke neuralne mreže

David Snyder i ostali u svojim istraživanjima [141–143], tražili su način da efektno upotrebe DNN osim za izdvajanje karakteristika i za samu klasifikaciju govornika. U tome su i uspeli uvođenjem x-vektora (tj ugrađenih karakteristika - emmbedings kako su ih prvobitno nazvali) [135, 141] i neuralne mreže specifične arhitekture i načina obuke. Njihov pristup donosi dve inovacije - izdvajanje x-vektora i klasifikaciju korišćenjem DNN.

X-vektori nastali su od ideje da informacije o govorniku treba izdvojiti na osnovu cele rečenice, a ne samo na osnovu frejmova signala, kao što se to radi sa i-vektorima. Na Slici 5.22 prikazana je šema obuke DNN koju su Snyder i ostali koristili za izdvajanje x-vektora, a kasnije i klasifikaciju govornika.



Slika 5.22: Izdvajanje x-vektora i klasifikacija korišćenjem DNN [141].

Mreža se sastoji iz sledećih slojeva: slojevi koji rade na nivou frejma, sloj koji izračunava statistiku, slojevi koji rade na nivou segmenta i izlazni sloj mreže (softmax sloj) [135]. Karakteristike koje su izdvajali na nivou frejma dužine $25ms$ bile su 20 MFCC koeficijenata, usrednjenih prozorom koji proklizava na $3ms$. Osim toga, za izdvajanje karakteristika primenjen je i detektor vokalne aktivnosti (VAD) [144] na osnovu energije govora.
Na ulaz neuralne mreže dovodi se pet uzastopnih vektora karakteristika spojenih u jedan ulazni vektor i to na način da se vektori $\{t-2, t-1, t, t+1, t+2\}$ nadovezuju jedan na drugi [141]. Sledeća dva sloja spajaju izlaze prethodnog sloja u trenutcima $\{t-2, t, t+2\}$ odnosno $\{t-3, t, t+3\}$ i na taj način spaja se devet uzastopnih frejmova - od $t-8$ do $t+8$ [141].
Izlaz poslednjeg sloja koji radi na nivou frejma postaje ulaz sloja koji izračunava srednju vrednost i standardnu devijaciju sabirajući sve što mu dolazi na ulaz duž jednog celog segmenta govora [141]. Eksperimentalno [135] je utvrđeno da segmenti govora na onovu koje se izračunavaju x-vektori su $30s$ ili cela rečenica ako je kraća od $30s$. Statistike segmenata zajedno se šalju na ulaz dva dodatna afina skrivena sloja [141] od kojih prvi za rezultat ima x-vektor $\boldsymbol{a}$, dok je x-vektor $\boldsymbol{b}$ rezultat sledećeg afinog sloja nakon primene ReLU aktivacione funkcije ($ReLU(x) = \max(x, 0)$) [145]. Ova funkcija jedna je od najpopularnijih aktivacionih funkcija korišćenih u dosadašnjem istraživanju dubokog učenja. U praksi, to znači da je x-vektor $\boldsymbol{b}$ nelinearna funkcija statistike. Specifičnost obuke mreže za izdvajanje x-vektora zasnovana je na činjenici da ulazne karakteristike na osnovu segmenata mogu biti različite dužine [141]. Rezultati poređenja x-vektora i i-vektora, sa PLDA klasifikatorom na ovako velikom broju snimaka i preko šest hiljada govornika [141] pokazali su da su x-vektori uporedivi sa i-vektorima i komplementarni u slučaju kombinacije ovih sistema. U slučaju dugačkih test rečenica, i-vektori su bolji, a kada su u pitanju kraći segmenti, x-vektori daju bolje rezultate [141]. Takođe, x-vektori su se pokazali bolje kada je u pitanju promena jezika, odnosno kada obučavanje na jednom, a

testiranje na drugom jeziku [141]. U opisanim eksperimentima, obučene neuralne mreže imaju između pet i osam miliona parametara. Ovaj podatak govori o kompleksnosti neuralnih mreža kao klasifikatora i numeričkoj zahtevnosti obuke jednog ovakvog sistema.

Uvećanje podataka

Zarad poređenja rezultata i-vektora i x-vektora na javno dostupnim setovima podataka za prepoznavanje govornika, u radu [142] iskorišćeno je uvećanje podataka izmenama originalnih snimaka dodavanjem žamora drugih govornika, muzike, šuma ili modifikacija signala radi stvaranja odjeka. Uvećanje podataka najpre je upotrebljeno prilikom obuke PLDA klasifikatora - pored originalnog snimka, korišćene su i još dve kopije snimka uvećane nasumično izabranim načinom [142]. Pokazalo se da uvećanje podataka daje poboljšanje za sve upoređene sisteme, a najviše za x-vektore [142, 146].

5.4.4 Analiza ostalih tehnika prepoznavanja govornika

Iako su se uspešno pokazali u drugim oblastima obrade signala govora, skriveni Markovljevi modeli ispostavlja se da nisu pogodni za modeliranje govornika, čak ni u uslovima emotivnog govora. Ovakav zaključak izvodimo na osnovu radova [24, 28–30], kao i samog istorijata prepoznavanja govornika u kome su HMM zamenjeni sa GMM modelima.
Početna istraživanja neuralnih mreža za zadatak prepoznavanja govornika imala su za cilj da zamene ulazni vektor karakteristika, kao što je i-vektor. Ovakva rešenja uspela su da dostignu performanse klasičnih sistema i-vektora i PLDA, međutim ne i da pokažu značajna poboljšanja i benefite. Istraživanje DNN i teme x-vektora pokazala su se superiornijim u odnosu na i-vektore, naročito kada se neuralna mreža koristi kao ekstraktor karakteristika (x-vektora), a uporedive rezultate u slučaju kada se DNN koristi kao klasifikator. Najveće poboljšanje dobijeno je kada su test rečenice kratke - tada su se DNN pokazale neprikosnovenim. Najveći nedostatak neuralnih mreža upravo je njihova kompleksnost. Naime jedna mreža može sadržati desetine miliona parametara, a obuka mreže može trajati jako dugo (i po nekoliko nedelja). Bez obzira, sistemi x-vektora, kao i i-vektor sistemi, postaju standardna tehnika za prepoznavanje govornika u ovoj oblasti.

5.5 Rezime i zaključci

U ovoj sekciji prikazani su i analizirani rezultati eksperimenata sa *Neutralnim*, *Malim miks*, *Malim tri*, *Velikim tri* i *Velikim miks modelom* korišćenjem Gausovih mešavina (GMM) i i-vektora kao tehnika klasifikacije i modeliranja govornika, na bazama ruskog, srpskog, italijanskog i engleskog emotivnog govora. Korišćene karakteristike govora bile su 13 MFCC koeficijenata. Korišćena je mala količina trening podataka - svega šest rečenica za *Neutralni* i *Male modele*, i 18 rečenica za *Velike modele*. Na osnovu rezultata, može se zaključiti sledeće:

- uspešnost sistema za prepoznavanje govornika manja je u prisustvu emotivnog govora,
- uključivanje emotivnog govora umesto neutralnog povećava procenat prepoznavanja govornika,
- *Miks modeli* daju bolje rezultate nego *Tri modeli*,
- kada su tehnike klasifikacije i modeliranja u pitanju, rezultati govore da, iako su i-vektori današnji standard za prepoznavanje govornika, GMM daje stabilnije rezultate za relativno mali broj govornika

Uticaj emocija je takođe analiziran. Ustanovljeno je sledeće:

- neutralni govor najlakše se prepoznaje i kada se modeliranje govornika vrši samo neutralnim govorom, i kada se za model koristi i neutralni i emotivni govor,
- bes je emocija koja najviše degradira prepoznavanje govornika,
- ostale emocije imaju uticaj koji se razlikuje u zavisnosti od eksperimenata i baze na kojoj je eksperiment sproveden

Zaključak o razlici u prepoznavanju govornika muškog i ženskog pola je da su ovako konstruisani sistemi praktično podjednako pogodni i za muškarce i za žene.

6. Određivanje broja Gausovih mešavina subtractive klasterizacijom

Modeli Gausovih mešavina čine osnovu modernih sistema za prepoznavanje govornika, što ovu tehnniku čini i izuzetno važnom. Uobičajen postupak obuke GMM podrazumeva da se broj mešavina zada unapred, mešavine se incijalizuju i na osnovu toga se vrši dalja obuka modela. Uobičajeno, primenjuje se tenika K najbližih suseda (KNN), koju smo i mi primenjivali u dosadašnjim eksperimentima.

K najbližih suseda (K Nearest Naighbours - KNN) je neparametarska tehnika klasterizacije [147]. Osnovna ideja koja stoji iza ove tehnike je da se vektori karakteristika grupišu u klastere na osnovu međusobne udaljenosti. Odbirci se najpre razvrstavaju u klastere nasumično. Za svaki od klastera izračunava se centar $\boldsymbol{\mu_i}$, $i = 1, \cdots, C$, gde je C broj klastera i $\boldsymbol{\Sigma_i}$ matrica rasipanja za tako formirani klaster. Nakon toga, vrši se preraspodela vektora koji su bliži nekim drugim centrima nego centru svog klastera. Postupak se ponavlja dok god postoji premeštanje vektora karakteristika.

Lee i ostali primenili su inkrementalni K-means algoritam za određivanje optimalnog broja GMM [38]. U njihovom pristupu, dodaje se jedna po jedna komponenta mešavine i meri se njena statistička korelacija sa već postojećim komponentama. Kada sledeća dodata mešavina postane zavisna, optimalan broj komponenti je određen. Eksperimentalno su pokazali da ova tehnika daje bolje rezultate u odnosu na broj komponenti dobijen na osnovu kriterijuma količine informacije [38]. Kombinacijom kriterijuma Wang i ostali [39] pokazali su da se broj komponenti raspodele može precizno odrediti, međutim, eksperimenti su sprovedeni samo na simuliranim podacima.

Dosadašnji radovi pristupali su unapređenju GMM ili UBM-GMM algoritma određivanjem optimalnog broja mešavina na osnovu kojih su kreirani modeli za svakog od govornika, međutim i dalje sa isitm brojem komponenti za svakog govornika. Naš cilj je da na osnovu uzorka govora jednog govornika, subtractive klasterizacijom odredimo broj mešavina za tog govornika, i da u praksi verifikujemo da je govornike moguće modelirati različitim brojem komponenti. Osim toga, cilj nam je i da broj komponenti odredimo na automatski način.

6.1 Subtractive klasterizacija

Subtractive klasterizacija ima za ideju da svako od postojećih merenja može biti centar klastera [148, 149]. Zahvaljujući ovoj polaznoj pretpostavci, kao centri klasera razmatraju se samo

stvarna merenja, pa je složenost ovog algoritma praktično linearna. Merenjima se dodeljuje određena gustina na osnovu koje se u iterativnom postupku određuju centri klasterizacije. Prema opisu algoritma u literaturi [148–150] počinje se određivanjem centra prvog klastera, a to je merenje koje ima najveću početnu gustinu. Početna gustina izračunava se po formuli:

$$\Upsilon_i = \sum_j e^{-\frac{\|x_i - x_j\|^2}{(r_a/2)^2}} \tag{6.1}$$

$$\Upsilon_c = \max\{\Upsilon_i\}. \tag{6.2}$$

Parametar r_a naziva se radijus klasterizacije i određuje se na osnovu ulaznih podataka korišćenjem sledeće formule [151]:

$$r_a = \frac{1}{4}\Big[\max\{\|x_i - x_j\|\} + \min\{\|x_k - x_l\|\}\Big], \tag{6.3}$$

$$r_a = r_a^{\beta}. \tag{6.4}$$

Dodatno, faktor β uveden je kako bi se amortizovalo postojanje eventualnih outlier merenja. Radu Cui i ostalih [151], ustanovljeno je da je optimalna vrednost faktora $\beta = 0.5$ za zadatak klasterizacije dužice oka. Nakon određivanja prvog centra klastera započinje se iterativni postupak određivanja ostalih centara klastera na onsovu modifikovane gustine svakog od merenja. Modifikovana gustina merenja data je sledećom formulom [150]:

$$\Upsilon_i^k = \Upsilon_i^{k-1} - \Upsilon_c^{k-1} \cdot e^{-\frac{\|x_i - x_c^*\|^2}{(r_b)^2/2}}, \tag{6.5}$$

$$r_b = \epsilon r_a, \tag{6.6}$$

gde Υ_i^k označava gustinu u trenutnoj iteraciji, a Υ_i^{k-1} u prethodnoj. Pri tome je $\Upsilon_c^{k-1} = max\{\Upsilon_i^{k-1}\}$, a r_b novi radijus klasterizacije. Parametar ϵ naziva se faktor zatezanja i prema Jingu i ostalima [150] $\epsilon = 1$, dok je u slučaju primena na fazi-logici [152] predlog da ovaj faktor bude u granicama $\epsilon \in [1.25, 1.5]$. Postupak se završava kada se ispuni uslov:

$$\Upsilon_c^k / \Upsilon_c < \delta, \tag{6.7}$$

gde je δ izabrani prag klasterizacije. Faktori β, ϵ i δ utiču na broj i veličinu klastera i njihovim podešavanjem utiče se na finalnu efikasnost klasterizacije. Ključni elementi primene algoritma subtractive klasterizacije koji su bili predmet našeg istraživanja su:

(1) određivanje radijusa klasterizacije,

(2) promena radijusa klasterizacije kroz iteracije algoritma,

(3) određivanje praga zaustavljanja algoritma.

6.2 Analiza Gausovih slučajnih promenljivih

Cilj nam je da na osnovu ponašanja i raspodele podataka odredimo radijus klasterizacije koji će omogućiti kvalitetnu klasterizaciju. Konkretno, interesuje nas raspodela minimuma i maksimuma Euklidskog rastojanja dva vektora. Modeliranje klastera u kasnijoj obradi biće urađeno GMM, te za početnu pretpostavku uzimamo da su podaci raspodeljeni Gausovski. Pošli smo od jednodimenzione i jednomodne gausove raspodele, zatim smo rezultat generalizovali na višedimenzionu raspodelu i na kraju na višedimenzionu i višemodnu raspodelu koja i odgovara našim podacima.

6.2.1 Raspodela rastojanja dve jednodimenzione Gausova promenljive

Neka su X i Y dve nezavisne slučajne promenljive sa Gausovom raspodela datom parametrima $\mathcal{N}(\mu_X, \sigma_X^2)$ i $\mathcal{N}(\mu_Y, \sigma_Y^2)$. Želimo da odredimo normu rastojanja ove dve varijable i njeno ponašanje. U slučaju jednodimenzione i jednomodne raspodele, $Z = |X - Y|$. Najpre možemo zaključiti da razlika $W = X - Y \sim \mathcal{N}(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2) = \mathcal{N}(\mu_W, \sigma_W^2)$ takođe uzima Gausovu raspodelu. Funkcija gustine verovatnoće i funkcija raspodele date su jednačinama:

$$f_W(x) = \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(x-\mu_W)^2}{2\sigma_W^2}} \tag{6.8}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_X^2 + \sigma_Y^2)}} e^{-\frac{(x-\mu_X+\mu_Y)^2}{2(\sigma_X^2+\sigma_Y^2)}} \tag{6.9}$$

$$F_W(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu_W}{\sqrt{2\sigma_W^2}}\right)\right] \tag{6.10}$$

$$= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu_X+\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right)\right], \tag{6.11}$$

$$\tag{6.12}$$

gde je $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. Apsolutna vrednost ove promenljive $Z = |W|$ dakle uzima vrednost presavijene Gausove raspodele (Folded Normal Distribution) [153]:

$$F_Z(x) = P(Z \leqslant x) = P(|W| \leqslant x) = \begin{cases} 0, & x < 0 \\ P(-x \leqslant W \leqslant x), & x \geqslant 0 \end{cases} \tag{6.13}$$

$$P(-z \leqslant W \leqslant x) = F_W(x) - F_W(-x) \tag{6.14}$$

$$= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu_X+\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right)\right] - \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{-x-\mu_X+\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right)\right] \tag{6.15}$$

$$= \frac{1}{2}\left[\operatorname{erf}\left(\frac{x-\mu_X+\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right) - \operatorname{erf}\left(\frac{-x-\mu_X+\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right)\right] \Rightarrow \tag{6.16}$$

$$F_Z(x) = \frac{1}{2}\operatorname{erf}\left(\frac{x-\mu_X+\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right) - \frac{1}{2}\operatorname{erf}\left(\frac{-x-\mu_X+\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right), x \geqslant 0. \tag{6.17}$$

U slučaju da X i Y pripadaju istoj raspodeli, tj da je $\mu_X = \mu_Y$ i $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ tada važi da je $W \sim \mathcal{N}(0, 2\sigma^2)$. Tada izraz za funkciju raspodele postaje:

$$F_Z(x) = \operatorname{erf}\left(\frac{x}{2\sigma}\right),\ x \geqslant 0. \tag{6.18}$$

Funkcija gustine verovatnoće izračunava se tada u obliku:

$$f_Z(x) = \frac{dF_Z(x)}{dz} = \frac{dF_W(x)}{dz} - \frac{dF_W(-x)}{dx} \tag{6.19}$$

$$= f_W(x) - f_W(-x)\frac{d}{dx}(-x) = f_W(x) + f_W(-x) \tag{6.20}$$

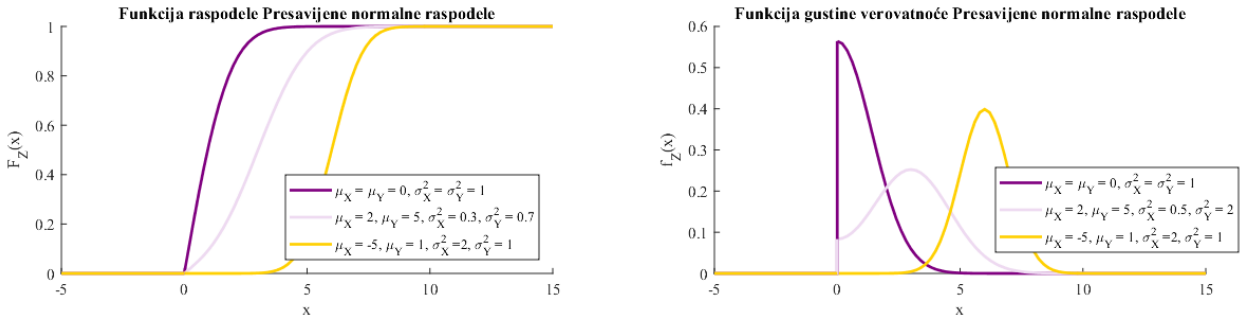
$$= \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(x-\mu_W)^2}{2\sigma_W^2}} + \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(x+\mu_W)^2}{2\sigma_W^2}} \tag{6.21}$$

$$f_Z(x) = \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}} e^{-\frac{(x-\mu_X+\mu_Y)^2}{2(\sigma_X^2+\sigma_Y^2)}} + \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}} e^{-\frac{(x+\mu_X-\mu_Y)^2}{2(\sigma_X^2+\sigma_Y^2)}},\ x \geqslant 0 \tag{6.22}$$

U slučaju da X i Y pripadaju istoj raspodeli, tj da je $\mu_X = \mu_Y$ i $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ tada važi:

$$f_Z(x) = \frac{1}{\sqrt{\pi\sigma^2}} e^{-\frac{x^2}{4\sigma^2}},\ x \geqslant 0 \tag{6.23}$$

Grafički prikaz funkcije raspodele i funkcije gustine verovatnoće presavijene normalne raspodele dat je na Slici 6.1. Očekivanje i varijansa ove raspodele su tada:



[A] [B]

Slika 6.1: Izgled [A] funkcije raspodele i [B] funkcije gustine verovatnoće presavijene normalne raspodele.

$$\mu_Z = \sigma_W \sqrt{\frac{2}{\pi}} e^{\left(\frac{-\mu_W^2}{2\sigma_W^2}\right)} + \mu_W \operatorname{erf}\left(\frac{\mu_W}{\sqrt{2\sigma_W^2}}\right), \tag{6.24}$$

$$\sigma_Z^2 = \mu_W^2 + \sigma_W^2 - \mu_Z^2 \Rightarrow \tag{6.25}$$

$$\mu_Z = \sqrt{\sigma_X^2+\sigma_Y^2}\sqrt{\frac{2}{\pi}} e^{\frac{-(\mu_X-\mu_Y)^2}{2(\sigma_X^2+\sigma_Y^2)}} + (\mu_X-\mu_Y)\operatorname{erf}\left(\frac{\mu_X-\mu_Y}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right), \tag{6.26}$$

$$\sigma_Z^2 = (\mu_X-\mu_Y)^2 + \sigma_X^2 + \sigma_Y^2 - \mu_Z^2, \tag{6.27}$$

gde je $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. Ove jednačine transformišu se kada X i Y imaju istu raspodelu u sledeće izraze:

$$\mu_Z = \frac{2\sigma}{\sqrt{\pi}},\ \sigma_Z^2 = 2\sigma^2 - \frac{4\sigma^2}{\pi} = 2(1-\frac{2}{\pi})\sigma^2. \tag{6.28}$$

6.2.2 Raspodela poretka rastojanja

Neka je N ukupan broj tačaka raspodele slučajne promenljive X. Tada je ukupan broj tačaka za slučajnu promenljivu Z, koja predstavlja apsolutnu distancu dve tačke, dat sa $N_Z = \frac{N(N-1)}{2}$. Poređajmo vrednosti realizacija promenljive Z u rastući niz $Z^{(1)} \leqslant Z^{(2)} \leqslant \cdots \leqslant Z^{(N_Z)}$. Interesuje nas da odredimo koja je raspodela (i)-te realizacije, pod uslovom da znamo da je ona baš (i)-ta. U konkretnom slučaju za $i = 1$ i za $i = N_Z$ dobićemo raspodele minimuma i maksimuma. Ovo nije ništa drugo do statistika poretka [154] promenljive Z. Funkciju raspodele

(i)-te realizacije izvodimo na sledeći način:

$$F_Z^{(i)}(x) = P(Z^{(i)} \leqslant x) \tag{6.29}$$
$$= P(Z \leqslant x | Z \text{ je } i-\text{to}) \tag{6.30}$$
$$= \frac{P(Z \leqslant x \wedge Z \text{ je } i-\text{to})}{P(Z \text{ je } i-\text{to})} \tag{6.31}$$
$$= \frac{P(Z \leqslant x) P(Z \text{ je } i-\text{to} | Z \leqslant x)}{P(Z \text{ je } i-\text{to})}. \tag{6.32}$$

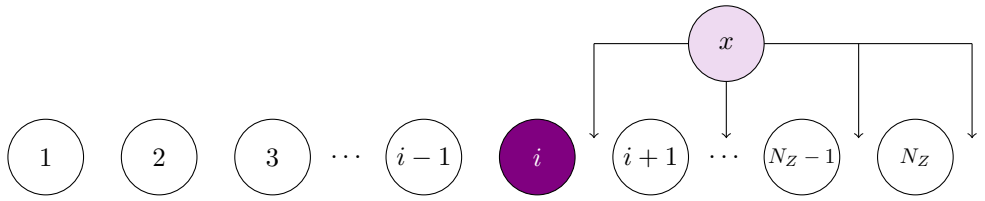
Verovatnoća da se Z nalazi na i-tom mestu jednaka je kao i za svaku drugu poziciju:

$$P(Z \text{ je } i-\text{to}) = \frac{1}{N_Z} \tag{6.33}$$

Verovatnoća da je $Z \leqslant x$ nije ništa drugo do funkcije raspodele promenljive Z:

$$P(Z \leqslant x) = F_Z(x). \tag{6.34}$$

Ostaje da odredimo verovatnoću da se Z nalazi na i-tom mestu pod uslovom da znamo da je $Z \leqslant x$. Grafički prikaz odbirka poređanih po veličini u odnosu na x dat je na Slici 6.2: Neka je



Slika 6.2: Odbirci poređani po veličini i vrednost x tako da je $Z^{(i)} \leqslant x$.

je ukupno j vrednosti Z takođe manje od x, pri čemu važi da je $i \leqslant j \leqslant N_Z$. Onim vrednostima koje su manje ili jednake x odgovara verovatnoća $F_Z(x)$, dok onim vrednostima koje su veće od x odgovara verovatnoća $[1 - F_Z(x)]$. Naravno, uslov je da su funkcije raspodele za sve Z jednake. Sledeće, želimo da odredimo broj načina da je tačno j realizacija manje ili jednako x, i da je Z tačno i-to. Pored Z, za koje znamo da je manje ili jednako x, želimo da od preostalih $N_Z - 1$ vrednosti, izaberemo još $j - 1$. Broj načina na koji ovo možemo uraditi je $\binom{N_Z-1}{j-1}$. Verovatnoća da Z bude baš na i-tom mestu tada je $\frac{1}{j}$. Ukupna verovatnoća je tada suma za sve moguće vrednosti $j \in \{i, i+1, \ldots, N_Z\}$: Izraz za verovatnoću dat je izrazom

$$P(Z \text{ je } i-\text{to} | Z \leqslant x) = \sum_{j=i}^{N_Z} \frac{1}{j} \binom{N_Z - 1}{j - 1} F_Z^{j-1}(x) \big[1 - F_Z(x)\big]^{N_Z - j}. \tag{6.35}$$

Funkcija raspodele tada dobija sledeći oblik:

$$\begin{aligned}
F_Z^{(i)}(x) &= \frac{F_Z(x)\sum_{j=i}^{N_Z}\frac{1}{j}\binom{N_Z-1}{j-1}F_Z^{j-1}(x)\big[1-F_Z(x)\big]^{N_Z-j}}{\frac{1}{N_Z}} \\
&= N_Z\sum_{j=i}^{N_Z}\frac{1}{j}\binom{N_Z-1}{j-1}F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j} \\
&= \sum_{j=i}^{N_Z}\frac{N_Z}{j}\binom{N_Z-1}{j-1}F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j} \\
&= \sum_{j=i}^{N_Z}\frac{N_Z}{j}\frac{(N_Z-1)!}{(N_Z-j)!(j-1)!}F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j} \\
&= \sum_{j=i}^{N_Z}\frac{N_Z!}{(N_Z-j)!j!}F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j} \\
&= \sum_{j=i}^{N_Z}\binom{N_Z}{j}F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j}. \qquad (6.36)
\end{aligned}$$

Time se dobija konačna forma za funkciju raspodele statistike poretka (i) slučajne promenljive Z [155]:

$$F_Z^{(i)}(x) = P(Z^{(i)} \leqslant x) = \sum_{j=i}^{N_Z}\binom{N_Z}{j}F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j} \qquad (6.37)$$

Funkcija gustine verovatnoće izvodi se na sledeći način:

$$\begin{aligned}
f_Z^{(i)} &= \frac{dF_Z^{(i)}(x)}{dx} \\
&= \sum_{j=i}^{N_Z}\binom{N_Z}{j}j\cdot F_Z^{j-1}(x)\big[1-F_Z(x)\big]^{N_Z-j}\frac{dF_Z(x)}{dx} \\
&\quad + \sum_{j=i}^{N_Z-1}\binom{N_Z}{j}(-1)(N_Z-j)\cdot F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j-1}\frac{dF_Z(x)}{dx} \\
&= \sum_{j=i}^{N_Z}\frac{N_Z!}{(N_Z-j)!j!}j\cdot F_Z^{j-1}(x)\big[1-F_Z(x)\big]^{N_Z-j}f_Z(x) \\
&\quad - \sum_{j=i}^{N_Z-1}\frac{N_Z!}{(N_Z-j)!j!}(N_Z-j)\cdot F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j-1}f_Z(x) \\
&= \sum_{j=i}^{N_Z}\frac{N_Z!}{(N_Z-j)!(j-1)!}\cdot F_Z^{j-1}(x)\big[1-F_Z(x)\big]^{N_Z-j}f_Z(x) \\
&\quad - \sum_{j=i}^{N_Z-1}\frac{N_Z!}{(N_Z-j-1)!j!}\cdot F_Z^{j}(x)\big[1-F_Z(x)\big]^{N_Z-j-1}f_Z(x) \qquad (6.38)
\end{aligned}$$

Da bi krajnja granica sumiranja bila ista u obe sume, u prvoj sumi uvodimo smenu $k = j - 1$. Time se izraz svodi na:

$$\begin{aligned} f_Z^{(i)} &= \sum_{k=i-1}^{N_Z-1} \frac{N_Z!}{(N_Z-k-1)!k!} \cdot F_Z^k(x)\big[1-F_Z(x)\big]^{N_Z-k-1} f_Z(x) \\ &\quad - \sum_{j=i}^{N_Z-1} \frac{N_Z!}{(N_Z-j-1)!j!} \cdot F_Z^j(x)\big[1-F_Z(x)\big]^{N_Z-j-1} f_Z(x) \\ &= \frac{N_Z!}{(N_Z-i)!(i-1)!} \cdot F_Z^{i-1}(x)\big[1-F_Z(x)\big]^{N_Z-i} f_Z(x) \\ &\quad + \sum_{k=i}^{N_Z-1} \frac{N_Z!}{(N_Z-k-1)!k!} \cdot F_Z^k(x)\big[1-F_Z(x)\big]^{N_Z-k-1} f_Z(x) \\ &\quad - \sum_{j=i}^{N_Z-1} \frac{N_Z!}{(N_Z-j-1)!j!} \cdot F_Z^j(x)\big[1-F_Z(x)\big]^{N_Z-j-1} f_Z(x) \end{aligned} \tag{6.39}$$

Sume u poslednjem redu jednakosti 6.39 su identične i potiru se. Izraz za funkciju gustine verovatnoće svodi se tada samo na prvi član:

$$\begin{aligned} f_Z^{(i)} &= \frac{N_Z!}{(N_Z-i)!(i-1)!} \cdot F_Z^{i-1}(x)\big[1-F_Z(x)\big]^{N_Z-i} f_Z(x) \\ &= N_Z \frac{(N_Z-1)!}{(N_Z-i)!(i-1)!} \cdot F_Z^{i-1}(x)\big[1-F_Z(x)\big]^{N_Z-i} f_Z(x) \\ &= N_Z \binom{N_Z-1}{i-1} F_Z^{i-1}(x)\big[1-F_Z(x)\big]^{N_Z-i} f_Z(x) \end{aligned} \tag{6.40}$$

Konačno, funkcija gustine verovatnoće statistike poretka i data je formulom [156]:

$$f_Z^{(i)}(x) = N_Z \binom{N_Z-1}{i-1} F_Z^{i-1}(x)\big[1-F_Z(x)\big]^{N_Z-i} f_Z(x). \tag{6.41}$$

6.2.3 Raspodela minimuma i maksimuma rastojanja

Minimum i maksimum rastojanja su statistike poretka u posebnim slučajevima kada je $i = 1$ i $i = N_Z$. Zamenom vrednosti $i = 1$, koja odgovara minimumu, u izraze 6.37 i 6.41 dobijamo

odgovarajuću funkciju raspodele $F_{Z_m}(x)$ i funkciju gustine verovatnoće $f_{Z_m}(x)$:

$$\begin{aligned} F_{Z_m}(x) = F_Z^{(i=1)}(x) &= \sum_{j=1}^{N_Z} \binom{N_Z}{j} F_Z^j(x)\big[1-F_Z(x)\big]^{N_Z-j} \\ &= \sum_{j=0}^{N_Z} \binom{N_Z}{j} F_Z^j(x)\big[1-F_Z(x)\big]^{N_Z-j} \\ &\quad - \binom{N_Z}{0}\big[1-F_Z(x)\big]^{N_Z} \\ &= (F_Z(x)+1-F_Z(x))^{N_Z} - \binom{N_Z}{0}\big[1-F_Z(x)\big]^{N_Z} \\ &= 1-\big[1-F_Z(x)\big]^{N_Z}. \end{aligned} \tag{6.42}$$

$$\begin{aligned} f_{Z_m}(x) = f_Z^{(i=1)}(x) &= N_Z\binom{N_Z-1}{0} F_Z^0(x)\big[1-F_Z(x)\big]^{N_Z-1} f_Z(x) \\ &= N_Z\big[1-F_Z(x)\big]^{N_Z-1} f_Z(x). \end{aligned} \tag{6.43}$$

Funkciju raspodelu $F_{Z_M}(x)$ i funkciju gustine verovatnoće $f_{Z_M}(x)$ za maksimum određujemo zamenom $i = N_Z$ u izraze 6.37 i 6.41:

$$\begin{aligned} F_{Z_M}(x) = F_Z^{(i-N_Z)}(x) &= \sum_{j=N_Z}^{N_Z} \binom{N_Z}{j} F_Z^j(x)\big[1-F_Z(x)\big]^{N_Z-j} \\ &= \binom{N_Z}{N_Z} F_Z^{N_Z}(x) = F_Z^{N_Z}(x). \end{aligned} \tag{6.44}$$

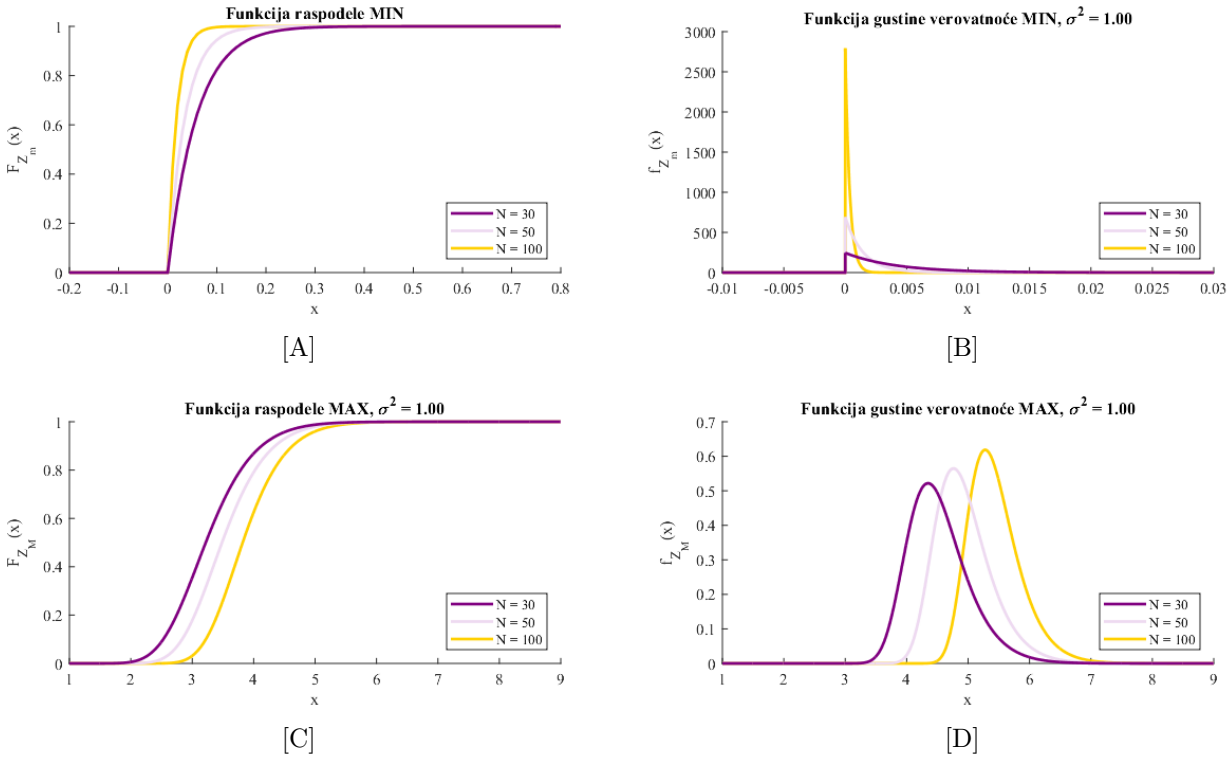
$$\begin{aligned} f_{Z_M}(x) = f_Z^{(i=N_Z)}(x) &= N_Z\binom{N_Z-1}{N_Z-1} F_Z^{N_Z-1}(x)\big[1-F_Z(x)\big]^{N_Z-N_Z} f_Z(x) \\ &= N_Z F_Z^{N_Z-1}(x) f_Z(x). \end{aligned} \tag{6.45}$$

Funkcija raspodele i funkcija gustine verovatnoće minimuma i maksimuma promenljive koja je raspodeljena sa presavijenom Gausovom raspodelom je:

$$F_{Z_m}(x) = 1-\left[1-\operatorname{erf}\left(\frac{x}{2\sigma}\right)\right]^{N_Z}, \quad f_{Z_m}(x) = \frac{N_Z}{\sqrt{\pi\sigma^2}}\left[1-\operatorname{erf}\left(\frac{x}{2\sigma}\right)\right]^{N_Z-1} e^{-\frac{x^2}{4\sigma^2}}, \tag{6.46}$$

$$F_{Z_M}(x) = \operatorname{erf}^{N_Z}\left(\frac{x}{2\sigma}\right), \qquad f_{Z_M}(x) = \frac{N_Z}{\sqrt{\pi\sigma^2}}\operatorname{erf}^{N_Z-1}\left(\frac{x}{2\sigma}\right) e^{-\frac{x^2}{4\sigma^2}}. \tag{6.47}$$

Ove funkcije grafički su prikazane na Slici(6.3) za vrednost $\sigma^2 = 1$. Podsećamo da N označava broj odbiraka X, dok je $N_Z = \frac{N(N-1)}{2}$ označava broj različitih distanci Z koje izračunaju na osnovu X.



Slika 6.3: Funkcija raspodele [A] minimuma i [C] maksimuma rastojanja i funkcija gustine verovatnoće [B] minimuma i [D] maksimuma rastojanja za $\sigma^2 = 1$.

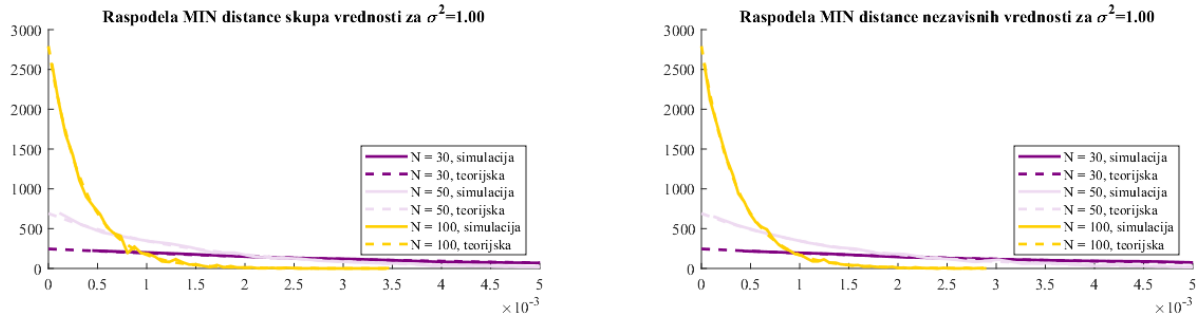
6.2.4 Eksperiment generisanja zavisnih i nezavisnih rastojanja

Eksperimentalno smo ispitali kako izgleda funkcija gustine verovatnoće za minimum i maksimum slučajne promenljive distance rastojanja dve jednako raspodeljene Gausovske varijable. Eksperiment smo sproveli na dva načina:

(1) Generisali smo jedan niz od N vrednosti Gausove raspodele zadate varijanse. Zatim smo izračunali sva međusobna rastojanja ovih vrednosti. Ukupan broj rastojanja nastalih na taj način je $N_Z = \frac{N(N-1)}{2}$. Ovako generisani odbirci rastojanja nisu nezavisni. Odredili smo maksimum i minimum rastojanja. Simulaciju smo ponovili $N_{sim} = 5000$ puta.

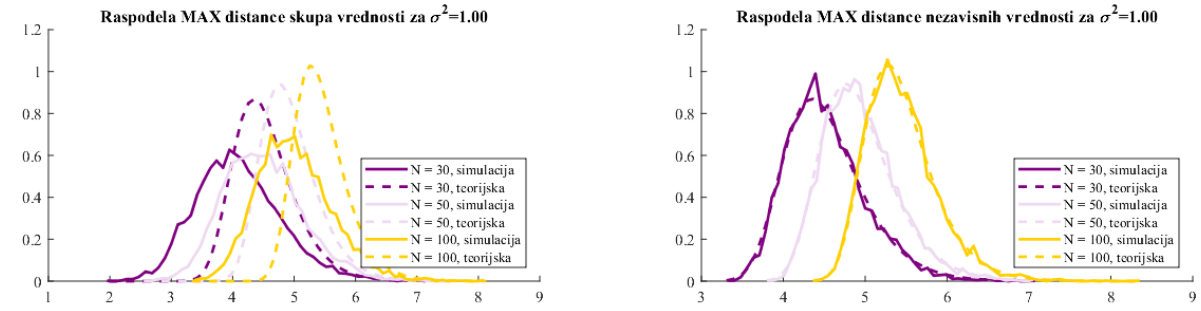
(2) Generisali smo dva niza od po $N_Z = \frac{N(N-1)}{2}$ vrednosti Gausove raspodele zadate varijanse. Zatim smo odredili N_Z distanci prema formuli $Z_i = |X_i - Y_i|$, odnosno našli smo distancu između prve vrednosti prvog niza i prve vrednosti drugog niza, zatim druge vrednosti prvog niza i druge vrednosti drugog niza itd. Distance generisane na ovaj način su međusobno nezavisne. Odredili smo minimum i maksimum. Ceo postupak ponovljen je $N_{sim} = 5000$ za zadatu vrednost N.

Rezultati simulacije za minimum rastojanja, za različite vrednosti N, upoređeni sa teorijskom raspodelom prikazani su na Slici 6.4. U slučaju maksimuma, ovi grafici prikazani su na Slici 6.5.



[A] [B]

Slika 6.4: Rezultat procene funkcije gustine verovatnoće minimuma za različite vrednost N kada su vrednosti distance [A] zavisne i [B] nezavisne.



[A] [B]

Slika 6.5: Rezultat procene funkcije gustine verovatnoće maksimuma za različite vrednost N kada su vrednosti distance [A] zavisne i [B] nezavisne.

6.2.5 Očekivanje i varijansa minimuma i maksimuma rastojanja

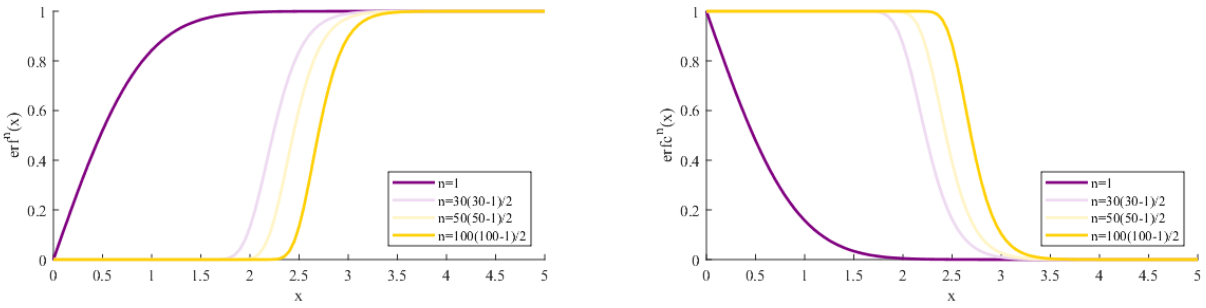
Očekivanje i varijansa slučajne promenljive X date funkcijom gustine verovatnoće $f(x)$ izračunavaju se po definiciji [157]:

$$\mu_X = E\{X\} = \int_{-\infty}^{+\infty} x f(x) dx \tag{6.48}$$

$$\sigma_X^2 = E\{(X - \mu_X)^2\} = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \tag{6.49}$$

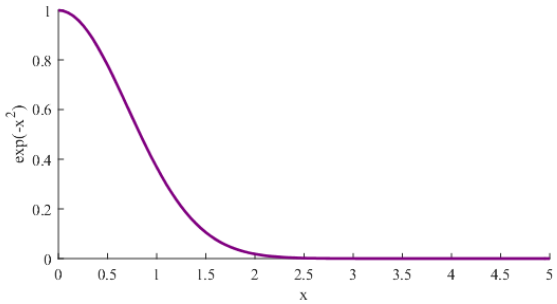
$$\tag{6.50}$$

Nije na odmet podsetiti se i skicirati funkcije e^{-x^2} i $\mathrm{erf}(x)$, jer će nam biti potrebne u kasnijoj analizi.



[A] [B]

Slika 6.6: [A] Funkcija greške i [B] komplementarna funkcija greške stepenovana na različite vrednosti.



Slika 6.7: Eksponencijalna funkcija e^{-x^2}.

Očekivanje minimuma rastojanja $E\{Z_m\}$ tada je jednako:

$$\begin{aligned} E\{Z_m\} &= \int_{-\infty}^{+\infty} x f_{Z_m}(x)dx \\ &= \int_{0}^{+\infty} x f_{Z_m}(x)dx \\ &= \int_{0}^{+\infty} x \frac{N_Z}{\sqrt{\pi\sigma^2}} \left[1 - \text{erf}\left(\frac{x}{2\sigma}\right)\right]^{N_Z-1} e^{-\frac{x^2}{4\sigma^2}} dx \end{aligned} \tag{6.51}$$

Aproksimaciju ove funkcije uradićemo numeričkom metodom integracije koristeći trapezno pravilo:

$$\int_a^b f(x)dx \approx \Delta\left[\frac{f(a)}{2} + \sum_{k=1}^{K} f(a+k\Delta) + \frac{f(b)}{2}\right], \tag{6.52}$$

gde su a i b granice integracije, K broj delova na koje je interval $[a, b]$ podeljen i Δ veličina jednog dela. Važi da je $\Delta = (b-a)/K$. Izraz za aproksimaciju očekivanja minimuma postaje:

$$\begin{aligned} E\{Z_m\} &\approx \Delta\left[\frac{0 \cdot f_{Z_m}(0)}{2} + \sum_{k=1}^{K-1} k\Delta \frac{N_Z}{\sigma\sqrt{\pi}} \left[1 - \text{erf}\left(\frac{k\Delta}{2\sigma}\right)\right]^{N_Z-1} e^{-\frac{k^2\Delta^2}{4\sigma^2}} + \frac{K\Delta \cdot f_{Z_m}(K\Delta)}{2}\right] \\ &= \frac{N_Z}{\sigma\sqrt{\pi}} \sum_{k=1}^{K-1} k\Delta^2 \left[1 - \text{erf}\left(\frac{k\Delta}{2\sigma}\right)\right]^{N_Z-1} e^{-\frac{k^2\Delta^2}{4\sigma^2}} + \frac{K\Delta \cdot f_{Z_m}(K\Delta)}{2}. \end{aligned} \tag{6.53}$$

Kada $K \to +\infty$, a $\Delta \to 0$, tada je poslednji član u izrazu teži nuli pa možemo pisati:

$$E\{Z_m\} \approx \frac{N_Z}{\sigma\sqrt{\pi}} \sum_{k=1}^{K-1} k\Delta^2 \left[1 - \text{erf}\left(\frac{k\Delta}{2\sigma}\right)\right]^{N_Z-1} e^{-\frac{k^2\Delta^2}{4\sigma^2}}. \tag{6.54}$$

Razmotrićemo sada kako se ponašaju članovi sume. Funkcija erf $\left(\frac{k\Delta}{2\sigma}\right)$ sa porastom k teži broju 1, a samim tim i za dovoljno veliku vrednost k, gde K može biti i konačan broj, važi da je:

$$\text{erf}\left(\frac{k\Delta}{2\sigma}\right) \to 1 \quad \Rightarrow \quad \left[1 - \text{erf}\left(\frac{k\Delta}{2\sigma}\right)\right] \to 0 \tag{6.55}$$

Slično važi i za eksponencijalni član:

$$e^{-\frac{k^2\Delta^2}{4\sigma^2}} \to 0. \tag{6.56}$$

Sa druge strane, za male vrednosti k, stepenovanje na $N_Z - 1$ člana koji je manji od 1 dovodi do brzog opadanja na nulu:

$$\left[1 - \text{erf}\left(\frac{k\Delta}{2\sigma}\right)\right]^{N_Z-1} \to 0, \quad e^{-\frac{k^2\Delta^2}{4\sigma^2}} < 1, \tag{6.57}$$

a da je ceo izraz proporcionalan sa Δ^2. Zaključujemo da je za dovoljno veliko:

$$E\{Z_m\} = 0. \tag{6.58}$$

Eksperimentalno smo utvrdili da se dovoljno velikim može smatrati N_Z koje odgovara međusobnim rastojanjima $N = 50$ vrednosti ($N_Z = 1225$). Na osnovu prethodno izvedenog i definicije varijanse, zaključujemo i da će varijansa minimalne distance težiti nuli odnosno:

$$var\{Z_m\} \to 0. \tag{6.59}$$

Na sličan način potražićemo i očekivanje maksimuma rastojanja:

$$\begin{aligned} E\{Z_M\} &= \int_{-\infty}^{+\infty} x f_{Z_M}(x)dx \\ &= \int_{0}^{+\infty} x f_{Z_M}(x)dx \\ &= \int_{0}^{+\infty} x \frac{N_Z}{\sqrt{\pi\sigma^2}} \,\text{erf}^{N_Z-1}\left(\frac{x}{2\sigma}\right) e^{-\frac{x^2}{4\sigma^2}} dx \end{aligned} \tag{6.60}$$

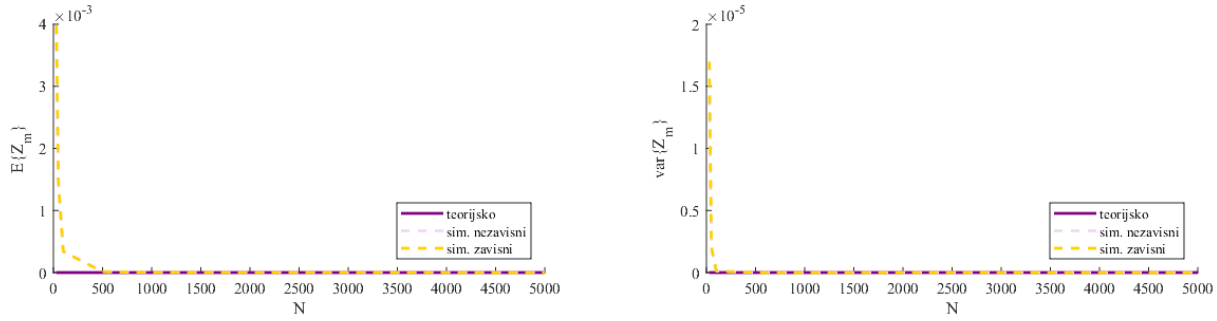
Primenom trapeznog pravila imamo da je:

$$E\{Z_M\} \approx \Delta\left[\frac{0 \cdot f_{Z_M}(0)}{2} + \sum_{k=1}^{K-1} k\Delta \frac{N_Z}{\sigma\sqrt{\pi}} \,\text{erf}^{N_Z-1}\left(\frac{k\Delta}{2\sigma}\right) e^{-\frac{k^2\Delta^2}{4\sigma^2}} + \frac{K\Delta \cdot f_{Z_M}(K\Delta)}{2}\right]. \tag{6.61}$$

Prvi član ovog izraza jednak je 0, dok je isti slučaj i za dovoljno velike vrednosti K, pa se izraz može svesti na:

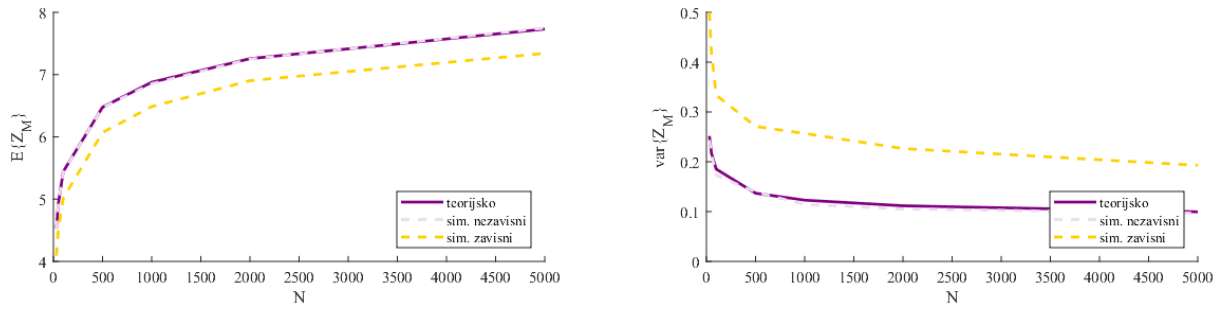
$$E\{Z_M\} \approx \sum_{k=1}^{K-1} k\Delta^2 \frac{N_Z}{\sigma\sqrt{\pi}} \,\text{erf}^{N_Z-1}\left(\frac{k\Delta}{2\sigma}\right) e^{-\frac{k^2\Delta^2}{4\sigma^2}}. \tag{6.62}$$

Zavisnost koja postoji između vrednosti $E\{Z_M\}$ i N_Z je kompleksna. Skiciraćemo takođe i vrednost varijanse maksimuma rastojanja, odkale možemo zaključiti da varijansa opada sa brojem vrednosti za koja se računaju rastojanja. Zavisnost očekivanja i varijanse od broja vrednosti N prikazani su na Slici 6.8 za minimum i na Slici 6.9 za maksimum rastojanja. Pored teorijske zavisnosti prikazali smo i rezultate simulacija sa zavisnim i nezavisnim rastojanjima, generisane prema opisu eksperimenta u sekciji 6.2.4.



[A] [B]

Slika 6.8: Zavisnost [A] očekivane vrednosti i [B] varijanse minimuma rastojanja od N broja vrednosti.



[A] [B]

Slika 6.9: Zavisnost [A] očekivane vrednosti i [B] varijanse maksimuma rastojanja od N broja vrednosti.

6.2.6 Raspodela rastojanja dve Gausova promenljive u više dimenzija

Sada ćemo potražiti raspodelu rastojanja za višedimenzione Gausove slučajne vektore. U jednodimenzionom slučaju, rastojanje dve vrednosti merili smo njihovom apsolutnom razlikom, što možemo smatrati specijalnim slučajem Euklidskog rastojanja u jednoj dimenziji. U opštem slučaju, $\boldsymbol{X} = (X_1, X_2, \ldots, X_D)$ i $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_D)$ su D-dimenzione nezavisne slučajne promenljive kojima odgovaraju $\mathcal{N}\big(\boldsymbol{\mu_X}, \boldsymbol{\Sigma_X^2}\big)$ i $\mathcal{N}\big(\boldsymbol{\mu_Y}, \boldsymbol{\Sigma_Y^2}\big)$. Razmatraćemo specijalan slučaj kada $\boldsymbol{X}$ i $\boldsymbol{Y}$ pripadaju istoj raspodeli, tj kada je $\boldsymbol{\mu_X} = \boldsymbol{\mu_Y} = \boldsymbol{\mu}$, $\boldsymbol{\Sigma_X} = \boldsymbol{\Sigma_Y} = \boldsymbol{\Sigma}$. Osim toga, dodatno ćemo uprostiti izračunavanje pretpostavkom da su varijanse po svim dimenzijama jednake i međusobno nekorelisane, odnosno da je kovarijaciona matrica data u obliku $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. Euklidska distanca promenljivih X i Y data je izrazom:

$$S = \sqrt{\sum_{d=1}^{D} Z_d^2} = \sqrt{\sum_{d=1}^{D} (X_d - Y_d)^2}. \tag{6.63}$$

Iz prethodnog izvođenja poznato nam je da je $Z_d \sim \mathcal{N}\big(0, 2\sigma^2\big)$, što nam daje mogućnost da upotrebimo rezultat iz literature koji za promenljivu S daje centralnu χ raspodelu [158, 159],

sa D stepeni slobode. Funkcija gustine verovatnoće i funkcija raspodele tada su dati sa:

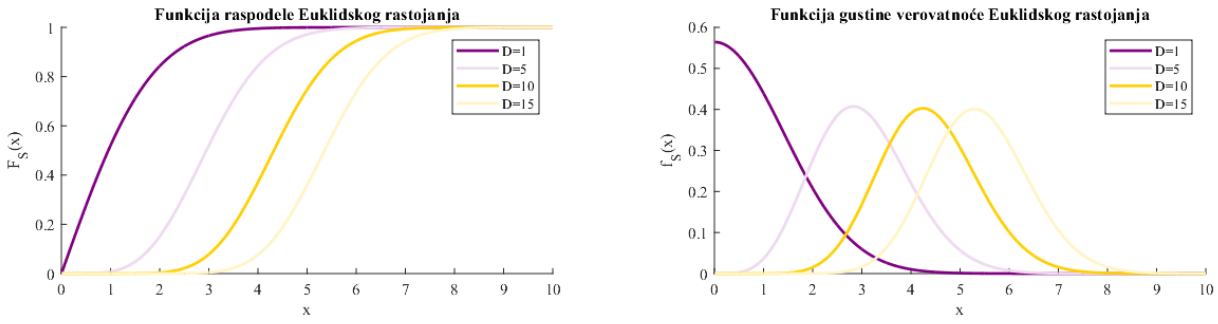
$$f_S(x) = \frac{x^{D-1}}{2^{D-1}\sigma^D \Gamma\left(\frac{D}{2}\right)} e^{-\frac{x^2}{4\sigma^2}} \tag{6.64}$$

$$F_S(x) = \frac{\gamma\left(\frac{D}{2}, \frac{x^2}{4\sigma^2}\right)}{\Gamma\left(\frac{D}{2}\right)}, \tag{6.65}$$

za $x \geqslant 0$, gde su $\Gamma\left(\frac{D}{2}\right)$ kompletna i $\gamma\left(\frac{D}{2}, \frac{x^2}{4\sigma^2}\right)$ donja nekompletna Γ funkcija [160]. Ove funkcije date su izrazima:

$$\Gamma(\alpha) = \int_0^{+\infty} \xi^{\alpha-1} e^{-\xi} d\xi, \quad \gamma(\alpha, x) = \int_0^{x} \xi^{\alpha-1} e^{-\xi} d\xi. \tag{6.66}$$

Pri tome je poznato da, ako je α ceo broj, važi da je $\Gamma(\alpha) = (\alpha - 1)!$.



[A] [B]

Slika 6.10: Funkcija [A] raspodele i [B] gustine verovatnoće za različite vrednosti dimenzije D.

6.2.7 Raspodela minimuma i maksimuma rastojanja

Korišćenjem izraza za funkciju raspodele minimuma 6.44 i funkcije gustine verovatnoće minimuma 6.45 imamo da je:

$$F_{S_m}(x) = 1 - \left[1 - F_S(x)\right]^{N_Z} = 1 - \left[1 - \frac{1}{\Gamma\left(\frac{D}{2}\right)} \gamma\left(\frac{D}{2}, \frac{x^2}{4\sigma^2}\right)\right]^{N_Z}, \tag{6.67}$$

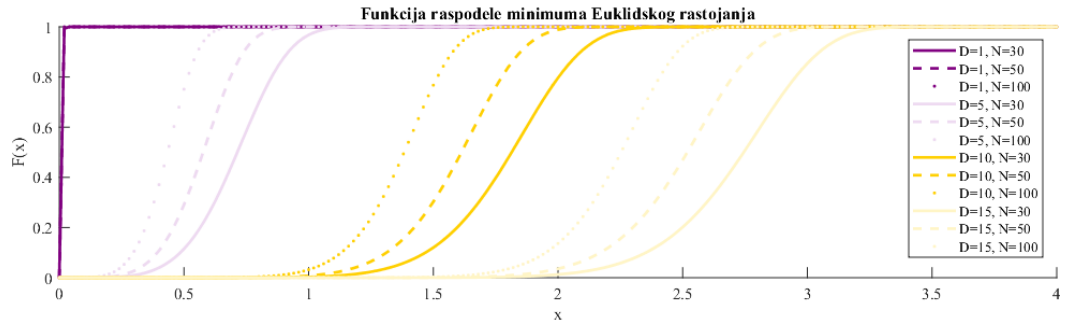
$$f_{S_m}(x) = N_Z \left[1 - \frac{1}{\Gamma\left(\frac{D}{2}\right)} \gamma\left(\frac{D}{2}, \frac{x^2}{4\sigma^2}\right)\right]^{N_Z - 1} \frac{x^{D-1}}{2^{D-1}\sigma^D \Gamma\left(\frac{D}{2}\right)} e^{-\frac{x^2}{4\sigma^2}}. \tag{6.68}$$

Korišćenjem izraza za funkciju raspodele maksimuma 6.46 i funkcije gustine verovatnoće maksimuma imamo da je 6.47:

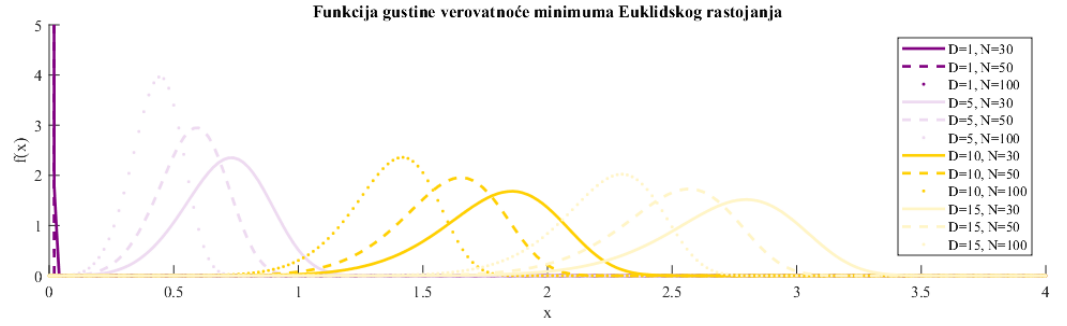
$$F_{S_M} = F_S^{N_Z}(x) = \Gamma^{-N_Z}\left(\frac{D}{2}\right) \gamma^{N_Z}\left(\frac{D}{2}, \frac{x^2}{4\sigma^2}\right), \tag{6.69}$$

$$f_{S_M}(x) = N_Z \Gamma^{-N_Z}\left(\frac{D}{2}\right) \gamma^{N_Z - 1}\left(\frac{D}{2}, \frac{x^2}{4\sigma^2}\right) \frac{x^{D-1} e^{-\frac{x^2}{4\sigma^2}}}{2^{D-1}\sigma^D} \tag{6.70}$$

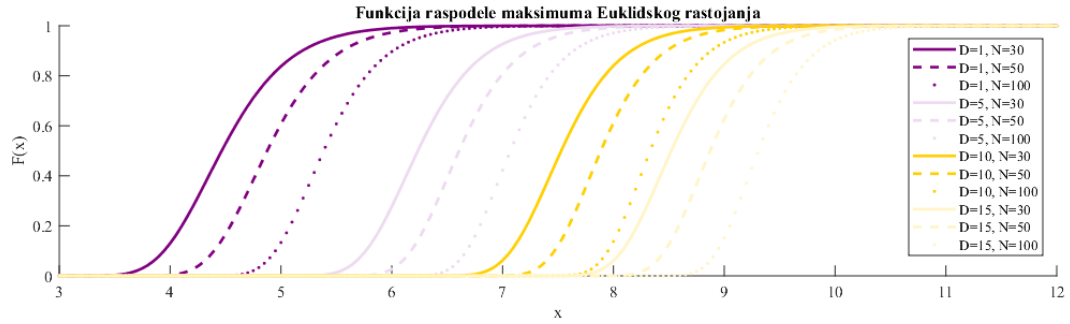
Izgled funkcije raspodele i funkcije gustine verovatnoće minimuma i maksimuma dati su na Slici 6.11.



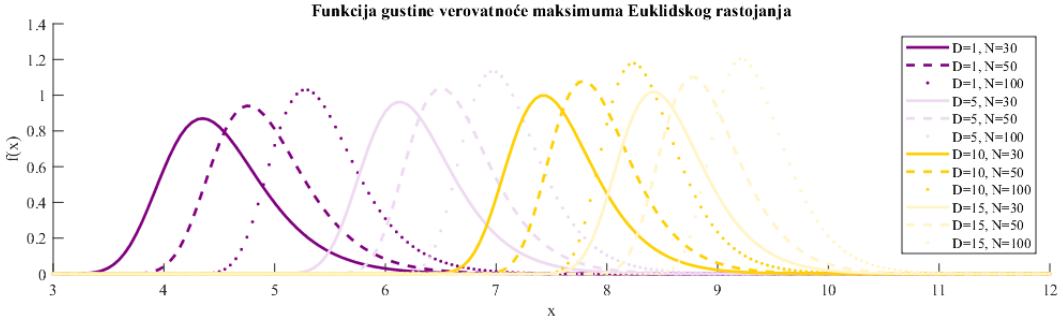
[A]



[B]



[C]



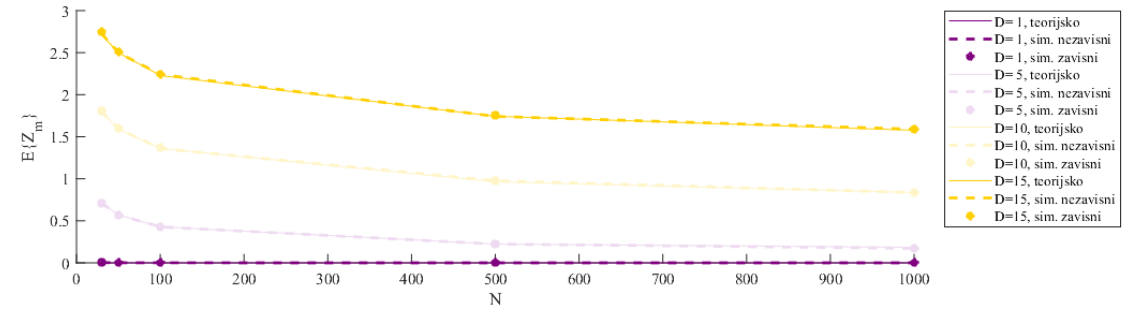
[D]

Slika 6.11: Funkcija [A] raspodele i [B] gustine verovatnoće minimuma i funkcija [C] raspodele i [D] gustine verovatnoće maksimuma Euklidskog rastojanja za različite vrednosti dimenzije D i broja vektora N, za $\sigma^2 = 1$.

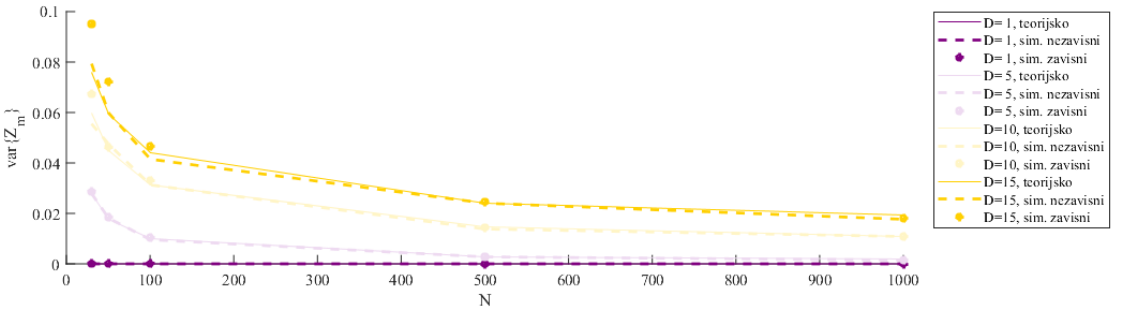
6.2.8 Očekivanje i varijansa minimuma i maksimuma rastojanja

Očekivanja minimuma i maksimuma rastu sa porsatom dimanzionalsnoti D. Varijansa minimuma raste, dok maksimuma opada. Očekivanje minimuma opada sa brojem vektora, dok u

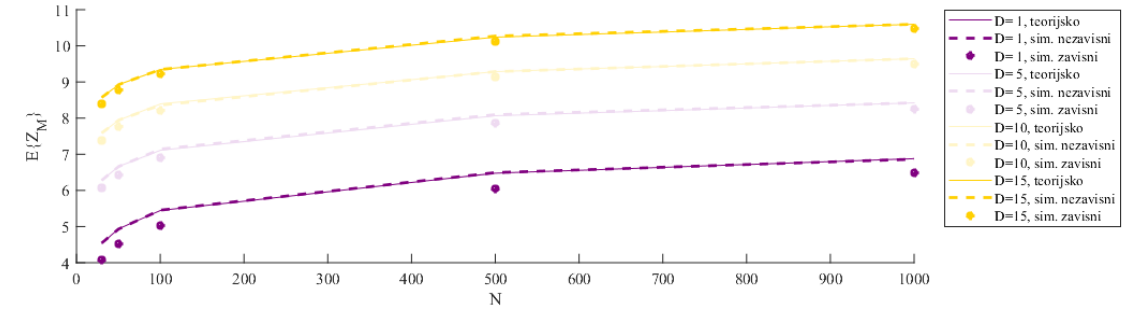
slučaju maksimuma očekivanje raste sa porastom broja vektora. Varijanse i minimuma i maksimuma opadaju sa porastom broja vektora. Zavisnost očekivanja minimuma i maksimuma od broja vektora i dimenzionalnosti nismo tražili u analitičkom obliku, već eksperimentalno. Teorijsku zavisnost smo izračunali numerički na osnovu formule funkcije gustine verovatnoće i opsega na kome je ona veća od nule, i eksperimentalno za međusobno zavisne i međusobno nezavisne vektore, kako je opisano u 6.2.4. Rezultate očekivanja i varijanse minimuma smo predstavili grafički na Slici 6.12[A] i 6.12[B] , a maksimuma na Slici 6.12[C] i 6.12[D].
Na kraju, zanimljivo je pogledati na koji način zavisi razlika očekivanja maksimuma $E\{S_M\}$ i minimuma $E\{S_m\}$ rastojanja od varijanse podataka σ^2 za različiti broj vektora. Fiksirali smo broj dimenzija $D = 13$, jer će nam kasnije biti od značaja u eksperimentima sa realnim podacima i poređenja radi za $D = 5$ - Slika 6.13. Nakon ove slike, potražili smo i izgled izraza $\frac{E\{S_M\}-E\{S_m\}}{\sigma}$ i rezultati su dati na Slici 6.14.



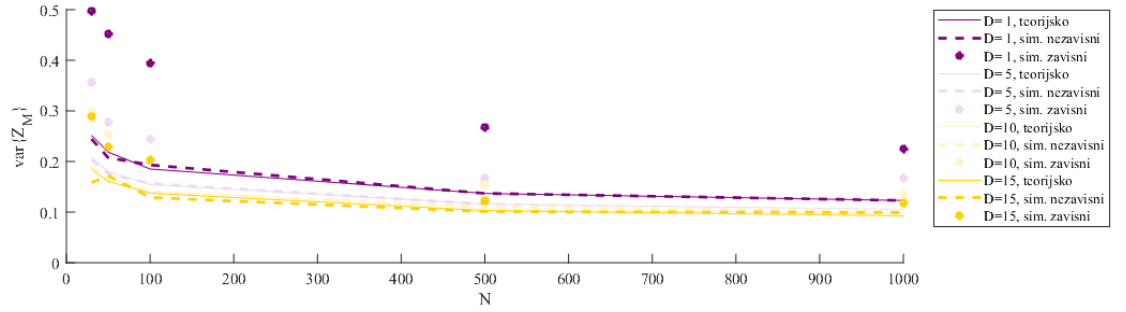
[A]



[B]

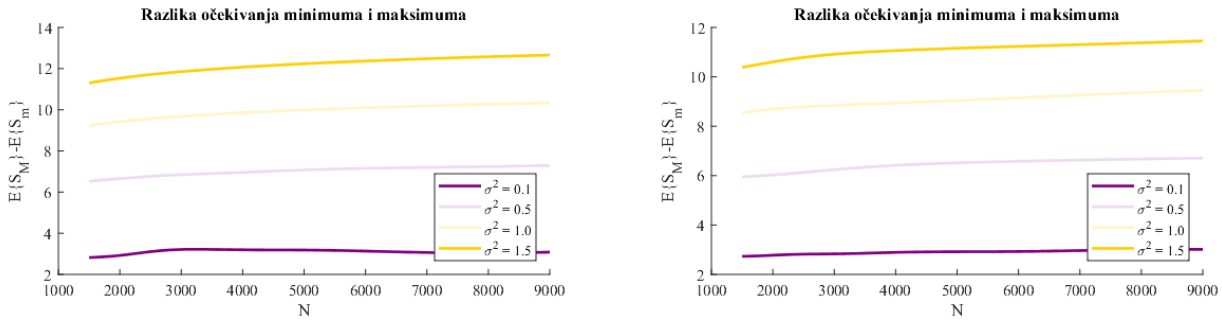


[C]



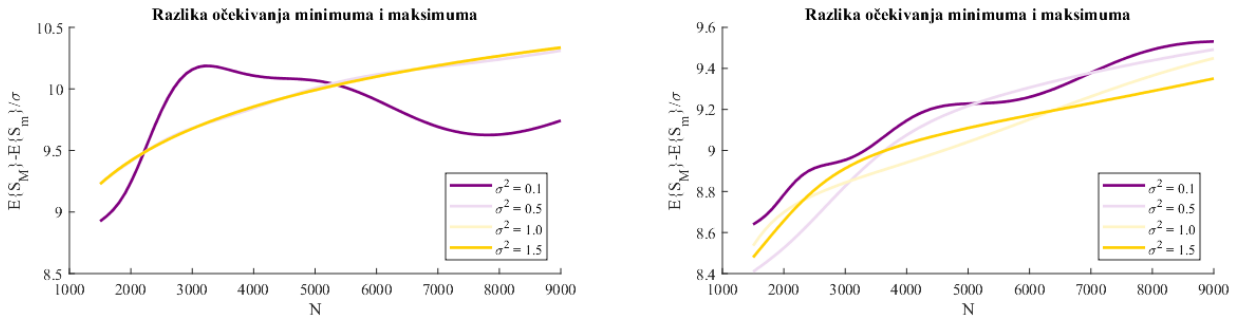
[D]

Slika 6.12: [A] Očekivanje i [B] varijansa minimuma Euklidskog rastojanja i [C] očekivanje i [D] varijansa maksimuma Euklidskog rastojanja za različite vrednosti dimenzije D za simulacije sa nezavisnim i zavisnim rastojanjima u onsnosu na teorijsku formulu, za $\sigma^2 = 1$.



[A] [B]

Slika 6.13: Razlika očekivanja maksimuma i minimuma za različite vrednosti varijanse podataka σ^2 i za [A] $D = 13$ i [B] $D = 5$.



[A] [B]

Slika 6.14: Razlika očekivanja maksimuma i minimuma podeljena sa σ za različite vrednosti varijanse podataka σ^2 i za [A] $D = 13$ i [B] $D = 5$.

6.2.9 Raspodela rastojanja dve višemodalne Gausove promenljive

Višemodlana, višedimenziona Gausova promenljiva data je funkcijom gustine verovatnoće:

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma_m}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_m})^T \boldsymbol{\Sigma_m}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_m})}, \tag{6.71}$$

gde je M broj modova, α_m verovatnoća pojavljivanja moda m, $\boldsymbol{\mu_m}$ i $\boldsymbol{\Sigma_m}$ vektor srednje vrednosti i kovarijaciona matrica m-tog moda. Pri tome važi da je $\sum_{m=1}^{M} \alpha_m = 1$. Raspodela razlike slučajnih promenljivih X i Y, koje su obe multimodalno Gausovski raspodeljene sa istom, jednodimenzionom raspodelom, takođe će biti multimodalna Gausova raspodela sa $\frac{M(M-1)}{2}$ modova. U više dimenzija $D > 1$, posmatra se Euklidsko rastojanje dva vektora. Od interesa je odrediti zavisnost očekivanja i varijanse minimuma i maksimuma tog rastojanja od broja modova u mešavini. To smo uradili eksperimentalno.

6.2.10 Raspodela minimuma i maksimuma Euklidskog rastojanja

Eksperiment je postavljen na sledeći način:

(1) Generisali smo niz koeficijenata α_m kao uniformno raspodeljene slučajne promenljive $\alpha_m \sim U[0, 1]$ i skalirali ih da ispunimo uslov $\sum_{m=1}^{M} \alpha_m = 1$.

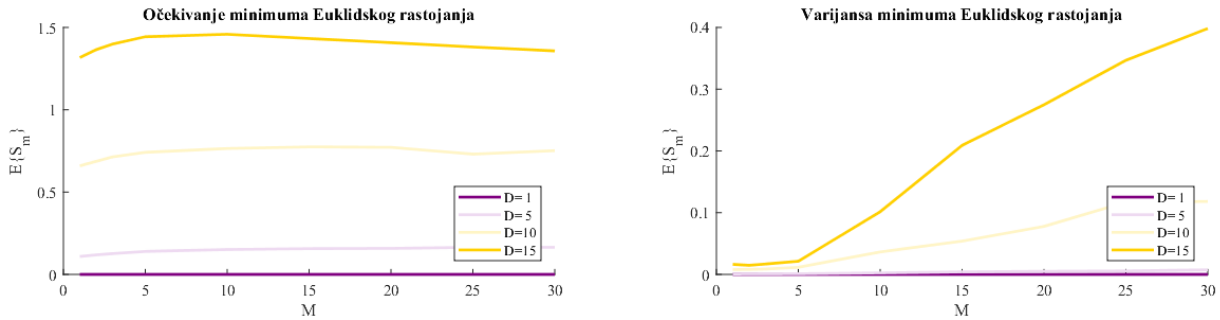
(2) Generisali smo vektore srednjih vrednosti, takođe kao uniformne slučajne promenljive sa raspodelom $\sim U[-10, 10]$.

(3) Generisali smo vektore varijanse, koji odgovaraju dijagonalnim kovarijacionim matricama, kao uniformne slučajne promenljive sa raspodelom $\sim U[0.5, 1.5]$.

(4) Za svaki od M modova generisali smo $N_m = \alpha_m N$ slučajnih vektora koji su dati raspodelom $\sim \mathcal{N}(\boldsymbol{\mu_m}, \boldsymbol{\Sigma_m})$. Tako je ukupan broj generisanih N.

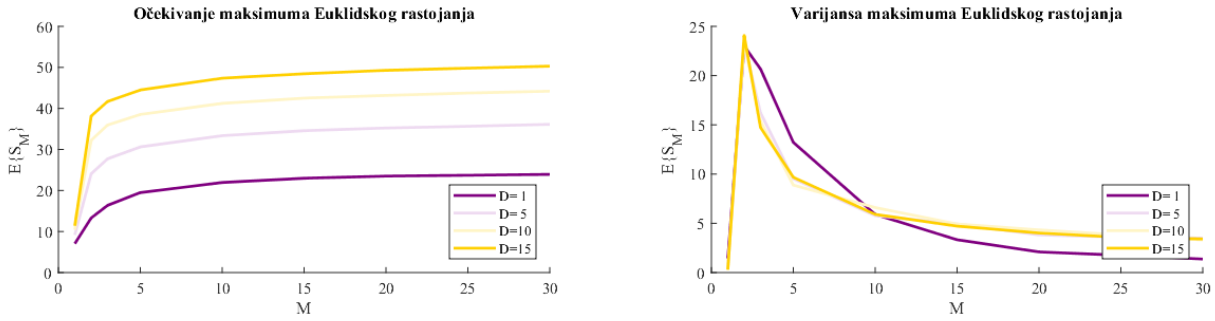
(5) Odredili smo sve međusobne distance ovako generisanih vektora i pronašli minimum i maksimum.

Ceo postupak ponovljen je $N_{sim} = 1000$ puta, za različite vrednosti M i D, a za ukupno $N = 3000$ vektora. Rezultati koje smo dobili prikazani su na Slikama 6.15 i 6.16.



[A] [B]

Slika 6.15: Zavisnost [A] očekivane vrednosti i [B] varijanse minimuma rastojanja od M broja komponenti Gausove mešavine i D, broja dimenzija.



[A] [B]

Slika 6.16: Zavisnost [A] očekivane vrednosti i [B] varijanse maksimuma rastojanja od M broja komponenti Gausove mešavine i D, broja dimenzija.

6.3 Parametri klasterizacije

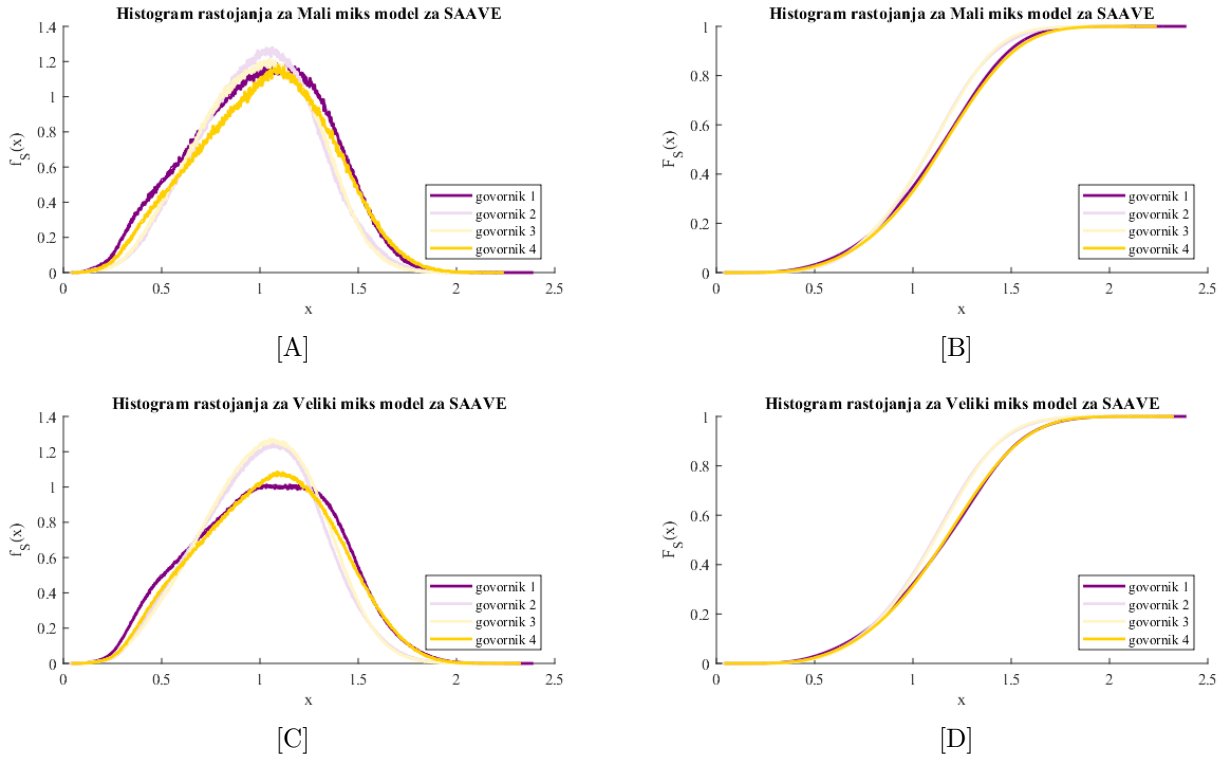
U prethodnim sekcijama izvedene su formule i skicirane zavisnosti očekivanja minimuma i maksimuma euklidskog rastojanja vektora koji uzimaju Gausovu raspodelu od dimenzija vektora D, broja vektora N_Z, broja modova M i varijanse podataka σ^2. Cilj nam je da odredimo stabilnu procenu za radijus klasterizacije r_a. Umesto formule 6.3 koja uključuje minimum i maksimum norme rastojanja merenja, predlažemo razliku njihovih očekivanja. U ovu formulu potrebno je uključiti efekat dimenzionalnosti i efekat varijanse koja se razlikuje po dimenzijama:

$$\begin{aligned} r_a &= \frac{1}{D/2}\frac{1}{\sqrt{\sum_{d=1}^{D}\sigma_d^2}}\Big[E\{S_M\} - E\{S_m\}\Big] \\ &= \frac{2}{D\sqrt{\sum_{d=1}^{D}\sigma_d^2}}\int_0^{+\infty}\Big[f_{S_M}(x) - f_{S_m}(x)\Big]xdx \\ &= \frac{2}{D\sqrt{\sum_{d=1}^{D}\sigma_d^2}}\int_0^{+\infty}\Big[N_Z F_S^{N_Z-1}(x)f_S(x) - N_Z\big[1-F_S(x)\big]^{N_Z-1}f_S(x)\Big]xdx \\ &= \frac{2N_Z}{D\sqrt{\sum_{d=1}^{D}\sigma_d^2}}\int_0^{+\infty}\Big[F_S^{N_Z-1}(x) - \big[1-F_S(x)\big]^{N_Z-1}\Big]f_S(x)xdx, \end{aligned} \tag{6.72}$$

gde je sa $F_S(x)$ označena funkcija raspodele Euklidskog rastojanja, a sa $f_S(x)$ funkcija gustine verovatnoće Euklidskog rastojanja. Ove dve funkcije nemamo u analitičkom obliku, već ćemo ih proceniti na osnovu normalizovanih merenja.
Brzina opadanja funkcije gustine subtractive klasterizacije zavisi i od faktora kompresije ϵ. Na osnovu procenjene prosečne varijanse, odnosno standardne devijacije po dimenzijama, testirali smo parametar ϵ sa vrednošću:

$$\epsilon = \sqrt{\sum_{d=1}^{D}\sigma_d^2}. \tag{6.73}$$

Algoritam se zaustavlja kada funkcija gustine dostigne donji prag u odnosu na početnu maksimalnu vrednost.



Slika 6.17: Zavisnost [A] očekivane vrednosti i [B] varijanse maksimuma rastojanja od M broja komponenti Gausove mešavine i D, broja dimenzija.

6.4 Opis eksperimenta

Evaluirali smo dva eksperimenta opisana u sekeciji 5.1 i to *Mali miks model* i *Veliki miks* model. Izmena koja uključuje subtractive klasterizaciju bila je u predobuci modela obuci - određen je optimalan broj M za svaki od GMM govornika i inicijalizacija modela GMM rezultatima klasterizacije.

6.4.1 Normalizacija podataka

Primena subtractive klasterizacije podrazumeva podatke koji su skalirani u jediničnu D-dimenzionu hiperkocku. Primenjena je **min-max** normalizacija:

$$\hat{x}_i = \frac{x_i - \min x_j}{\max x_j - \min x_j}. \tag{6.74}$$

U našem slučaju, skalirali smo $D = 13$ MFCC koeficijenata, svaku od komponenti vektora zasebno. Normalizaciju smo vršili na podacima za obuku i to na nivou jedne rečenice.

6.4.2 Procena radijusa klasterizacije

Na osnovu normalizovanih podataka, za svaki od modela izračunali smo radijus klasterizacije prema formuli 6.72. Pri tome smo koristili histogram sa 1000 delova da procenimo vrednosti funkcija raspodele, funkcija gustine verovatnoće i varijanse podataka za svaku od dimenzija. Na osnovu varijanse podataka izračunali smo ϵ prema formuli 6.73. Vrednost praga zaustavljanja algoritma postavljena je na $\delta = 0.001$. Osnovni eksperimenti sprovedeni su sa ovako postavljenim parametrima, a zatim su i ponovljeni za $\beta \in [0.3, 1]$. U sledećoj sekciji prikazani

su rezultati eksperimenata za osnovne parametre i za najbolji rezultat postignut subtractive klasterizacijom varijacijom parametra β.

6.4.3 Formiranje klastera i obuka GMM

Nakon određivanja radijusa klasterizacije r_a i r_b, na osnovu izračunatih maksimuma funkcije gustine Υ, određen je broj i centri klastera. Zatim su vektori raspodeljeni u klastere na osnovu blizine najbližem centru klastera u normalizovanom prostoru. Na kraju su klasteri sa manje od $D = 13$ vektora izbačeni iz skora. Ovi klasteri poslužili su kao inicijalni klasteri za obuku modela GMM, na standardni način. Vredi naglasiti da su vektori u obuci GMM korišćeni sa osnovnim vrednostima, pre normalizacije.

6.5 Rezultati eksperimenta

[B] Mali miks model

Prepoznavanje ostvareno *Malim miks* modelom za GMM sa brojem komponenti i klasterima određenim subtractive klasterizacijom dato je u Tabeli 6.1. Najbolji rezultat ostvaren za bazu RUSLANA je baš onaj sa osnovnim parametrima subtracive klasterizacije - uspešnost od 93.69%. Rezultati za GEES bazu su jedini ispod 90% i iznose 77.5% za osnovne parametre i 84.17% za $\beta = 0.6$. Rezultat od 95.00% prepoznavanja najbolji je za osnovne parametre za EMOVO bazu, dok kod SAVEE baze osnovni parametri daju 96.25%, a najbolji rezultat je 98.75% za $\beta = 0.5$.

Tabela 6.1: Rezultati testiranja *Malog miks modela*.

[B] *Mali miks* model						
Baza	Neutralno	Radost	Bes	Tuga	Strah	Sve
Predložena formula						
RUSLANA	**98.36**	88.93	92.21	96.72	92.21	**93.69**
GEES	**91.67**	87.75	62.5	62.5	83.33	**77.50**
EMOVO	**100.0**	100.0	83.33	100.0	91.67	**95.00**
SAVEE	**100.0**	93.75	87.50	100.0	100.0	**96.25**
Najbolji rezultat						
RUSLANA	**98.36**	88.93	92.21	96.72	92.21	**93.69**
GEES	**95.65**	83.33	79.17	70.83	91.67	**84.17**
EMOVO	**100.0**	100.0	83.33	100.0	91.67	**95.00**
SAVEE	**100.0**	100.0	100.0	100.0	93.75	**98.75**

[E] Veliki miks model

Veliki miks model ima tri puta više obučavajućih rečenica u odnosu na *Mali miks* model i to proporcionalno po emotivnim stanjima. Prepoznavanje na RUSLANA bazi ovog modela je 95.24% za osnovne parametre i 97.30% najbolji rezultat za parametar $\beta = 0.8$ subtractive klasterizacije. U slučaju GEES baze, upotrebom osnovnih parametara, model daje 93.33% tačnog prepoznavanja, što je ujedno i najbolji rezultat. Slična situacija je i sa EMOVO bazom, gde je prepoznavanje uspešno za sve test rečenice. Rezultat na SAAVE bazi nepromenjen u odnosu na *Mali miks* model 96.25%.

Tabela 6.2: Rezultati testiranja *Velikog miks modela.*

[E] *Valiki miks model*						
Baza	Neutralno	Radost	Bes	Tuga	Strah	Sve
Predložena formula						
RUSLANA	**96.72**	90.16	96.72	95.90	96.72	**95.24**
GEES	**95.83**	91.67	83.33	100	95.93	**93.33**
EMOVO	**100.0**	100.0	100.0	100.0	100.0	**100.0**
SAVEE	**100.0**	100.0	93.75	87.50	100.0	**96.25**
Najbolji rezultat						
RUSLANA	**98.77**	94.67	97.95	97.54	97.54	**97.30**
GEES	**95.83**	91.67	83.33	100	95.93	**93.33**
EMOVO	**100.0**	100.0	100.0	100.0	100.0	**100.0**
SAVEE	**100.0**	100.0	93.75	87.50	100.0	**96.25**

6.5.1 Analiza broja komponenti GMM

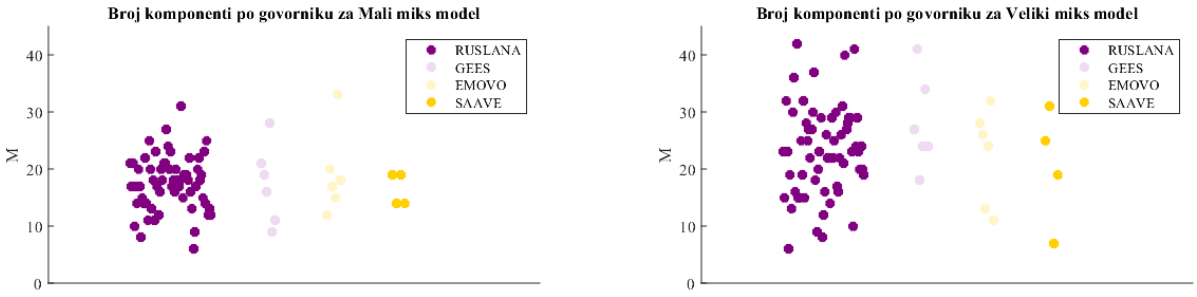
Primenom subtractive klasterizacije postigli smo da je broj komponenti GMM određen za svakog od govornika posebno. Broj komponenti po govorniku za *Mali miks* model prikazan je na Slici 6.18[A], a za *Veliki miks* model na Slici 6.18[B], dok su u Tabeli 6.3 prikazan minimalan, maksimalan broj komponenti po govorniku po bazi za oba eksperimenta, kao i prosečan broj komponenti po govorniku, po bazi za eksperimente [B] i [E]. Minimalan broj komopnenti po modelu govornika je 6, dok je maksimalan broj je 42. Prosečan broj komponenti $M = 17$, odnosno $M = 19$ za EMOVO bazu i za *Mali miks* model, dok je za *Veliki miks* model ovaj broj nešto veći i kreće se od $M = 20$ do $M = 28$.

Tabela 6.3: Minimalan i maksimalan broj komponenti po bazi podataka, kao i prosečan broj komponenti za *Mali miks* i *Veliki miks* model

Baza	min	maks	[B] prosek	[E] prosek
RUSLANA	6	42	17	23
GEES	9	41	17	28
EMOVO	11	33	19	22
SAAVE	7	31	17	20

.

Ovakav rezultat slaže se sa rezultatima koje su ostvarili Li i ostali [38], i možemo ga interpretirati kao sublimaciju informacija o različitim glasovima koje govornik izgovara, što je izvorna ideja koja stoji iza GMM. Broj mešavina je veći za *Veliki miks* model, što se može opravdati količinom podataka koja je dostupna za obuku modela. Ipak, i za jedan i za drugi model broj mešavina je svakako veći od $M = 30$ koji se zadaje za sve govornike u osnovnoj verziji algoritma.

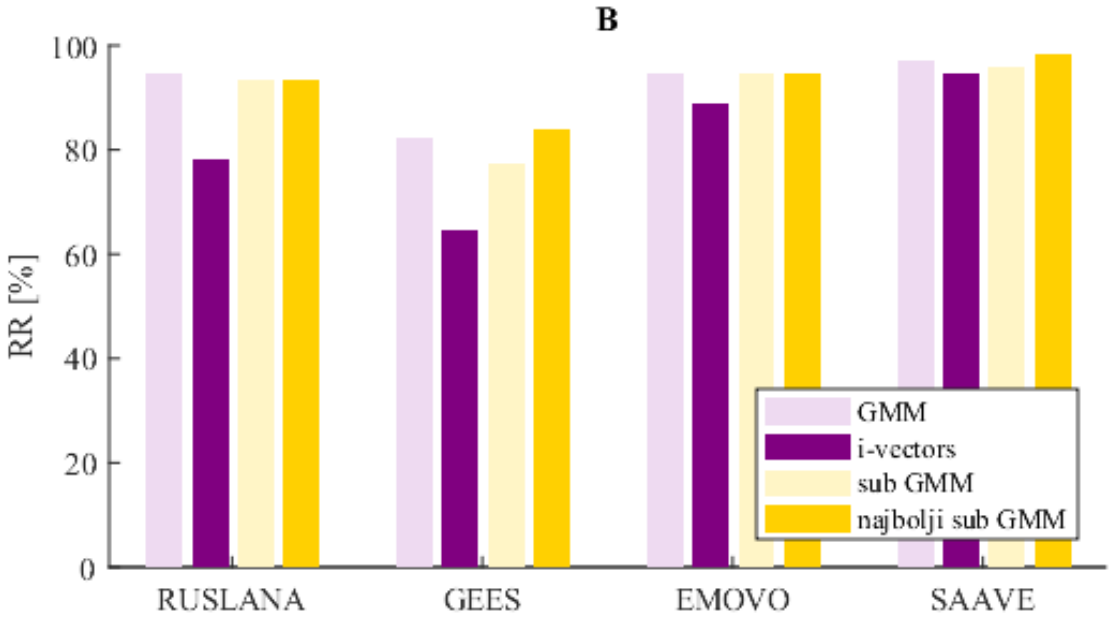


[A] [B]

Slika 6.18: Broj komponenti M u Gausovim mešavinama za govornike za [A] *Mali miks* i [B] *Veliki miks* model.

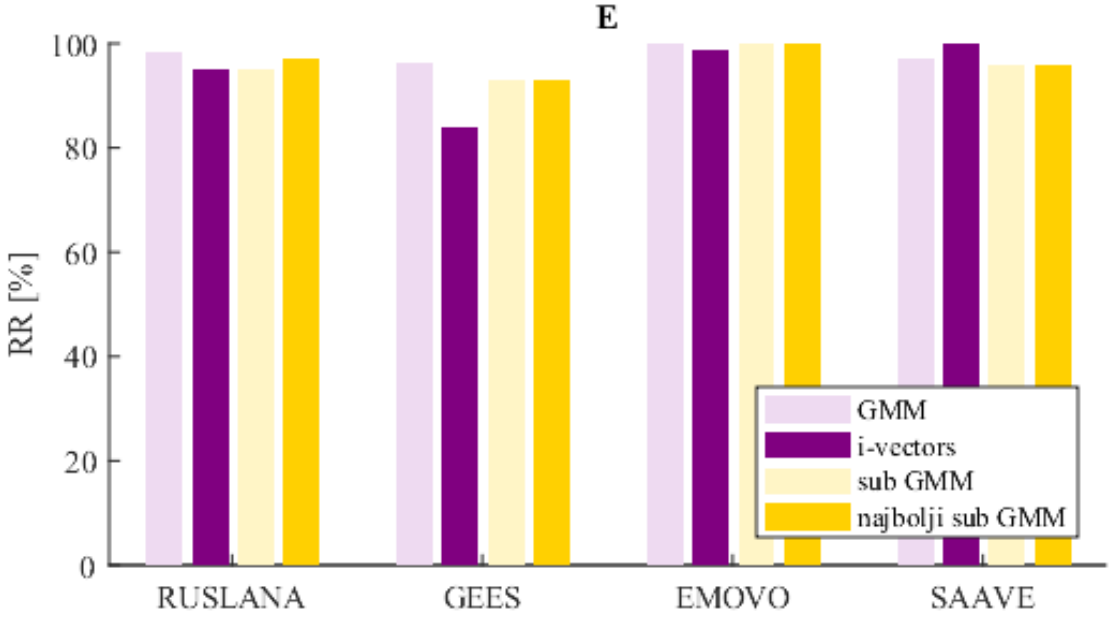
6.5.2 Poređenje sa standardnim GMM i i-vektorima

Poređenje rezultata dobijenih na osnovu subtractive klasterizacije sa standardnim GMM i i-vektorima za *Mali miks* model dato je na Slici 6.19, a za *Veliki* na Slici 6.20. Ovi modeli daju rezultate koji su u nivou sa standardnim GMM, pri čemu za opis govornika uglavnom koriste manji broj mešavina.



Slika 6.19: Rezultati *Malog miks modela* za GMM i i-vektore za sve emocije.

Interesantno je da u evaluaciji modela sa različitim vrednostima parametra β, broj komponenti GMM nije u direktnoj vezi sa kvalitetom modela - modeli sa većim brojem komponenti nisu nužno bili i najbolji za modeliranje govornika.



Slika 6.20: Rezultati *Velikog miks modela* za GMM i i-vektore za sve emocije.

6.6 Rezime i zaključci

U dosadašnjim istraživanjima parametri klasterizacije birani su apriori, bez uzimanja u obzir prirode podataka, dimenzionalnosti i količine podataka za klasterizaciju. Parametar koji dominantno utiče na ponašanje algoritma subtractive klasterizacije je poluprečnik klasterizacije r_a. U ovom delu rada, odredjemo je r_a u odnosu na skup ulaznih podataka na osnovu teorijske zavisnosti koje su prethodno izvedene za statistike gausovski raspodeljenih podataka. Konkretno, od initeresa bila je raspodela očekivanja minimuma i maksimuma rastojanja vektora, na osnovu kojih je predložena formula za r_a. Izvedene su funkcije raspodele, funkcije gustine verovatnoće i očekivanja za minimum i maksimum Euklidskog rastojanje gausovski raspodeljenih slučajnih promenljivih. Polazna tačka bio je jednodimenzioni unimodlani slučaj, zatim je analizom obuhvaćen višedimenzioni slučaj i na kraju i višemodalni. Hipoteze su verifikovane eksperimentima na veštački generisanim podacima, testiranjem i rastojanja koja su međusobno nezavisna i koja su međusobno zavisna, kao i na realnim podacima govora. Na osnovu analize može se zaključiti sledeće:

- Raspodele i očekivanja minimuma i maksimuma Euklidskog rastojanja vektora karakteristika imaju kompleksnu zavisnost od broja vektora N, dimenzije vektora D, kao i varijanse podataka σ_d.

- Očekivanje minimuma i maksimuma Euklidskog rastojanja dva vektora karakteristika skoro i da ne zavisi od broja Gausovih mešavina M koje opisuju podatke. Očekivanje minimuma praktično je konstantno za bilo koju vrednost broja mešavina, dok očekivanje maksimuma raste za vrednosti $M < 10$, nakon čega sporo raste.

Eksperimenti na realnim podacima govora za predloženu formulu r_a pokazali su sledeće:

- Primenom subtractive klasterizacije, broj komponenti u GMM bio je prosečno, u zavisnosti od baze govora, između 17 i 19 za eksperimente sa *malim miks* modelom, i od 20 do 28 za eksperimente sa *velikim miks modelom.*

- Broj komponenti po govorniku ima veću varijansu za veći broj rečenica u uzorku.

- Govornici se mogu modelirati sa različitim brojem Gausovih komponenti.
- Veći broj komponenti ne znači i bolji procenat prepoznavanja.
- Rezultati dobijeni na osnovu modela sa različitim brojem komponenti GMM dostižu najbolje rezultate za GMM i i-vektore.

Na osnovu izvođenja i rezultata eksperimenata na generisanim i realnim podacima može se tvrditi da je subtractive klasterizacija dala obećavajuće rezultate kada je reč o broju komponenti Gausove mešavine. Prostor za buduće istraživanje predstavlja način određivanja centara klastera, kao i raspodela vektora po klasterima, kroz iteracije klasterizacije kako bi što bolje modelovali gausovski raspodeljene podatke. U praksi, to bi bio postupak za preciznije određivanje parametara klasterizacije, kao i modifikacija samog iterativnog postupka kako bi centri klastera dobijenih subtractive klasterizacijom odgovarali centrima Gausovih komponenti. Značajan pomak takođe bi bilo potencijalno određivanje kvaliteta samog modela pre verifikacije u praksi.

Deo III

Zaključak, buduće istraživanje i korišćena literatura

7. Zaključak

Osnovni cilj identifikacije govornika je odgovor na pitanje "Ko je to izgovorio?". Poslednjih godina, ova oblast ponovo je dobila na značaju zahvaljujući razvoju pametnih, personalizovanih sistema [1, 2]. Identifikacija govornika jedan je od načina autentifikacije korisnika za korišćhenje servisa u svakodnevnom životu kao što su telefonsko bankarstvo, pretraga interneta i preuzimanje zaštićenih informacija [3]. Osim primene u svakodnevnom životu, identifikacija govornika je od interesa u forenzičkim istraživanjima, telefonskim servisnim centrima, centrima za hitne slučajeve, u transferu poverljivih podataka itd. U tim slučajevima, govornici uglavnom ne koriste uobičajeni ton glasa. U muzici, jedna od tema automatske analize snimaka je i prepoznavanje pevača [5, 6]. Kada je govornik uznemiren, pod stresom, plače, smeje se, šapuće, viče ili peva, karakteristike njegovog glasa su izmenjene [7]. Fenomen emotivnog govora retko je modelovan u dosadašnjem istraživanju prepoznavanja govornika. Gledano iz perspektive emotivnog stanja govornika, ključni izazov u sistemima za prepoznavanje govornika je razlika u emotivnom stanju govornika prilikom obuke sistema i trenutka kada je potrebno izvršiti prepoznavanje. Karakteristike glasa menjaju se pod uticajem emocija, što u sistemima za identifikaciju govornika, može dovesti do greške koja dalje izaziva frustraciju korisnika, a u hitnim slučajevima, naročito kada su vojne i bezbednosne primene u pitanju, može dovesti do ozbiljnih posledica. Ovi primeri ilustruju važnost zadatka prepoznavanja govornika i razvoj pouzdanih sistema koji na njega mogu odgovoriti. Cilj u dizajnu ovih sistema je robusnost, a da se pri tome koristi što manje podataka za obuku.

Ovo istraživanje, obuhvatilo je četiri nivoa zadatka prepoznavanja govornika. Analizom su obuhvaćeni do sada poznati algoritmi za modeliranje i klasifikaciju govornika. Teorijski su obrađene tehnike Gausovih mešavina, skrivenih Markovljevih modela, mašina potpornih vektora, i-vektora, dubokih neuralnih mreža i x-vektora. Eksperimentalno su evaluirane metoda standardnog modela Gausovih mešavina, kao tehnika koja je osnov modernog prepoznavanja govornika, i tehnika i-vektora koja se smatra savremenom tehnikom rasprostranjenom u komercijalnoj primeni. Zatim su sprovedeni eksperimenti sa varijacijama sadržaja govora za obuku modela govornika, koja se odnosi na korišćenje i emotivnog govora za obuku modela govornika - različit broj rečenica izgovorenih u određenom emotivnom stanju i različit ukupan broj rečenica. Jedan model govornika obučavan je sa rečenicama neutralnog govora, ali i sa rečenicama emotivnog govora - radosti, besa, tuge i straha. Nakon toga isprobane su i varijaciju konfiguracije modela govornika u smislu modeliranja govornika sa više od jednog modela, takođe na osnovu grupisanja emotivnog govora i kasnije prepoznavnja ne osnovu ovako distribuiranog modela. Na kraju su izvršene varijacije strukture modela govornika određivanjem broja mešavina za svakog od govornika na osnovu inicijalne klasterizacije. Modeli Gausovih mešavina čine os-

novu modernih sistema za prepoznavanje govornika, što ovu tehnniku čini i izuzetno važnom. Dosadašnji radovi pristupali su unapređenju GMM ili UBM-GMM algoritma određivanjem optimalnog broja mešavina na osnovu kojih su kreirani modeli za svakog od govornika, međutim i dalje sa isitm brojem komponenti za svakog govornika. U ovom istraživanju subtractive klasterizacijom određivan je broj mešavina za određenog govornika na osnovu uzoraka njegovog glasa i to na automatski način. Osim toga, teorisjki su izvedene zavisnosti parametra subtractive klasterizacije od broja vektora obučavajućeg uzorka, njihove dimenzionalnosti, kao i raspodele.

Osnovni eksperimenti koje su sprovedeni bili su sa *Neutralnim*, *Malim miks*, *Malim tri*, *Velikim tri* i *Velikim miks modelom* korišćenjem Gausovih mešavina (GMM) i i-vektora kao tehnika klasifikacije i modeliranja govornika, na bazama ruskog, srpskog, italijanskog i engleskog emotivnog govora. Korišćene karakteristike govora bile su 13 MFCC koeficijenata. orišćena je mala količina trening podataka - svega šest rečenica za *Neutralni* i *Male modele*, i 18 rečenica za *Velike modele*. Analiza je proširena određivanjem broja komponenti u GMM na osnovu subtractive klasterizacije. Na osnovu dobijenih rezultata, može se zaključiti da je uspešnost sistema za prepoznavanje govornika značajno manja u prisustvu emotivnog govora. Uključivanje emotivnog govora umesto neutralnog govora u fazi obuke sistema povećalo je procenat prepoznavanja govornika. Bolje rezultate u obuci emotivnim govorom dali su *miks modeli* u odnosu na *tri modele* i to kada za algoritma klasifikacije koriste Gausove mešavine. Iako je primena i-vektora današnji standard za prepoznavanje govornika, GMM daje stabilnije rezultate. Što se uticaja emocija tiče, ustanovljeno je da neutralni govor najmanje utiče na prepoznavanje govornika, i kada se modeliranje govornika vrši samo neutralnim govorom, i kada se za model koristi i neutralni i emotivni govor. Od ostalih emocija, bes najviše degradira prepoznavanje govornika, dok ostale emocije imaju uticaj koji se razlikuje u zavisnosti od konfiguracije korišćenog modela i baze na kojoj je eksperiment sproveden. Kada je pol u pitanju, zaključak je da su ovako konstruisani sistemi praktično podjednako pogodni i za muškarce i za žene.

U dosadašnjim istraživanjima parametri subtractive klasterizacije birani su apriori, bez uzimanja u obzir prirode podataka, dimenzionalnosti i količine podataka za klasterizaciju. Parametar koji dominantno utiče na ponašanje algoritma subtractive klasterizacije je poluprečnik klasterizacije. U ovom delu rada, odredili smo poluprečnik klasterizacije u odnosu na skup ulaznih podataka na osnovu teorijske zavisnosti koje su izvedene za statistike gausovski raspodeljenih podataka. Konkretno, od interesa je bila raspodela očekivanja minimuma i maksimuma rastojanja vektora, na osnovu kojih je predložena formula za poluprečnik klasterizacije. Izvedene su funkcije raspodele, funkcije gustine verovatnoće i očekivanja za minimum i maksimum Euklidskog rastojanje gausovski raspodeljenih slučajnih promenljivih. Polazna tačka bio je jednodimenzioni unimodlani slučaj, zatim je analizom obuhvaćen višedimenzioni slučaj i na kraju i višemodalni. Hipoteze su verifikovane eksperimentima na veštački generisanim podacima, testirajući i rastojanja koja su međusobno nezavisna i koja su međusobno zavisna, kao i na realnim podacima govora. Na osnovu analize, zaključak je da raspodele i očekivanja minimuma i maksimuma Euklidskog rastojanja vektora karakteristika imaju kompleksnu zavisnost od broja vektora, dimenzije vektora, kao i varijanse podataka. Teorijski i eksperimentalno utvrđeno je da da očekivanje minimuma i maksimuma Euklidskog rastojanja dva vektora karakteristika skoro i da ne zavisi od broja Gausovih mešavina koje opisuju podatke. Očekivanje minimuma praktično je konstantno za bilo koju vrednost broja mešavina, dok očekivanje maksimuma raste za manje od deset komponenti u mešavini, a nakon toga sporo raste. Eksperimenti na realnim podacima govora za predloženu formulu poluprečnika klasterizacije pokazali su da primenom subtractive klasterizacije, broj komponenti u GMM bio je prosečno manji u odnosu na uobičajeno zadatih trideset komponenti i u eksperimentu sa *malim miks* modelom i u eksperimentu sa *velikim miks modelom*. Zaključak je i da je broj komponenti po govorniku ima veću varijansu za veći broj rečenica u uzorku, kao i da se gvornici mogu modelirati sa različitim brojem Gausovih komponenti. Rezultati eksperimenata pokazali su da veći broj komponenti ne znači i bolji procenat

prepoznavanja, a rezultati dobijeni na osnovu modela sa različitim brojem komponenti GMM dostižu najbolje rezultate za GMM i i-vektore.
Na osnovu izvođenja i rezultata eksperimenata na generisanim i realnim podacima može se tvrditi da je subtractive klasterizacija dala obećavajuće rezultate kada je reč o broju komponenti Gausove mešavine. Prostor za buduće istraživanje predstavlja način određivanja centara klastera, kao i raspodela vektora po klasterima, kroz iteracije klasterizacije kako bi što bolje modelovali gausovski raspodeljene podatke. U praksi, to bi bio postupak za preciznije određivanje parametara klasterizacije, kao i modifikacija samog iterativnog postupka kako bi centri klastera dobijenih subtractive klasterizacijom odgovarali centrima Gausovih komponenti. Značajan pomak takođe bi bilo potencijalno određivanje kvaliteta samog modela pre verifikacije u praksi.

Literatura

[1] Z. Kozhirbayev, B. A. Erol, A. Sharipbay, and M. M. Jamshidi, "Speaker recognition for robotic control via an iot device," *2018 World Automation Congress (WAC)*, pp. 1–5, 2018.

[2] M. Milošević, Ž. Nedeljković, U. Glavitsch, and Ž. Đurović, "Speaker modeling using emotional speech for more robust speaker identification," *Journal of Communications Technology and Electronics*, vol. 64, no. 11, pp. 1256–1265, 2019.

[3] A. Alarifi, I. Alkurtass, and A. Alsalman, "Svm based arabic speaker verification system for mobile devices," *2012 International Conference on Information Technology and e-Services*, pp. 1–6, 2012.

[4] A. K. Jain, A. Ross, S. Prabhakar, *et al.*, "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, 2004.

[5] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proceedings of the 3rd international conference on music information retrieval*, vol. 13, p. 17, 2002.

[6] A. Mesaros, *Singing voice recognition for music information retrieval*. PhD thesis, Tampere University of Technology, 2012.

[7] K. R. Scherer, T. Johnstone, G. Klasmeyer, and T. Bänziger, "Can automatic speaker verification be improved by training the algorithms on emotional speech?," in *Sixth International Conference on Spoken Language Processing*, 2000.

[8] P. Univaso, "Identificación forense de hablantes en argentina: un tutorial," 2016.

[9] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *ISCA & IEEE workshop on spontaneous speech processing and recognition*, 2003.

[10] I. Pollack, J. M. Pickett, and W. H. Sumby, "On the identification of speakers by voice," *the Journal of the Acoustical Society of America*, vol. 26, no. 3, pp. 403–406, 1954.

[11] J. Shearme and J. Holmes, "An experiment concerning the recognition of voices," *Language and Speech*, vol. 2, no. 3, pp. 123–131, 1959.

[12] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 354–358, 1963.

[13] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker verification over long distance telephone lines," in *International Conference on Acoustics, Speech, and Signal Processing,*, pp. 524–527, IEEE, 1989.

[14] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 293–296, IEEE, 1990.

[15] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[16] O. Ghahabi Esfahani, *Deep learning for i-vector speaker and language recognition*. PhD thesis, Universitat Politécnica de Catalunya, 2018.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[18] K. Patrick, "Joint factor analysis of speaker and session variability: Theory and algorithms," tech. rep., 2005.

[19] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, IEEE, 2007.

[20] M. Milošević and U. Glavitsch, "Combining gaussian mixture models and segmental feature models for speaker recognition," *ISCA. Proceedings of Interspeech*, pp. 2042–2043, 2017.

[21] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[22] G. Klasmeyer, T. Johnstone, T. Bänziger, C. Sappok, and K. R. Scherer, "Emotional voice variability in speaker verification," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 213–218, 2000.

[23] M. V. Ghiurcau, C. Rusu, and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," in *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 4944–4947, IEEE, 2011.

[24] A. Mansour and Z. Lachiri, "A comparative study in emotional speaker recognition in noisy environment," in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pp. 980–986, IEEE, 2017.

[25] A. Mansour and Z. Lachiri, "Svm based emotional speaker recognition using mfcc-sdc features," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 4, pp. 538–544, 2017.

[26] T. Wu, Y. Yang, and Z. Wu, "Improving speaker recognition by training on emotion-added models," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 382–389, Springer, 2005.

[27] D. Li and Y. Yang, "Emotional speech clustering based robust speaker recognition system," in *2009 2nd International Congress on Image and Signal Processing*, pp. 1–5, IEEE, 2009.

[28] I. Shahin, "Using emotions to identify speakers," in *The 5th international workshop on signal processing and its applications (WoSPA 2008)*, 2008.

[29] I. Shahin, "Speaker identification in the shouted environment using suprasegmental hidden markov models," *Signal Processing*, vol. 88, no. 11, pp. 2700–2708, 2008.

[30] I. Shahin, "Speaker identification in emotional environments," *Iranian Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 41–46, 2009.

[31] L. Chen and Y. Yang, "Emotional speaker recognition based on model space migration through translated learning," in *Chinese Conference on Biometric Recognition*, pp. 394–401, Springer, 2013.

[32] L. Chen and Y. Yang, "Applying emotional factor analysis and i-vector to emotional speaker recognition," in *Chinese Conference on Biometric Recognition*, pp. 174–179, Springer, 2011.

[33] D. Li, Y. Yang, Z. Wu, and T. Wu, "Emotion-state conversion for speaker recognition," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 403–410, Springer, 2005.

[34] Z. Wu, D. Li, and Y. Yang, "Rules based feature modification for affective speaker recognition," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, IEEE, 2006.

[35] S. R. Krothapalli, J. Yadav, S. Sarkar, S. G. Koolagudi, and A. K. Vuppala, "Neural network based feature transformation for emotion independent speaker identification," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 335–349, 2012.

[36] D. Li, Y. Yang, and T. Huang, "Pitch envelope based frame level score reweighed algorithm for emotion robust speaker recognition," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pp. 1–4, IEEE, 2009.

[37] D. Li, Y. Yang, and W. Dai, "Cost-sensitive learning for emotion robust speaker recognition," *The Scientific World Journal*, vol. 2014, 2014.

[38] Y. Lee, K. Y. Lee, and J. Lee, "The estimating optimal number of gaussian mixtures based on incremental k-means for speaker identification," *International Journal of Information Technology*, vol. 12, no. 7, pp. 13–21, 2006.

[39] J. Wang, Z. Wang, C. Yang, N. Wang, and X. Yu, "Optimization of the number of components in the mixed model using multi-criteria decision-making," *Applied Mathematical Modelling*, vol. 36, no. 9, pp. 4227–4240, 2012.

[40] R. de Luis-García, C. Alberola-López, O. Aghzout, and J. Ruiz-Alzola, "Biometric identification systems," *Signal Processing*, vol. 83, no. 12, pp. 2539–2557, 2003.

[41] H. Beigi, *Fundamentals of Speaker Recognition*. Springer Science & Business Media, 2011.

[42] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification.," in *Interspeech*, pp. 925–929, 2013.

[43] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[44] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6, IEEE, 2015.

[45] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, pp. 3442–3446, 2017.

[46] P. López-Otero, *Improved strategies for speaker segmentation and emotional state detection*. PhD thesis, Universidade de Vigo, 2015.

[47] L. Docio-Fernandez and C. Garcia-Mateo, *Speaker Segmentation*, pp. 1277–1284. Boston, MA: Springer US, 2009.

[48] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge," in *Interspeech*, pp. 7–11, 2017.

[49] "English oxford living dictionaries."

[50] J. Bock, *Sentence production. From mind to mouth.*, pp. 181—-216. Academic Press, 1995.

[51] H. Schriefers and G. Vigliocco, "Psychology of speech production," *International Encyclopedia of the Social and Behavioral Sciences*, vol. 22, pp. 14879–14882, 2001.

[52] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Speech Prosody 2004, International Conference*, 2004.

[53] G. Hickok, "Computational neuroanatomy of speech production," *Nature Reviews Neuroscience*, vol. 13, no. 2, p. 135, 2012.

[54] V. A. Fromkin, *Speech errors as linguistic evidence*, vol. 77. Walter de Gruyter, 1984.

[55] M. F. Garrett, "Syntactic processes in sentence production," *New approaches to language mechanisms*, vol. 30, pp. 231–256, 1976.

[56] W. J. Levelt, "Producing spoken language: A blueprint of the speaker," in *The neurocognition of language*, pp. 83–122, Oxford University Press, 1999.

[57] G. S. Dell, F. Chang, and Z. M. Griffin, "Connectionist models of language production: Lexical access and grammatical encoding," *Cognitive Science*, vol. 23, no. 4, pp. 517–542, 1999.

[58] J. L. Flanagan, "Speech synthesis," in *Speech Analysis Synthesis and Perception*, pp. 204–276, Springer, 1972.

[59] S. T. Jovičić, *Govorna Komunikacija: fiziologija, psihoakustika i percepcija*. Izdavačko preduzeće NAUKA, 1999.

[60] "Acoustic theory of speech production 8."

[61] Ž. Đurović, "Beleške sa predavanja obrade i prepoznavanja govora," 2011.

[62] E. Cataldo, R. Sampaio, J. Lucero, and C. Soize, "Modeling random uncertainties in voice production using a parametric approach," *Mechanics Research Communications*, vol. 35, no. 7, pp. 454–459, 2008.

[63] E. Cataldo, C. Soize, and R. Sampaio, "Uncertainty quantification of voice signal production mechanical model and experimental updating," *Mechanical Systems and Signal Processing*, vol. 40, no. 2, pp. 718–726, 2013.

[64] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell system technical journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[65] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[66] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[67] T. F. Zheng and L. Li, *Robustness-related issues in speaker recognition*. Springer, 2017.

[68] F. Nolan, *The phonetic bases of speaker recognition*. PhD thesis, University of Cambridge, 1980.

[69] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[70] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[71] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, vol. 14. PTR Prentice Hall Englewood Cliffs, 1993.

[72] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[73] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[74] V. Dellwo, M. Huckvale, and M. Ashby, *How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification*. 2007.

[75] E. Zetterholm, "Prosody and voice quality in the expression of emotions," in *International Conference on Spoken Language Processing*, no. 1043, 1998.

[76] R. B. Blackman and J. W. Tukey, "The measurement of power spectra from the point of view of communications engineering—part i," *Bell System Technical Journal*, vol. 37, no. 1, pp. 185–282, 1958.

[77] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[78] H. Farsi and R. Saleh, "Implementation and optimization of a speech recognition system based on hidden markov model using genetic algorithm," in *2014 Iranian Conference on Intelligent Systems*, pp. 1–5, 2014.

[79] T. Nwe, S. Foo, and S. L.C.D., "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[80] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions Acoustics, Speech, Signal Processing*, vol. 28, no. 4, pp. 357—-366, 1980.

[81] I. Jokic, V. Delic, S. Jokic, and Z. Peric, "Automatic speaker recognition dependency on both the shape of auditory critical bands and speaker discriminative mfccs," *Advances in Electrical and Computer Engineering*, vol. 15, no. 4, pp. 25–33, 2015.

[82] H. Hermansky, "Perceptual linear predictive(plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[83] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[84] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, pp. 1–7, 2008.

[85] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[86] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[87] N. N. Lokhande, N. S. Nehe, and P. S. Vikhe, "Voice activity detection algorithm for speech recognition applications," in *IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012), vol. iccia*, no. 6, pp. 1–4, 2012.

[88] D. J. Broad and F. Clermont, "Formant estimation by linear transformation of the lpc cepstrum," *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 2013–2017, 1989.

[89] T. Bäckström, "Lecture notes on speech processing, fundamental frequency modelling and estimation," October 2015.

[90] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[91] D.-J. Liu and C.-T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 609–621, 2001.

[92] B. Kotnik, H. Höge, and Z. Kacic, "Evaluation of pitch detection algorithms in adverse conditions," in *Proc. 3rd international conference on speech prosody*, pp. 149–152, 2006.

[93] I. Jolliffe, *Principal component analysis*. Springer, 2011.

[94] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[95] Q. Jin and A. Waibel, "Application of lda to speaker recognition," in *Sixth International Conference on Spoken Language Processing*, 2000.

[96] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157–183, 2003.

[97] A. Paeschke, M. Kienast, W. F. Sendlmeier, *et al.*, "F0-contours in emotional speech," in *Proc. ICPhS*, vol. 99, pp. 929–933, 1999.

[98] E. Abadjieva, I. R. Murray, and J. L. Arnott, "Applying analysis of human emotional speech to enhance synthetic speech," in *Third European Conference on Speech Communication and Technology*, 1993.

[99] K. R. Alluri, S. Achanta, R. Prasath, S. V. Gangashetty, and A. K. Vuppala, "A study on text-independent speaker recognition systems in emotional conditions using different pattern recognition models," in *International Conference on Mining Intelligence and Knowledge Exploration*, pp. 66–73, Springer, 2016.

[100] A. Mansour and Z. Lachiri, "Speaker recognition in emotional context," in *Proceedings of Engineering Technology of Second International Conference on Automation, Control, Engineering and Computer Science*, pp. 122–126, 2016.

[101] V. Makarova and V. A. Petrushin, "Ruslana: A database of russian emotional utterances," in *Seventh International Conference on Spoken Language Processing*, 2002.

[102] S. T. Jovicic, Z. Kasic, M. Dordevic, and M. Rajkovic, "Serbian emotional speech database: design, processing and evaluation," in *9th Conference Speech and Computer*, 2004.

[103] V. Delic, M. Bojanic, M. Gnjatovic, M. Secujski, and S. Jovicic, "Discrimination capability of prosodic and spectral features for emotional speech recognition," *Elektronika ir Elektrotechnika*, vol. 18, no. 9, pp. 51–54, 2012.

[104] S. Haq and P. J. Jackson, "Multimodal emotion recognition," in *Machine audition: principles, algorithms and systems*, pp. 398–423, IGI Global, 2011.

[105] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in *International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 3501–3504, European Language Resources Association (ELRA), 2014.

[106] V. Dellwo, A. Leemann, and M.-J. Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[107] K. R. Scherer, "What are emotions? and how can they be measured?," *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[108] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[109] D. Reynolds, *Gaussian Mixture Models*, pp. 827–832. Boston, MA: Springer US, 2015.

[110] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[111] T. Wu, Y. Yang, Z. Wu, and D. Li, "Masc: a speech corpus in mandarin for emotion analysis and affective speaker recognition," in *2006 IEEE Odyssey-the speaker and language recognition workshop*, pp. 1–5, IEEE, 2006.

[112] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, pp. 1–32, 2013.

[113] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," 1997.

[114] P. Nayana, D. Mathew, and A. Thomas, "Comparison of text independent speaker identification systems using gmm and i-vector methods," *Procedia computer science*, vol. 115, pp. 47–54, 2017.

[115] M. Milošević and Ž. Đurović, "Emotion feature extraction for emotion classification from speech signal," in *1st International Conference on Electrical, Electronic and Computing Engineering*, 2014.

[116] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[117] M. Liberman, "Emotional prosody speech and transcripts," 2002.

[118] Z. Shan and Y. Yang, "Learning polynomial function based neutral-emotion gmm transformation for emotional speaker recognition," in *2008 19th International Conference on Pattern Recognition*, pp. 1–4, IEEE, 2008.

[119] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "Iitkgp-sesc: speech database for emotion analysis," in *International conference on contemporary computing*, pp. 485–492, Springer, 2009.

[120] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "Iitkgp-sehsc: Hindi speech corpus for emotion analysis," in *2011 International conference on devices and communications (ICDeCom)*, pp. 1–5, IEEE, 2011.

[121] Ž. Nedeljković and Ž. Đurović, "Automatsko prepoznavanje emocija na osnovu govora upotrebom skrivenih markovljevih modela," *Zbornih radova sa 59. konferencije ETRAN*, 2015.

[122] Ž. Nedeljković, M. Milošević, and Ž. Đurović, "Analysis of features and classifiers in emotion recognition systems: Case study of slavic languages," *Archives of Acoustics*, vol. 45, no. 1, p. 129–140, 2020.

[123] J.-F. Mari, J.-P. Haton, and A. Kriouile, "Automatic word recognition based on second-order hidden markov models," *IEEE Transactions on speech and Audio Processing*, vol. 5, no. 1, pp. 22–25, 1997.

[124] T. S. Polzin and A. Waibel, "Detecting emotions in speech," in *Proceedings of the CMC*, vol. 16, Citeseer, 1998.

[125] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1995.

[126] J. F. Pierna, V. Baeten, A. M. Renier, R. Cogdill, and P. Dardenne, "Combination of support vector machines (svm) and near-infrared (nir) imaging spectroscopy for the detection of meat and bone meal (mbm) in compound feeds," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 7-8, pp. 341–349, 2004.

[127] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[128] S. Besbes and Z. Lachiri, "Multi-class svm for stressed speech recognition," in *2016 2nd international conference on advanced technologies for signal and image processing (ATSIP)*, pp. 782–787, IEEE, 2016.

[129] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE signal processing letters*, vol. 13, no. 5, pp. 308–311, 2006.

[130] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[131] N. A. Hendy and H. Farag, "Emotion recognition using neural network: A comparative study," *Proceedings of World Academy of Science, Engineering and Technology*, vol. 7, no. 3, p. 791, 2013.

[132] S. Lange and M. Riedmiller, "Deep auto-encoder neural networks in reinforcement learning," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2010.

[133] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5688–5691, IEEE, 2011.

[134] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4814–4818, IEEE, 2015.

[135] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 165–170, IEEE, 2016.

[136] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.

[137] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923*, 2015.

[138] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.

[139] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1695–1699, IEEE, 2014.

[140] P. Matějka, O. Glembek, O. Novotný, O. Plchot, F. Grézl, L. Burget, and J. H. Cernocký, "Analysis of dnn approaches to speaker identification," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5100–5104, IEEE, 2016.

[141] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, pp. 999–1003, 2017.

[142] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

[143] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5796–5800, IEEE, 2019.

[144] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust speech recognition and understanding*, IntechOpen, 2007.

[145] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[146] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4253–4256, IEEE, 2012.

[147] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[148] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent & fuzzy systems*, vol. 2, no. 3, pp. 267–278, 1994.

[149] K. Hammouda and F. Karray, "A comparative study of data clustering techniques," *University of Waterloo, Ontario, Canada*, vol. 1, 2000.

[150] X.-x. Jing, L. Zhan, H. Zhao, and P. Zhou, "Speaker recognition system using the improved gmm-based clustering algorithm," in *2010 International Conference on Intelligent Computing and Integrated Systems*, pp. 482–485, IEEE, 2010.

[151] X. Cui, S. Liu, and L. Jia, "An improved method of semantic driven subtractive clustering algorithm," in *2015 IEEE 5th International Conference on Electronics Information and Emergency Communication*, pp. 232–235, IEEE, 2015.

[152] K. Demirli, S. Cheng, and P. Muthukumaran, "Subtractive clustering based modeling of job sequencing with parametric search," *Fuzzy Sets and Systems*, vol. 137, no. 2, pp. 235–270, 2003.

[153] F. Leone, L. Nelson, and R. Nottingham, "The folded normal distribution," *Technometrics*, vol. 3, no. 4, pp. 543–550, 1961.

[154] H. A. David and H. N. Nagaraja, "Order statistics," *Encyclopedia of Statistical Sciences*, 2004.

[155] K. Border, "Lecture 14: Order statistics; conditional expectation." online.

[156] J. Pitman, *Probability*. New York, Berlin and Hilderberg: Springer, 1993.

[157] B. Kovacevic and Z. Durovic, *Fundamentals of stochastic signals, systems and estimation theory: with worked examples*. Springer Publishing Company, Incorporated, 2008.

[158] S. Aja-Fernńdez and G. Vegas-Sánchez-Ferrero, "Statistical analysis of noise in mri," *Switzerland: Springer International Publishing*, 2016.

[159] N. L. Johnson, S. Kotz, and N. Balakrishnan, "Continuous univariate distributions," 1994.

[160] A. M. Legendre, *Memoire sur les integrations par arcs d'ellipse*. Imprimerie royale, 1786.

[161] J. Pittermann, A. Pittermann, and W. Minker, *Handling emotions in human-computer dialogues*. Springer, 2010.

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

[163] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[164] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for nist 2012 speaker recognition evaluation," tech. rep., SRI INTERNATIONAL MENLO PARK CA SPEECH TECHNOLOGY AND RESEARCH LAB, 2013.

[165] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, and J. H. Hansen, "Crss systems for 2012 nist speaker recognition evaluation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6783–6787, IEEE, 2013.

[166] M. Falcone and A. Gallo, "The siva speech database for speaker verification: Description and evaluation," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3, pp. 1902–1905, IEEE, 1996.

[167] A. Higgins, "Yoho speaker verification," in *Speech Research Symposium, Baltimore, MD*, 1990.

[168] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[169] B. Dropuljić, M. T. Chmura, A. Kolak, and D. Petrinović, "Emotional speech corpus of croatian language," in *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 95–100, IEEE, 2011.

[170] J. Cichosz, "Database of polish emotional speech," 2008.

[171] M. Igras and B. Ziółko, "Baza danych nagrań mowy emocjonalnej," *Studia Informatica*, vol. 34, no. 2B, pp. 67–77, 2013.

[172] P. Staroniewicz and W. Majewski, "Polish emotional speech database–recording and preliminary validation," in *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, pp. 42–49, Springer, 2009.

[173] M. Milošević, U. Glavitsch, L. He, and V. Dellwo, "Segmental features for automatic speaker recognition in a flexible software framework," in *The 25th annual conference of the International Association for Forensic Phonetics and Acoustics, at York, United Kingdom*, 2016.

Deo IV

Skraćenice, liste slika, tabela i dodaci

A. Lista skraćenica

FFT	Brza Furijeova transformacija
LPCC	Linearni prediktivni kepstralni koeficijenti
LFPC	Log-frekvencijski koeficijenti
MFCC	Mel-frekvencijski kepstralni koeficijenti
PLP	Perceptualni linearni prediktivni kepstralni koeficijenti
GMM	Gaussian Mixture Models - Gausove mešavine
HMM	Hidden Markov Models - Skriveni Markovljevi modeli
SVM	Support Vector Machines
PCA	Principal component analysis - Analiza glavnih komponeneata
JFA	Joint Factor Analysis - Analiza faktora
LDA	Linear Discriminant Analysis - Linarna disckriminaciona analiza
EM	Expectation maximization
UBM	Universal Backgrund Model - Univerzalni pozdinski model
BRG	Broj govornika
AGH	Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Poljska
ETF	Elektrotehnički fakultet Univerziteta u Beogradu
LDC	Linguistic Data Consortium - Konzorcijuma lingvističkih podataka
PLDA	Probabilistička linearna diskriminaciona analiza
VAD	Voice Activity Detector - detektor vokalne aktivnosti
BN	Bottleneck - Karakteristike uskog grla koje se izdvajaju upotrebom neuralnih mreža
RR	Procenat prepoznavanja
AIC	Akaike Information Criterion - Akaike kriterijum količine informacije
BIC	Bayesian Information Criterion - Bajesov kriterijum količine informacije
NIST	National Institute of Standards and Technology

B. Lista slika

1.1 Multidisciplinarnost istraživanja o glasu. 3
1.2 Emotivni govor i pristupi prepoznavanju govornika 4

2.1 Opšta šema kreiranja modela govornika u sistemu za prepoznavanje govornika . 10
2.2 Opšta šema testiranja sistema za prepoznavanje govornika 10

3.1 Šema generisanja govora na osnovu tri tipa informacije [52]. 14
3.2 Koraci u psiholingvističkoj fazi generisanja govora (serijski model) [56] 15
3.3 Šematski prikaz govornih organa [58] 16
3.4 Mehanički model generisanja govornog signala [63]. 18

5.1 Šema obuke modela govornika u eksperimentima [A]-[E] [2]. 37
5.2 Rezultati *Neutralnog modela* za GMM i i-vektore za sve emocije. 43
5.3 Rezultati *Malog miks modela* za GMM i i-vektore za sve emocije. 44
5.4 Rezultati *Malog tri modela* za GMM i i-vektore za sve emocije. 45
5.5 Rezultati *Velikog tri modela* za GMM i i-vektore za sve emocije. 46
5.6 Rezultati *Velikog miks modela* za GMM i i-vektore za sve emocije. 47
5.7 Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Neutralni model* govornika. 48
5.8 Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Mali miks model* govornika. 48
5.9 Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Mali tri model* govornika. 49
5.10 Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Veliki tri model* govornika. 49
5.11 Uporedni rezultati za testiranje neutralnim rečenicama i svim rerečenicama za *Veliki miks model* govornika. 50
5.12 Rezultati testiranja po eksperimentima za različite baze za [A] GMM i [B] i-vektore. 50
5.13 Rezultati po emocijama za bazu RUSLANA. 51
5.14 Rezultati po bazama, emocijama i tehnikama. 52
5.15 Rezultati po bazama, emocijama i tehnikama. 52
5.16 Rezultati po bazama, emocijama i tehnikama. 52
5.17 Rezultati po bazama i polu. 53
5.18 Tranzicije skrivenih stanja kod ergodičkih skrivenih Markovljevih modela [122]. . 56

5.19 Primer (a) linearno neseparabilnih klasa, (b) linearno separabilnih klasa (c) optimalne hipperravni koja razdvaja dve linearno separabilne klase. 58
5.20 Proces obučavanja duboke neuralne mreže [122] 60
5.21 Izdvajanje suženog (BN) seta karakteristika korišćenjem DNN [137]. 61
5.22 Izdvajanje x-vektora i klasifikacija korišćenjem DNN [141]. 62

6.1 Izgled [A] funkcije raspodele i [B] funkcije gustine verovatnoće presavijene normalne raspodele. 68
6.2 Odbirci poređani po veličini i vrednost x tako da je $Z^{(i)} \leqslant x$. 69
6.3 Funkcija raspodele [A] minimuma i [C] maksimuma rastojanja i funkcija gustine verovatnoće [B] minimuma i [D] maksimuma rastojanja za $\sigma^2 = 1$. 73
6.4 Rezultat procene funkcije gustine verovatnoće minimuma za različite vrednost N kada su vrednosti distance [A] zavisne i [B] nezavisne. 74
6.5 Rezultat procene funkcije gustine verovatnoće maksimuma za različite vrednost N kada su vrednosti distance [A] zavisne i [B] nezavisne. 74
6.6 [A] Funkcija greške i [B] komplementarna funkcija greške stepenovana na različite vrednosti. 75
6.7 Eksponencijalna funkcija e^{-x^2}. 75
6.8 Zavisnost [A] očekivane vrednosti i [B] varijanse minimuma rastojanja od N broja vrednosti. 77
6.9 Zavisnost [A] očekivane vrednosti i [B] varijanse maksimuma rastojanja od N broja vrednosti. 77
6.10 Funkcija [A] raspodele i [B] gustine verovatnoće za različite vrednosti dimenzije D. 78
6.11 Funkcija [A] raspodele i [B] gustine verovatnoće minimuma i funkcija [C] raspodele i [D] gustine verovatnoće maksimuma Euklidskog rastojanja za različite vrednosti dimenzije D i broja vektora N, za $\sigma^2 = 1$. 79
6.12 [A] Očekivanje i [B] varijansa minimuma Euklidskog rastojanja i [C] očekivanje i [D] varijansa maksimuma Euklidskog rastojanja za različite vrednosti dimenzije D za simulacije sa nezavisnim i zavisnim rastojanjima u onsnosu na teorijsku formulu, za $\sigma^2 = 1$. 81
6.13 Razlika očekivanja maksimuma i minimuma za različite vrednosti varijanse podataka σ^2 i za [A] $D = 13$ i [B] $D = 5$. 82
6.14 Razlika očekivanja maksimuma i minimuma podeljena sa σ za različite vrednosti varijanse podataka σ^2 i za [A] $D = 13$ i [B] $D = 5$. 82
6.15 Zavisnost [A] očekivane vrednosti i [B] varijanse minimuma rastojanja od M broja komponenti Gausove mešavine i D, broja dimenzija. 83
6.16 Zavisnost [A] očekivane vrednosti i [B] varijanse maksimuma rastojanja od M broja komponenti Gausove mešavine i D, broja dimenzija. 84
6.17 Zavisnost [A] očekivane vrednosti i [B] varijanse maksimuma rastojanja od M broja komponenti Gausove mešavine i D, broja dimenzija. 85
6.18 Broj komponenti M u Gausovim mešavinama za govornike za [A] *Mali miks* i [B] *Veliki miks* model. 88
6.19 Rezultati *Malog miks modela* za GMM i i-vektore za sve emocije. 88
6.20 Rezultati *Velikog miks modela* za GMM i i-vektore za sve emocije. 89

C. Lista tabela

2.1 Poređenje parametara različitih vrsta biometrijskih podataka [40]. 9

3.1 Prosečna, minimalna i maksimalna vrednost fundamentalne frekvencije [60]. . . 17

4.1 Promene karakteristika govora za emotivna stanja u odnosu na neutralno stanje. 31

5.1 Rezultati sistema za prepoznavanje govornika u uslovima emotivnog govora. . . 36
5.2 Baze emotivnog govora 36
5.3 Broj i emocije rečenica za obuku svakog od pet modela [2]. 38
5.4 Struktura skupa rečenica za testiranje. BRG je broj govornika. 38
5.5 Rezultati testiranja *Neutralnog modela*. 43
5.6 Rezultati testiranja *Malog miks modela*. 44
5.7 Rezultati testiranja *Malog tri modela*. 45
5.8 Rezultati testiranja *Velikog tri modela*. 46
5.9 Rezultati testiranja *Velikog miks modela*. 47

6.1 Rezultati testiranja *Malog miks modela*. 86
6.2 Rezultati testiranja *Velikog miks modela*. 87
6.3 Minimalan i maksimalan broj komponenti po bazi podataka, kao i prosečan broj komponenti za *Mali miks* i *Veliki miks* model 87

D.1 Najpoznatije baze podataka za prepoznavanje govornika 117
D.2 Baze emotivnog govora 121

E.1 Koeficijenti segmentalnih karakteristika za foneme i grupe fonema 124
E.2 Procenat prepoznavanja govornika na osnovu svake od korišćenih karakteristika, usrednjenih na nivou foneme i grupe fonema, kao i broj konfuzija muških i ženskih govornika. 124
E.3 Rezultati sistema za prepoznavanje govornika upotrebom segmentalnih karakteristika, GMM i hibridnog pristupa 125

D. Baze govora

Tokom sada već nekoliko decenija dugog istraživanja prepoznavanja govornika i nešto mlađe oblasti prepoznavanja emocija u govoru, korišćeni su različiti setovi podataka [161], [162]. Predstavljanje baza podataka je važno jer se rezultati ne mogu tumačiti nezavisno od skupa podataka na kome su dobijeni - broja govornika, emocija i uopšte različitosti snimaka u bazi. Neki od tih setova korišćeni su i tokom ovog istraživanja - preuzeto je i korišćeno više od deset setova podataka, sa licencom za korišćenje na lično ime i na ime Elektrotehničkog fakulteta, Univerziteta u Beogradu (ETF).

U istraživanju sistema za prepoznavanje govornika koji su robusni u odnosu na emotivno stanje govornika, evaluacija ovih sistema rađena je uglavnom na bazama emotivnog govora. Baze emotivnog govora (Tabela D.2) relativno su male u odnosu na baze podataka namenjene za prepoznavanje govornika (Tabela D.1), medjutim, u pripremi baza podataka za prepoznavanje govornika, emocije nisu uzimane u obzir. U sledećoj sekciji ova tema je objašnjena sa više detalja, a u nastavku je dat pregled i opis relevantnih setova podataka.

D.1 Baze za prepoznavanje govora i govronika

Tabela D.1: Najpoznatije baze podataka za prepoznavanje govornika

Akronim	Jezik	BRG	Tip snimka	Ref
TIMIT	Engleski	630	laboratorijski	[163]
NIST SRE	Engleski	2000+	laboratorijski	[164, 165]
SIVA	Italijanski	500	telefon	[166]
YOHO	Japanski	138	laboratorijski	[167]
RSR2015	Engleski	300	mobilni telefon	[168]
TEVOID	Nemački	50	laboratorijski	[106]

D.1.1 TIMIT - Acoustic-Phonetic Continuous Speech Corpus

Jedna od standardnih baza govora je TIMIT baza [163]. Osnovna namena podataka ove baze su akustičko-fonetička istraživanja kao i razvoj i evaluacija sistema za automatsko prepoznavanje

govora i govornika. Ova baza sastoji se od snimaka govora 630 govornika američkog engleskog jezika. Svaki od govornika čitao je 10 fonetski bogatih rečenica. Osim samih snimaka koji su 16-to bitnom .wav formatu, baza sadrži i fonetsku transkripciju. Autori baze su Massachusetts Institute of Technology (MIT), SRI International (SRI) i Texas Instruments, Inc. (TI). Govor je sniman u TI, transkripcija je rađena na MIT-u, a verifikovana i finalno pripremljena na National Institute of Standards and Technology (NIST).

D.1.2 TEVOID - Temporal Voice Idiosyncrasy baza podataka

TEVOID[1] [106] baza podataka sastoji se od neutralnog govora 25 muških i 25 ženskih govornika. Svim govornicima je maternji jezik ciriški-švajcarski nemački. Snimano je 256 pročitanih rečenica po govorniku i još nekoliko dodatnih rečenica spontanog govora. Snimci su kreirani u zvučno izolovanoj prostoriji na Univerzitetu u Cirihu, pod istim uslovima za sve govornike.

D.2 Baze emotivnog govora

D.2.1 Berlin - Baza nemačkog emotivnog govora

Baza nemačkog emotivnog govora[2] [116] poznatija kao Berlin baza ili Emo DB među prvim je i ujedno i najkorišćenijim bazama govora ovog tipa. Baza se sastoji od glasova pet ženskih i pet muških govornika. Svako od govornika izgovara po 10 rečenica (pet kraćih i pet dužih), predefinisanog sadržaja, a koje su uobičajene u svakodnevnoj komunikaciji. Govornici su rečenice izgovorali u svakom od emotivnih stanja: neutralno, bes, strah, radost, tuga, gađenje i dosada. Snimanje je vršeno sa frekvencijom odabiranja od $48kHz$, koja je finalno smanjena na $16kHz$.

D.2.2 RUSLANA - Russian Language Affective Speech Database

Baza podataka ruskog emotivnog govora, RUSLANA[3] [101] sastoji se od glasova 12 muških i 49 ženskih govornika ruskog jezika. Govornici su bili studenti fakulteta lingvistike, Univerziteta u Sankt Peterburgu. Svaki od govornika izgovarao je 10 unapred definisanih, fonetski izbalansiranih rečenica, različitih dužina. Svaka od rečenica izgovarana je u šest emotivnih stanja: bes, radost, strah, tuga, neutralno stanje i iznenađenje. Snimanje je vršeno u zvučno izolovanom studiju Odseka za fonetiku Univerziteta u Sankt Petesburgu. Korišćeni su isti uslovi za sve govornike.

D.2.3 MASC - Mandarin Affective Speech Corpus

Baza mandarinskog-kinsekog emotivnog govora [111] (MASC - Mandarin Affective Speech Corpus) Ova baza sadrži snimke 23 ženska i 45 muških govornika u pet emotivnih stanja: neutralno, bes, radost, panika i tuga. Govornici su maternji govornici kineskog jezika. Svaki od govornika izgovara pet fraza, 10 rečenica po tri puta za svaku od emocija i dva pasusa samo za neutralno stanje. Sadržaj ovog teksta pokriva sve foneme u kineskom jeziku. Snimci imaju frekvenciju odabiranja $22.05kHz$, a snimanje je rađeno u tihoj kancelariji.

[1] Baza podataka i pravo na korišćenje u istraživačke svrhe dobijeno je od autora baze prof. dr Volker Dellwo-a

[2] Baza podataka je javno dostupna

[3] Baza podataka i pravo na korišćenje u istraživačke svrhe dobijeno je od autora baze prof. dr Valery-a A. Petrushin-a.

D.2.4 GEES - Baza snimaka govorne ekspresije emocija i stavova

GEES baza[4] [103] kreirana je za potrebe istraživanja akustičkih obeležja ekspresija u govoru. Bazu čine snimci govora tri ženska i tri muška govornika, studenata završne godine Fakulteta dramskih umetnosti (FDU) u Beogradu. Svaki od govornika izgovarao je 32 reči, 30 kratkih, 30 dugih rečenica i jedan paragraf u pet emotivnih stanja: bes, radost, strah, tuga i neutralno stanje. Snimanje je rađeno u antisonornoj sobi studija FDU. Frekvencija odabiranja bila je $48kHz$ a digitalizacija je vršena na nivou 16 bita.

D.2.5 CrES - Croatian emotional speech corpus

Baza hrvatskog emotivnog govora (CrES - Croatian emotional speech corpus) [5] [169] sadrži spontani i odglumljeni govor za ukupno 341 govornika u 5 emotivnih stanja: bes, radost, strah i neutralno stanje. Rečenice su duge i kratke, pri čemu svaki od govornika izgovara drugačiju rečenicu. Snimaci imaju frekvenciju odabrianja $11.025kHz$.

D.2.6 DPES - Database of Polish Emotional Speech

Najstarija baza poljskog emotivnog govora (DPES - Database of Polish Emotional Speech) [6] [170] sastoji se od govora četiri ženska i četiri muška govornika. Matrnji jezik govornika je poljski. Svako od govornika izgovara pet rečenica dužine pet do šest reči u svakom od šest emotivnih stanja: bes, radost, strah, tuga, dosada i neutralno stanje. Snimci su napravljeni u auli Poljskog nacionalnog fakulteta za film, televiziju i pozorište u Lodzu. Frekvencija odabiranja snimaka je $44.100kHz$.

D.2.7 AGH DB - baza emotivnog govora

AGH Baza emotivnog govora [7] [171] kreirana je na AGH Univerzitetu nauke i tehnologije u Krakovu, Poljska. Sastoji se od snimaka kojima se izražava šest različitih emocija: radost, tuga, strah, iznenađenje, ironija i neutralno stanje. Glasovi u bazi su od šest žena i šest muškaraca od kojih su neki profesionalni glumci, amateri i studenti dobrovokjci. Svm govornicima je poljski jezik maternji. Svaki od govornika čitao je po 24 reči, 46 rečenica i jedan pasus za svaku od emocija. Snimci su frekvencije odabiranja $44.100kHz$, rezolucije 16 bita.

D.2.8 PESD - Polish Emotional Speech Database

Još jedna baza poljskog emotivnog govora (Polish Emotional Speech Database - PESD) [172] sadrži glasove sedam muških i šest ženskih govornika, kojima je poljski maternji jezik. Svako od govornika izgovara po 10 rečenica u svakoj oj sedam emocija: bes, radost, strah, tuga, neutralno stanje, iznenađenje i gađenje. Snimci su frekvencije odabiranja $44.100kHz$, a kreirani su u laboratorisjkom studiju.

[4] Baza podataka i pravo na korišćenje u istraživačke svrhe dobijeno je od autora baze prof. dr Slobodana Jovičić-a.

[5] Baza podataka je dostupna na zahtev uz besplatnu licencu. Elektrotehnički fakultet u Beogradu je vlasnik licence.

[6] Baza podataka je dostupna na zahtev, uz besplatnu licencu. Milana M. je vlasnik licence.

[7] Baza podataka je dostupna na zahtev, uz licencu. AGH je ustupio licencu i bazu podataka Elektrotehničkom fakultetu u Beogradu.

D.2.9 IEMOCAP - interactive emotional dyadic motion capture database

IEMOCAP baza (IEMOCAP - interactive emotional dyadic motion capture database)[8] [130]. Ova baza snimljena je u interaktivnim sesijama sa govorom, markerima lica, glave i ruku tokom spontanih i odglumljenih scenarija komunikacije. 10 učesnika snimanja su odglumili određene scenarije definisanih za različite emocije: radost, bes, tuga, frustracija i neutralno stanje. Baza sadrži oko 12 sati materijala. Jezik baze podataka je engleski.

D.2.10 EPS - Emotional Prosody Speech and Transcripts

Baza emocionalne prozodije sa transkriptom govora (EPS - Emotional Prosody Speech and Transcripts) [117] razvijena je od strane Konzorcijuma lingvističkih podataka (LDC). Ova baza sadrži audio snimke i odgovarajuće transkripte sakupljene sa ciljem da podrže istraživanja emocionalne prozodije. Glasovi pripadaju profesionalnim glumcima koji izgovaraju sematički neutralan sadržaj (datumi i brojevi) u 14 različitih emocionalnih kategorija: vrući bes, hladni bes, radost, panika, tuga, dosada, gađenje, frustracija, nervoza, očaj, interesovanje, stid, ponos, prezir. Svaki od govornika izgovara 15 rečenica u svim emotivnim stanjima. Svakom snimku pridružen je i transkript. Frekvencija odabiranja je $22.05kHz$.

D.2.11 IITKGP-SESC - Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus

Ime baze podataka dolazi po institutu na kome je nastala: Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus(IITKGP-SESC) [119]. Baza je snimljena na Telugu jeziku od glasova dvoje glumaca sa All India Radio (AIR), Vijayawada, Indija. Baza je oformljena od rečenica odlumljenih u osam emotivnih stanja. Sadržaj rečenica je emotivno neutralan.

D.2.12 SAVEE - Surrey Audio-Visual Expressed Emotion

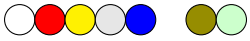
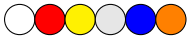
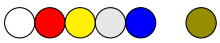
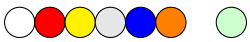
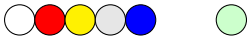
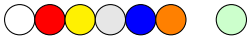
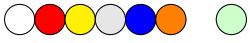
Sari baza audio-vizualnih izraza emocija (SAVEE - Surrey Audio-Visual Expressed Emotion) snimljena je sa namenom razvoja sistema za prepoznavanje emocija [104]. Bazu sačinjavaju snimci četiri muška govornika u sedam emotivnih stanja (neutralno, bes, radost, strah, tuga, iznenađenje, gađenje). Za svako do emotivnih stanja, govornik je čitao 15 rečenica, osim za neutralno za koje postoji 30 rečenica. Rečenice su bile odabrane iz standardne TIMIT [163] baze. Jezik baze je engleski. Snimanje je izvršeno u laboratoriji za vizuelne medije sa audio-vizuelnom opremom visokog kvaliteta.

D.2.13 EMOVO - Italian Emotional Speech Database

EMOVO [105] baza emotivnog govora italijanskog jezika sastoji se od glasova šest govornika/glumaca koji izgovaraju 14 rečenica u svakom od sedam emotivnih stanja (neutralno, bes, radost, strah, tuga, iznenađenje, gađenje). Snimanje je izvršeno y prostoriji laboratorija Fondazione Ugo Bordoni u Rimu. Snimci su zabeleženi sa frekvencijom odabiranja od 48 kHz, 16-bit stereo, u .wav format.

[8]Baza je javno dostupna uz licencu. Milana M. je nosilac licence.

Tabela D.2: Baze emotivnog govora

Baza	Jezik	Emocije	BRG	Tekst po govorniku	Ref.
Berlin	Nemački		10	10 predefinisanih rečenica	[116]
	Ruski		61	10 predefinisanih rečenica	[101]
MASC	Kineski		68	5 reči, 10 rečenica i 2 pasusa (neutralno stanje)	[111]
GEES	Srpski		6	30 reči, 30 kratkih, 30 dugih rečenica i jedan praragraf	[102]
CrES	Hrvatski		341	spontani govor	[169]
DPES	Poljski		8	5 rečenica	[170]
AGH DB	Poljski		12	46 rečenica, 24 reči i 1 pasus	[171]
PSED	Poljski		13	10 rečenica	[172]
IEMOCAP	Engleski		10	12 sati ukupno	[130]
EPS	Engleski		8	12 sati ukupno	[117]
IITKGP-SESC	Telugu		2		[119]
Shahin*	Engleski		40	4rečenice	[28]
EMOVO	Italijnski		6	14 rečenica	[105]
SAVEE	Engleski		4	4rečenice	[104]

Znak	Emocija	Znak	Emocija	Znak	Emocija	Znak	Emocija	Znak	Emocija
○	Neutralno	●	Iznenađenje	●	Tuga	●	Prezir	●	Ponos
●	Bes	●	Panika	●	Frustracija	●	Ironija	●	Nervoza
●	Hladni bes	●	Strah	●	Očaj	●	Gađenje	●	Stid
●	Radost	●	Dosada	●	Nervoza	●	Interesovanje	●	Saosećanje

*Baza je korišćena u datom istraživanju, nema posebnu publikaciju.

E. Hibridna klasifikacija

Gausove mešavine su danas osnova za kreiranje kompleksnijih modela za preopznavanje govornika. U našem sučaju, spajali smo ih sa klasifikacijom na osnovu segmentalnih karakteristika (SF) u kreiranju hibridnog klasifikatora [20].

E.1 Segmentalne karakteristike

Segmentalne karakteristike su usrednjenen vrednosti karakteristika na svim pojavama jedne foneme ili svim fonemamam iz određene grupe fonema (na primer zatvoreni vokali, otvoreni vokali, frikativi itd) [173]. Karakteristike koje smo upotrebljavali su:

- fundamentalna frekvencija F_0,
- promena fundamentalne frekvencije ΔF_0,
- prve četiri formantne učestanosti F_1 do F_4,
- njihove širine B_1 do B_4,
- razlika prva dva formanta $|F_2 - F_1|$,
- energija E,
- promena energije ΔE,

Ovaj pristup podrazumeva poznavanje leksičkog sadržaja govora, tj podpada pod prepoznavanje govornika u zavisnosti od teksta. Sa druge strane, za GMM to nije slučaj, već se MFCC karakteristike koje su bile ulaz u GMM koriste cele rečenice. Takođe smo eksperimentisali i sa izračunavanjima MFCC samo na delovima rečenice koji su markirani kao neka od fonema, kao i samo na vokalima, međutim nismo dobili poboljšanje u odnosu na osnovno izračunavanje [20]. Klasifikacija govornika na osnovu segmentalnih karakteristika vrši se izračunavanjem udaljenosti svake od karakteristika u test rečenicama od krarakteristike u modelu:

$$\Delta d(x, kar) = \frac{|mean_{model}(x, kar) - mean_{test}(x, kar)|}{mean_{model}(x, kar)}, x \in \{foneme, grupefonema\}. \quad \text{(E.1)}$$

Ukupan rezultat udaljenosti test rečenice od modela govornika dat je formulom:

$$D = \sum_{kar \in karakteristike} cp(kar) \cdot \sum_{p \in foneme} \Delta d(p, kar) + \tag{E.2}$$

$$+ \sum_{kar \in karakteristike} cg(kar) \cdot \sum_{g \in grupe fonema} \Delta d(g, kar) \tag{E.3}$$

gde su $cp(kar)$ i $cg(kar)$ eksperimentalno utvrđeni koeficijenti, dati tabelom E.1:

Tabela E.1: Koeficijenti segmentalnih karakteristika za foneme i grupe fonema

Karakteristika	*cp*	*cg*
E	0.18	0.10
ΔE	0.19	0.09
$F_1 - F_4$	0.55	0.15
$B_1 - B_4$	0.12	0.05
F_0	0.15	0.5
ΔF_0	0.20	0.10
$\|F2 - F1\|$	0.30	0.05

E.2 Rezultati eksperimenata

Eksperimente smo sprovodili na bazi govora TEVOID [106] u osnovnoj varijanti sa 16 govornika i manualno obeleženim granicama fonema, kao i na proširenoj varijanti ove baze sa 50 govornika i automatskim obeležavanjem granica fonema. Rezultati koje smo dobili korišćenjem svake pojedinačne karakteristike prikazani su u tabeli E.2, dok su rezultati klasifikatora na osnovu segmentalnih karakteristika, GMM i hibridne klasifikacije prikazani u tabeli E.3.

Tabela E.2: Procenat prepoznavanja govornika na osnovu svake od korišćenih karakteristika, usrednjenih na nivou foneme i grupe fonema, kao i broj konfuzija muških i ženskih govornika.

	Fonema		Grupa Fonema	
Karakteristika	PP[%]	m/ž konf.	PP[%]	m/ž konf.
E	25.88	902	25.78	926
ΔE	16.75	889	18.26	933
$F_1 - F_4$	56.69	137	45.56	323
$B_1 - B_4$	37.60	392	33.50	483
F_0	33.59	12	34.86	12
ΔF_0	14.94	502	14.65	434
$\|F2 - F1\|$	28.61	551	20.07	750

Tabela E.3: Rezultati sistema za prepoznavanje govornika upotrebom segmentalnih karakteristika, GMM i hibridnog pristupa

	TEVOID16		TEVOID50	
	[%]	m/ž konfuzije	[%]	m/ž konfuzije
SF	84.23	11	68.69	26
GMM	91.75	6	95.84	7
GMM+SF	95.12	0	95.84	1

E.3 Rezime i zaključci

Ovi rezultati pokazali su da je moguće poboljšati sisteme za prepoznavanje govornika korišćenjem leksičkog sadržaja govora. U ovom konkretnom slučaju klasifikacija na osnovu segmentalnih karakteristika poboljšava rezultate koji se dobijaju na osnovu GMM.
Pristup modeliranju govornika analizom segmentalnih karakteristika fonema značajan je kao fundamentalno istraživanje glasa. Značaj se sastoji u transparentnosti dobijenog modela. Vrednosti segmentalnih karakteristika govornika imaju fizičko značenje i samim tim bliske su ljudskom razumevanju i opažanju. Modeli govornika kao što su GMM, i-vektori, DNN, HMM, itd, u tom smislu su kao crna kutija - modeli govornika opisani hiljadama apstraktnih parametara koji nemaju fizičko značenje.

Biografija

Milana Milošević rođena je 23.05.1988. đak generacije i Vukovac u osnovnoj školi, Vukovac u Matematičkoj gimnaziji, osvajala je nagrade na takmičenjima iz matematike, fizike i informatike i predstavljala Srbiju na Međunarodnoj Olimpijadi iz Fizike 2007. godine u Iranu. Elektrotehnički fakultet, odsek za Signale i sisteme, završila je u roku – osnovne studije sa prosečnom ocenom 9.91 i master studije sa prosečnom ocenom 10. Njen diplomski rad nagrađen je trećom nagradom od strane udruženja BAFA. Aktivno je učestvovala u radu Fakulteta kao predstavnica studenata u Nastavno-naučnom veću, Savetu fakulteta i kao Studentkinja prodekanka. Doktorske studije je upisala 2012 godine na Elektrotehničkom fakultetu na odseku Upravljanje sistemima i obrada signala. Položila je sve ispite sa prosečnom ocenom 10 i trenutno je posvećena izradi doktorske disertacije. Tečno govori engleski, španski i nemački jezik. Milana Milošević usmerena je na istraživanja u oblasti mašinske obrade govora – detekciju emotivnog stanja govornika i prepoznavanje govornika u uslovima emotivnog govora. Neke od primena ovih istraživanja su u sistemima za autentifikaciju, bezbednosnim sistemima, pametnim automobilima, u forenzičke svrhe. Perspektiva za primenu je u pouzdanosti sistema – u monitoringu raspoloženja i psihofizičkog stanja inženjera i drugih uključenih u procese proizvodnje, nadzor sistema, programiranje sistema i ostalo.

Izjave

Изјава о ауторству

Име и презиме аутора Милана Милошевић

Број индекса 5026/2012

Изјављујем

да је докторска дисертација под насловом

Идентификација говорника у условима емотивног говора

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 30.05.2020.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Милана Милошевић

Број индекса 5026/2012

Студијски програм Управљање системима и обрада сигнала

Наслов рада Идентификација говорника у условима емотивног говора

Ментор др Жељко Ђуровић, редовни професор

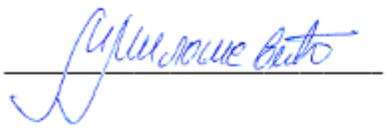
Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 30.05.2020.



Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Идентификација говорника у условима емотивног говора

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 30.05.2020.

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.