UNIVERZITET U BEOGRADU

ELEKTROTEHNIČKI FAKULTET

Đorđe T. Grozdić

PRIMENA NEURALNIH MREŽA U PREPOZNAVANJU ŠAPATA

doktorska disertacija

Beograd, 2017

UNIVERZITET U BEOGRADU

ELEKTROTEHNIČKI FAKULTET

Đorđe T. Grozdić

PRIMENA NEURALNIH MREŽA U PREPOZNAVANJU ŠAPATA

doktorska disertacija

Beograd, 2017

UNIVERSITY OF BELGRADE

SCHOOL OF ELECTRICAL ENGINEERING

Đorđe T. Grozdić

APPLICATION OF NEURAL NETWORKS IN WHISPERED SPEECH RECOGNITION

Doctoral Dissertation

Belgrade, 2017

PODACI O MENTORU I ČLANOVIMA KOMISIJE:

MENTOR:

dr Dragana Šumarac Pavlović, vanredni profesor
(Univerzitet u Beogradu, Elektrotehnički fakultet)

ČLANOVI KOMISIJE:

dr Dragana Šumarac Pavlović, vanredni profesor
(Univerzitet u Beogradu, Elektrotehnički fakultet)

dr Miomir Mijić, redovni profesor
(Univerzitet u Beogradu, Elektrotehnički fakultet)

dr Miško Subotić, naučni saradnik
(Centar za unapređenje životnih aktivnosti)

dr Goran Kvaščev, docent
(Univerzitet u Beogradu, Elektrotehnički fakultet)

DATUM ODBRANE: _____

ZAHVALNOST

Želim da izrazim zahvalnost svojim najbližim kolegama i saradnicima, koji su doprineli izradi ove disertacije.

Pre svih, veliku zahvalnost dugujem profesoru Slobodanu Jovičiću, svom dugogodišnjem mentoru, pored koga sam se prvi put upustio u multidisciplinarna istraživanja govora i digitalnu obradu govornih signala, na nesebičnoj pomoći i korisnim sugestijama, kao i na prenetom znanju tokom svih ovih godina. Profesor Jovičić je kao mentor rukovodio izradom mog diplomskog rada, master teze, a do svoje penzije i najvećim delom ove doktorske disertacije.

Izuzetno sam zahvalan sadašnjem mentoru, profesorki Dragani Šumarac, i profesoru Miomiru Mijiću, na stručnoj pomoći i brojnim korisnim savetima sa kojima su me usmeravali tokom doktorskih studija, kao i na ukazanom poverenju tokom svih ovih godina u radu u Laboratoriji za akustiku na Elektrotehničkom fakultetu.

Na kraju, zahvaljujem se svojoj porodici, pre svega roditeljima i sestri na pruženoj ljubavi, razumevanju i podršci koja mi je u mnogome pomogla da istrajem u svojim ciljevima.

Naslov: Primena neuralnih mreža u prepoznavanju šapata

REZIME

Nedavno postignuti uspesi dubinskih neuralnih mreža u različitim zadacima mašinskog učenja su doprineli da vestačke neuralne mreze ponovo zauzmu bitnu ulogu u automatskom prepoznavanju govora. U ovom doktoratu je ispitana primena vestačkih neuralnih mreza u prepoznavanju šapata. Šapat kao specifican oblik govora predstavlja ozbiljan problem za aktuelne sisteme automatskog prepoznavanja govora. Naime, neusaglašeni scenariji u kojima je sistem obučen sa neutralnim govorom a testiran sa šapatom daju veoma loše rezultate prepoznavanja. U eksperimentima ovog doktorata su demonstrirani različiti pristupi sa kojima se ovaj degradirani uspeh u automatskom prepoznavanju izolovanih reči može poboljšati primenom neuralnih mreža. Predlozena su dva nova sistema bazirana na neuralnim mrezama – jedan na višeslojnim perceptronima, (MLP-IF sistem) a drugi na dubinskim neuralnim mrežama (DNN-HMM sistem). Oba sistema u svojim *front-end* delovima imaju zadatak da reše ključni problem smanjenja akustičkih razlika između neutralnog govora i šapata. MLP-IF sistem u tu svrhu koristi inverzno filtriranje čime se potiskuje zvučnost iz govora koji nakon obrade po svojim spektralnim karakteristikama postaje sličniji šapatu. Drugi sistem, koji je u formi tandem DNN-HMM sistema, isti zadatak rešava na suprotan način tako što iz šapata rekonstruiše karakteristike neutralnog govora. Naime, u osnovi *front-end* dela DNN-HMM sistema je poseban tip dubinskih neuralnih mreža, poznat kao dubinski *denoising* autoenkoder koji zahvaljujući svojoj dubinskoj strukturi na efikasan način vrši rekonstrukciju kepstralnih karakteristika neutralnog govora iz šapata. Oba sistema su testirana u različitim obuka/test scenarijima sa tri različita tipa kepstralnih koeficijenata: MFCC (*Mel-Frequency Cepstral Coefficients*), TECC (*Teager-Energy Cepstral Coefficients*) i TEMFCC (*Teager-based Mel-Frequency Cepstral Coefficients*), pri čemu su se TECC govorna obeležja pokazala kao najpodesnija za prepoznavanje šapata. MLP-IF sistem zahvaljujući inverznom filtriranju i TECC obeležjima ostvaruje poboljšanje uspeha prepoznavanja izolovanih reči u šapatu

za 12,5% u odnosu na klasični MLP sistem sa MFCC obeležjima. Najbolje performanse je poseduje tandem DNN-HMM sistem koji sa TECC obeležjima poboljšava prepoznavanje šapata za 31% u poređenju sa tradicionalnim HMM sistemom i ostvaruje tačnost od 92,81% u prepoznavanju reči izgovorenih u šapatu.

Ključne reči: automatsko prepoznavanje govora, šapat, veštačke neuralne mreže
Naučna oblast: elektrotehnika
Uža naučna oblast: akustika
UDK broj: 621.3

Title: Application of neural networks in whispered speech recognition

SUMMARY

The recent success of Deep Neural Networks (DNN) in different machine learning tasks has significantly contributed to the rise in the popularity of artificial neural networks (ANN) and their today's role in Automatic Speech Recognition (ASR). This thesis examines how artificial neural networks can benefit in automatic whispered speech recognition. Whisper, as a specific form of verbal communication, represents one of the most challenging problems in current automatic speech recognition systems. Namely, the performance of traditional ASR systems trained on neutral speech degrades significantly when a whisper is applied. The experiments of this thesis present different approaches based on the application of neural networks that improve isolated word recognition under whispered speech condition. Two systems based on neural networks are proposed – one that is based on multilayer perceptrons (MLP-IF system), and another that is based on a deep neural network (DNN-HMM system). Both systems have the same task to solve in their *front-ends* and that is to alleviate the acoustic mismatch between neutral speech and whisper. For that purpose, MLP-IF system uses inverse filtering which suppresses sonority from a neutral speech which after such treatment becomes more similar to whisper in terms of their spectral characteristics. The second system, which is in the form of tandem DNN-HMM system, solves the same task in an opposite way by reconstructing characteristics of neutral speech from whispered speech. Namely, the basis of the *front-end* part of the DNN-HMM system is a special type of deep neural network, known as deep *denoising* autoencoder, which thanks to its deep structure effectively performs reconstruction of cepstral characteristics of neutral speech from whispered samples. Both systems were tested in different train/test scenarios with three types of cepstral features: MFCC (*Mel-Frequency Cepstral Coefficients*), TECC (*Teager-Energy Cepstral Coefficients*) and TEMFCC (*Teager-based Mel-Frequency Cepstral Coefficients*). The TECC features proved to be the most suitable for whisper recognition. Due to inverse filtering and the

application of TECC features, MLP-IF system improves isolated word recognition in whispered speech by 12.5% compared to the traditional MLP system with MFCC features. The best performance has the tandem DNN-HMM system which with TECC features improves whisper recognition by 31% in comparison with the traditional HMM system, and achieves 92.81% accuracy in isolated word recognition under whisper conditions.

Keywords: automatic speech recognition (ASR), whispered speech, artificial neural networks (ANN)
Scientific area: electrical engineering
Scientific subarea: acoustics
UDC number: 621.3

SADRŽAJ

SPISAK SLIKA:

SPISAK TABELA

SPISAK SKRAĆENICA

Skraćenica:	Puni naziv:
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
CD	Cepstral Distance
DAE	Denoising Autoencoder
DARPA	Defense Advanced Research Projects Agency
DCT	Discrete Cosine Transform
DDAE	Deep Denoising Autoencoder
DNN	Deep Neural Network
DTW	Dynamic Time Warping
DTW- FF	Dynamic Time Warping - Frame Fixing algorythm
ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FLDA	Fisher's Linear Discriminant Analysis
GDR	Generalized Delta Rule
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
LPCC	Linear Prediction Cepstral Coeffcicients
LTASS	Long Term Average Speech Spectrum
LVQ	Learning Vector Quantization
MAP	Maximum a posteriori probability
MFC	Mel-Frequency Cepstrum
MFCC	Mel-Frequency Cepstral Coefficients
MIT	Massachusetts Institute of Technology
MLE	Maximum Likelihood Estimate
MSE	Mean Squared Error
PCA	Principal Component Anlaysis
PCM	Pulse Code Modulation
RBM	Restricted Boltzmann Machine
RBF	Radial Bias Function
SEM	Standard Error of the Mean
SEO	Standard Energy Operator
SRN	Simple Recurent Neural Network
TDNN	Time Delay Neural Network
TECC	Teager Energy Cepstral Coefficients
TEMFCC	Teager Energy based Mel-Frequency Cepstral Coefficients
TEO	Teager Energy Operator
TI	Texas Instruments
WAV	Waveform Audio File Format
WER	Word Error Rate
Whi-Spe	Whispered Speech database

1 UVOD

1.1 MOTIVACIJA

Zahvaljujući konstantnom razvoju i usavršavanju tokom poslednjih dvadeset godina, moderni sistemi za automatsko prepoznavanje govora (*Automatic Speech Recognition - ASR*) su postigli zadovoljavajuće performanse i značajne komercijalne primene. Ipak, performanse ovih sistema su i dalje skromne u poređenju sa čovekovim sposobnostima percepcije govora i u tom pogledu postoji širok prostor za njihovo dalje usavršavanje. Sve masovnija komercijalna upotreba govornih tehnologija je ukazala na značajan broj nedostataka i slabih tačaka ASR sistema, koji predstavljaju ozbiljne probleme u praktičnoj primeni. U cilju postizanja što efikasnije i kvalitetnije interakcije između čoveka i mašine, predstoji i neophodno je buduće rešavanje ovih problema. Uprkos velikom uloženom istraživačkom naporu i dosadašnjem radu na polju usavršavanja ASR sistema, u pojedinim situacijama oni su ostali i dalje prilično osetljivi i nepouzdani. Na performanse ASR sistema utiču razni faktori, uključujući: kvalitet i tip ulaznog govornog signala, individualne karakteristike govornika poput dijalekata, brzine i načina govora, anatomije vokalnog trakta, psiho-fizičkog stanja govornika, itd. Osim navedenih, pojavljuju se i drugi uticaji prvenstveno iz okolne sredine, pod kojima se podrazumevaju: ambijentalni šum, reverberacija, buka, itd. Pored standardnog modaliteta govora, poznatog kao normalan ili neutralan govor, postoji i čitav niz drugih modaliteta govora. Naime, govorna ekspresija se može realizovati kroz različite

modalitete, počevši od šapata i tihog govora, preko neutralnog govora, ekspresivnog (afektivnog i emotivnog) govora, pevanog govora, pa sve do govora sa Lombardovim efektom i vike [Zhang at al., 2007]. Govorne tehnologije, pre svega prepoznavanje i sinteza govora, su razvijane na normalnom govoru. Međutim, upotreba i drugih modaliteta govora u uobičajnoj govornoj komunikaciji nije retka pojava i značajno umanjuje procenat uspeha prepoznavanja, odnosno kvalitet sintetizovanog govora. Prema tome, svi navedeni nestandardni modaliteti govora predstavljaju ozbiljan problem za konvencionalne ASR sisteme, pri čemu je njihovo prepoznavanje u velikoj meri degradirano i u većini slučaja neizvodljivo. Zbog toga se posledenjih godina fokus istraživanja u govornim tehnologijama pomera ka ovim modalitetima, pre svega ka prepoznavanju i sintezi emotivnog govora [Ayadi et al., 2011]. Od svih modaliteta, šapat se razlikuje po obezvučenosti i najviše se razlikuje u odnosu na normalan govor. Činjenica da šapat ima veoma veliku razumljivost i da se često koristi u govornoj komunikaciji nametnula je istraživanje i u ovom pravcu. Pronalazak načina poboljšanja uspeha automatskog prepoznavanja šapata predstavlja osnovnu motivaciju istraživačkog rada ove disertacije.

1.2 O ŠAPATU

Šapat je specifični oblik verbalne komunikacije koji je u čestoj upotrebi. Koristi se u različitim situacijama radi stvaranja diskretne i intimne atmosfere, kao i zarad prikrivanja poverljivih i privatnih informacija od okolnih slušalaca. Čest je u telefonskim razgovorima kada se diktiraju brojevi kreditinih kartica, pasoša, adrese, datumi rođenja, lekarske dijagnoze, računi i druge poverljive informacije [Fan et al., 2011]. U ovakvim situacijama govornici nastoje da tiše govore ili prosto šapuću jer nemaju drugog izbora da zaštite svoju privatnost na javnom mestu. Takođe, govornici šapuću kada ne žele da ometaju ljude u svom okruženju, na primer u biblioteci ili tokom poslovnog sastanka, ali i u kriminalnim aktivnostima kada se za cilj ima prikrivanje identiteta. Pored svesne produkcije šapata, on se može javiti i kao posledica zdravstvenih problema, na primer u slučaju ozbiljnog rinitisa, laringitisa, ili može biti posledica hroničnog oboljenja laringealnih sturktura [Jovičić et al., 2008 a].

Sa istraživačkog aspektra, šapat je veoma interesantna tema raznih studija iz oblasti forenzike i identifikacije govornika [Jin et al., 2007; Fan et al., 2008; Fan et al.,

2009]. Priroda šapata, mehanizmi njegovog generisanja, kao i razlike u odnosu na normalan govor su analizirane u sledećim radovima [Meyer-Eppler, 1957; Thomas, 1969; Holmes et al., 1983; Sugito et al., 1991; Leggetter et al., 1995; Jovičić, 1998; Matsuda et al., 1999]. Akustičke osobine šapata i bezvučnog izgovora vokala su takođe bile predmet istraživanja [Eklund et al., 1996; Konno et al., 1996; Jovičić, 1998] kao i karakterisitike konsonanata [Jovičić et al, 2008]. Šapat je ispitivan i kao specifični modalitet govora [Wenndt et al., 2002; Zhang et al., 2007]. Bio je predmet raznih multidisciplinarnih studija u kojima su za potrebe fizike, medicine, defektologije analizirane: aerodnamika vokalnog trakta [Sundberg et al., 2009], moždane aktivnosti u afoniji [Tsunoda et al., 2012], šapat kod post-leringektomisanih pacijenata [Sharifzadeh et al., 2009], laringealna hiperfunkcija tokom šapata [Rubin et al., 2006], itd.

1.3 OSVRT NA DOSADAŠNJA ISTRAŽIVANJA PREPOZNAVANJA ŠAPATA

Šapat je postao aktuelna istraživačka tema, od bitnog značaja za govorne tehnologije, pre svega zbog svoje specifičnosti i velike razlike u odnosu na normalan govor. Ta razlika se prvenstveno ogleda u pogledu odsustva glotalnih vibracija, šumne strukture i niskog SNR (*Signal to Noise Ratio*) [Morris, 2003]. Takođe, potreba za istraživanjem automatskog prepoznavanja šapata je dodatno motivisana sve većim zahtevima za usavršavanjem aktuelnih ASR sistema. Ipak, istraživanja šapata na ovom polju su još uvek u svom povoju, a do trenutka pisanja ove doktorske teze postojao je veoma mali broj objavljenih naučnih studija.

U jednoj od prvih studija automatskog prepoznavanja šapata [Ito et al., 2005] ispitano je prepoznavanje fonema u japanskom jeziku, kao i prepoznavanje kontinualnog govora u šapatu korišćenjem nekoliko HMM (*Hidden Markov Model*) modela i MFCC (*Mel Frequency Cepstral Coefficients*) obeležja. Ispitana su tri posebna govorna moda: šapat, tihi govor i normalan govor. Kao govorna baza korišćeni su snimci šapata tokom telefonskih razgovora. Baza je formirana tako što su govornici čitali unapred pripremljeni tekstualni materijal, dok je njihov razgovor sniman putem mobilnog telefona. Rezultati eksperimenata su pokazali da se prekrivanjem usta i telefona rukom postiže bolji SNR u bučnim ambijentalnim uslovima, čime se donekle poboljšava prepoznavanje tihog govora i šapata. Takođe je pokazano da se ASR sistemi koji su obučeni na govoru mogu adaptirati na šapat, dodavanjem malog uzorka šapata u

procesu obuke prepoznavača. Na ovaj način je takođe moguće poboljšati automatsko prepoznavanje šapata. Maksimalni postignuti uspeh u prepoznavanju šapata sa ovakvim *speaking-style-independent* modelom iznosti 66% tačnosti.

U kasnijim radovima [Lim, 2011; Mathur et al., 2012; Yang et al., 2012] autori su pokušali da ublaže akustičku neusaglašenost između neutralnog govora i šapata i poboljšaju performanse prepoznavanja šapata koristeći različite metode adaptacije ASR modela na šapat i transformacijom govornih obeležja [Yang et al., 2012]. Pojedine studije su se fokusirale na dizajn *front-end* dela ASR sistema i posebne metode ekstrakcije govornih obeležja [Zhang et al., 2010; Ghaffarzadegan et al., 2014 a] kao i na izmene u bankama filtara [Ghaffarzadegan et al., 2014 a]. U pomenutim radovima kao osnova ASR sistema su korišćeni HMM prepoznavač i MFCC obeležja. Koristeći isti sistem i obeležja, analizirano je prepoznavanje kontekstno nezavisnih monofona, konteksto zavisnih trifona i izolovanih reči u šapatu [Galić et al., 2014]. U navedenom slučaju ispitan je uspeh sistema zavisnog od govornika (*speaker-dependent system*) u različitim obuka/test scenarijima. Maksimalni postignuti uspeh prepoznavanja šapata sa ovakvim sistemom obučenim na normalnom govoru, u slučaju monofona, trifona i izolovanih reči iznosi: 64.80%, 28.32% i 36.24%, respektivno. Dakle, ostvareni rezultati potvrđuju da automatsko prepoznavanje štapata još uvek nije dostiglo zadovoljavajuće performanse koje bi zadovoljile potrebe komercijalne upotrebe.

Ozbiljni problemi se javljaju i u sistemima za verifikaciju i identifikaciju govornika. Prepoznavanje govornika na osnovu šapata je analizirano u sledećoj studiji [Fan et al., 2011]. Dve baze podataka su upotrebljene, jedna sa snimcima spontano izgovornih rečenica u šapatu i druga sa čitanim rečenicama u šapatu. Kao prepoznavač govornika korišćen je GMM (*Gaussian Mixture Model*) sistem, sa PLP (*Perceptual Linear Prediction*) i MFCC obeležjima. Posebna pažnja je posvećena analizi neusaglašenih obuka/test scenarija u kojima je prepoznavanje šapata veoma degradirano. U tim slučajevima je pokazano da upotreba linearne ili eksponencijalne frekvencijske skale umesto tradicionane Mel-logaritamske skale povećava tačnost identifikacije govornika u šapatu. Na taj način postignut je maksimalni uspeh u identifikaciji koji iznosi 83.84% u slučaju testiranja sistema sa spontanim šapatom.

1.4 CILJEVI DOKTORSKE DISERTACIJE

Ovaj kratak pregled dosadašnjih istraživanja ukazuje da je automatsko prepoznavanje šapata do sada testirano uglavnom sa HMM [Ito et al., 2005; Galić et al., 2014] i GMM sistemima [Fan et al., 2011] kao i sa MFCC i PLP obeležjima. Prema tome, veštačke neuralne mreže (ANN), kao predstavnici jedne popularne alternativne tehnike prepoznavanja govora, nisu do sada analizirane u problematici automatskog prepoznavanja šapata. Zbog toga, jedan od primarnih ciljeva ove disertacije je da se ispitaju mogućnosti i performanse ANN u prepoznavanju izolovanih reči izgovorenih u šapatu. U ovoj studiji, su korišćena dva tipa ASR sistema koji su bazirani na neuralnim mrežama – MLP (*Multilayer Perceptron*) *feed-forward* neuralne mreže i tandem DNN-HMM sistem čiju osnovu čini dubinska neuralna mreža (DNN), tačnije njen specijalni tip poznat kao dubinski *denoising* autoenkoder (DDAE). MLP sistem je poznat po svojim performansama i kompromisu koji realizuje između uspeha prepoznavanja, brzine prepoznavanja i neophodnih memorijskih resursa [Siniscalchi et al., 2013]. Dosadašnja istraživanja su dokazala da *feed-forward* neuralne mreže mogu aproksimirati bilo koju funkciju, što ih čini univerzalnim aproksimatorima [Bishop, 1995]. Sa druge strane, DNN sistemi predstavljaju najnoviju generaciju neuralnih mreža koje su pronalaskom novog i efikasnog algoritma za njihovu obuku 2006. godine [Hinton et. al; 2006] našle široku primenu u raznim zadacima mašinskog učenja. Moć DNN sistema pre svega leži u njihovoj dubinskoj arhitekturi koja im omogućava brzu obradu velike količine podataka i rešavanje najkompleksnih *pattern-recognition* problema. Kao ulazni vektori podataka u MLP i tandem DNN-HMM sisteme, pored tradicionalnih MFCC (*Mel Frequency Cepstral Coefficients*) obeležja, u ovoj disertaciji su testirana još dva tipa kepstralnih koeficijenata - TECC (*Teager Energy Cepstral Coefficients*) i TEMFCC (*Teager Energy based Mel Frequency Cepstral Coefficients*) [Dimitriadis et al., 2005]. U pitanju su novija i unapređena kepstralna obeležja zasnovana na nelinearnom *Teager* energetskom operatoru (*Teager-Energy Operator - TEO*) [Quatieri, 2002; Dimitriadis et al., 2005] koji služi za izračunavanje *Teager* energije. Za razliku od tradicionalnog načina izračunavanja energije, koji podrazumeva kvadriranje amplituda odbiraka signala, *Teager* energija opisuje "stvarnu" energiju zvučnog izvora i pored amplitudskih obuhvata i frekvencijske informacije. Superiornost TEO se ogleda u sposobnosti praćenja modulacije energije što omogućava bolju

predstavu formantnih informacija u vektorima obeležja u poređenju sa tradicionalnim MFCC obeležjima. U dosadašnjim istraživanjima, karakteristike nelinearnih TEO obeležja su pokazale obećavajuće rezultate u automatskom prepoznavanju tihog i takozvanog nemuštog govora (*murmured speech*) kao i u slučaju prepoznavanja drugih oblika govora pod stresom (*stressed speech*) [Zhou et al., 1998; Heracleous, 2009]. U prilog TEO obeležja idu i rezultati studije [Dimitriadis et al., 2005] koja potvrđuje značajna poboljšanja u prepoznavanju govora u šumnim uslovima prilikom korišćenja *Teager* energije. Sa druge strane, šapat je po svojoj prirodi dosta sličan govoru pod stresom i šumnim uslovima, poput nemuštog govora (*non-audible murmur*). Zbog navedenih prednosti, u ovoj disertaciji je postavljena i ispitana pretpostavka da TECC i TEMFCC obeležja mogu biti dobri deskriptori šapata.

Još jedan cilj ove disertacije je analiza neusaglašenih obuka/test scenarija, poput govor/šapat i šapat/govor scenarija. Posebno su interesantni rezultati govor/šapat scenarija, koji odgovara realnoj situaciji u kojoj govornik tokom interakcije sa ASR sistemom iz normalnog govora pređe u šapat. U ovakvoj situaciji, prepoznavanje šapata je dosta degradirano. U svrhu boljeg razumevanja neusaglašenih scenarija i sa ciljem razvijanja sitema koji omogućava bolje prepoznavanje šapata, u disertaciji su detaljno opisani eksperimenti sa neuralnim mrežama i analize konfuzije u prepoznavanju reči u neusaglašenim obuka/test scenarijima. Takođe su ispitane i akustičke karakteristike šapata, kepstralne distance između stimulusa u normalnom govoru i šapatu, matrice konfuzije, itd. Na osnovu rezultata sprovedenih eksperimenta, ova teza predlaže dva nova pristupa predobrade govornih signala, koji smanjuju akustičke razlike između normalnog govora i šapata i na taj način poboljšavaju uspeh prepoznavanja reči u neusaglašenim obuka/test scenarijima. Prvi pristup se zasniva na inverznom filtriranju i implementiran je u sklopu *front-end* dela MLP sistema, pri čemu se iz normalnog govora potiskuje zvučnost i smanjuje spektralna razlika u poređenju sa šapatom. Drugi pristup se bazira na rekonstrukciji karakteristika normalnog govora iz šapata pomoću dubinskog *denoising* autoenkodera. Ovaj pristup je korišćen u *front-end* delu tandem DNN-HMM sistema koji se pokazao kao najbolje rešenje za automatsko prepoznavanje šapata.

1.5 KRATAK OPIS SADRŽAJA DISERTACIJE

Poglavlje 2 daje kratak uvod u osnove automatskog prepoznavanja govora. Najpre su definisani: osnovni koncept automatskog prepoznavanja govora, neophodni koraci u kreiranju jednog ASR sistema, kao i statistička formulacija problema donošenja odluke. Opisane su dve široko primenjene tehnike u ASR sistemima: trenutno aktuelni skriveni Markovljevi modeli (HMM) i nešto starija ali još uvek prisutna tehnika od velikog teorijskog značaja - dinamičko vremensko usklađivanje (DTW). U poglavlju su objašnjene prednosti i mane svake od ovih tehnika, čime je napravljena osnova za dalje poređenje sa neuralnim mrežama.

Poglavlje 3 pruža najbitnije informacije o veštačkim neuralnim mrežama, počevši od njihovog istorijskog razvoja, različitih tipova i arhitektura mreža, preko njihvog načina procesiranja, obuke, pa sve do veze sa statističkim modelima. Posebna pažnja je posvećena *feedforward* tipu neuralnih mreža, tačnije višeslojnim perceptronima (MLP), *Backpropagation* metodi obuke i dubinskim autoenkoderima koji su korišćeni u eksperimentima ove disertacije. Neuralne mreže su komentarisane i sa aspekta njihovih performansi, mogućnosti i ograničenja u prepoznavanju reči, sa posebnim osvrtom na dodirne tačke i prednosti u odnosu na ostale tehnike automatskog prepoznavanja govora.

Poglavlje 4 opisuje osnovne karakteristike i prirodu šapata. Navedene su i objašnjene razlike između šapata i normalnog govora, kao i način na koji te razlike utiču na automatsko prepoznavanje šapata. Sa fiziološkog aspekta su detaljno opisani način generisanja šapata, specifičan rad artikulacionih organa i njihov uticaj na akustička svojstva šapata, pre svega na: energiju, spektralni nagib i položaj formanata. Diskutovane su mogućnosti prenosa informacija šapatom i ograničenja u percepciji šapata. Na kraju poglavlja je dat pregled rezultata i mogućnosti do sada testiranih ASR sistema u prepoznavanju šapata.

Poglavlje 5 objašnjava postupak kreiranja prvog i trenutno jedinog korpusa šapata za srpski jezik, Whi-Spe, na kome su vršeni svi eksperimenti ove disertacije. Najpre je dat pregled postojećih i do sada u literaturi poznatih korpusa šapata (za japanski, engleski i kineski jezik), a zatim dizajn, snimanje i obrada Whi-Spe korpusa. Opisane su

specifične manifestacije šapata tokom snimanja, kao i metode kontrole kvaliteta snimaka. Poglavlje sadrži i detaljno opisuje postupak evaluacije i akustičke analize Whi-Spe korpusa, pre svega u pogledu: analize talasnih oblika signala i njihovih spektrograma, spektralnog nagiba i kepstralnih karakteristika. U nastavku poglavlja je objašnjena uloga inverznog filtriranja u kreiranju baze pseudo-šapata, kao i kepstralna analiza postignutih rezultata.

Poglavlje 6 sistematski prikazuje postupak kreiranja MLP sistema za automatsko prepoznavanje izolovanih reči u normalnom govoru i šapatu. U prvom delu poglavlja opisan je takozvani *front-end* deo ASR sistema, zadužen za predobradu govornih signala (segmentacija, vremensko usklađivanje, filtriranje i prozorovanje govornih signala). Zatim se opisuje detaljan matematički postupak ekstrakcije MFCC, TECC i TEMFCC govornih obeležja, uz poseban akcenat i tumačenje karakteristika i postupka izračunavanja *Teager* energije. Sledi objašnjenje formiranja delova baze za obuku, validaciju i testiranje neuralnih mreža. Na samom kraju poglavlja opisano je kreiranje veštačke neuralne mreže (MLP), određivanje njene optimalne strukture, kao i postupak njenog obučavanja.

Poglavlje 7 opisuje postupak kreiranja tandem DNN-HMM sistema za automatsko prepoznavanje izolovanih reči u normalnom govoru i šapatu. Na početku poglavlja je prezentovan *front-end* deo sistema čiju osnovu čini dubinska neuralna mreža (DNN) – tačnije dubinski *denoising* autoenkoder (DDAE). U ovom delu su pored predobrade signale i ekstrakcije govornih obeležja detaljno opisani: struktura, kreiranje i obuka dubinskog *denoising* autoenkodera, koji u *front-end* delu tandem DNN-HMM sistema ima ulogu sekundarnog ekstraktora robustnih govornih obeležja. U nastavku teksta je opisan *back-end* deo sistema koga čini HMM prepoznavač, kao i postupak obuke tako formiranog tandem sistema.

Poglavlje 8 daje pregled eksperimentalnih rezultata automatskog prepoznavanja izolovanih reči u normalnom govoru i šapatu ostvarenih pomoću MLP sistema. Prezentovane su analize različitih obuka/test scenarija, poređenje performansi višeslojnih perceptrona i njihovih uspeha u klasifikaciji reči u zavisnosti od korišćenja tri tipa kepstralnih koeficijenata. Prikazana je analiza konfuzije u prepoznavanju reči zajedno sa spektralnim tumačenjem konfuzija posebno istaknutih kritičnih parova reči.

Na osnovu dobijenih rezultata, u ovom poglavlju je postavljena hipoteza o maskiranju određenih govornih obeležja usled zvučnosti, koja predstavlja glavni uzrok degradiranog prepoznavanja reči u neusaglašenim obuka/test scenarijima. Radi dokaza hipoteze i poboljšanja uspeha prepoznavanja reči, pre svega u šapatu, predložena je izmena *front-end* sistema gde je u fazi predobrade signala implementirano inverzno filtriranje. Tako izmenjen sistem je nazvan MLP-IF, a ostvareni eksperimentalni rezultati u poboljšanju uspeha prepoznavanja reči su opisani na kraju poglavlja.

Poglavlje 9 predstavlja eksperimente sa tandem DNN-HMM sistemom u automatskom prepoznavnaju izolovanih reči. Na početku poglavlja su tabelarno prikazani usrednjeni rezultati muških i ženskig govornika u usaglašenim obuka/test scenarijima pri korišćenju sva tri tipa keptralnih koeficijenata, a zatim su u nastavku poglavlja prezentovani i postignuti rezultati u neusaglašenim obuka/test scenarijima.

Poglavlje 10 vrši komparaciju maksimalnih ostvarenih rezultata prepoznavanja reči u usaglašenim i neusaglašenim obuka/test scenarijima pomoću MLP, MLP-IF i DNN-HMM sistema. Takođe, poglavlje daje i dodatno poređenje performansi ovih sistema sa još nekim od postojećih ASR sistemima iz literature (DTW i HTK-HMM) koji su testirani na istoj bazi podataka (Whi-Spe).

Poglavlje 11 sumira postignute rezultate, diskutuje ih, objašnjava naučni doprinos ove disertacije i predlaže moguće pravce daljeg istraživanja.

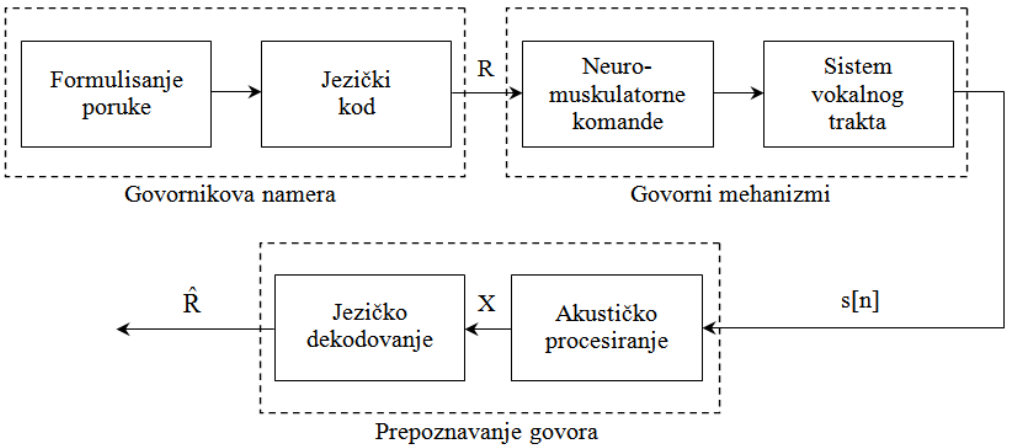
2 UVOD U AUTOMATSKO PREPOZNAVANJE GOVORA

Prva bitnija istraživanja na polju automatskog prepoznavanja govora su se pojavila uporedo sa idejom o novim korisničkim servisima u kojima korisnici direktno interaguju sa mašinama i na taj način pristupaju servisu. Pored automatizacije servisa, povećanja njegove brzine i efikasnosti, glavni motiv za ovakvim vidom komunikacije je bio smanjenje ljudskih resursa a time i finansijskih troškova. Danas je automatsko prepoznavanje govora u širokoj komercijalnoj primeni, međutim mogućnosti ovakvih mašina (govornih automata) u prepoznavanju govora su još uvek skromne u poređenju sa čovekovim sposobnostima percepcije. Zbog toga su aktuelna istraživanja automatskog prepoznavanja govora zacrtala mnogo više ciljeve, prvenstveno u pogledu usavršavanja ovih sistema i postizanja što prirodnije i lakše komunikacije između čoveka i mašine. U ovom poglavlju je priložen kratak pregled osnovnog koncepta automatskog prepoznavanja govora (*Automatic Speech Recognition - ASR*). Ukratko su opisane samo neke od najznačajnijih tehnika koje se koriste u ovakvim sistemima, poput Dinamičkog usklađivanja u vremenu (*Dynamic Time Warping - DTW*) i Skrivenih Markovljevih modela (*Hidden Markov model - HMM*), kao i njihove prednosti i ograničenja u prepoznavanju govora.

2.1 OSNOVE AUTOMATSKOG PREPOZNAVANJA GOVORA

Osnovni cilj jednog ASR sistema je tačno i efikasno konvertovanje govornog signala u tekstualnu poruku, nezavisno od uređaja kojim je govor snimljen, govornikovog akcenta, njegovog psiho-fizičkog stanja, uticaja telekomunikacionog kanala, akustičkog okruženja, itd [Rabiner et al., 2007]. Ovaj cilj još uvek nije u potpunosti dostignut i aktuelna je tema brojnih studija. ASR sistem je osmišljen da oponaša čovekovu percepciju govora, a u teorijskom pogledu idealan ASR sistem treba da u potpunosti parira ljudskim sposobnostima.

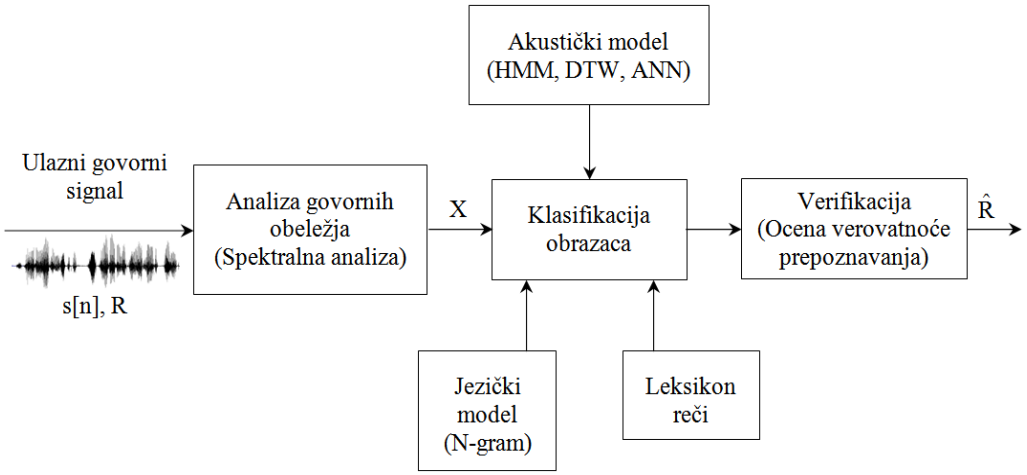
Uprošćeni model generisanja i prepoznavanja govora je prikazan na Slici 2.1. Proces generisanja govora počinje željom govornika da izrazi svoje misli i prenese neku poruku sagovorniku, odnosno mašini. Kako bi ta poruka bila izgovorena, neophodno je da govornik prvo u svojoj glavi formuliše ideju (tekst) i da je konvertuje u jezički kod, odnosno u simboličku predstavu niza glasova, R, koje ta poruka sadrži. U ovom stadijumu se donosi i odluka o načinu govora poput: brzine izgovora, naglašavanja određenih segmenata, i drugih prozodijskih obeležja. Sledeći korak u generisanju govora je aktiviranje neuro-mišićnih centara i artikulacionih organa pomoću kojih se ta poruka dalje transformiše u akustički oblik. Generisani akustički signal se tokom prenosa kroz vokalni trakt dodatno uobličava. Po izlasku iz vokalnog trakta, polazna informacija je enkodovana u akustički signal, s[n], koji se dalje u vidu govornog signala prenosi kroz telekomunikacioni kanal i stiže do prepoznavača.



Slika 2.1 Uprošćeni model generisanja i prepoznavanja govora.

U osnovi bloka za prepoznavanje je akustički procesor koji analizira primljeni govorni signal i pretvara ga u niz akustičkih obeležja, X, koja opisuju različite govorne zvuke i karakterišu njihove spektralne i/ili vremenske osobine. Sledi blok za jezičko dekodovanje koji obično koristi statističku procenu verovatnoće u pronalaženju niza reči koje su akustički najsličnije obeležjima na ulazu, X, a potom ih uređuje u prepoznatu rečenicu $\hat{R}$.

Detaljniji prikaz procesa automatskog prepoznavanja govora je ilustrovan na Slici 2.2 [Rabiner et al., 2007]. Govorni signal, s[n], koji dolazi na ulaz ASR sistema se prvo segmentira u niz jednakih vremenskih okvira (proces poznat kao *frame blocking*). Sledi spektralna analiza u kojoj se iz ovih kratkih vremenskih okvira izdvajaju odgovarajuća akustička obeležja X. Govorni signal predstavljen nizom obeležja $X=\{x_1,x_2,...,x_t\}$, stiže u blok za klasifikaciju obrazaca, u kome se vrši njihovo akustičko upoređivanje sa postojećim setom akustičkih modela u cilju pronalaska najsličnijeg para obrazaca (*pattern recognition*).



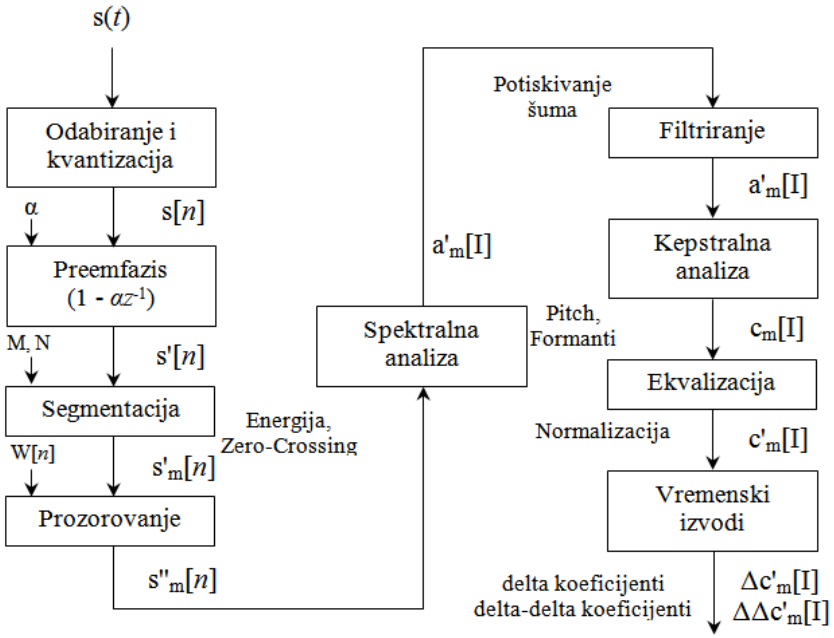
Slika 2.2 Generalizovana blok šema prepoznavnaja govora.

Pronalaskom akustičkog modela koji se najbolje poklapa sa sekvencom akustičkih obeležja, ulazni govorni signal, s[*n*], se dekoduje u simboličku predstavu, $\hat{R}$. U sprezi sa leksikonom reči, blok za klasifikaciju daje ocenu akustičkog podudaranja ulaznog signala sa prepoznatim rečima. Prepoznati niz reči takođe dobija ocenu jezičkog modela (*N-gram*). Objedinjavanjem ove dve ocene definiše se verovatnoća prepoznavanja ulaznog govornog signala i vrši se verifikacija prepoznate sekvence reči, $\hat{R}$, čime se proces automatskog prepoznavanja završava.

2.1.1 KORACI U KREIRANJU ASR SISTEMA

Prilikom planiranja i kreiranja sistema za automatsko prepoznavanje govora neophodni su sledeći koriaci:

1) Prvi korak u dizajniranju ASR sistema je odabir odgovarajućih obeležja za prepoznavanje govora. U tu svrhu se koriste razne vrste i kombinacije akustičkih, govornih i slušnih obeležja, koja imaju za cilj modelovanje spektralnih karakteristika govora sa aspekta čovekove percepcije zvuka. Jedna od najzastupljenijih akustičkih obeležja u ASR sistemima su Mel-frekvencijski kepstralni koeficijenti (*Mel-frequency cepstral coefficients - MFCC*), koji pomoću nelinearne Mel-frekvencijske skale aproksimiraju karakteristike čovekovog auditornog sistema. Na Slici 2.3 je prikazan postupak ekstrakcije MFCC obeležja iz govornog signala.



Slika 2.3 Procedura ekstrakcije govornih obeležja (MFCC) i njihovih prvih i drugih vremenskih izvoda.

Ulazni govorni signal s(t) se najčešće digitalizuje sa frekvencijom odabiranja od 8000Hz do 22050Hz. Na diskretni signal s[n] se potom primenjuje preemfazis radi isticanja harmonika na višim učestanostima i kompenzacije sa njihovim slabljenjem na višim frekvencijama. Posle preemfazisa, signal s'[n] se segmentira na m kratkih vemenskih okvira trajanja 15-40ms, koji se međusobno preklapaju

najčešće 10-20ms. Svaki od okvira govornog signala se množi nekom prozorskom funckijom W[n], najčešće tipa Hamming, čime se obezbeđuje prevencija diskontinuiteta između signala iz susednih prozora. Posle ovakve predobrade govornih signala sledi njihova spektralna analiza. Sa prelaskom u frekvencijski domen, na samom početku opciono se vrši filtriranje signala i potiskivanje šuma. Sledeći korak u ekstrakciji MFCC je kepstralna analiza signala i izdvajanje koeficijenata. Uobičajeno se izdvaja prvih 10-20 MFCC (uključujući i nulti koeficijent - c0). Ponekad se vrši ekvalizacija MFCC obeležja u vidu normalizacije ili oduzimanja njihove srednje vrednosti. Takođe, opciono se vrši izračunavanje prvih i drugih vremenskih izvoda kepstralnih koeficijenata, takozvanih delta (Δ) i delta-delta (ΔΔ) koeficijenta, radi boljeg opisivanja dinamike signala. Na taj način se kao rezultat ekstrakcije dobija vektor MFCC obeležja, koji sadrži niz kepstralnih koeficijenata i njihovih prvih i drugih vremenskih izvoda. Najčešće se jedan govorni okvir predstavlja sa 13 MFCC, 13 ΔMFCC i 13 ΔΔMFCC što daje vektor obeležja dužine D=39 koeficijenata.

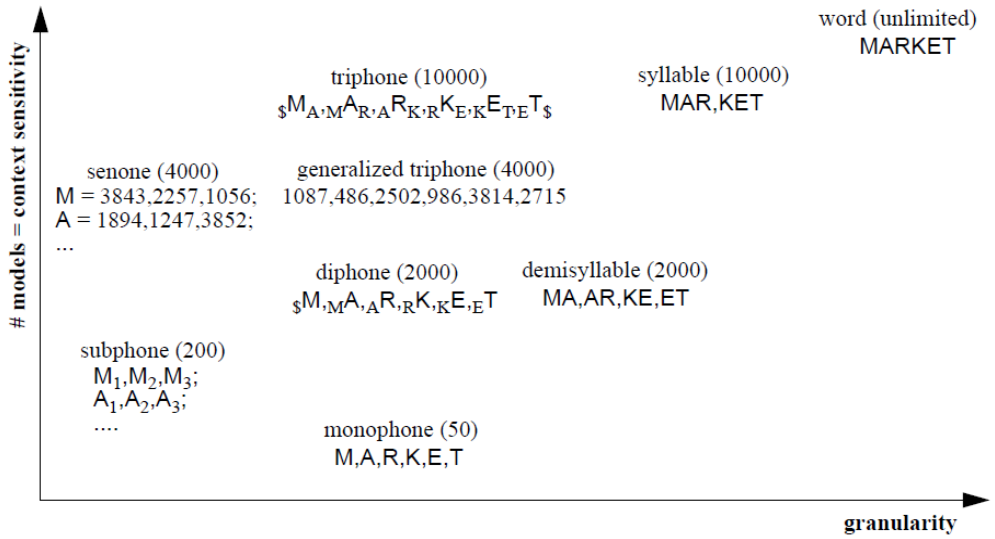
2) Drugi korak u kreiranju ASR sistema je definisanje zadatka automatskog prepoznavanja. Taj zadatak može biti relativno jednostavan, poput prepoznavanja izgovora svega deset cifara, do veoma složenog prepoznavanja kontinualnog spontanog govora sa velikim leksikonom reči. U zavisnosti od broja reči, korpusi se dele na male (do par stotina reči), srednje (nekoliko hiljada reči) i velike (reda stotina hiljada reči) [Itakura, 1975; Miyatake et al., 1990; Kimura, 1990]. Iako su jednostavniji, mali rečnici mogu biti teški za prepoznavanje u slučaju kada sadrže veliki broj sličnih reči. Takav primer su rečnici engleskog alfabeta u kojima postoji često izražena konfuzija između takozvanih E-set glasova: "B, C, D, E, G, P, T, V, Z" [Hild et al., 1993].

Takođe je bitno da se na samom početku kreiranja ASR sistema definiše da li je sitem zavisan ili nezavisan od govornika. Sistemi zavisni od govornika su dizajnirani i prilagođeni govoru pojedinca i imaju veći uspeh u prepoznavanju reči u odnosu sa sistemime koji su nezavisni od govornika i čije su greške prepoznavanja uglavnom 3-5 puta veće [Lee, 1988]. Postoje i takozvani multi-speaker sistemi koji su namenjeni korišćenju manjoj grupi ljudi. Što se tiče

snimaka govora, oni mogu biti izolovani, diskontinualni i kontinualni. Primer snimka izolovanog govora su odvojeno snimljeni izgovori reči, dok diskontinualni govor podrazumeva snimke izgovora rečenica sa veštački napavljenim pauzama između reči. Prepoznavanje izolovanog i diskontinualnog govora u poređenju sa kontinualnim govorom je dosta jednostavnije zbog jasnih granica između reči. Sa druge strane u kontinualnom govoru, usled koartikulacije ove granice često nisu dovoljno jasne što rezultuje lošijim prepoznavanjem koje je obično 6% lošije u odnosu na prepoznavanje izolovanih reči [Bahl et al., 1981].

U pogledu obuke ASR sistemi se mogu trenirati na snimcima čitanog ili spontanog govora. Spontani govor sadrži uzdahe, prekide, nedovršene rečenice, iznenadne pauze, lažne početke rečenica, kašalj, zamuckivanje, smeh itd., što ga čini dosta kompleksnijim i komplikuje njegovo prepoznavanje. Zbog svega nabrojanog, kreiranje kvalitetne baze snimaka spontanog govora je veoma zahtevno i često nemoguće. Postojeće baze spontanog govora su najčešće kreirane prikupljanjem dovoljnog broja audio snimaka radio ili TV-emisija.

3) Treći korak je akustičko i jezičko modelovanje govornih jedinica. Na Slici 2.4. su ilustrovani primeri akustičkih modela koji se razlikuju u pogledu veličine govornih jedinica koje modeluju i njihove osetljivosti na kontekst.



Slika 2.4 Akustičko modelovanje govornih jedinica [Tebelskis, 1995]. Sa povećanjem granulacije govornih jedinica raste i njihova osetljivost na kontekst. Brojevi u zagradama predstavljaju broj govornih jedinica za engleski jezik.

Prema složenosti ili granularnosti, govorne jedinice se dele na: reči, slogove, polu-slogove, trifone, generalizovane trifone, difone, monofone, senone i pod-glasove. Kao što se vidi sa Slike 2.4, složene govorne jedinice (poput reči i slogova) su više osetljive na kontekst. Takođe važi da modeli sa većom osetljivošću obezbeđuju tačnije prepoznavanje, pod uslovom da su modeli dobro obučeni. Međutim, sa porastom granulacije, odnosno složenosti govornih jedinica, obuka modela je teža a njihovo prepoznavanje lošije pre svega zbog manjeg broja uzoraka koji su na raspolaganju u procesu obuke.

4) U tom pogledu, modelovanje izolovanih reči ili slogova ima smisla samo za manje govorne korpuse, i tada obezbeđuju veoma visok uspeh u prepoznavanju. Sa druge strane, u slučaju obimnih rečnika, najčešće se vrši modelovanje prostijih govornih jedinica poput trifona. Mnogi sistemi su zasnovani na modelima monofona, pre svega zbog njihove jednostavnosti.

Što se tiče načina akustičkog modelovanja govornih jedinica, postoje dva tipa - modelovanje obrazaca (*Template-based models*) i statističko modelovanje (*Statistical-based models*). U ASR sistemima čiji su akustički modeli zasnovani na obrascima, prepoznavanje se svodi na poređenje nepoznatog govornog stimulusa sa svim ostalim prethodno snimljenim rečima (*template*) u cilju pronalaska najsličnijih parova reči. Prednost ovakvih modela je korišćenje skoro savršenih modela reči, što je ujedno i mana jer su takvi obrasci fiksni i slabo modeluju pojavu bilo koje varijabilnosti u govoru. Ovaj problem se donekle može rešiti korišćenjem velikog broja snimaka koji modeluju različite varijabilnosti u izgovoru reči, što ipak na kraju postaje nepraktično rešenje. Primer ovakvog akustičkog modelovanja su DTW sistemi o kojima će biti više reči u narednom poglavlju. Drugi tip ASR sistema je zasnovan na statističkom modelovanju akustičkih varijacija u govoru (primer HMM sistemi). U procesu obuke, za svaki od modela se na osnovu prethodnih snimaka govornih jedinica kreira njihova statistička raspodela, pomoću kojih se opisuju različita stanja ovih modela. Ovakvi ASR sistemi koji u osnovi *pattern recognition* procesa sadrže statističke modele su danas najzastupljeniji. Najveća mana ovih sistema je neophodno postavljanje *a priori* pretpostavki tokom modelovanja govora, čime se u startu

unosi izvesna greška koja dodatno ograničava njihove performanse i krajnje domete u prepoznavanju.

Svaki od akustičkih modela zahteva odgovarajuću obuku sa uzorcima govora. Kroz iteracije u procesu obuke tačnost modela se povećava sve do postizanja nekih željenih performansi, nakon čega se trening zaustavlja i sistem se može testirati. Pored obuke akustičkih modela, postoji i obuka jezičkih modela[1]. Obuka jezičkih modela zahteva dovoljnu količinu teksta na osnovu koga ASR sistem iz različitih nizova i kombinacija reči "uči" sintaksu.

5) Završni korak u kreiranju ASR sistema je evaluacija njegovih performansi. Performanse ASR sistema se najčešće opisuju u vidu uspeha prepoznavanja reči, ili kao procenat pogrešno prepoznatih reči (word error rate - WER).

2.1.2 STATISTIČKA FORMULACIJA PROBLEMA AUTOMATSKOG PREPOZNAVANJA GOVORA

Automatsko prepoznavanje govora se može formulisati kao matematički problem donošenja odluke kojim se bavi Statistička teorija odlučivanja. Ovaj problem se svodi na traženje Bajesove maksimalne *a posteriori* verovatnoće (*Maximum a posteriori probability - MAP*) koja se odnosi na niz reči $\hat{R}$ koji najviše odgovara nizu akustičkih obeležja X sa ulaza prepoznavača. Matematički formulisano, potrebno je pronaći maksimum *a posteriori* verovatnoće P(R|X), odnosno niz reči $R = \{r_1, r_2 \dots r_N\}$ (nepoznate dužine N) koji najverovatnije potiče od niza akustičkih vektora $X = \{x_1, x_2 \dots x_T\}$ dobijenih u procesu akustičke analize tokom predobrade govornog signala:

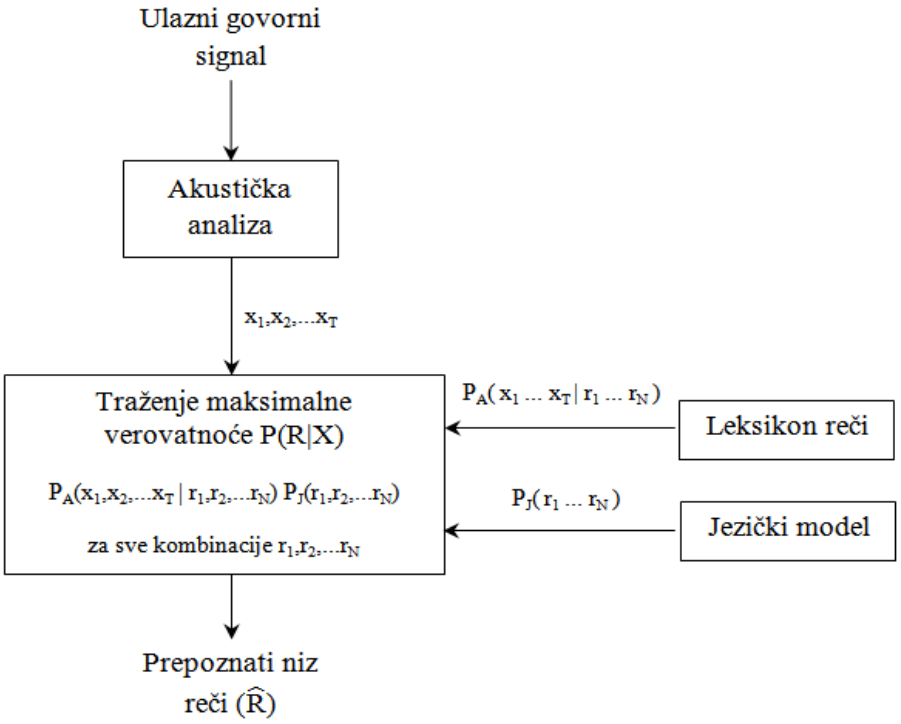
$$\hat{R} = \arg\max_{R} P(R|X) . \tag{2.1}$$

Korišćenjem Bajesovog pravila, izraz (2.1) se može zapisati u sledećoj formi:

$$\hat{R} = \text{asarg}\max_{R} \frac{P(X|R)P(R)}{P(X)} , \tag{2.2}$$

[1] Obuka jezičkih modela nije uvek obavezni sastavni deo ASR sistema.

gde je nepoznata *a posteriori* verovatnoća P(R|X) rastavljena na dve *a priori* verovatnoće P(R) i P(X) i jednu *a posteriori* verovatnoću P(X|R). Ove komponente u izrazu (2.2) su poznate i definisane su odgovarajućim raspodelama verovatnoće koje su dobijene procesom treninga akustičkog i jezičkog modela. *A priori* verovatnoća P(R) predstavlja verovatnoću da se određene reči nađu zajedno u uređenom nizu R = {$r_1, r_2 \dots r_N$} i definisene su jezičkim modelom prepoznavača. Iz tog razloga ova verovatnoća se obično obeležava sa $P_J(R)$. *A posteriori* verovatnoća P(X|R) je određena akustičkim modelom i leksikonom reči, i predstavlja uslovnu verovatnoću da je niz obeležja X = {$x_1, x_2 \dots x_T$} dobijen akustičkom analizom niza reči R = {$r_1, r_2 \dots r_N$}. Ona se obično označava sa $P_A(X|R)$ kako bi se naznačila akustička priroda termina.



Slika 2.5 Bajesov princip odlučivanja u automatskom prepoznavanju govora.

Prilikom izračunavanja i traženja maksimalne *a posteriori* verovatnoće P(R|X) u izrazu (2.2) se zanemarijuje *a priori* verovatnoća P(X), pošto je ona nezavisna od niza reči R po kome se vrši optimizacija. Iz tog razloga se opisana procedura prepoznavanje govora uobičajno piše u sledećoj formi:

$$\hat{R} = \underbrace{\text{asarg}\max_{R}}_{\text{Korak 3}} \underbrace{P(X|R)}_{\text{Korak 1}} \; P(R) \;, \qquad (2.3)$$

gde prvi korak predstavlja izračunavanje verovatnoće koja se tiče akustičkog modelovanja govora u rečenici R, dok se drugi korak odnosi na izračunavanje verovatnoće kod jezičkog modelovanja reči u rečenici R. Treći korak je pretraga svih validnih rečenica u cilju pronalaženja najsličnije ulaznoj. Pronalaskom takve rečenice završava se proces automatskog prepoznavanja govora. Opisani Bajesov princip odlučivanja je prikazan na Slici 2.5.

2.2 DINAMIČKO USKLAĐIVANJE U VREMENU (DTW)

Dinamičko usklađivanje u vremenu (*Dynamic Time Warping - DTW*) je jedan od najstarijih i najznačajnijih algoritama u automatskom prepoznavanju govora [Vintsyuk, 1971; Itakura, 1975; Sakoe et al., 1978]. Spada u klasu algoritama baziranih na tehnici dinamičkog programiranja koja se koristi za pronalaženje optimalne putanje u postupku utvrđivanja vremenske neusaglašenosti dva signala.

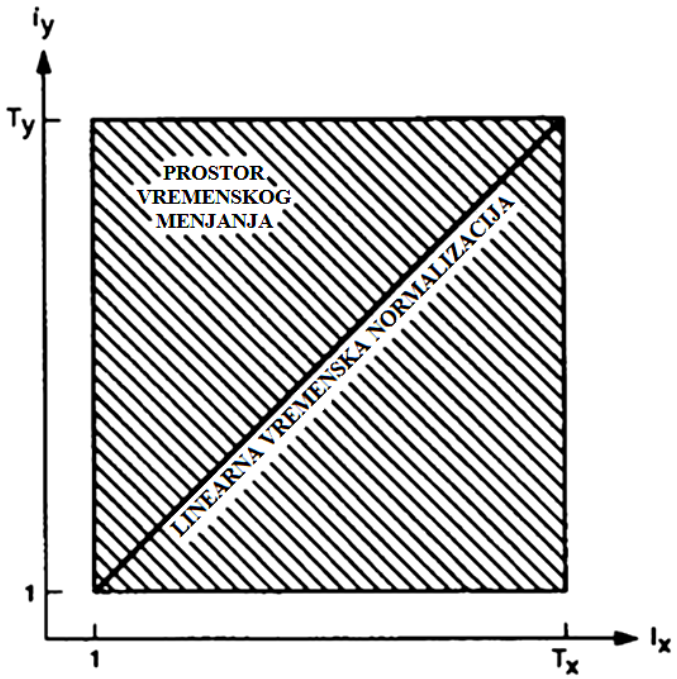
Najjednostavniji način prepoznavanja određene reči je njeno upoređivanje sa svim drugim obrascima reči sa kojima ASR sistem raspolaže u cilju pronalaženja najsličnijeg para. Ovaj zadatak ipak nije tako lak iz više razloga. Prvi razlog je vremenska razlika u trajanju različitih snimaka govora što dodatno otežava njihovo upoređivanje. Tehnikom linearne vremenske normalizacije (Slika 2.6), ovi snimci se mogu svesti na isto trajanje. U ovom slučaju optimalna putanja je predstavljena pravom dijagonalnom linijom (Slika 2.6). Međutim, linearnim usklađivanjem signala problem nije u potpunosti rešen, jer i dalje postoje rezlike u trajanju izgovora pojedinih glasova, što je posledica neujednačene brzine izgovora određenih segmenata reči. To je drugi razlog koji u velikoj meri komplikuje poređenje dva signala.

Uzmimo u obzir govorne signale koje poredimo, x i y, i njihove predstave u vidu vektora akustičkih obeležja $X=\{x_1,x_2,...,x_{Tx}\}$ i $Y=\{y_1,y_2,...,y_{Ty}\}$. Usled različitog trajanja signala, ovi vektori imaju razlite dužine, odnosno različite brojeve okvira - T_x i T_y. Posle linearnog svođenja njihovih dužina na referentnu dužinu T_x, razlika između signala x i y se može jednostavno izračunati u vidu sume Euklidskih distanci:

$$d(x,y)=\sqrt{\sum_{i=1}^{T_x}(x_i-y_j)^2}\ , \tag{2.4}$$

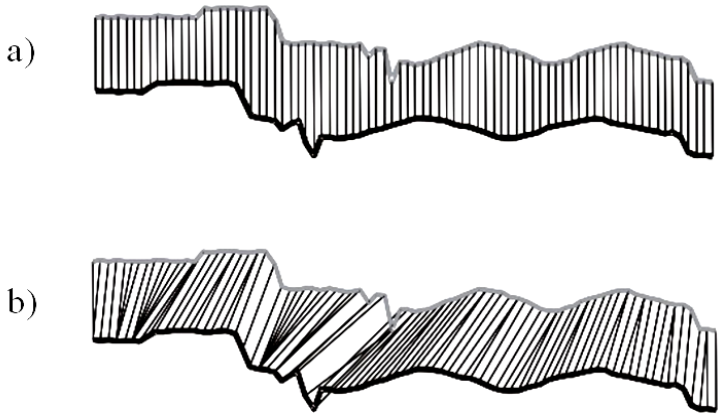
gde su i=1,2,...,T_x i j=1,2,...,T_y indeksi okvira x i y signala, i pritom važi:

$$j=\frac{T_y}{T_x}i\ . \tag{2.5}$$



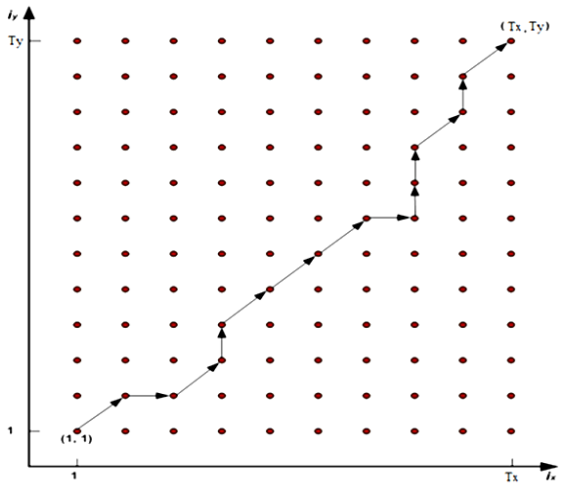
Slika 2.6 Linearno usklađivanje u vremenu dve sekvence različitog trajanja.

Sa Slike 2.7 a) se vidi da ovakav način poređenja dva signala ne modeluje dobro realno stanje u vremenu i zahteva drugačiji prirodniji način usklađivanja i normalizacije dva signala. Za razliku od linearnog usklađivanja gde se okviri govornih signala porede po principu "jedan na jedan" i gde je vremenska osa fiksna, u DTW algoritmu, zasnovanom na nelinearnom usklađivanju, analiziraju se sve moguće kombinacije govornih okvira i vrši se njihovo poređenje u potrazi za najsličnim parovima. Na ovaj način se postiže mnogo bolje poređenje i vremensko poravnanje signala, što čini vremensku osu elastičnom (Slika 2.7 b)) i pogodnom za modelovanje promenljive brzine izgovora pojedinih segmenata reći.



Slika 2.7 Primer: a) linearnog i b) nelinearnog usklađivanja dva signala u vremenu.

Prema tome, u osnovi DTW algoritma stoji izračunavanje matrice lokalnih distanci za sve kombinacije okvira signala x i y (dimenzija $T_x \times T_y$), na osnovu koje se određuje optimalna putanja duž koje se signal "rasteže" (ekspanzija) ili "sabija" (kompresija), i svodi na zadatu referentnu dužinu T_x, Slika 2.8.



Slika 2.8 Primer matrice distanci za dva niza govornih obeležja X (apscisa) i Y (ordinata) i pronalaska optimalne putanje.

DTW algoritm je određen sa nekoliko pravila:

- Prilikom računanja matrice lokalnih distanci svi okviri moraju biti uključeni u proračun optimalne putanje.
- Tokom određivanja optimalne putanje granični uslovi se moraju poštovati, odnosno putanja u matrici počinje u položaju (1,1) i završava se u (T_x,T_y).

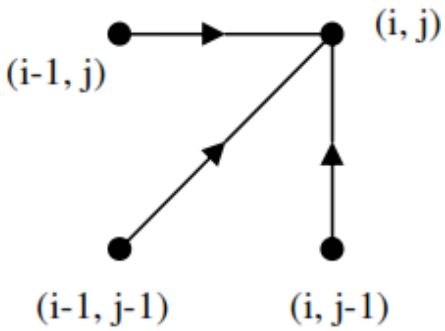
- Optimalna putanja se određuje u koracima, pri čemu se indeksi *i* i *j* mogu uvećati samo za 1 u svakom koraku. (Uslov kontinualnosti).

- Važi uslov monotonosti, tj. optimalna putanja uzima elemente matrica tako da su njeni indeksi monotono rastući i ne opadajući (ili su konstantni ili se povećavaju).

- Lokalne distance *d(i,j)* između određenih okvira test i referentnog singla učestvuju u formiranju kumulativne sume *D*(*i*,*j*) koja se nazova funkcijom distance (*distance function*) ili funckijom cene (*cost function*). Izračunava se na sledeći način:

$$D(i, j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i.j), \qquad (2.6)$$

Kumulativna suma *D*(*i*,*j*) se često naziva i globalnom distancom i koristi se kao mera sličnosti dva signala u postupku prepoznavanja govora. Ona se izračunava za sve parove reči (test i referentne reči). Prilikom određivanja optimalne putanje za svaku reč, vrši se inicijalizacija matrice lokalnih distanci, a vrednosti *D*(*i*,*j*) se postavlja na *d*(1,1). U osnovnoj, takozvanoj simetričnoj verziji DTW algoritma, dozvoljena su tri koraka prilikom biranja optimalne putanje, a to su: horizontalni, vertikalni i dijagonalni korak (Slika 2.9), pri čemu se bira korak sa najmanjom distancom (cenom).



Slika 2.9 Dozvoljeni koraci u DTW algoritmu pri određivanju optimalne putanje.

Distance u ovim koracima se izračunavaju prema formulama: *D*(*i*-1,*j*) + *d*(*i*,*j*), *D*(*i*,*j*-1) + *d*(*i*,*j*) i *D*(*i*-1,*j*-1) + *d*(*i*,*j*), za horizontalni, vertikalni i dijagnonalni korak respektivno. U slučaju horizontalnog koraka, nagib optimalne putanje je jednak nuli a test signal se komprimuje. Prilikom vertikalnog koraka nagib je 90˚ pri čemu se signal proširuje

("rasteže"), dok u dijagonalnom koraku optimalna kriva ima ugao od 45˚ i tada nema promene u strukturi signala. Optimalna putanja se posle izvesnog broja iteracija završava u položaju (T_x,T_y). Dobijena konačna globalna distanca $D(T_x,T_y)$ predstavlja cenu za dati par reči i koristi se kao mera sličnosti dva signala u postupku prepoznavanja govora. Putanja sa najnižom cenom oslikava najbolje poklapanje između test signala i referentne reči.

2.2.1 OGRANIČENJA DTW ALGORITMA U PREPOZNAVANJU GOVORA

Uprkos svojoj jednostavnosti, automatsko prepoznavanje govora zasnovano na DTW algoritmu je ograničeno sa nekoliko nedostataka:

- Veliki hendikep ovih sistema je nemogućnost modelovanja glasova i manjih govornih segmenata, zbog čega se ovi sistemi ne mogu koristiti u prepoznavanju kontinualnog govora.

- DTW sistemi mogu modelovati samo izolovane reči, što ih čini nepraktičnim za prepoznavanje velikih korpusa reči. Modelovanje varijabilnosti u govoru je u direktnoj zavisnosti od broja snimaka reči sa kojima DTW sistem raspolaže, usled čega ovi sistemi imaju problem sa postizanjem visokih performansi u prepoznavanju govora.

- DTW sistemi su uglavnom *speaker-dependent* tipa i ne koriste se u prepoznavanju govora nezavisnog od govornika.

- Procesorski su zahtevni, jer je za svako testiranje neophodno ponovno izračunavanje matrice lokalnih distanci i optimalne putanje.

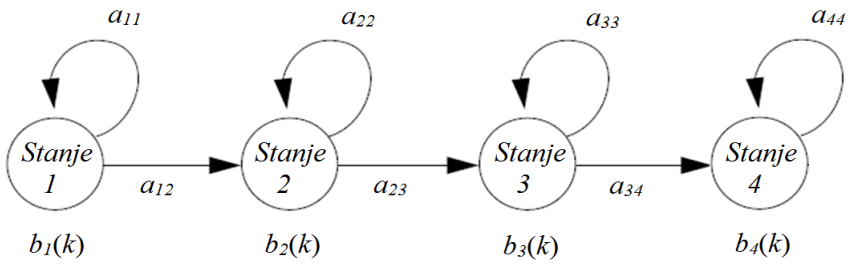
Bez obzira na svoje nedostatke, DTW algoritam se u jednom obliku i dalje koristi u složenim hibridnim sistemima za automatsko prepoznavanje govora kao efikasan i praktičan način nelinearne normalizacije trajanja reči i usaglašavanja različite brzine izgovora.

2.3 SKIRENI MARKOVLJEVI MODELI (HMM)

Pored determinističkog modelovanja govora, postoji i stohastički pristup u kome se govor može okarakterisati kao slučajan Markovljev proces. Ovakav način

sagledavanja govora je rezultovao primenom skrivenih Markovljevih modela (*Hidden Markov Models - HMMs*) u akustičkom modelovanju, predikciji i prepoznavanju govora. Prvi teorijski rezultati o skrivenim Markovljevim modelima su publikovani kasnih šezdesetih i ranih sedamdesetih [Baum et al., 1966-1972], dok su prve implementacije ovih rezultata u oblasti prepoznavanja govora izvršene sedamdesetih godina [Baker, 1975; Jelinek, 1975; Bakis, 1976]. Iako je ova tehnika poznata već skoro 50 godina, ona se pojavila u masovnoj upotrebi tek tokom posledenje decenije sa razvojem procesorske moći računara i danas predstavlja sastavni deo većine ASR sistema.

Stohastičko modelovanje govora je zasnovano na statističkom opisivanju modela definisanim sa određenim brojem stanja. Posmatrajmo HMM sistem koji je opisan tako da se u svakom trenutku može naći u nekom od N mogućih stanja, S_1,S_2, ... S_N, Slika 2.10. Sistem menja stanja u regularnim ekvidistantnim vremenskim intervalima shodno unapred definisanim verovatnoćama. Svako stanje je opisano funkcijama gustine verovatnoće, najčešće mešavinama Gausovih raspodela, koje opisuju ponašanje opservacionih vektora akustičkih obeležja. Otuda sličnost sa GMM (*Gaussian Mixture Model - GMM*) sistemima, koji se često nazivaju i skrivenim Markovljavim modeloma sa jednim stanjem. Pored raspodela verovatnoće, HMM je takođe opisan sa skupom verovatnoća prelaska (tranzicije) iz stanja *i* u stanje *j* čime se opisuje vremenski sled događaja. Prema tome, skriveni Markovljevi modeli se sastoje iz dva procesa - prvog u kome se vrši tranzicija kroz skup stanja i drugog u kome se za svako stanje generiše niz opservacija. U ovakvom sistemu nema jasne predstava o tome u kome je stanju sistem trenutno, odnosno stanja su "sakrivena" pa otuda i potiče naziv ovih modela.

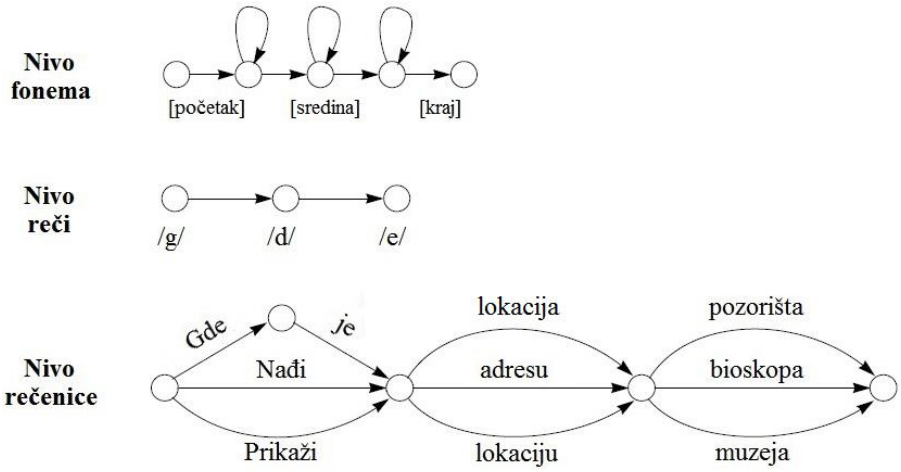


Slika 2.10 Jednostavan HMM model reči sa četiri stanja.

Na slici 2.10 je prikazan HMM model reči sa N=4 stanja koji će poslužiti kao primer definisanja elemenata HMM-a i generisanja sekvenci opservacija. Skriveni Markovljev model se definiše pomoću sledećih parametara:

1) *N* je broj stanja u modelu. Iako su stanja skrivena, za veliki broj praktičnih problema postoji jasna fizička interpretacija stanja i njihovih fizičkih značenja. Na primer u slučaju govora, jasna je veza između HMM stanja i segmenata reči (rečenice). Najčešće se tako modeluju akustičke predstave različitih govornih jedinica (trifoni, difoni, monofoni, subfonemi...) i njihov položaj u reči (rečenici). U opštem slučaju, stanja HMM sistema se definišu tako da se u svako stanje može stići iz bilo kog drugog stanja (to je takozvani ergodični model), međutim u slučaju govora između stanja se koristi *left-to-right* tip tranzicije [Rabiner et al., 1993; Rabiner et al., 2007]. Obično se sa $S = \{ S_1, S_2, \ldots S_N \}$ označava skup svih mogućih stanja, a sa q_t stanje modela u nekom vremenskom trenutku *t*. Na slici 2.11. je prikazana hijerarhiski uređena struktura stanja i njihovih tranzicija prilikom modelovanja fonema, reči i rečenice.

2) *M* je broj različitih opservacija koje se mogu realizovati (generisati) iz stanja. Ovaj parametar na primer može predstavljati dimenziju nekog diskretnog alfabeta, odnosno broj simbola, broj reči itd... Opservacioni simboli odgovaraju fizičkom izlazu sistema koji se modeluje. Skup ovih simbola se obično označava na sledeći način $V = \{ V_1, V_2, \ldots V_M \}$.



Slika 2.11 Hijerarhiska struktura HMM modela.

3) Skup raspodela verovatnoća tranzicija, $A = \{a_{ij}\}$, gde je:

$$a_{ij} = P[q_{t+1} = S_j \mid q_t = S_i], \qquad 1 \le i,\ j \le N. \tag{2.7}$$

4) Pošto se radi o verovatnoćama važi da je $a_{ij} > 0$ za svako i i j. Obično su verovatnoće a_{ii} tranzicije veće od verovatnoća a_{ij} tranzicija. Za neke druge tipove HMM može se apriori usvojiti da je $a_{ij} = 0$ ili $a_{ij} = 1$ za neke specifične parove ij.

5) Skup raspodela verovatnoće pojave opservacionih simbola u određenim stanjima, $B = \{b_j(k)\}$, gde je:

$$b_j(k) = P[o_t = v_k \mid q_t = S_j], \qquad 1 \le j \le N$$
$$1 \le k \le M, \tag{2.8}$$

raspodela verovatnoće generisanja simbola u stanju j.

6) Inicijalna raspodela verovatnoće stanja $\pi = \{\pi_i\}$, gde je:

$$\pi_i = P[q_1 = S_i], \quad 1 \le j \le N. \tag{2.9}$$

Sa datim parametrima N, M, A, B i π, HMM sistem je u potpunosti opisan i može da generiše odgovarajuću sekvencu opservacija:

$$O = (o_1, o_2, o_2 \ldots o_{T-2}, o_{T-1}, o_T), \tag{2.10}$$

gde je svaka od opservacija o_t neki od simbola iz skupa V, dok T predstavlja ukupan broj opservacija u sekvenci. Postupak generisanja sekvenci opservacije se realizuje u sledećim koracima:

1) Bira se početno stanje $q_1 = S_j$ prema inicijalnoj raspodeli verovatnoća π.

2) Vremenski trenutak se postavlja na $t = 1$.

3) Bira se opservacija $o_t = v_k$ shodno raspodeli verovatnoća pojave simbola u stanju S_j, $b_j(k)$.

4) Vrši se prelazak u novo stanje $q_{t+1} = S_i$ na osnovu raspodela verovatnoća za tranziciju stanja, odnosno a_{ij}.

5) Vreme se postavlja na $t = t + 1$ i vraća se na korak 3, ukoliko je $t < T$. U obratnom, za slučaj da je $t = T$, procedura se se završava.

Ovako opisana procedura se može koristiti kao generator opservacija a istovremeno kao model koji opisuje sekvencu opservacija za zadati HMM. Na osnovu gornje diskusije postaje jasno da je skriveni Markovljev model opisan sa tri parametra A, B i π, osnosno tripletom zapisanim u skraćenoj notaciji:

$$\lambda = (A, B, \pi). \tag{2.11}$$

Sa željom da se formira HMM koji će efikasno opisivati određeni fizički proces ili u pogledu primene već postojećeg HMM-a, javljaju se tri ključna problema:

1) Prvi problem je efikasno izračunavanje verovatnoće $P(O|\lambda)$, koja predstavlja verovatnoću da je neku datu opservaciju $O = (o_1, o_2, \ldots o_T)$ generisao model λ. U osnovi ovog problema je postupak izračunavanja verovatnoće generisanja neke opservacije iz poznatog modela. Ovaj problem se može tretirati i kao problem evaluacije u kojoj se ocenjuje koliko neka sekvenca obeležja odgovara usvojenom modelu. Rešavanje ovog zadatka je od krucijalnog značaja za rešavanje problema prepoznavanja oblika. Matematičko rešavanje ovog problema se svodi na primenu takozvanog *forward* algoritma.

2) Drugi problem je određivanje odgovarajuće sekvence stanja $Q = (q_1, q_2, \ldots q_t)$ koja najbolje opisuje i odgovara poznatoj sekvenci opservacija $O = (o_1, o_2, \ldots o_T)$ i modelu $\lambda = (A, B, \pi)$. Ovaj problem se svodi na rešavanje "skrivenosti" u skrivenim Markovljevim modelima, čime se pokušava da se da odgovor na pitanje koja sekvenca stanja "tačno" odgovara poznatoj sekvenci opservacija. Koristi se u cilju ispitivanja statistike pojedinih stanja, kao i za oređivanje

strukture modela radi optimizacije sekvenci stanja u prepoznavanju kontinualnog govora.

3) Treći problem razmatra postupak određivanja parametara HMM sistema, $\lambda = (A, B, \pi)$ ukoliko su na raspolaganju poznate sekvence opservacija. Problem se svodi na određivanje maksimalne verovatnoće $P(O|\lambda)$, što se matematički postiže upotrebom *forward-backward* algoritma. Polazi se od skupa uzoraka sa kojima se raspolaže, tj. sa sekvencama opservacija koje se koriste u *forward-backward* algoritmu u cilju optimizacije parametra HMM sistema radi što boljeg fitovanja ulaznih podataka. Ovaj proces je neophodan i obavlja se na samom početku u fazi obuke HMM sistema.

U sistemima za prepoznavanje izolovanih reči, navedeni problemi se svode na sledeće. Za svaku od *R* reči iz rečnika formira se poseban HMM sa *N* stanja. Govorni signal se predstavlja kao sekvenca okvira akustičkih obeležja. Dobijene sekvence obeležja se koriste i procesu formiranja baza sa kojom će se vršiti obuka HMM, pri čemu je svaka reč predstavljena sa odgovarajućim brojem uzoraka odnosno sekvenci akustičkih obeležja (generisanih od strane jednog ili više govornika). Prvi zadatak je da se formiraju pojedini modeli za svaku reč ponaosob. Ovaj problem je definisan problemom broj 3. Da bismo razumeli fizički postupak generisanja dobijene opservacione sekvence, koristimo se problemom 2 kako bi se izvršila segmentacija sekvence u pojedina stanja. Konačno, kada se formira *R* različitih modela za svaku reč ponaosob, može se vršiti prepoznavanje izgovorenih reči tako što se za snimljenu reč izvrši rešavanje problema 1 za svaki od generisanih modela. Model koji dobije najveće poverenje (najveću odgovarajuću verovatnoću, verodostojnost) biće prepoznat kao izgovorena reč.

2.3.1 OGRANIČENJA HMM SISTEMA U PREPOZNAVANJU GOVORA

HMM sistemi i pored svojih naprednih performansi koje ih čine trenutno najzasutpljenijim sistemima za prepoznavanje govora, imaju nekoliko poznatih slabosti:

- Pretpostavka prvog reda (*The First-Order Assumtion*) podrazumeva da su sve verovatnoće zavisne isključivo i jedino od trenutnog stanja, što je netačno. Posledica ove pretpostavke na kojoj su zasnovani HMM sistemi je otežano

modelovanje koartikulacije koja je pod snažnim uticajem događaja koji su prethodili. Druga posledica je pogrešno modelovanje trajanja usled korišćenja eksponencijalno opadajućih raspodela umesto Puasonove raspodele (*Poison*) ili neke druge pogodnije raspodele zvonastog oblika [Tebelskis, 1995].

- Pretpostavka nezavisnosti (*Independance Assumtion*) nalaže da nema korelacije između susednih ulaznih frejmova. Ova pretpostavka je takođe pogrešna u pogledu govornih signala. Zbog toga HMM sistemi analiziraju samo po jedan okvir govora u trenutku i ne koriste prednosti analize susednih okvira.

- Modelovanje raspodela verovatnoća nije toliko tačno. Usled apriori izbora statističkih raspodela (mešavina Gausovih raspodela) za svako stanje modela, postoji slabo poklapanje sa realnim stanjem akustičkog prostora.

- Kriterijum procene maksimalne verodostojnosti (*Maximum Likelihood Estimate - MLE*) i njegovo forsiranje umeseto *a posteriori* verovatnoća u procesu treninga dovodi do slabije diskriminacije između akustičkih modela. (ograničen skup trening podataka i njima odgovarajući akustički modeli).

Zbog opisanih nedostataka, HMM sistemi mogu imati dobre performanse samo u modelovanju kontekstualno zavisnih monofona [Hwang et al., 1993].

2.4 REZIME

U ovom poglavlju je ukratko opisan osnovni koncept ASR sistema kao i matematička formulacija problematike donošenja odluke u procesu automatskog prepoznavanja govora. Dat je kratak osvrt na dve *pattern recognition* tehnike - jedne zasnovane na *tamplate matching* algoritmu (DTW) i druge utemeljene na statističkom modelovanju (HMM). Dinamičko vremensko usklađivanje, kao najstarija i najjednostavnija *pattern recognition* tehnika, je i dalje u upotrebi u raznim praktičnim aplikacijama nelinearne normalizacije trajanja signala. Pomoću ovog algoritma se na jednostavan način postiže usklađivanje brzine izgovora u govornim signalima. Ipak, nedostaci DTW algoritma poput: procesorski zahtevnog proračuna, lošeg modelovanja varijabiliteta, nemogućnosti prepoznavanja kontinualnog govora i primene samo u *speaker-dependent* zadacima, su u velikoj meri ograničile upotrebu ovog algoritma u

ozbiljnijim ASR zadacima. Sa druge strane, današnji sofisticirani ASR sistemi su uglavnom zasnovani na statističkoj klasifikaciji. Međutim i HMM sistemi imaju svoje nedostatke, pre svega u pogledu: *The First-Order* i *Independance* pretpostavki, definisanja *a priori* verovatnoća i loše diskriminacije usled korišćenja MLE (*Maximum Likelihood Estimation*). U narednom poglavlju biće detaljno opisana još jedna popularna *pattern recognition* tehnika zasnovana na veštačkim neuralnim mrežama (*Artificial Neural Networks - ANN*), koja u izvesnoj meri nadomešćuje nedostatke prethodno opisanih sistema.

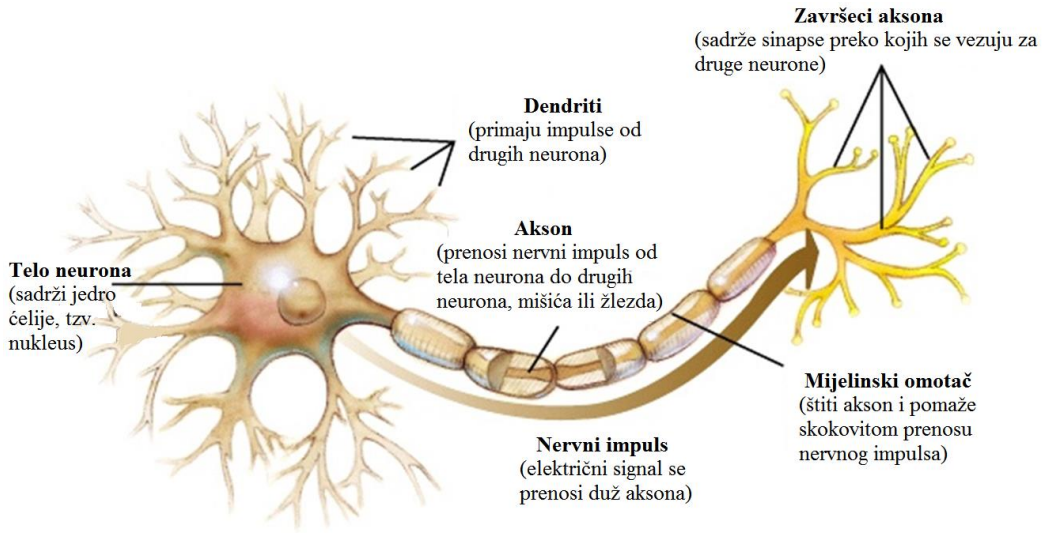
3 UVOD U VEŠTAČKE NEURALNE MREŽE

U ovom poglavlju je opisana posebna *pattern recognition* tehnika poznata kao veštačke neuralne mreže (*Artificial Neural Networks - ANN*). Veštačke neuralne mreže predstavljaju relativno novu generaciju sistema za informaciono procesiranje koji imaju mogućnost učenja, memorisanja i procesiranja na osnovu podataka sa kojima se obučavaju. Inspirisane su biološkim nervnim sistemom čoveka i koriste se u različitim zadacima mašinskog učenja poput: uparivanja ("mečovanja") i klasifikacije oblika, aproksimacije funkcija, optimizacije, vektorske kvantizacije i klasterizacije podataka. Ovo poglavlje sadrži informacije o istorijskom razvoju i osnovnim teorijskim aspektima veštačkih neuralnih mreža, o njihovim najbitnijim tipovima, topologijama i metodama obuke. Poseban akcenat je stavljen na *feedforward* mreže i višeslojne perceptrone (*Multi Layer Perceptron - MLP*), *Backpropagation* proceduru obuke mreža, kao i na dubinsko učenje (*Deep learning*) i dubinske autoenkodere (*Deep autoencoder*). Opisane su njihove mogućnosti, ograničenja, prednosti i dodirne tačke sa konvencionalnim statističkim metodama.

3.1 ISTORIJSKI RAZVOJ VEŠTAČKIH NEURALNIH MREŽA

U ovom Istraživanja neuralnih mreža su započeta još u 19. veku sa utemeljivanjem moderne naučne discipline neurobiologije i sa prvim ozbiljnijim studijama čovekovog nervnog sistema [Cajal, 1892]. Tada je po prvi put ustanovljeno

da se nervni sistem sastoji iz velikog broja fizičkih nervnih jedinica, odnosno ćelija, poznatih kao neuroni, koji su međusobno povezani i komuniciraju prenosom električnih impulsa duž specifičnih nervnih vlakana, takozvanih aksona (*axons*). Aksoni se granaju u vidu brojnih nastavaka, dendrita (*dendrons*), i povezuju se sa hiljadama drugih neurona preko posebnih međućelijskih veza tj. sinapsi (*synapsis*), Slika 3.1.

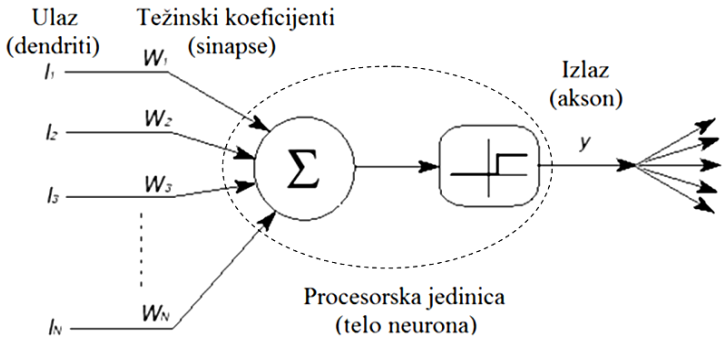


Slika 3.1 Anatomija (građa) multipolarnog neurona.

Ovaj biološki model neurona je detaljno ispitivan i elaboriran narednih decenija. Otkrivene su različite vrste neurona, njihove karakteristike, analizirane su pobude i odzivi neurona na električne signale, utvrđeno je i mapirano njihovo povezivanje i grupisanje u određenim nervnim područjima itd. Iako je sa stanovišta neurobiologije ispitivanje izolovanog neurona relativno jednostavno, pitanje analize međusobnog funkcionisanja većeg broja neurona koje omogućava kompleksne funkcije poput percepcije i kognicije je dugo bilo neostvarivo bez savremenih računarskih sistema. Sa razvojem procesorske snage računara, analiza i simulacija veće grupe neurona je postala izvodljiva što je dovelo do novih eksperimenata i boljeg razumevanja njihovih karakteristika.

Sa novim saznanjima i spoznajom do tada neistraženog čovekovog nervnog sitema, pojavile su se i ideje o njegovom veštačkom modelovanju. Tako su McCulloc i Pits 1943. godine uspešno napravili prvi kompjuterski model neurona, čime počinje era veštačkih neuralnih mreža. Ovaj model je osmišljen kao binarni matematički model koji

na osnovu ulaznog sitimulusa i odgovorajuće definisane funkcije praga (aktivacione funkcije) određuje izlaz veštačkog neurona - nulu ili jedinicu, Slika 3.2.



Slika 3.2 Matematički model neurona koji su predložili McCulloc i Pits.

Ubrzo je ustanovljeno da sistem ovakvih neurona sa dodavanjem odgovarajućih težinskih koeficijenta može da modeluje bilo koju proizvoljnu matematičku funkciju [Minsky, 1967]. Kasnija istraživanja su se bavila automatskim pronalaskom ovih težinskih koeficijenata, odnosno procedurom obuke neuralnih mreža. Razvijen je poseban tip neuralnih mreža, poznat kao jednoslojni perceptron (*single-layer perceptrone*) i postupak obuke mreže kroz iterativnu proceduru [Rosenblatt, 1962]. Pokazano je da tokom ovakvog procesa obuke neuralna mreža konvergira ka određenom skupu težinskih koeficijenata koji daju odgovarajuću funkciju sve dok je ta funkcija rešiva sa odgovarajućom topologijom mreže [Tebelskis, 1995]. Međutim, Minsky i Papert su u svojim kasnijim analizama 1969. godine [Minsky et al., 1969] izrazili skepticizam u pogledu mogućnosti ANN i istakli da pored jednoslojnih i višeslojni perceptroni (*multi-layer perceptrone*) imaju ozbiljna ograničenja u modelovanju pojedinih funkcija. Iako ovi zaključci nisu bili ispravni, oni su u velikoj meri uticali na zastoj i prekid daljih istraživanja na polju veštačkih neuralnih mreža. Ova pauza je trajala skoro 15 godina sve do 1982. godine kada je Hopfild ponovo oživeo istraživanja na ovu temu. On je sugerisao da se neuralne mreže mogu analizirati u pogledu funkcije energije, što je rezultovalo pojavom Bolcmanove mašine (*Boltzmann Machine*) odnostno tzv. Hopfildove mreze [Ackley et al., 1985] koja predstavlja stohastičku rekurentnu mrežu koja se može obučiti tako da izvršava različite *pattern recognition* zadatke od mapiranja proizvoljnih obrazaca (*pattern mapping*) do njihove nadogradnje (*pattern completion*). Ubrzo potom je razvijen i popularizovan mnogo brži način obuke neuralnih mreža, poznat kao *backpropagation* metod ili metod propagacije

u nazad, pomoću koga se višeslojni perceptroni mogu obučiti tako da aproksimiraju bilo koju željenu funkciju [Rumelhart et al, 1986]. Sa ovim otkrićem premošćene su sve do tada poznate prepreke neuralnih mreža a istraživanja su ponovo zaživela. Iako su ANN imale neobičan tok istorije, one su još uvek u fazi razvoja. Danas ANN, posebno dubinske neuralne mreže (*Deep Neural Networks – DNN*), nalaze veoma širok spektar primena u različitim praktičnim zadacima, među kojima i u automatskom prepoznavanju govora/govornika.

3.2 POREĐENJE NEURALNIH MREŽA I KONVENCIONALNIH RAČUNARA

Skup veštačkih neurona koji su gusto povezani i formiraju složenu strukturu mreže ispoljavaju neke od bitnih osobina bioloških neuralnih mreža. Pošto se ovakve neuralne mreže implementiraju na kompjuterima valja uporediti njihove karakteristike i performanse pre svega u pogledu: brzine procesiranja podataka, veličine i složenosti sistema, zauzeća memorijskih resursa, tolerancije na greške i kontrole celokupnog procesa.

- Brzina. Sa stanovišta brzine procesiranja informacija, neuralne mreže su prilično spore a vremenski ciklus za izračunavanje pojedinih matematičkih operacija se meri u milisekundama. Savremeni računari, pak, izvršavaju matematičke operacije veoma brzo za svega nekoliko nanosekundi. Prema tome, konvencionalni računari su u ovom pogledu milion puta brži od neuralnih mreža.

- Procesiranje. Računari su zasnovani na sekvencijalnom procesiranju podataka, što znači da obavljaju jednu po jednu instrukciju u nizu. Sa druge strane, neuralne mreže mogu sprovoditi masovne paralelne operacije, što objašnjava superiornost čovekovog načina procesiranja podataka koji i pored sporih procesorskih jedinica mnogo brže obavlja određene zadatake u poređenju sa računarima.

- Složenost. Ljudski mozak se sastoji od velikog broja procesorskih jedinica, otprilike od 10^{11} različitih tipova neurona i 10^{15} interkonekcija, pri čemu procesiranje podataka nije ograničeno na samo par neurona. Ovako veliki broj neurona i kompleksnost njihovih veza daje neuralnim mrežama izuzetno veliku moć u procesiranju podataka. Veštačke neuralne mreže za razliku od bioloških

imaju znatno manji broj neurona ali su i pored toga daleko moćnije u rešavanju *pattern recognition* problema od konvencionalnih računara.

- Memorisanje. Neuralne mreže skladište podatke u vidu snaga veza koje spajaju neurone (tj. u vidu težinskih koeficijenta). U računarima informacije su određene svojom lokacijom u memoriji. Skladištenje svake nove informacije na istoj lokaciji rezultuje brisanjem prethodne, dok se kod neuralnih mreža sa dolaskom nove informacije samo podešavaju i uobličavaju težinski koeficijenti, bez brisanja starih (prethodnih) informacija. Na taj način se informacija u mozgu adaptira, dok se kod računara striktno briše ili memoriše (čuva u memoriji).

- Tolerancija na greške. U slučaju greške ili otkaza pojedinih neurona, informacije u neuralnoj mreži su i dalje očuvane. Kompjuteri nisu tolerantni na slične otkaze, pri čemu se u slučaju greške informacija gubi i ne može se povratiti iz memorije.

- Kontrola. Mozak ne poseduje centralnu jedinicu koja kontroliše procese, dok su računari centralizovani i poseduju kontrolnu jedinicu koja nadgleda i upravlja procesima i aktivnostima sistema. Kod neuralnih mreža svaki neuron u zavisnosti od lokalno dostupnih informacija zasebno obrađuje i prosleđuje podatke.

3.3 OSNOVE VEŠTAČKIH NEURALNIH MREŽA

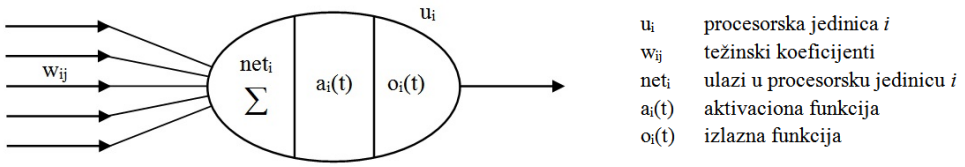
U ovoj sekciji biće ukratko opisani osnovni koncepti veštačkih neuralnih mreža. Postoje različiti tipovi ANN, međutim sve njih karakterišu neke zajedničke osobine poput:

- Posedovanja procesorske jedinice,
- Povezanosti neurona u vidu neuronskih veza,
- Procesiranja podataka, i
- Procedure obuke.

U nastavku teksta svaka od ovih karakteristika neuralnih mreža biće detaljno opisana.

3.3.1 PROCESORSKE JEDINICE

Neuralne mreže se sastoje iz velikog broja gusto povezanih procesorskih jedinica, takozvanih čvorova (*nodes*), koji uobičajeno rade simultano i u paraleli u nekoj od uređenih regularnih arhitektura neuralnih mreža. Sve računske operacije u neuronima se obavljaju u ovoj jedinici. Opšti model procesorske jedinice veštačkog neurona je prikazan na Slici 3.3. U literaturi i šemama poput ove, procesorske jedinice se obično predstavljaju krugom.



Slika 3.3 Matematički model procesorske jedinice neurona.

Model procesorske jedinice se sastoji iz tri dela. Prvi, takozvani ulazni deo, služi za primanje ulaznih stimulusa, njihovo sumiranje i množenje odgovarajućim težinskim koeficijentima w_{ij} (sinaptičkim težinama). Ovako sumirana vrednost se naziva interna aktivaciona vrednost procesorske jedinice. Stanje neuralne mreže je u svakom trenutku opisano skupom ovakvih aktivacionih vrednosti. To stanje se menja u vremenu, od trenutka do trenutka, u zavisnosti od promene ulaznih podataka. Drugi deo procesorske jedinice predstavlja aktivacionu funkciju koja na osnovu sumirane skalarne vrednosti odlučuje da li se radi o pobudi (pozitivne vrednosti) ili inhibiciji (negativne vrednosti). U zavisnosti od donete odluke treći izlazni deo procesorske jedinice generiše signal koji se prosleđuje ostalim neuronima u okolini. Svi signali, kako ulazni tako i izlazni, u zavisnosti od tipa mreže mogu biti kontinualni ili diskretni i deterministički ili stohastički.

U zavisnosti od lokacije unutar neuralne mreže, procesorske jedinice ili čvorovi se dele na ulazne (nalaze se u prvom sloju mreže i primaju stimuluse), skrivene (nalaze se unutar mreže i obrađuju podatke) i izlazne (prosleđuju donete odluke, npr. motorne komande).

3.3.2 NEURONSKE VEZE (SINAPSE)

Procesorske jedinice su međusobno povezane skupom veza koje su opisane odgovarajućim težinskim koeficijentima. Svaki težinski koeficijent ima realnu vrednost koja se kreće od $-\infty$ do $+\infty$, mada se u nekim slučajevima koriste ograničeni opsezi (čime se smanjuju sposobnosti mreže). Snaga konekcije ili težina veze definiše meru u kojoj će ta veza uticati na susedne jedinice. Veće vrednosti težinskih koeficijenata karakterišu snažnije i informaciono bitnije veze. Pozitivne vrednosti deluju na pobuđivanje mreže, dok negativne vrednosti utiču da neuroni inhibiraju jedni druge. Neuronske veze su uglavnom jednosmerne, mada mogu biti i dvosmerne što je slučaj sa specifičnim topologijama mreža u kojima ne postoje striktno definisani ulazni i izlazni čvorovi (neuroni).

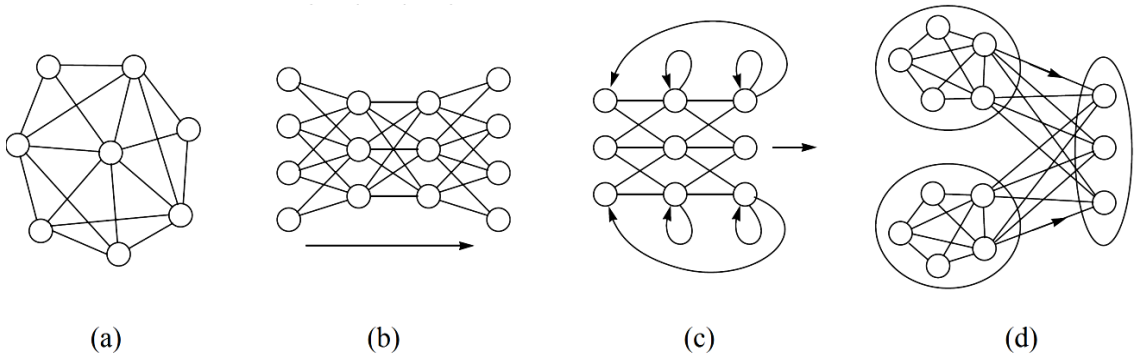
Težinski koeficijenti predefinišu reakciju neuralne mreže na ulazni stimulus. U tom pogledu težine veza predstavljaju dugotrajnu memoriju (*long-term memory*) odnosno znanje mreže [Tebelskis, 1995]. Tokom procesa obuke težine veza se menjaju, pri čemu su te promene spore i postepene, baš kao i proces učenja. Posle završene obuke ovi koeficijenti ostaju fiksni. Sa druge strane, u neuralnim mrežama postoji i kratkoročna memorija (*short-term memory*) koja je okarakterisana aktivacionim stanjem mreže u vidu prenosne (aktivacione) funkcije trenutnog ulaza u mrežu. U zavisnosti od načina povezivanja neurona, razlikuju se sledeće topologije mreža:

a) Nestrukturirane (neuređene) mreže, koje su najpogodnije za *pattern completion* zadatke u kojima je cilj generalizovanje i rekonstruisanje ulaznih parcijalnih (nepotpunih) podaka sa kojima mreža nije obučavana;

b) Slojevite mreže, koje su pogodne za *pattern association* zadatke (klasifikacija, klasterovanje, dijgnostika, asocijacija), u kojima je zadatak mreže da napravi asocijaciju između određenih ulaznih i izlaznih podataka (mapiranje ulaznih podataka u izlazne);

c) Rekurentne mreže, ili mreže sa povratnim spregama, koje su uspešne u *pattern sequencing* zadacima tj. zadacima u kojima se prate aktivnosti mreže tokom vremena, i

d) Modularne mreže, koje su sastavljene od više jednostavnijih mreža.



Slika 3.4 Različite topologije neuralnih mreža: (a) nestruktuirana, (b) slojevita, (c) rekurentna i (d) modularna.

Nestrukturirane mreže mogu sadržati kružne veze, što ih automatski čini rekurentnim. Slojevite mreže mogu, a i ne moraju biti rekurentne, dok modularne mreže mogu sadržati više različitih topologija mreža. Nestrukturirane mreže imaju dvosmerne veze, dok svi ostali tipovi neuralnih mreža imaju samo jednosmerne veze.

Veze između dve grupe neurona, kao u slučaju povezivanja dva sloja neurona, mogu biti:

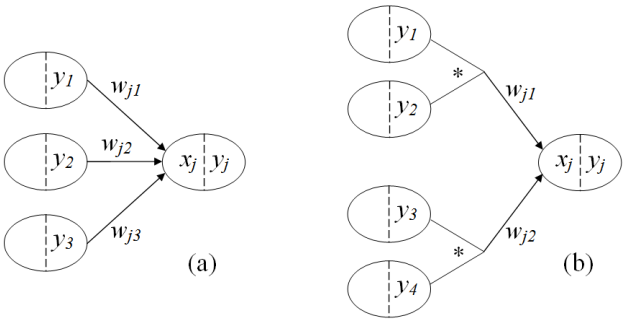
1) Potpune, kada je svaki neuron povezan sa svakim neuronom,
2) Slučajne, kada su samo neki neuroni povezani, i
3) Susedne, kada su povezani samo najbliži susedni neuroni.

Potpuno povezane mreže imaju najveći stepen slobode, pa teorijski mogu da nauče mnogo više funkcija od ograničenih mreža. Ipak ovakva topologija mreže nije uvek najbolje rešenje. Ukoliko mreža ima suviše slobode, ona može jednostavno memorisati skup podataka sa kojim se obučava, bez upuštanja u proces učenja i potrage za skrivenim obrascima koji diskriminišu podatke. Mana ovakvih mreža je loša generalizacija novih ulaznih podataka. Ova pojava je poznata pod terminom *overfitting*. Ograničavanjem složenosti mreže u pogledu povezanosti neurona, sloboda mreže se ograničava, njihova struktura se pojednostavljuje, mreže postaju ekonomičnije u pogledu računskih i memorijskih zahteva, a pri tome se postiže mnogo bolja generalizacija podataka.

3.3.3 PROCESIRANJE

Ova Procesiranje, odnosno obrada podataka, uvek počinje sa dolaskom stimulusa na ulaz mreže pri čemu se prvo izračunavaju aktivacione vrednosti ulaznih procesorskih jedinica pa onda preostalih koje slede. To izračunavanje aktivacionih vrednosti može biti sinhrono (istovremeno i u paraleli kod svih jedinica) ili asinhrnono (odvija se jedno po jedno u vremenu, ili po nekom slučajnom, prirodnom redosledu) [Tebelskis, 1995]. U nestruktuiranim mrežama ovaj proces izračunavanja se naziva širenje aktivacione vrednosti (*spreading activation*), dok se u slojevitim mrežama zove propagacija unapred (*forward propagation*), zbog prirodnog smera propagacije impulsa od ulaznog ka izlaznom sloju mreže. U slojevitim mrežama aktivacione vrednosti se stabilizuju u svim procesorskim jedinicama čim se dostigne izlazni sloj mreže. U rekurentnim mrežama usled postojanja povratnih sprega postoji mogućnost da se ove vrednosti nikada ne stabilizuju, već se kontinualno ažuriruaju zajedno sa procesorskim jedinicama.

Proces ažuriranja aktivacione vrednosti u procesorskoj jedinici se odvija u dva koraka. Prvo se izračunava takozvana integraciona funkcija, koja sumira ulazne signale u procesorskoj jedinici, a potom tako izračunatu internu aktivacionu vrednost prosleđuje tzv. transfer ili aktivacionoj funkciji. Na Slici 3.5. su ilustrovana dva moguća slučaja izračunavanja interne aktivacione vrednosti.



Slika 3.5 Izračunavanje interne aktivacione vrednosti: (a) tipični oblik i (b) slučaj "sigma-pi".

Prvi slučaj predstavlja najčešću situaciju, gde se interna aktivaciona vrednost, x_j, izračunava kao suma ulaznih stimulusa koji su prethodno pomnoženi težinskim koeficijentima:

$$x_j = \sum_i y_i w_{ji} \,, \tag{3.1}$$

gde su y_i izlazne aktivacione vrednosti prethodnih neurona, w_{ji} težinski koeficijenti veze između *i* i *j* procesorskih jedinica. U drugom slučaju, koji se ređe javlja, izlazne aktivacione vrednosti prethodnih procesorskih jedinica se prvo međusobno množe pa se tek onda skaliraju težinskim koeficijentima. Ovakav tip veze se zove "sigma-pi" konekcija i pritom se koristi sledeća integralna funkcija:

$$x_j = \sum_j w_{ji} \prod_{k \in k(i)} y_k \ . \tag{3.2}$$

Čest je slučaj da integralne funkcije pored ulaznih stimulusa i težinskih koeficijenata zavise i od stanja samog neurona. To stanje se modeluje realnom veličinom θ_j i nosi naziv "*bias*" ili prag (*treshold*). U ovom slučaju formula (3.1) se zapisuje u sledećem obliku:

$$x_j = \sum_i y_i w_{ji} + \theta_j \, . \tag{3.3}$$

Ovaj izraz se može pojednostaviti ukoliko pored postojećih *N* ulaza uvede još jedan ulaz (*N*+1 ulaz) koji će se smatrati da je uvek jednak 1 ($y_0 = 1$). Na taj način prag se može modelovati težinskim koeficijentom w_0 i u nastavku analize se smatra da se radi o neuronu sa nultim pragom.

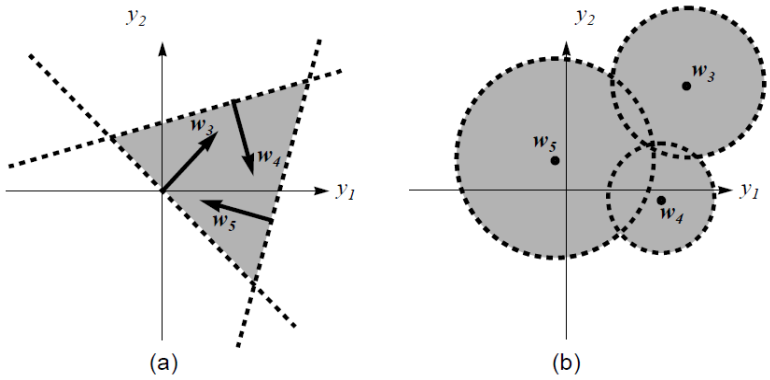
Pored linearnih funkcija, opisanih izrazima (3.1) i (3.3), u upotrebi su i drugi tipovi integralnih funkcija poput: kvadratne, sferične i polinomske funkcije. Ove funkcije su redom definisane sledećim jednačinama:

$$x_j = \sum_i y_i^2 w_{ji} + \theta_j \, , \tag{3.4}$$

$$x_j = \rho^{-2} \sum_i (y_i - w_{ji})^2 + \theta_j \, , \tag{3.5}$$

$$x_j = \sum_i \sum_k y_i y_k w_{ji} + y_i^{\alpha_i} + y_k^{\alpha_k} + \theta_j \, . \tag{3.6}$$

Za razliku od lineanih funkcija koje omogućavaju procesorskim jedinicama modelovanje hiperravni u procesu klasifikacije podataka, sferična funkcija (3.5) modeluje hipersfere, Slika 3.6, i koristi se u *Learned Vector Quantization* (LKV) i *Radial Bias Function* (RBF) tipovima neuralnih mreža.

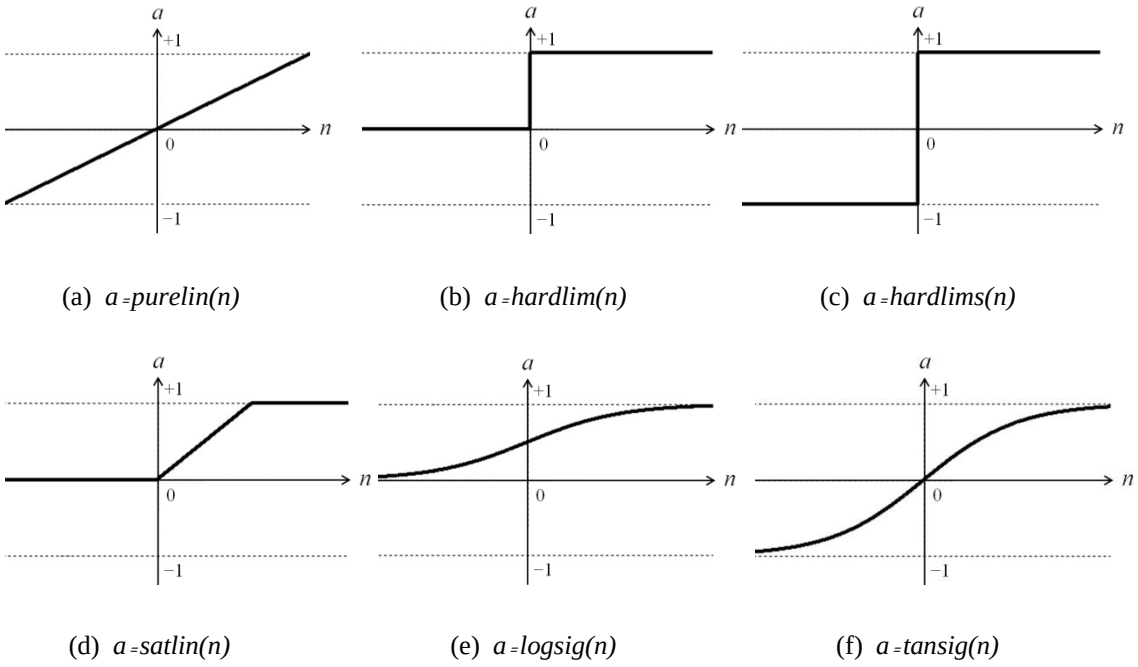


Slika 3.6 Formiranje složenih funkcija na osnovu: (a) hiper-ravni i (b) hipersfrea. (Tebelskis, 1995).

Posle izračunavanja interne aktivacione vrednosti, x_j, sledi izračunavanje izlazne aktivacione vrednosti, y_j, koja je funkcija od promenljive x_j. Ona predstavlja transfer funkciju i ima zadatak da internu aktivacionu vrednost x_j transformiše u neku prihvatljivu vrednost izlaza, to jest u aktivacionu vrednost. Najčešće se zahteva da te izlazne veličine budu ograničenog opsega, na primer (0,1), (-1,1) ili u vidu nekog binarnog skupa vrednosti. Transfer funkcije mogu biti determinističke ili stohastičke. Na Slici 3.7 su prikazani najčešće korišćeni oblici determinističkih transfer funkcija: linearna, jedinična odskočna funkcija, signum funkcija, rampa funkcija, unipolarna i bipolarna sigmoid funkcija.

Linearna funkcija oblika $y=x$ je u retkoj upotrebi jer nije efikasnosna i degradira funkcionalnost i smisao višeslojnih mreža. Za svrhu modelovanja nelinearnih funkcija, potrebne su procesorske jedinice sa nelinearnim karakteristikama. Najjednostavnija nelinearna funkcija je step funkcija, definisana sledećim izrazom:

$$y = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}. \tag{3.7}$$



(a) $a = purelin(n)$ (b) $a = hardlim(n)$ (c) $a = hardlims(n)$

(d) $a = satlin(n)$ (e) $a = logsig(n)$ (f) $a = tansig(n)$

Slika 3.7 Različiti tipovi aktivacionih funckija i njihove MATLAB oznake: (a) linearna funkcija, (b) step funkcija, (c) hard limiter (signum) funkcija, (d) rampa, (e) unipolarna sigmoid funkcija i (f) bipolarna sigmoid funkcija.

Sa ovom funkcijom višeslojne neuralne mreže imaju dosta veće mogućnosti nego sa linearnom funkcijom i mogu modelovati bilo koju operaciju Bulove algebre. Slična hard limiter funkcija je signum funkcija za koju važi:

$$y = \begin{cases} +1, x \geq 0 \\ -1, x < 0 \end{cases}. \tag{3.8}$$

Ipak, mana ovih funkcija je postojanje diskontinuiteta koji otežava pronalaženje poželjnog skupa težinskih koeficijenata. Koriste se obično u jednostavnijim mrežama sa jednim slojem neurona. Kompromis između hard limiter i linearnih funkcija je takozvana funkcija rampe:

$$y = \begin{cases} 1, x > 1 \\ x, 0 \leq x \leq 1 \\ 0, x < 0 \end{cases}. \tag{3.9}$$

Sa druge strane, postoji veliki broj slučajeva kada se zahtevaju kontinualni izlazi umesto binarnih. U te svrhe se najčešće koriste unipolarne i bipolarne sigmoid funkcije:

$$y = \frac{1}{1+\mathrm{e}^{-\lambda x}}, \tag{3.10}$$

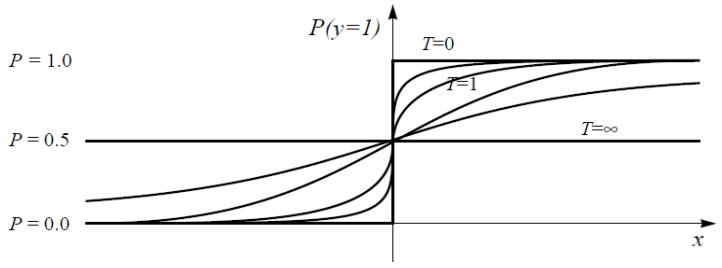
$$y = \frac{2}{1+\mathrm{e}^{-\lambda x}} - 1. \tag{3.11}$$

Ove funkcije imaju odlike nelinearnosti, kontinualnosti i diferencijabilnosti što im omogućava da u višeslojnim neuralnim mrežama aproksimiraju bilo koju složenu funkciju, a pritom nisu zahtevne u procesu obuke. Zbog svojih dobrih karakteristika, one su u najmasovnijoj upotrebi.

Takođe postoje i nedeterminističke transfer funkcije koje su po prirodi probabilističke. Ovakve funkcije obično generišu izlaz u opsegu (0,1) koji opsuje verovatnoću nekog događaja. Primer takve transfer funkcije je prikazan na Slici 3.8 i definisan je na sledeći način:

$$P(y=1) = \frac{1}{1-\mathrm{e}^{-x/T}}, \tag{3.12}$$

gde je T npr. temperatura koja je promenljiva u vremenu.



Slika 3.8 Nedeterministička transfer funkcija. Zavisnost izlzne verovatnoće P od promenljive temperature T.

U slučaju beskonačno visoke temperature, verovatnoća na izlazu je uniformna i iznosi P = 0.5. Sa smanjivanjem temperature, transfer funkcija poprima oblik sigmoid funkcije, dok za T = 0 ima formu signum funkcije. Ovakav postupak postepenog menjanja oblika

transfer funkcije se koristi u nekim tehnikama obuke neuralnih mreža (*simulated annealing*), čime se postiže izbegavanje "upadanja" u lokalni minimum i sigurniji pronalazak globalnog minimuma u *Backpropagation* tehnici o kojoj će kasnije biti više reči.

3.3.4 OBUKA NEURALNIH MREŽA

Još jedna zajednička karakteristika svih tipova neuralnih mreža je njihova mogućnost obuke, odnosno učenja. Obuka ili trening mreže u najopštijem smislu, prestavlja pažljivo podešavanje neuronskih veza tako da mreža bude osposobljena za obavljanje odgovarajućih zadataka. Ovaj proces modifikacije težinskih koeficijenata (pomeranje hiperravni/hipersfera) predstavlja takozvano parametarsko obučavanje mreža. Postoji i strukturno obučavanje mreža koje podrazumeva menjanje trenutne tolologije mreže u vidu dodavanja ili izbacivanje pojedinih veza između neurona (dodavanje i izbacivanje hiperravni/hipersfera). Menjanjem topologije mreže i ograničavanjem skupa funkcija koja ta mreža može da modeluje, moguće je poboljšati sposobnosti generalizacije i brzine učenja mreže. Na isti način se vrši i prevencija pojave *overfitting* problema.

Ipak, uobličavanje težinskih koeficijenata je osnovni način modifikacije mreže, kojim se ujedno može uticati i na topologiju mreže. Na primer, davanje vrednosti nula pojedinim težinskim koeficijenatima ima isti efekat kao i brisanje istih konekcija. Sama problematika pronalaženje skupa težinskih koeficijenata koji treba da omoguće neuralnoj mreži da izračuna neku funkciju nije uopšte laka. Jednostavno i analitičko rešenje ovog problema postoji samo za najjednostavnije *pattern recognition* probleme, npr. kada je mreža linearna a potrebno je da se mapira skup ortogonalnih ulaznih vektora u izlazne vektore. U ovom slučaju, težine veza su određene sa:

$$w_{ij} = \sum_{p} \frac{y_i^p t_j^p}{\left\| y^p \right\|^2}, \tag{3.13}$$

gde je y ulazni vektor, t je izlazni željeni vektor (tzv. *target* vektor) a p je broj obrazaca.

U opštem slučaju, mreže su nelinearne i višeslojne, a njihovi težinski koeficijenti se mogu obučiti jedino kroz iterativnu proceduru, poput procedure spuštanja gradijenta (*gradient descent*) tokom procesa merenja globalnih performansi mreže (procedura minimizacije najmanje greške) [Hinton, 1989]. Ovaj način obuke zahteva višestruko ponavljanje treninga i prolaska kroz bazu podataka na kojoj se mreža obučava što je slično čovekovom postupku učenju i savladavanju novih veština [Tebelskis, 1995]. Svaki prolazak kroz bazu podataka se naziva iteracijom ili epohom. Tokom skladištenja novog znanja u vidu modifikacija težinskih koeficijenata, izmene koeficijenata moraju biti veoma blage kako ne bi došlo do uništavanja prethodno stečenog znanja. U tu svrhu se koristi mala konstanta, poznata kao mera ili brzina učenja (*leanrning rate* (ε)). Ova konstanta kontroliše veličinu promene težinskih koeficijenata. Pronalazak odgovarajuće brzine učenja je veoma važno jer ukoliko je ta vrednost mala, učenje može trajati u nedogled, dok sa druge strane ako je suviše velika vrednost, u procesu učenja tokom iteracija se remeti prethodno stečeno znanje [Tebelskis, 1995]. Nažalost, ne postoji analitički metod za pronalazak optimalne brzine učenja. Brzina učenja se obično empirijski određuje i optimizuje, isprobavanjem različitih vrednosti.

Postoji više različitih procedura obuke, pri čemu većina njih uključujući i onu opisanom jednačinom (3.13), predstavlja varijaciju Hebovog principa učenja [Hebb, 1949]. Ovaj princip je zasnovan na pojačavanju veze između dve jedinice ukoliko su izlazne aktivacione vrednosti međusobno korelisane:

$$\Delta w_{ji} = \varepsilon y_i y_j , \tag{3.14}$$

Posle ovakvog postupka obuke, mreža može aktivirati drugu procesorsku jedinicu na osnovu poznatog stanja prve jedinice.

Jedna bitna varijacija Hebovog pravila je takozvano Delta pravilo ili *Widrow-Hoff* pravilo, koje se primenjuje kada postoji poznata izlazna vrednost jedne od dve procesorske jedinice. Ovo pravilo pojačava vezu između dve jedinice ukoliko postoji korelacija između aktivacione vrednosti prve jedinice, y_i, i greške druge procesorske jedinice koja predstavlja razliku između aktivacione vrednosti y_j i željene izlazne vrednosti t_j.

$$\Delta w_{ji} = \varepsilon y_i (t_j - y_j) . \quad (3.15)$$

Cilj ovog pravila je minimizacija te greške sa izračunavanjem izlaza y_j koji treba da bude sličniji t_j. U slučaju jednoslojnih mreža sa procesorskim jedinicama koje daju binarne izlaze, Delta pravilo je poznato pod nazivom *Perceptron Learning Rule* i garantuje pronalazak najboljeg skupa težinskih koeficijenata koji postoji [Rosenblatt, 1962]. U kontekstu višeslojnih mreža, Delta pravilo je osnova *Backpropagation* postupka obuke, koji će detaljno biti opisan u sekciji 3.5.

Postoji još jedna varijacija Hebovog principa učenja, u slučaju sfernih funkcija kod LVQ (*Learned Vector Quantization*) i RBF (*Radial Bias Function*) tipova mreža:

$$\Delta w_{ji} = \varepsilon (y_i - w_{ji}) y_j . \quad (3.16)$$

Ovo pravilo se svodi na pomeranje centra sfere (w_{ji}) bliže ulaznom obrascu (y_i) ukoliko je izlazna klasa y_j aktivna [Tebelskis, 1995].

3.4 TIPOVI NEURALNIH MREŽA

Postoje različiti tipovi neuralnih mreža koji se mogu klasifikovati u nekoliko grupa u zavisnosti od načina njihove obuke. Razlikuju se tri osnovne metode obuke neuralnih mreža:

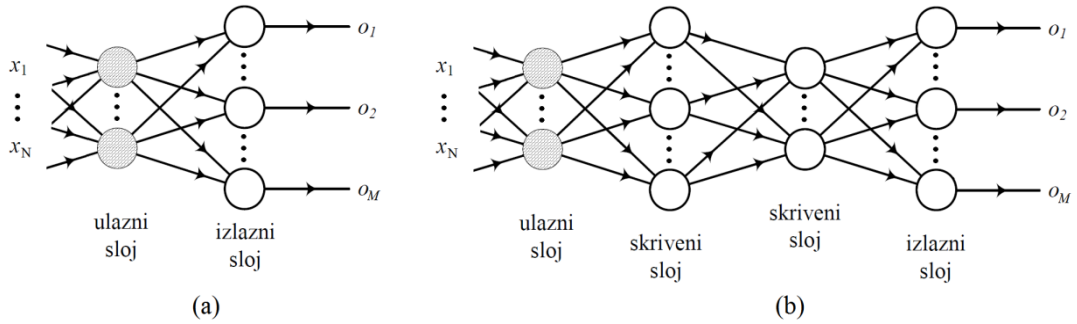
1) Supervizirano obučavanje (*Supervised learning*) ili obučavanje sa nadgledanjem je metoda obuke u kojoj mreža raspolaže trening bazom koja se sastoji iz uređenih ulaz-izlaz parova podataka. U ovakvoj obuci "učitelj" prati proces obuke neuralne mreže i eksplicitno ispravlja greške koje mreža čini tokom treninga. Tipovi mreža koje koriste ovakvu vrstu obuke su pre svega *feedforward* mreže, od kojih su najpoznatije: jednoslojni perceptroni (*Single-layer Perceptrons - SLP*), višeslojni perceptroni (*Multi-layer Perceptrons - MLP*), mreže sa vremenskim kašnjenjem (*Time Delay Neural Network - TDNN*) i mreže koje koriste učenje kvantizacije vektora (*Learned Vector Quantization - LVQ*). Pored *feedforward* mreža supervizirano obučavanje se primenjuje i u rekurentnim mrežama (*Hopfield network*, *Boltzman Machine* i *Elman network*).

2) Polu-supervizirano obučavanje (*Semi-supervised learning*) je metoda učenja u kojoj "učitelj" tokom treninga vrši samo evaluaciju ponašanja neuralne mreže kao "dobro" ili "loše". Prema tome, analogno superviziranom učenju, ova metoda obučavanja se može tretirati na isti način sa tim što umesto učitelja, koji egzaktno ukazuje kakav odziv neuralne mreže treba da bude, u ovom slučaju imamo "kritičara" koji daje grublju ocenu odziva neuralne mreže. Iz tog razloga se ovakvo učenje naziva i obučavanje sa kritikom ili podsticanjem (*reinforcement learning*).
3) Nesupervizirano obučavanje (*Unsupervised learning*), ili obučavanje bez nadgledanja, je metoda učenja neuralne mreže u kojoj nema "učitelja" i mreža mora sama da pronađe pravila među obrascima iz trening baze podataka. Ovakve mreže se mogu koristiti u kompresiji, klasterovanju, kvantizaciji, klasifikaciji i mapiranju ulaznih podataka. Primer ovakve mreže je Kohenenova mreža (*Kohenen network*), autoenkoderi itd.

Većina mreža se može svrstati u neku od ove tri grupe. Ipak, postoje i hibridne mreže koje kombinuju nesupervizirano i supervizirano obučavanje. Poseban slučaj su dinamičke neuralne mreže (*Dynamic Neural Networks*) koje imaju sposobnost menjanja svoje arhitekture tokom vremena. U nastavku teksta biće opisane samo najpopularniji tipovi neuralnih mreža.

3.4.1 FEEDFORWARD NEURALNE MREŽE

Jedna od najpopularnijih arhitektura neuralnih mreža je *feedforward* topologija, u kojoj su neuroni uređeni po slojevima a veze između njih ne formiraju zatvorene putanje. U ovakvim mrežama svaki neuron iz jednog sloja je povezan sa ulazom neurona iz sledećeg sloja i tako redom ukoliko postoji više slojeva. Informacije teku u jednom smeru, unapred od ulaznog sloja ka izlaznom sloju, bez pravljenja povratnih sprega. Najjednostavniji tip *feedforward* mreža su Perceptroni [Rosenblatt, 1962] koji koriste supervizirano učenje. Perceptroni se sastoje iz binarnih aktivacionih funkcija i organizovani su u slojevima, kao na primerima na slici 3.9.



Slika 3.9 *Feedforward* neuralne mreže: (a) jednoslojni perceptroni, (b) višeslojni perceptroni.

Obučavaju se pomoću Delta pravila, definisanog jednačinom (3.15), ili nekim od njenih varijacija. U slučaju jednoslojnih perceptrona, prikazanih na slici 3.9 (a), Delta pravilo se može direktno primeniti. Pošto su aktivacione vrednosti perceptrona binarne vrednosti, radi se o uopšćenom Delta pravilu poznatom kao *Perceptrone Learning rule*. Integraciona funkcija perceptrona (3.1) je linearna i u primeru sa Slike 3.9 (a) se može zapisati u sledećem obliku:

$$u(x_1, x_2, ..., x_N) = \sum_{k=1}^{N} w_k x_k \,, \tag{3.17}$$

gde je pretpostavljeno da neuron ima N ulaza. Težinski koeficijenti, w_k, predstavljaju znanje koje mreža treba da stekne tokom procesa obuke i kasnije primeni u realnim uslovima. Ulazni podaci i težinski koeficijenti se mogu vektorski zapisati:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}_{N\times 1} , \tag{3.18}$$

$$W = \begin{bmatrix} w_1 & w_2 & \cdots & w_N \end{bmatrix}_{1\times N} . \tag{3.19}$$

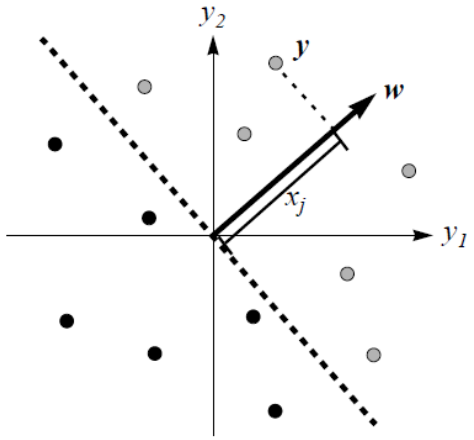
Tada se integraciona funkcija (3.17) može skraćeno napisati u sledećoj formi:

$$u = W \cdot X \,, \tag{3.20}$$

a izlaz neurona:

$$o = f(u) = (\mathrm{W} \times \mathrm{X}), \tag{3.21}$$

gde je *f(u)* transfer funkcija. Kada se u postupku obuke mreže za zadati ulaz X_k dobije izlaz mreže o_k, potrebno je da se on uporedi sa željenim izlazom t_k. U slučaju da je $o_k = t_k$ ne treba modifikovati težinske koeficijente. U obratnom, kada je $o_k \neq t_k$, koeficijente treba povećati ili smanjiti za neku vrednost Δw definisanu izrazom (3.15). Ovakva procedura obuke garantuje pronalazak skupa težinskih koeficijenata koji će tačno klasifikovati obrasce u bilo kom skupu trening podataka, pod uslovom da su obrasci linearno separabilni, tj. ukoliko se mogu podeliti u dve klase povlačenjem prave kao što je ilustrovano na Slici 3.10. U većini slučajeva baza podataka na kojoj se vrši obuka ipak nije linearno sparabilna i tada jednoslojni perceptroni ne mogu modelovati određene funkcije (na primer XOR funkciju iz Bulove algebre). Za rešavanje ovakvih zadataka neophodno je više slojeva, odnosno težinskih koeficijenata. *Feedforward* neuralne mreže sa više skrivenih slojeva, ili MLP, se u literaturi često spominju kao dubinske neuralne mreže (*Deep Neural Networks – DNN*) [Bengio, 2009; Deng et al., 2014].



Slika 3.10 Linearna separabilnost. [Tebelskis, 1995]

Višeslojni perceptroni (MLP) pored ulaznog i izlaznog sloja poseduju dodatne slojeve poznate kao skrivene slojeve, Slika 3.9 (b). Težinski koeficijenti pojedinih neurona, W_l, l=1,2,...,M, se mogu zapisati u vidu matrice:

$$\mathrm{W}=\begin{bmatrix}\mathrm{W}_1\\ \mathrm{W}_2\\ \vdots\\ \mathrm{W}_M\end{bmatrix}=\begin{bmatrix}w_{11} & w_{12} & \vdots & w_{1N}\\ w_{21} & w_{22} & \vdots & w_{2N}\\ \vdots & \vdots & \ddots & \vdots\\ w_{M1} & w_{M2} & \vdots & w_{MN}\end{bmatrix}_{M\times N}, \tag{3.22}$$

gde je N broj slojeva mreža, a M broj neurona u sloju. Teorijski, ovakve mreže sa takozvanom dubinskom arhitekturom mogu da nauče bilo koju funkciju, ali su zato nešto komplikovanije za obuku. Naime, Delta pravilo se ne može direktno primeniti na MLP jer ne postoje definisane ciljane izlazne vrednosti (*targets*) za skrivene slojeve. Zahvaljujući korišćenju kontinualnih umesto diskretnih aktivacionih funkcija, MLP mrežama je omogućeno korišćenje parcijalnih izvoda i izračunavanje uticaja svakog težinskog koeficijenta na bilo koju izlaznu aktivacionu vrednost. Ovi parametri pomažu određivanju načina i mere modifikacije težinskih koeficijenta u cilju smanjenja greške mreže. Ova generalizacija Delta pravila je poznata kao *Backpropagation* algoritam i biće razmatran u poglavlju 3.5.

Iako višeslojni perceptroni mogu imati bilo koji broj skrivenih neurona, jedan skriveni sloj je dovoljan za većinu potreba. Dodatno skriveni slojevi usporavaju obuku mreže. MLP se takođe u pogledu arhitekture mogu ograničiti na različite načine. Na primer u pogledu njihove geometrijske povezanosti, ograničavanjem i grupisanjem težinskih koeficijenata.

Poseban vid ograničavanja MLP mreža je dodavanje vremenskog kašnjenja težinskim koeficijentima. Ovaj princip se koristi u mrežama sa vremenskim kašnjenjem (*Time Delay Neural Networks - TDNN*). TDNN su prvobitno razvijene za potrebe prepoznavanja fonema [Lang, 1989; Waibel et al., 1989], ali su takođe korišćene i u prepoznavanju rukopisa [Idan et al., 1992; Bodenhausen et al., 1993], čitanja govora sa usana [Bregler et al, 1993] i u drugim zadacima. TDNN su bazirane na dvodimenzionalnim ulazima, gde je jedan od njih vreme. Konekcije su zakašnjene u vremenu, kako bi se istaklo da njihove povezane jedinice nisu vremenski usaglašene [Tebelskis, 1995]. Za obuku TDNN mreža se koristi standardni *Backpropagation* algoritam, pri čemu je jedina razlika to što se vezani težinski koeficijenti modifikuju prema njihovoj srednjoj vrednosti greške umesto prema pojedinačnoj greški.

Još jedan tip neuralnih mreža koristi superviziranu obuku a to su mreže koje koriste učenje kvantizacije vektora (*Learned Vector Quantization - LVQ*), [Kohonon, 1989]. LVQ mreža je jednoslojna mreža koja na izlazu daje klase, a težinski koeficijenti sa ulaza predstavljaju centre hipersfera, Slika 3.6. Trening ovih mreža se zasniva na pomeranju hipersfera radi tačnijeg modelovanja klasa.

3.4.2 REKURENTNE NEURALNE MREŽE

Posebnu vrstu mreža sa povratnim spregama predstavljaju rekurentne neuralne mreže. Njihov najpoznatiji tip su Hopfildove mreže koje imaju nestrukturiranu arhitekturu, procesorske jedinice sa binarnim izlazima i simetrične neuronske veze $w_{ji}=w_{ij}$ čiji se težinski koeficijenti asinhrono ažuriraju [Hopfield, 1982]. Hopfild je u svojim istraživanjima pokazao da ovakve mreže trenirane Hebovim pravilom, imaju sposobnost generalizovanja nepotpunih podataka (obrazaca) kada se pojave na ulazu mreže. Naime, ukoliko bi se mreži dovela oštećena ili nepotpuna verzija nekog od obrazaca, aktivacione vrednosti mreže bi se ažurirale na slučajni asinhroni način (koristeći prethodno stečene težinske koeficijente u procesu obuke) i tada bi mreža postepeno rekonstruisala ceo obrazac aktivacionih vrednosti koji je najsličniji stanju prostora i stabilizovao bi se na tom obrascu. Hopfildov osnovni koncept se zasniva na analizi dinamike mreže u pogledu funkcije energije,

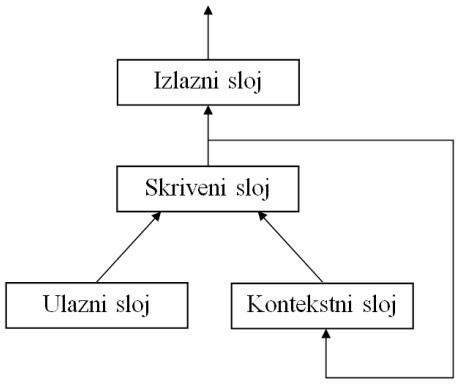
$$E = -\frac{1}{2}\sum_{i}\sum_{j\neq i} w_{ji} y_i y_j \,, \tag{3.23}$$

koja može samo da opada ili ostaje ista. Kada aktivaciona vrednost neurona na izlazu promeni znak, energija opada dok u obratnom ostaje nepromenjena. Za obrasce koji odgovaraju već postojećem znanju mreže, postiže se minimum. Ovo implicira da mreža uvek dolazi u stabilno stanje (osim ako dospe u lokalni minimum (lažni globalni minimum) kao posledica interferencije između skladištene memorije).

Drugi tip rekurentnih neuralnih mreža je *Boltzman Machine* mreža. To je Hopfildova mreža sa skrivenim procesorskim jedinicama, stohastičkim aktivacionim funkcijama i procedurom obuke zasnovanoj na simuliranom hlađenju (*simulated annealing*), objašnjenoj na Slici 3.8. Skriveni slojevi doprinose Bolcmanovoj mašini da

bolje koreliše podatke u poređenju sa Hopfildovom mrežom, pa zato može da nauči složenije obrasce. Stohastičke aktivacione funkcije, poput 3.12, omogućavaju pouzdaniji pronalazak globalnog maksimuma i izbegavanje lokalnih minimuma. Simuliranje hlađenja, odnosno postepenog opadanja temperature u procesu obuke pomaže mreži da mnogo efektnije uči.

Poznat je još jedan tip rekurentnih mreža koji ima slojevitu strukturu i povratne sprege sa drugim slojevima. U pitanju je Elmanova mreža [Elman, 1990]. Ova mreža je jednostavna rekurentna neurlana mreža (*Simple Recurent Neural Network - SRN*) koja se sastoji iz jednog ulaznog, jednog skrivenog i jednoz izlaznog sloja. U ovom obliku podseća na troslojnu *feedforward* mrežu. Ipak, ona poseduje dodatni sloj neurona poznat kao kontekstni sloj koji se preko povratne sprege snabdeva podacima sa izlaza skrivenog sloja bez njihovog modifikovanja u vidu otežavanja težinskih koeficijenata, kao što je prikazano na Slici 3.11. Elmanova mreža pamti ove vrednosti i šalje ih na izlaz pri sledećoj aktivaciji mreže. Ove vrednosti se potom šalju nazad u skriveni sloj uz modifikovanje težinskih koeficijenata. Elmanove mreže su korisne za predikciju sekvenci, pošto imaju ograničenu *short-term* memoriju.

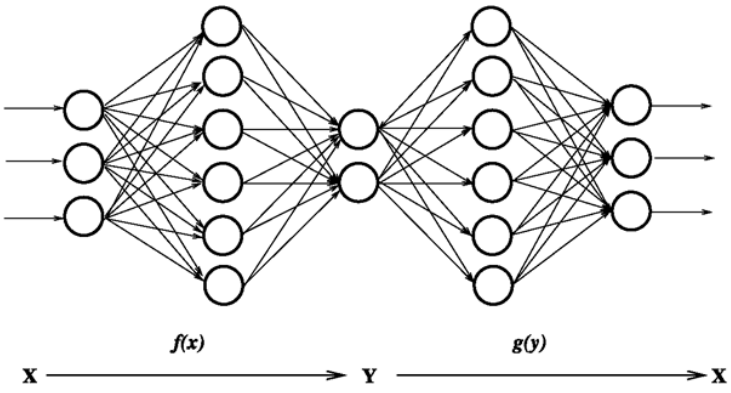


Slika 3.11 Princip rada Elmanove rekurentne mreže.

3.4.3 DUBINSKI AUTOENKODERI

Dubinski autoenkoder (*Deep autoencoder*) predstavlja poseban tip dubinskih neuralnih mreža (*Deep Neural Network - DNN*) koji na prvi pogled ima veoma jednostavan zatatak – da na svom izlazu reprodukuje, odnosno rekonstruiše, podatke sa svog ulaza. Međutim, svrha autoenkodera leži u njegovim skrivenim slojevima u kojima se formira interna reprezentacija ulaznog podatka, na osnovu koje se tokom dalje

propagacije kroz ostale slojeve mreže na izlazu dobija što je moguće istovetnija estimacija ulaznog podatka. Ova odlika autoenkodera je našla primenu u različitim zadacima kompresije i ekstrakcije obeležja (*feature extraction*) [Bengio, 2009] o čemu će biti više reči u nastavku ove teze. Naime, autoenkoder ima jedan ulazni sloj koji prima originalne ulazne vektore obeležja, jedan ili više skrivenih slojeva koji predstavljaju transformisana obeležja i jedan izlazni sloj koji treba da rekonstruiše i oponaša vektore podataka sa ulaza. Kada je broj skrivenih slojeva autoenkodera veći od jednog, autoenkoder se smatra dubinskim autoenkoderom [Deng et al., 2014]. Ulazni i izlazni slojevi autoenkodera imaju jednaku dimenzionalnost. Veličina skrivenih slojeva može biti manja od ulaznog sloja, kada je cilj autoenkodera formiranje uskog grla i kompresija ulaznih obeležja u takozvana *bottleneck* obeležja (*bottleneck features*), ili veća kada autoenkoder ima zadatak da mapira ulazna obeležja u neki veći dimenzionalni prostor.



Slika 3.12 Primer arhitekture autoenkodera sa skrivenim slojem koji formira usko grlo (*bottleneck*).

Postupak mapiranja se može analizirati u dva koraka i predstavljen je na Slici 3.12. U prvom koraku autoenkoder vrši preslikavanje ulaznog vektora obeležja $\boldsymbol{x}$ u svoje skrivene slojeve kroz nelinearnu funkciju $f_\theta(\mathbf{x})$:

$$\boldsymbol{y} = f_\theta(\boldsymbol{x}) = f_1(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \tag{3.24}$$

gde su: **W** matrica težinskih koeficijenata (dimenzije $d \times d'$), **b** *bias* vektor, $f_1()$ nelinearna funkcija poput *sigmoid* ili *tanh*. Ovaj korak preslikavanja se naziva enkodovanje i obavlja se u ulaznom delu autoenkodera, poznatog kao enkoder. Tako dobijena skrivena reporezentacija ulaznog vektora se u vidu vektora **y** potom mapira

“nazad” u cilju rekostruisanja ulaznog vektora obeležja:

$$\boldsymbol{z} = f_{\theta'}(\boldsymbol{y}) = f_2(\boldsymbol{W'y} + \boldsymbol{b'}), \tag{3.25}$$

gde su: **W'** matrica težinskih koeficijenata, dimenzija $d' \times d$, pri čemu važi $\mathbf{W'} = \mathbf{W^T}$, **b'** je *bias* vektor, $f_2()$ je ili nelinearna funkcija poput *sigmoid* ili *tanh* funkcija, ili je linearna funkcija. Ovaj korak u preslikavanju se naziva dekodovanje i izvršava se u delu autoenkodera poznatog kao dekoder. Kao i kod ostalih tipova neuralnih mreža, cilj procesa obuke je minimizovanje srednje kvadratne greške:

$$L(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|^2. \tag{3.26}$$

U cilju sprečavanja situacije u kojoj autoenkoder uči trivijalnu funkciju mapiranja, obično se tokom treninga uvode neke vrste ograničenja. Na primer, dodaje se Gausov šum ulaznim signalima odnosno vektorima (što je primenjeno i u eksperimentima ovog dokotrata) ili se koristi takozvani "*dropout*" trik kojim se nasumice pojedninim vrednostima vektora dodeljuje vrednost nula. Ovakvi autoenkoderi su poznati kao *denoising* autoenkoderi (DAE) [Vincent et al., 2008; Mimura et al., 2015]. Denoising autoenkoderi imaju istu strukturu kao i obični autoenkoderi, a jedina razlika je ta što *denoising* autoenkoderi na svom ulaznom sloju primaju iskvarenu verziju vektora sa izlaza (signal sa šumom ili signal degradiran nekim drugim tipom interfernecije). Drugim rečima, DAE koristi mapiranje kroz svoje skrivene slojeve za konvertovanje iskvarene verzije vektora ($\hat{\mathbf{x}}$ – ulazni signal) u njegovu čistu verziju na izlazu (**x** - *teacher* signal).

Obuka autoenkodera se može podeliti na dva dela: na proces pred–obuke i završne obuke. Pred−obuka ili *pre-training* podrazumeva obuku skupa *restricted Boltzmann Machines* (RBMs), posebnog tipa ANN odnosno Bolcmanovih mašina, pri čemu svaki RBM predstavlja jedan sloj autoenkodera. Nakon obuke jednog RBM, njegov izlaz se koristi kao ulaz sledećeg RBM u procesu njegove obuke. Na ovakav način, poznat kao *contrastive divergence* (CD) [Hinton, 2002], vrši se obuka drugog RBM i tako redom (izlaz skrivenog sloja drugog RBM se koristi kao ulazni vektor za trići RBM, itd.). Ovaj postupak obuke po principu sloj-po-sloj se može ponoviti više

puta. Po završetku ove faze pred–obuke, skup obučenih RBM formira dubinski autoenkoder koji se potom obučava u završnoj fazi poznatoj kao *fine-tuning*. U ovoj završnoj fazi, kao algoritam obuke se koristi *Backpropagation,* koji ima za cilj da fino podesi težinske koeficijente neuralne mreže kako bi smanjio grešku, odnosno razliku između izlaznog signala i *teacher* signala (čist signal).

3.5 BACKPROPAGATION ALGORITAM

Algoritam propagacije greške unazad ili *Backpropagation* algoritam je najpopularnija metoda supervizirane obuke neuralnih mreža. U literaturi se takođe javlja pod nazivom Propagacija greške (*Error propagation*) ili Generalizovano delta pravilo (*Generalized Delta Rule - GDR*). *Backpropagation* algoritam se primenjuje uglavnom u obuci *feedforward* neuralnih mreža, pa će iz tog razloga biti objašnjen na primeru MLP mreže sa nelinearnim transfer funkcijama (npr. *sigmoid*). Neka su integralne i transfer funkcije neurona ove mreže definisane sa:

$$x_j^p = \sum_i w_{ji} y_i^p \,, \tag{3.24}$$

$$y_j^p = f(x_j^p) = \frac{1}{1+e^{-x_j^p}} \,, \tag{3.25}$$

gde su *i* i *j* oznake neurona (procesorskih jedinica) a *p* broj trening obrazaca. Prema tome x_j^p predstavlja ulaz u procesorsku jedinicu *j* koji odgovara nekom obrascu *p,* dok je y_j^p generisani izlaz te jedinice na pobudu istim obrascem. w_{ji} je težinski koeficijent koji odgovara sinapsi koja povezuje *i* i *j* jedinicu, dok je t_j^p željena izlazna vrednost jedinice *j* na neku pobudu *p*.

Kao što je u prethodnim poglavljima objašenjno, cilj svakog treninga neuralnih mreža je pronalazak odgovarajućeg skupa težinskih koeficijenata koji će na najbolji mogući način omogućiti modelovanje odgovarajuće željene funkcije. Prednost *Backpropagation* algoritma je u tome što pored ostalih funkcija omogućava izračunavanje i nelinearnih funkcija, koje se analitičkim putem ne mogu modelovati.

Backpropagation primenjuje proceduru spuštanja (opadanja) gradijenta (*gradient descent*) na funkciji greške *E* koja je definisana sledećim formulama:

$$E^p = \frac{1}{2}\sum_j (y_j^p - t_j^p)^2 , \tag{3.26}$$

$$E = \sum_p E^p , \tag{3.27}$$

gde je sa E^p označena greška prilikom prepoznavanja nekog obrasca *p,* a sa *E* globalna greška za čitav skup obrazaca, dok *j* pripada nekom konačnom skupu izlaznih procesorskih jedinica *O*. Težinski koeficijenti se modifikuju (umanjuju ili povećvaju) proprcionalno funkciji greške E^p sa krajnjim ciljem smanjenja globalne greške *E*. Veličina promena pojedinačnih težinskih koeficijenta se izračunava diferenciranjem funkcije E^p:

$$\Delta^p w_{ji} = -\varepsilon \frac{\partial E^p}{\partial w_{ji}} = -\varepsilon \frac{\partial E^p}{\partial y_j^p}\frac{\partial y_j^p}{\partial x_j^p}\frac{\partial x_j^p}{\partial w_{ji}} = -\varepsilon \gamma_j^p f'(x_j^p), \tag{3.28}$$

gde je ε brzina obuke mreže. U zavisnosti od toga da li *j* pripada nekoj od procesorskih jedinica iz izlaznog sloja neuralne mreže ($j \in O$) ili ne ($j \notin O$), γ_j^p se izračunava:

$$\gamma_j^p = \frac{\partial E^p}{\partial y_j^p} = \begin{cases} (y_j^p - t_j^p) & \text{ако } j \in O \\ \sum_k \gamma_k^p \cdot f'(x_k^p) \cdot w_{kj} & \text{ако } j \notin O \end{cases}, \tag{3.29}$$

Zahvaljujući rekurziji, u višeslojnoj mreži moguće je direktno izračunati sve nepoznate γ_j^p (a time i $\Delta^p w_{ji}$) na osnovu vrednosti iz sledećih slojeva. Prema tome vrednosti γ se mogu izračunati počevši od izlaznog sloja mreže (primenom formule $(y_j^p - t_j^p)$) krećući se unazad, sloj po sloj, sve do ulaznog sloja (primenom formule $\sum_k \gamma_k^p \cdot f'(x_k^p) \cdot w_{kj}$).

Zbog toga se ova metoda obuke naziva "*Backpropagation*" odnosno propagacija greške unazad.

Obuka mreža pomoću *Backpropagation* algoritma je znatno brža nego trening sa Bolcmanovom mašinom, ali i pored toga zahteva određeno vreme da mreža konvergira ka optimalnim vrednostima težinskih koeficijenata. Brzina obuke se može ubrzati povećanjem vrednosti ε, odnosno brzine učenja, ali sve do neke granice kada je dalje povećanje kontraproduktivno. Iz tog razloga su istraživane druge mogućnosti ubrzavanja procesa treninga mreže. Ove tehnike su uglavnom inspirisane procedurom *gradient descent* u kojoj se funkcija greške *E* posmatra kao reljefna površ u kojoj je cilj naći globalni minimum spuštajući se inkrementalnim koracima Δw_{ji}. Posebna tehnika obuke poznata kao *momentum* [Rumelhart, 1986] obezbeđuje da ovi koraci budu mali i da se polako (ne naglo) povećavaju. Još moćnija tehnika, zasnovana na drugim izvodima, je *conjugent gradient* [Barnard, 1992].

Procedura *gradient descent*, a time i *Backpropagation* algoritam je podložan konvergenciji ka suboptimalnom rešenju, odnosno ka lokalnim minimumima funkcije greške. Ovaj problem se rešava ponavljanjem obuke neuralne mreže ili dodavanjem šuma tokom modifikacije težinskih koeficijenata.

3.6 VEZA SA STATISTIKOM

Automatsko prepoznavanje govora je jedan vid problema statističke klasifikacije podataka. Pored HMM sistema, koji su predstavnici statističkih modela u prepoznavanju govora, veštačke neuralne mreže takođe imaju tesnu vezu sa mnogim standardnim statističkim metodama što će biti pokazano kroz naredne primere.

Osnovna statistička formulacija problema prepoznavanja govora, odnosno formulacija donošenja odluke o tačno prepoznatoj reči (slogu, fonemu...), se svodi na Bajesovo pravilo odlučivanja. Ovo pravilo se bazira na određivanju statističkih raspodela *a posteriori* i *a priori* verovatnoća (2.2) koje su u praksi nepoznate. U tom slučaju najčešće je na raspolaganju samo skup trening podataka koje je potrebno analizirati i modelovati u vidu raspodele verovatnoća. Ovaj zadatak se može rešiti odgovarajućim statističkim procedurama ali i pomoću neuralnih mrežama. Za potrebe Bajesove klasifikacije na raspolaganju stoji veliki broj statističkih tehnika koje se dele

na parametarske ili neparametarske. U slučaju parametarskog pristupa, pretpostavlja se da su raspodele verovatnoća (P(R|X) ili P(X|R) i P(R)) date u parametarskom obliku (npr. Gausova raspodela), a zadatak je odrediti njihove parametre. To se najčešće postiže upotrebom statističkog kriterijuma maksimalne verodostojnosti (*Maximum Likelihood Estimation - MLE*), koji pronalazi parametre koji najbolje odgovaraju raspolažućim trening podacima. U neparametarskom pristupu se koriste statističke tehnike poput *Parzen windows* i *k-nearest neighbor rule*. Neuralne mreže, sa druge strane takođe omogućavaju Bajesovu klasifikaciju. Višeslojni perceptroni se asimptotski treniraju koristeći srednju kvadratnu grešku (*Mean Squared Error - MSE*), ili neku drugu vrstu funkcije greške (*Sum Squared Error, McClelland Error, Cross Entropy Error, Mean Absolute Error*, itd...), pri čemu izlazne aktivacione vrednosti uče da aproksimiraju *a posteriori* raspodele verovatnoća, a preciznost modelovanja raste sa veličinom trening baze podataka.

Drugi način klasifikacije podataka je pronalazak granica koje razdvajaju klase. U statistici se ovo postiže nekom od tehnika poznatih kao Diskriminantna analiza (*discriminant analysis*), poput Fišerove linearne diskiminantne analize (*Fisher's Linear Discriminant Analysis - FLDA*) koja pronalazi granicu koja najbolje razdvaja podatke na dve klase. Slično se postiže sa jednoslojnim perceptronima koji imaju jedan izlaz i koji su obučeni Delta pravilom da prave razliku između dve klase i formiraju hiperravan kao na Slici 3.10 [Tebelskis, 1995].

Ukoliko podaci u bazi za obuku nisu obleženi, tj. ukoliko ne postoje uređeni ulaz-izlaz parovi podataka, klasterizacija podataka se vrši nekom od statističkih tehnika poput: *nearest neigbor clustering*, *minimum squared error clustering* ili *k-means clustering*. Isto se postiže npr. sa Kohenenovom neuralnom mrežom sa kompetetivnim procesom obuke.

Neuralne mreže takođe mogu poslužiti za smanjenje dimenzionalnosti podataka poput statističke analize glavnih komponenata (*Principal Component Analysis - PCA*). Na primer jednoslojni perceptroni sa posebnim tipom obuke, poznatim kao *Sanger's Rule*, daju težinske koeficijente koji su jednaki glavnim komponentama podataka, tako da se na izlazu mreže dobijaja komprimovana predstava ulaznih vektora. Na sličan način višeslojni perceptroni koji su obučeni standardnim *Backpropagation* algoritmom

tako da na izlazu daju vektore sa ulaza, u skrivenom slojevima kreiraju komprimovanu predstavu ulaznih podataka.

Prema tome, neuralne mreže predstavljaju novu generaciju sistema za procesiranje koji između ostalog imaju sposobnost i da jednostavno i na potpuno nov način formulišu stare statističke metode [Tebelskis, 1995].

3.7 REZIME

Veštačke neuralne mreže iako inspirisane biološkim neuronima predstavljaju pre svega matematički model i imaju jedino par dodirnih tačaka sa čovekovim nervnim sistemom. Ipak, zajedničke karakteristike poput paralelnog procesiranja podataka i masovih sprega između neurona, omogućavaju veštačkim neuralnim mrežama impresivne mogućnosti u sprovođenju različitih statističkih metoda i rešavanju zahtevnih *pattern recognition* problema. Ovi zadaci se sa druge strane veoma teško i sporo rešavaju pomoću konvencionalnih računara, zbog čega su neuralne mreže veoma brzo stekle popularnost i privukle pažnju širokih naučnih krugova. Do sada je razvijeno i u praksi primenjivo više različitih arhitektura neuralnih mreža i metoda njihove obuke, od kojih su trenutno najzastupljenije *feedforward* neuralne mreže. Posebno su se istakli Višeslojni perceptroni, koji sa svojim nelinearnim aktivacionim funkcijama, najčešće sigmoid tipa, i sa *Backpropagation* algoritmom u procesu obuke, mogu aproksimirati bilo koju kompleksnu funkciju. Takođe, sve popularnije dubinske neuralne mreže zahvaljujući svojoj dubinskoj arhitekturi i impresivnim mogućnostima dubinskog učenja omogućavaju rešavanje najsloženijih *pattern recognition* problema današnjice. Zbog svojih sposobnosti učenja, generalizovanja, otpornosti na šum i paralelnog procesiranja podataka, neuralne mreže su našle primenu i u govornim tehnologijama gde predstavljaju alternativu HMM sistemima. Za razliku od HMM sistema koji imaju svoje nedostatke i probleme u vidu greški kvantizacije i parametarskog modelovanja funkcija raspodela verovatnoća, neuralne mreže nemaju takvih problema i dosta su uspešnije u akustičkom modelovanju. ANN su pokazale visoke performanse u statičkim, odnosno vremenski nezavisnim *pattern recognition* zadacima, poput automatskog prepoznavanja izolovanih slogova, reči itd. Međutim još uvek nije rešen problem kako se ANN mogu efikasno primeniti u dinamičkim uslovima. Zbog toga je aktuelni trend da se ANN koriste u akustičkom modelovanju, a HMM sistemi u

modelovanju vremena. Kombinacijom prednosti ANN i HMM sistema, kreirani su različiti hibridni i tandem sistemi. Eksperimentalno je dokazano da se ANN mogu primeniti na velikim korpusima reči, u prepoznavanju govora nezavisnog od govornika, u zadacima prepoznavanju kontinualnog govora itd. Ipak, zbog kompleksnosti ovakvih sistema i složenosti procedure treninga, u praksi su još uvek dominantni jednostavniji HMM sistemi.

Iako su u upotrebi već duži niz godina, smatra se da krajnji dometi veštačkih neuralnih mreža još uvek nisu u potpunosti dostignuti. Zbog svojih velikih i relativno neistraženih mogućnosti, neuralne mreže su i dalje u povoju i predstavljaju aktuelnu istraživačku temu.

4 ŠAPAT

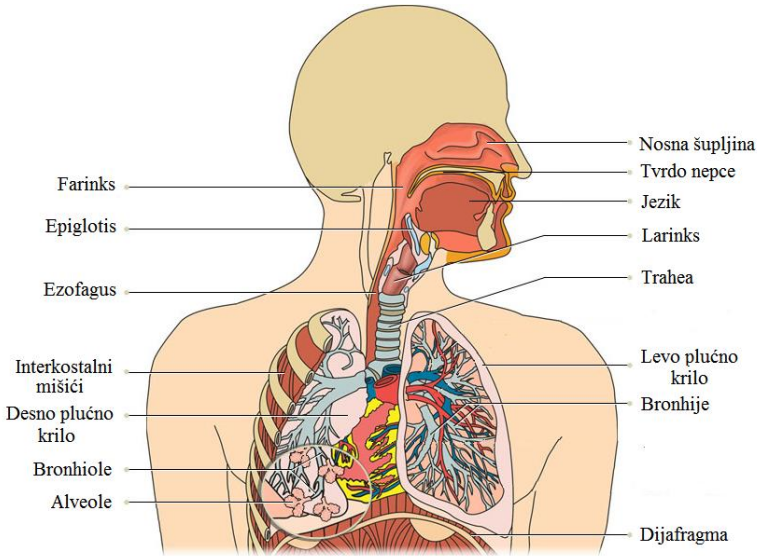
U cilju poboljšanja ASR sistema u prepoznavanju šapata, neophodno je prvo shvatiti njegove prirodne karakteristike i razlike u odnosu na normalan govor. U tu svrhu su u ovom poglavlju opisane razlike između govora i šapata, objašnjeno je kako te razlike utiču na prepoznavanje šapata, koje sve informacije šapat nosi sa sobom, itd. Prvo je obrađena tema načina generisanja šapata i specifičnog rada artikulacionih organa, zatim su nabrojane akutičke karakteristike šapata i ograničenja percepcije šapata. Na samom kraju je dat pregled do sada testiranih ASR sistema kao i njihove mogućnosti u prepoznavanju šapata.

4.1 FIZIOLOGIJA GOVORNOG MEHANIZMA U ŠAPATU

Šapat je specifičan oblik verbalne komunikacije koji se tokom čovekove evolucije razvio u paraleli sa govorom. Slično govoru, u osnovi generisanja šapata je govorni mehanizam koji je ilustrovan na Slici 4.1 i koji se može uprostiti na tri podsistema: na sistem respiratornih organa, fonatorni sistem i vokalni trakt [Jovičić, 1999].

Respiratorni sistem čine pluća, bronhije i trahee i predstavlja izvor energije kojim se generiše vazdušna struja kao osnova pobude govornog sistema. Fonatorni deo govornog mehanizma predstavlja larinks sa složenom strukturom hrskavica međusobno povezanih mišićnim i vezivinim tkivom koja omogućava njihovu veliku pokretljivost [Jovičić, 1999]. Osnovnu funkciju larinksa obavljaju glasnice, odnosno glasne žice, koje

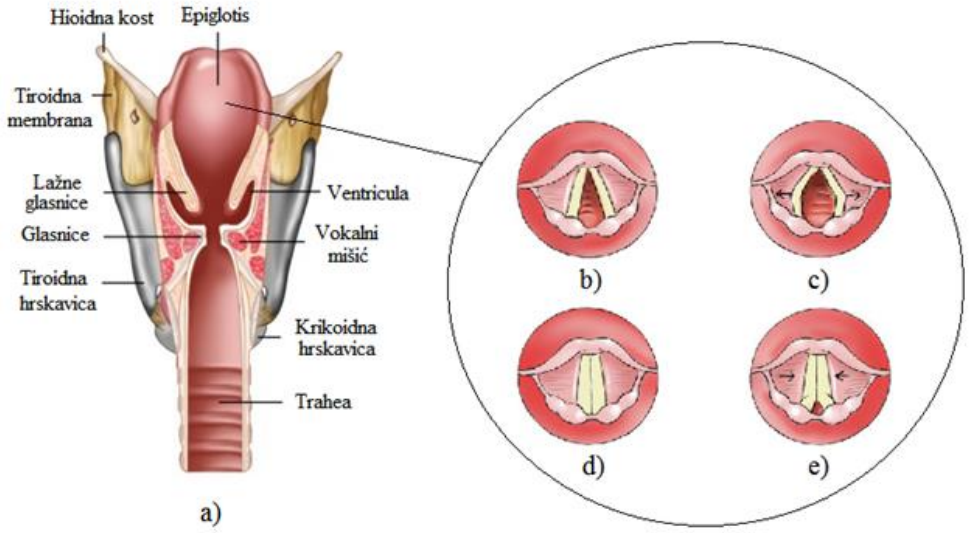
predstavljaju dva bočna mišića smeštena unutar larinksa, Slika 4.2 a). Glasne žice imaju dvostruku ulogu: biološku i akustičku. Njihova prva i osnovna funkcija je zaštita trahee, tj. dušnika, od upadanja hrane prilikom gutanja. Tokom disanja glasnice su razmaknute i formiraju trouglasti otvor koji se naziva glotis. Druga akustička funkcija glasnica se razvila kasnije tokom čovekove evolucije i ogleda se u njihovom treperenju koje daje vazdušnoj struji periodičnu i zvučnu sturkturu. Tako isprekidano vazdušno strujanje, odnosno zvučni talas, prolazi kroz vokalni trakt i dodatno se uobličava radom artikulacionih organa na čijem izlazu se dobija govorni signal.



Slika 4.1 Anatomija govornog mehanizma.

Međutim, u fiziološkom pogledu generisanja šapata postoje razlike u odnosu na produkciju govora. Prva razlika se javlja u fonatornom sistemu nakon pobude vazdušnom strujom iz resporatornih organa, a druga u vokalnom traktu i ogleda se u specifičnom radu artikulacionih organa. U slučaju šapata, prilikom prodiranja vazdušne struje kroz larinks glasnice miruju i ne vibriraju a gotalni otvor (*glotis*) ostaje delom otvoren [Lashley et al., 1980], kao što je prikazano na Slici 4.2 e). Na slici je ilustrovano nekoliko različitih položaja glotisa, u zavisnosti od toga da li se radi o zvučnom izgovoru, odnosno govoru, ili o stanju mirovanja, dubokom disanju ili šapatu. U stanju mirovanja, glasnice su razmaknute i kreiraju otvor oblika obrnutog slova 'V'. Sa udisajem glasnice se dodatno razmiču i obezbeđuju slobodan protok vazduha kroz otvor nalik obrnutom slovu 'U'. Kada se govori, glasnice su međusobno primaknute ali ne i previše zategnute, tako da po nailasku vazdušne struje one vibriraju. U šapatu su

glasnice takođe primaknute, ali ne čitavom dužinom i na svom kraju formiraju trouglasti otvor u vidu obrnutog slova 'Y'. Taj otvor je obično manji od 25% maksimalne površine otvora glotisa [Catford, 1964]. Glasnice su u ovom slučaju čvrsto stegnute (nisu labave kao u govoru), čime je sprečeno njihovo podrhtavanje. Sličan oblik glotisa sa nešto više razmaknutim glasnicama se formira u izgovoru bazvučnih glasova. Zahvaljujući endoskopiji, ustanovljeno je da je larinks u šapatu nešto više izdignut nego u govoru, zbog čega je dužina vokalnog trakta skraćena [Mills, 2009].



Slika 4.2 a) Uzdužni presek larinksa; Izgled glotisa gledano odozgo: b) tokom govora, c) prilikom dubokog udisaja, d) u mirnom stanju, e) tokom izgovora vokala u šapatu.

Pojedini autori u svojim istraživanjima [Monoson et al., 1984; Solomon et al., 1989] prave razliku između forsiranog šapata (*hard whisper* ili *forced whisper*) i slabog ili mekog šapata (*weak whisper* ili *soft whisper*). Pored toga što forsirani šapat nosi više energije i može dalje dopreti, u poređenju sa mekim šapatom izvesne su razlike u načinu generisanja i u akustičkim osobinama. Poređenje ova dva oblika šapata se može analizirati merenjem oblika glotisa i protoka vazduha [Sundberg et al., 2009]. Slabi šapat obično ima formu glotisa u obliku obrnutog slova 'V' [Laver, 1980]. Mills je korišćenjem video-endoskopa u svojim istraživanjima otkrio da je fiziologija govornog mehanizma tokom izgovora zvučnih i bezvučnih konsonanata veoma slična u šapatu [Mills, 2009]. Takođe je primetio postojanje razlike u veličini glotalnog otvora između govora i šapata. Merenjem protoka vazduha i njegovog pritiska u određenim područjima govornog mehanizma je pokazano da je u šapatu pritisak znatno veći nego u govoru i

najviši je u transglotalnom delu [Monoson et al., 1984]. Zbog toga protok vazduha u šapatu nije treperav poput onog u govoru, već je turbulentan, aperiodičan i ima šumnu strukturu. U normalnom govoru transglotalni pritisak je posledica razlike subglotalnog i intraoralnog pritiska. Međutim, kako su subglotalni i intraoralni pritisci u šapatu slični, veruje se da je povišeni transglotalni pritisak posledica smanjenja dimenzija glotisa, odnosno smanjenja glotalne impedanse [Klich, 1982]. Takođe je ustanovljeno da je ovaj pritisak veći u forsiranom šapatu nego u mekom.

Šapat i govor se razlikuju i u pogledu disanja. Otkriveno je da se u procesu govora i šapata udiše ista količina vazduha, međutim prilikom izdisaja više se vazuha istiskuje u šapatu [Schwartz, 1970; Stathopoulos, 1991]. Prema tome potrebe za udisanjem su veće tokom šapata, što dovodi do češćih pojava pauza i sporijeg izgovora.

Što se tiče razlika u artikulaciji, šapat je karakterističan po hiperartikulaciji, odnosno po povećanom naporu koji se javlja u cilju povećanja njegove razumljivosti. Iz tog razloga šapat odlikuju sporija brzina artikulacije i pažljivija artikucalija (npr. pažljiviji položaj jezika i nepca), duži izgovor pojedinih fonema (npr. duža artikulacija bezvučnih konsonanata radi isticanja razlika u dužini zvučno-bezvučnih glasova koja je prisutna u normalnom govoru), manja varijabilnost šapata, odnosno njegova veća stabilnost u odnosu na govor, itd. [Osfar, 2011]. Istraživanja su pokazala da su pokreti usana nešto brži u šapatu nego u normalnom govoru [Highashikawa et al., 2003]. Takođe, merenjem kontrakcija mišića usana i vilice je ustanovljeno da je njihovo anticiparno (preuranjeno) signaliziranje duže u šapatu nego pri normalnom govoru [Bonnot et al., 1991].

Na kraju, potrebno je istaći da šapat ne predstavlja uvek željenu radnju, već može biti i posledica zdravstvenih problema. Na primer šapat se može javiti kao propratni oblik ozbiljnijeg laringitisa i rinitisa, ili može biti posledica hroničnih oboljenja larinksa [Jovičić et al., 2008].

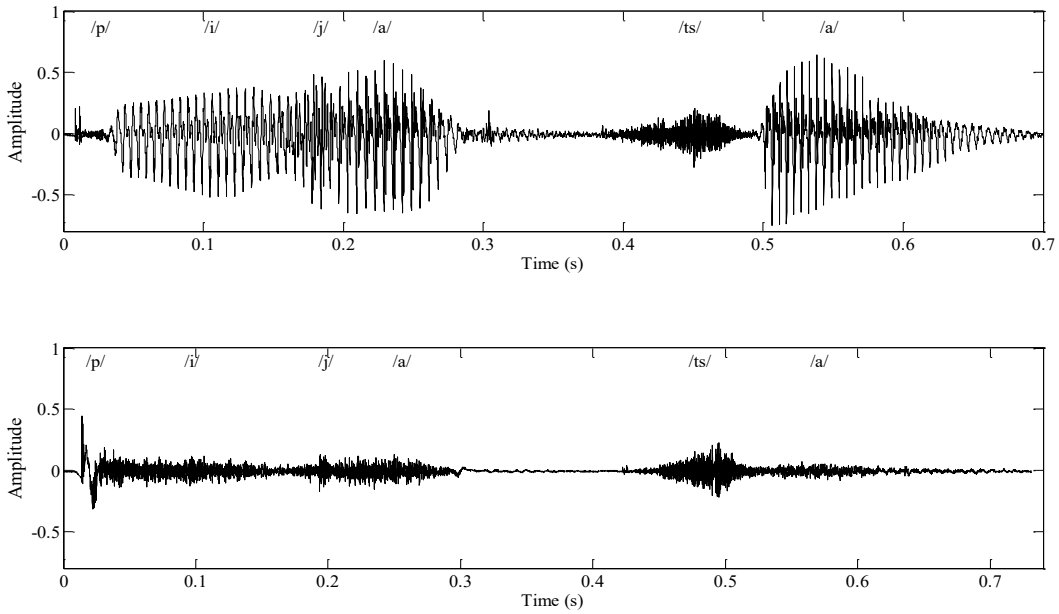
4.2 AKUSTIČKE KARAKTERISTIKE ŠAPATA

Usled razlika u načinu generisanja, akustičke karakteristike šapata su dosta drugačije od normalnog govora. Te razlike su primetne kako u vremenskom, tako i u

frekvencijskom domenu. U nastavku teksta uporedićemo: talasne oblike, spektre, spektrograme i intenzitete govora i šapata.

4.2.1 TALASNI OBLIK

Na Slici 4.3 su ilustrovani primeri talasnih oblika govornih signala u dva govorna moda. U pitanju je izgovor jedne reči od strane istog govornika u normalnom govoru i šapatu.

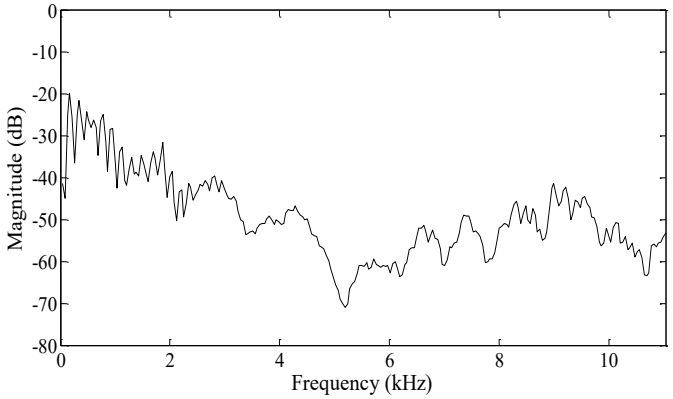


Slika 4.3 Poređenje talasnih oblika izgovora reči "pijaca" u govoru (slika gore) i šapatu (slika dole). [Grozdić et al., 2013 b]

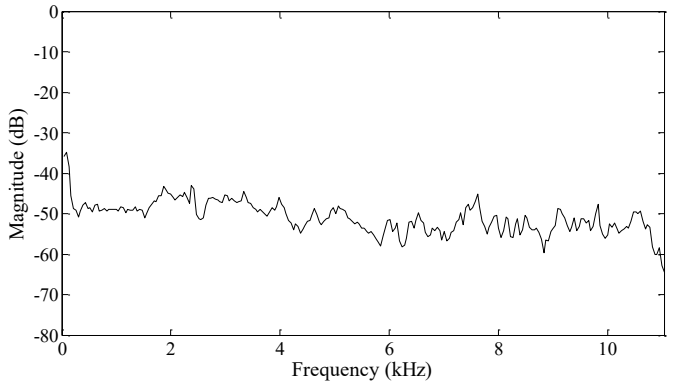
Upoređujući signale prva primetna razlika je znatno niža energija i šumna struktura šapata u odnosu na normalan govor. Usled nedostatka zvučnosti u šapatu, amplitude zvučnih glasova, pre svega vokala, su dosta niže, dok su amplitude bezvučnih glasova sličnog intenziteta kao u govoru [Ito, 2005]. Sa druge strane primetna je pojava iznenadnih i velikih skokova amplituda koja se javlja nakon okluzije kod ploziva (u slučaju sa slike kod ploziva /p/). U vremenskom domenu šapat karakteriše nešto duže trajanje, koje je u slučaju izgovora jedne reči reda 50ms. Iako je ovakvo produženo trajanje izgovora reči primetno na Slici 4.3, ono je dosta očiglednije na dužim snimcima rečenica.

4.2.2 *SPEKTRALNI NAGIB*

Razlike u procesu generisanja govora i šapata su se odrazile i na spektralni domen šapata. Prvo, usled nepostojanja periodične ekscitacije i harmonijske strukture, u šapatu ne postoji osnovna frekvencija a time ni čitav niz drugih spektralnih obeležja. Informacije o zvučnosti su izgubljene, a turbulentni i aperiodični izvor šapata se ponaša kao izvor šuma. U tom pogledu druga karakteristika šapata je dosta ravniji spektralni nagib [Jovičić, 1998; Ito et al., 2005; Zhang et al., 2007], ili niža spektralna gustina snage, koja se jasno može uočiti poređenjem slika 4.4 i 4.5.



Slika 4.4 Prikaz spektralnog nagiba u govoru. [Grozdić et al., 2013 b]

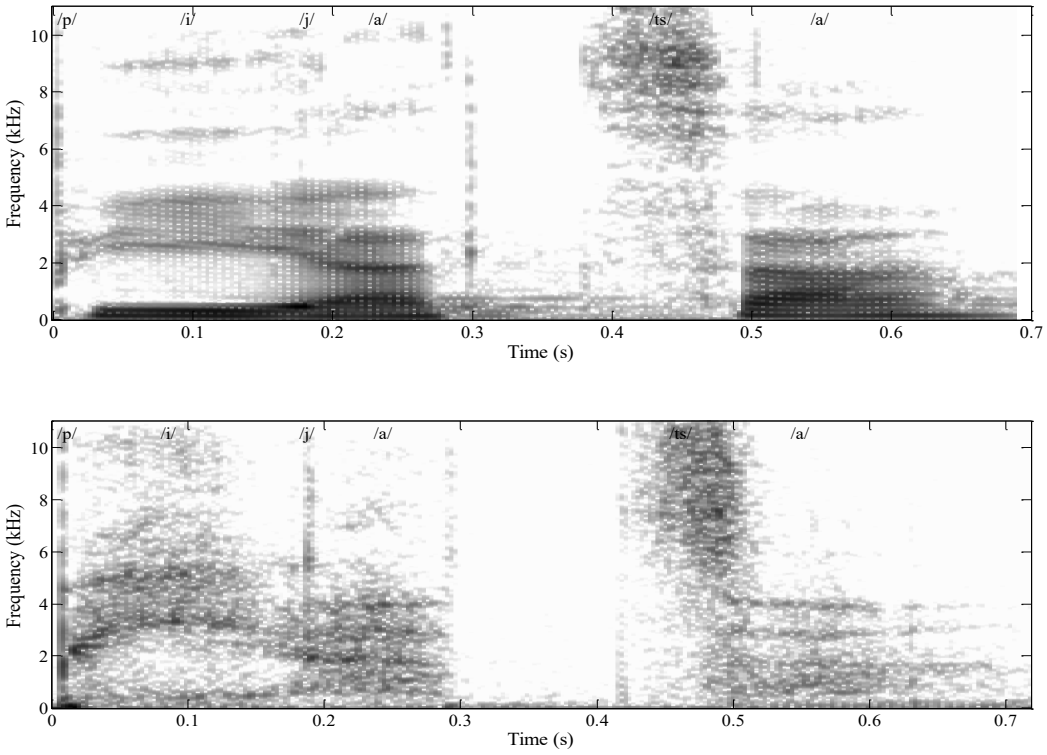


Slika 4.5 Prikaz spektralnog nagiba u šapatu. [Grozdić et al., 2013 b]

4.2.3 *FORMANTI*

Analiza spektrograma je omogućila otkrivanje drugih spektralnih odlika šapata i novih razlika u odnosu na govor. Te razlike se pre svega ogledaju u izmenjenim lokacijama formanata. Najizraženije promene formanata su u izgovoru vokala kod kojih

su izdignute frekvencije prva tri formanta. Frekvencijski pomeraj je najveći kod F1 i tumači se kao posledica karakterističnog položaja larinksa i skraćene dužine vokalnog trakta u šapatu. Pomeraji formanata F2 i F3 su manje izraženi. Pretpostavlja se da je manji pomeraj F2 posledica slabije promenjenog oblika oralne duplje i položaja jezika koji je veoma sličan onome u govoru [Kallail et al., 1984 b]. U srpskom jeziku [Jovičić, 1998; Jovičić et al., 2008] spektrogrami svih vokala ukazuju na povišene F1 vrednosti osim u slučaju vokala /u/. Drugi formant je izdignut kod svih vokala osim kod muškaraca u slučaju glasova /u/ i /e/, dok F3 i F4 nemaju pravilne promene. Ito, [Ito, 2005], je takođe ustanovio povišene frekvencije F1 i F2 u japanskom jeziku, ali nije analizirao više formante. Slično je i sa engleskim jezikom [Kallail et al., 1984 b]. Sa Slike 4.6. se može videti da spektrogram šapata ima prilično šumnu energetsku strukturu koja se za razliku od govora prostire i na više frekvencije. U tom pogledu spektrogram šapata dosta asocira na spektrogram dubokog disanja [Baken et al., 1991].



Slika 4.6 Poređenje spektorgrama izgovora reči "pijaca" u govoru (slika gore) i šapatu (slika dole).

4.2.4 INTENZITET

Jedna od glavnih karakteristika šapata je njegova niska energija zbog koje je snimanje šapata sa dobrim odnosom signal-šum (SNR) veoma teško ostvarivo pomoću konvencionalnih mikrofona i u lošim akustičkim ambijentalnim uslovima. Takođe, dodatni problem pravi ravni spektar šapata koji je dosta sličan šumu i otežava mogućnosti primene standardnih tehnika za potiskivanje šuma iz snimka.

4.3 PERCEPCIJA ŠAPATA

Uprkos velikim razlikama šapata u odnosu na govor, njegova percepcija je iznenađujuće visoka i nije mnogo lošija od razumljivosti govora. Usled nepostojanja zvučnosti i osnovne frekvencije, od šapata se ne očekuje da prenosi *pitch* informacije. Međutim, veliki broj istraživanja dokazuje upravo suprotno i tvrdi da šapat pored verbalnih informacija nosi dosta drugih neverbalnih informacija među kojima i *pitch* informacije. Tako je utvrđeno da se u šapatu u izvesnoj meri može prepoznati pol govornika, njegove emocije i druge prozodijske informacije [Hultsch et al., 1992].

Istraživanja auditornog sistema i percepcije šapata [Stevens et al., 1999) su pokazala da se distinkcija između zvučnih i bezvučnih glasova u šapatu obavlja na ranom stupnju u unutrašnjem uhu. Zvučni i bezvučni glasovi u šapatu se posebno tretiraju u pogledu drugačijeg enkodovanja informacija u slušnom nervu. Ipak, zbog izmenjenih frekvencija formanata, kod zvučnih glasova, pre svega vokala, se pojavljuje problem njihove diskriminacije [Tartter, 1991; Higashikuwa et al., 1999]. Kallail je u svojim eksperimentima istraživao i upoređivao prepoznavanje vokala u govoru i šapatu [Kallail, 1984 a]. U njegovim analizama ispitanici su u 85% slučajeva uspešno prepoznavali vokale u govoru, dok je njihovo prepoznavanje u šapatu degradirano i iznosi 63%. Tartter je sa druge strane u svojim eksperimentima izmerila nešto više vrednosti i to 80-99% uspeha u prepoznavanju vokala u govoru i 72%-99% u šapatu [Tartter, 1991]. U oba slučaja, upoređujući pojedinačne rezultate po vokalima, njihov uspeh prepoznavanja u šapatu nije pao ispod 20% u odnosu na govor, što ukazuje na visok stepen prepoznavanja vokala u šapatu.

Sa druge strane, razlike između konsonanata u govoru i šapatu su dosta manje. Slično kao i u govoru u šapatu se javlja konfuzija između ploziva /p/ i /b/.

Prepoznavanje CV (*consonant-vocal*) difona je dobro u slučaju ploziva, ali je slabije u slučaju frikativa. Interesantani su rezultati istraživanja matrica konfuzije prepoznavanja CV difona u šapatu gde su mnogo češće greške tipa [bezvučni glas]→[zvučni glas] nego obrnuti slučaj. Prema tome zvučni glasovi u šapatu imaju veću tendenciju da predstavljaju svoje zvučne parnjake nego obratno [Tartter, 1989].

Eksperimenti su pokazali da je uprkos nedostatku glotalnih vibracija u šapatu prisutna izvesna informacija o zvučnosti [Dannenbring, 1980]. Nekoliko studija je pokazalo mogućnost slušalaca da ocene *pitch*, mada nije objašnjeno na koji se način to tačno postiže [Thomas, 1969; Highashikawa et al., 1996]. U govoru percepcija *pitch* frekvencije zavisi od prva tri formanta, dok je u šapatu nešto drugačije. Veruje se da je njena percepcija u šapatu kompleksnija i da se bazira na korelaciji sa promenama prva dva formanta.

Percepcija pola govornika je sastavni deo identifikacije govornika. Istraživanja su pokazala da se pol govornika može prepoznati sa 95% uspeha na osnovu zvučnog izgovora vokala, dok sa 75% uspeha u šapatu [Lass et al., 1976]. Što se tiče percepcije identiteta govornika, ona je nešto slabija. U eksperimentu [Tartter, 1991] većina slušalaca je mogla da identifikuje šapat jednog od 10 govonika sa 46,2% − 62,5% uspeha, dok su pojedinci imali uspeh u čak 96,3% slučajeva. Ti slušaoci, koji su skoro "nepogrešivo" identifikovali govornike, su tvrdili da su koristili pojedine karakteristike govora, poput dužine izgovora pojedinih slogova, kao dobre identifikatore govornika.

U šapatu postoje neki pokazatelji mogućnosti percepcije emocija. Na primer, Tartter je u svom istraživanju zatražio od govornika da sa osmehom na licu govore i šapuću, ali da "ne zvuče srećno" [Tartter et al., 1994]. Slično je urađeno sa mrštenjem. Rezultati su pokazali da su slušaoci mogli da prepoznaju da li se govornik mrštio u govoru i šapatu, dok to nije bilo moguće utvrditi za osmeh. Slično govoru [Grozdić et. al. 2011 a; Grozdić et al., 2011 b] i u šapatu su uočeni problemi u diskriminaciji snažnih emocija, poput radosti i straha ili radosti i ljutnje [Scherer et al., 1988; Tartter et al., 1994].

Zbog drugačije osmišljenih eksperimentata, istraživačkih metoda i razlika u maternjim jezicima kojima su ispitanici govorili, u istraživanjima percepcije šapata i

njihovim rezultatima su primetne razlike. Ipak, svi eksperimenti su usaglašeni oko jedne stvari a to je veoma dobra razumljivost šapata. Međutim, postoji još dosta nerazjašnjenih stvari u pogledu percepcije *pitch* frekvencije, emocija i identifikacije govornika u šapatu, čime je prostor za buduća istraživanja i dalje otvoren.

4.4 AUTOMATSKO PREPOZNAVANJE ŠAPATA (PREGLED DOSADAŠNJIH ISTRAŽIVANJA)

Primena govornih tehnologija u automatskom prepoznavanju šapata je relativno nova ideja koja je i dalje u svom povoju. Postoji veoma mali broj radova na ovu temu, pri čemu se u literaturi uglavnom spominju istraživanja koja se tiču automatske detekcije šapata [Carlin et al., 2006], identifikacije govornika [Fan et al., 2008; Fan et al., 2009; Fan et al., 2011], različitih pokušaja adaptacije i poboljšanja ASR sistema za prepoznavanje šapata [Itoh et al., 2002; Morris, 2003; Ito et al., 2005; Ghaffarzadegan et al., 2014] itd. Takođe postoje istraživanja koja se tiču resinteze, odnosno rekonstrukcije zvučnog govora iz šapata [Morris, 2003; Sharifzadeh, 2009]. Ova istraživanja su dobrim delom motivisana poboljšanjem kvaliteta životnih aktivnosti post-laringektomisanih pacijenata, sa ciljem da im se omogući komunikacija "normalnim" glasom, uprkos nemogućnostima larinksa. U nastavku teksta biće opisani do sada testirani ASR sistemi i njihove performanse u automatskom prepoznavanju šapata.

4.4.1 PRIMENA DTW U AUTOMATSKOM PREPOZNAVANJU ŠAPATA

DTW algoritam je u govornim tehnologijama već duže vremena u senci trenutno aktuelnih i po performansama dosta nadmoćnijih sistema poput konvencionalnih HMM i hibridnih (“tandem”) HMM/DNN rešenja, te kao takav nije bio od većeg interesa u analizi automatskog prepoznavanja šapata. Ipak, zbog svog velikog teorijskog značaja, za potrebe istraživanja šapata u srpskom jeziku, DTW je testiran u prepoznavanju izolovanih reči, a preliminarni rezulati ovih istraživanja su publikovani u nekoliko radova [Marković et al., 2013 a; Marković et al., 2013 b; Marković et al., 2014; Marković et al, 2015; Marković et al, 2016]. Testiranje je obavljeno na specijalno kreiranoj bazi snimaka izgovora izolovanih reči u dva govorna moda − u normalnom govoru i šapatu. Struktura ove govorne baze, koja je korišćena i u eksperimentima ove disertacije, je detaljno opisana u petom poglavlju.

Preliminarni rezultati DTW istraživanja prezentuju analizu automatskog prepoznavanja reči jednog dela korpusa, tačnije 20 različitih reči (6 boja i 14 brojeva). U eksperimentima su korišćena MFCC obeležja, a rezultati analize usaglašenih trening/test scenarija za 4 govornika su pokazali 98,74% uspeha u prepoznavanju reči u normalnom govoru, i 95,05% uspeha u prepoznavanju reči u šapatu [Marković et al., 2013 a]. Dodavanje *delta* i *delta-delta* obeležja nije imalo bitnijeg efekta na poboljšanje rezultata. U kasnijim eksperimentima sa usrednjavanjem rezultata većeg broja govornika (10 govornika) došlo je do očekivanog smanjenja uspeha prepoznavanja reči, koji u govoru iznosi 96,37%, a u šapatu 89,95% [Marković et. al., 2014]. Primena LPCC i dinamičkih obeležja nije imala značajnijeg uticaja.

Sudeći prema rezultatima u usaglašenim trening/test scenarijima, DTW algoritam je pokazao visok uspeh u prepoznavanju reči izgovorenih u šapatu, koji je za korpus ovog obima sasvim poredljiv sa performansama HMM sistemima. Ono u čemu DTW sistemi zaostaju je prepoznavanje reči u takozvanim neusaglašenim trening/test scenarijima, odnosno u scenarijima u kojima se sistem trenira na govoru a testira sa šapatom ili obrnuto. Zbog ovog nedostatka, pepoznavanje šapata pomoću DTW sistema u neusaglašenim scenarijima je jedino moguće ako se u procesu obuke pored normalnog govora koriste i snimci izgovora reči u šapatu. Ovaj proces je poznat kao adaptacija ASR sistema sa uzorcima šapata.

4.4.2 PRIMENA HMM U AUTOMATSKOM PREPOZNAVANJU ŠAPATA

Za razliku od DTW sistema, u literaturi se može naći znatno više radova u kojima su HMM sistemi primenjeni u prepoznavanju šapata. U jednoj od prvih studija automatskog prepoznavanja šapata [Ito et al., 2005] ispitana je mogućnost prepoznavanja fonema u šapatu pomoću HMM modela i MFCC obeležja. Cilj ovog istraživanja je bio razvoj posebnog ASR sistema za mobilne telefone koji bi bio robustan na bučne ambijentalne uslove i različite govorne modove, među kojima i na šapat. Tri govorna moda su analizirana: šapat, tihi govor i normalan govor. Upotrebljena govorna baza je sačinjena od snimaka telefonskih razgovora u šapatu na japanskom jeziku. U pitanju je bio unapred pripremljeni tekst koji su govornici čitali tokom uspostavljene telefonske veze. Rezultati eksperimenta su pokazali da se prikrivanjem usta i mobilnog telefona rukom može donekle poboljšati SNR u bučnom okruženju. U

usaglašenim trening/test scenarijima uspeh prepoznavanja govora je 82%, a šapata 68%. U neusaglašenim trening/test scenarijima uspeh u prepoznavanju je dosta degradiran. Na primer uspeh u prepoznavanju šapata pomoću HMM modela koji je obučen sa normalnim govorom iznosi svega 27%. U slučaju testiranja govorom, HMM model obučen na šapatu je imao 53% uspeha u prepoznavanju reči. U dobijenim rezultatima primetna je pojava da se normalan govor sa većim uspehom prepoznaje kod HMM modela obučenim na šapatu nego što je to slučaj sa prepoznavanjem šapata kod HMM modela koji je obučen na govoru. Sličan fenomen je zabeležen i u radovima [Grozdić et al., 2012; Grozdić et al., 2013; Grozdić et al., 2014 a; Galić et al., 2014 b; Grozdić et al., 2016] a detaljna analiza i tumačenje ove interesantne pojave je obrazloženo u ovoj tezi. U istom istraživanju je dokazano da se prepoznavanje šapata pomoću HMM modela koji je obučen sa normalnim govorom može poboljšati adaptacijom, koja podrazumeva dodavanje malog broja uzoraka snimaka šapata u procesu obuke. Na taj način se takozvani *speaking-style-independent* model adaptira na prepoznavanje šapata, te može prepoznati oba modaliteta govora. Maksimalni postignuti uspeh ovakvog modela u prepoznavanju šapata je 66%, a govora 80%.

Nakon ovog pionirskog rada, ubrzo su usledila istraživanja automatskog prepoznavanja šapata na engleskom govornom području. U ovim radovima autori su pokušali da ublaže akustičku neusaglašenost između neutralnog govora i šapata i poboljšaju performanse prepoznavanja šapata koristeći različite metode adaptacije ASR modela na šapat [Lim, 2011; Mathur et al., 2012; Yang et al., 2012] i transformacijom govornih obeležja [Yang et al., 2012]. Postoje studije koje su se fokusirale na dizajn *front-end* dela ASR sistema i posebne metode ekstrakcije govornih obeležja [Zhang et al., 2010; Ghaffarzadegan et al., 2014 a] kao i na izmene u bankama filtara podešavajući propustni opseg i položaj određenih filtara [Ghaffarzadegan et al., 2014 a]. Takođe, ispitana je i efikasnost takozvane metode normalizacije dužine vokalnog trakta (*Vocal Tract Length Normalization – VTLN*) [Ghaffarzadegan et al., 2014 b], kao i šift transformacije (*shift transformation*) [Boril et al., 2010; Ghaffarzadegan et al., 2014 b]. Nekoliko studija je obavljeno sa ciljem adaptacije ASR modela korišćenjem malog uzorka šapata [Ghaffarzadegan et al., 2014 b, 2015]. U radu [Ghaffarzadegan et al., 2014 b] je ispitana i metoda *Vector Tailor Series* (VTS) za adaptaciju modela sa

uzorcima pseudo-šapata. Testiran je i audio-vizuelni pristup automatskog prepoznavanja izgovora izolovanih reči (brojeva) u šapatu [Tao et al., 2014].

Pored analize prepoznavanja šapata u japanskom i engleskom jeziku, HMM sistemi su testirani i u prepoznavanju izolovanih reči, monofona i trifona u šapatu i normalnom govoru na srpskom jeziku [Galić et al., 2013 a; Galić et al., 2013 b; Galić et al., 2014; Grozdić et al., 2015; Grozdić et al. 2017]. Korišćenjem ranije spomenute govorne baze šapata za srpski jezik, HTK (*Hidden Markov Toolkit*) softvera [Young et al., 2001] i MFCC obeležja u analizi usaglašenih trening/test scenarija postignuti su maksimumi od 99,75% uspeha u prepoznavanju reči u govoru i 99,52% uspeha u prepoznavanju reči u šapatu. U neusaglašenim trening/test scenarijima uspesi prepoznavanja reči su znatno niži i iznose 51,86% u prepoznavanju govora i 36,24% u prepoznavanju šapata [Galić et al., 2014]. Nešto bolji rezultati u neusaglašenim scenarijima su dobijeni u prepoznavanju monofona i trifona i to 74,88% (monofoni) i 64,94% (triifoni) u govoru i 64,8% (monofoni) i 28,32% (trifoni) u šapatu.

Zbog sličnosti problematike sa automatskim prepoznavanjem šapata korisno je spomenuti i studiju [Fan et al., 2011] koja je ispitivala primenu skrivenih Markovljevih modela sa jednim stanjem (GMM) u automatskoj identifikaciji govornika na osnovu šapata. Za potrebe ovog istraživanja korišćene su dve govorne baze na engleskom jeziku, jedna sa snimcima spontano izgovorenih rečenica u govoru i druga sa rečenicama u šapatu. GMM model sa PLP (*Perceptual Linear Prediction*) i MFCC obeležjima je korišćen kao sistem za automatsko porepoznavanje govornika. U istraživanju je posebna pažnja posvećena neusaglašenim trening/test scenarijima u kojima je uspeh identifikacije govornika u velikoj meri bio degradiran. Eksperimenti ove studije su pokazali da upotreba linearne ili eksponencijalane frekvencijske skale umesto standardne Melove logaritamske skale doprinosi poboljšanju prepoznavanja govornika. Na taj način, maksimalni zabeleženi uspeh u identifikaciji govornika na osnovu njegovog spontanog govora u šapatu iznosi 83.84%.

Šapat je akustički modelovan najviše sa HMM sistemima, a automatsko prepoznavanje šapata je do sada testirano na japanskom, engleskom i srpskom jeziku. Rezultati svih istraživanja primene HMM modela u automatskom prepoznavanju šapata su pokazali visok uspeh u usaglašenim trening/test scenarijima i dosta lošije rezultate

prepoznavanja u neusaglašenim trening/test scenarijima. Ipak, iako degradirani rezultati HMM sistema u neusaglašenim scenarijima su daleko bolji od onih koji su postignuti sa DTW sistemima, i daju osnovu za njihovo dalje eventualno poboljšanje.

4.4.3 PRIMENA ANN U AUTOMATSKOM PREPOZNAVANJU ŠAPATA

Za razliku od skrivenih Markovljevih modela, koji su još pre dvanaest godina prvi put primenjeni u analizi šapata, veštačke neuralne mreže do pre pet godina uopšte nisu testirane u automatskom prepoznavanju šapata. Prva i do nedavno jedina istraživanja primene ANN u prepoznavanju šapata su izvršena za srpski jezik [Grozdić et al., 2012; Grozdić et al., 2013 a; Grozdić et al., 2013 b; Grozdić et al., 2013 c; Grozdić et al., 2014; Grozdić et al., 2016 a; Grozdić et al., 2016 b]. Analizirane su pre svega MLP neuralne mreže, koje su poznate po svojim dobrim performansama i kompromisu koji ostvaruju između tačnosti, brzine prepoznavanja obrazaca i trošenja memorijskih resursa [Siniscalchi et al., 2013]. Kao govorna obeležja pored tradicionalnih MFCC su testirani i drugi tipovi kepstralnih koeficijenata. Detaljno su ispitani svi neusaglašeni trening/test scenariji [Grozdić et al., 2013 b; Grozdić et al., 2013 c] i predložena je nova metoda predobrade govornih signala [Grozdić et al., 2014] koja poboljšava prepoznavanje izolovanih reči u šapatu. Upoređene su performanse MLP i HMM sistema u prepoznavanju šapata [Grozdić et al., 2016 a]. Takođe, analizirane su i dubinske neuralne mreže, koje u ulozi autoenkodera imaju zadatak da filtriraju i potisnu efekte šapata iz kepstralnih koeficijenata i na taj način u sklopu sa *back-end* HMM delom formiraju hibridni odnosno tandem HMM/DNN ASR sistem robustan na šapat [Grozdić et al., 2016 b]. Rezultati svih ovih istraživanja su tokom poslednjih pet godina publikovani u više radova [Grozdić et al., 2012; Grozdić et al., 2013 a; Grozdić et al., 2013 b; Grozdić et al., 2013 c; Grozdić et al., 2014; Grozdić et al., 2016 a; Grozdić et al., 2016 b] i sumirani u ovoj doktorskoj disertaciji.

Do trenutka pisanja ove teze, u literaturi je poznat samo još jedan rad koji je analiziralo primenu ANN u automatskom prepoznavanju šapata. U pitanju je rad [Lee et al., 2014] koji je upotrebio DNN tip neuralnih mreža za potrebe prepoznavanja šapata u mandarinskom dijalektu kineskog jezika.

4.5 REZIME

U ovom poglavlju su opisane neke od osnovnih karakteristika i razlika šapata u odnosu na govor, kako u fiziološkom načinu generisanja tako i u pogledu akustičkih osobina. Usled nedostatka glotalnih vibracija, odnosno zvučnosti, šapat ima dosta ravniji spektar od govora i nižu energiju, koja je glavni uzrok niskog SNR u audio snimcima. Analiza spektrograma potvrđuje izmenjene lokacije prva tri formanta, kao i šumnu spektralnu strukturu šapata sličnu onoj pri dubokom disanju. Iako je eksperimentalno dokazano da se šapat može uz minimalne napore sasvim dobro razumeti i da pored verbalnih nosi dosta neverbalnih informacija, automatsko prepoznavanje šapata uopšte nije jednostavan zadatak. Nizak SNR, izmenjene lokacije formanata, ravan spektar i šumna karakteristika šapata predstavljaju ozbiljne prepreke u prepoznavanju šapata pomoću tradicionalnih ASR sistema. Istraživanje automatskog prepoznavanja šapata su još uvek na početnom stadijumu, a do sada su prvenstveno testirani HMM sistemi. Aktuelne studije i dalje pokazuju skromni uspeh u prepoznavanju šapata u neusaglašenim obuka/test scenarijima. Iz tog razloga je započet niz novih istraživanja sa ciljem pronalaska novih govornih obeležja, metoda obuke ASR sistema i predobrade govornih signala. Poslednjih godina, veštačke neuralne mreže su ponovo u fokusu mnogih istraživanja među kojima i onih koje se tiču automatskog prepoznavanja šapata. Predstojeći deo ove teze, prezentovaće do sada ostvarene autorske rezultate i pomake u automatskom prepoznavanju šapata primenom ANN.

5 KREIRANJE I ANALIZA KORPUSA ŠAPATA

Na samom početku ovog poglavlja ukratko su predstavljene neke od postojećih govornih baza šapata, a zatim je detaljno opisano kreiranje i akustička analiza prve i trenutno jedine govorne baze tog tipa za srpski jezik, Whi-Spe, koja je korišćena kao osnova svih eksperimenata ove doktorske disertacije. Takođe, u nastavku poglavlja je opisan postupak inverznog fitriranja i generisanja baze pseudo-šapata, koja je takođe korišćena u eksperimentima ove teze.

5.1 POSTOJEĆI KORPUSI ŠAPATA

Istraživanja šapata su u velikoj meri ograničena malim brojem kvalitetno snimljenih i javno dostupnih korpusa šapata. Trenutno najviše eksploatisane baze u istraživanjima šapata su wTIMIT (*whispered TIMIT*), wMRT (*whispered Modified Rhyme Test*), UT-VE I (*UT Vocal Effort I*) i UT-VE II (*UT Vocal Effort II*) korpusi na enegleskom jeziku, koji su pre svega korišćeni u eksperimentima prepoznavanja šapata i ispitivanjima njegove razumljivosti.

Korpus wTIMIT [Lim, 2011] je razvijen na Univerzitetu Ilinois i kreiran je da zadovolji potrebe automatskog prepoznavanja velikog korpusa reči u šapatu. Ovaj korpus je fonetski izbalansiran i dovoljno veliki da zadovolji sve statističke zahteve treninga akustičkih modela. Snimljen je po ugledu na poznati TIMIT korpus normalnog govora [Zue et al., 1990], na čijem razvoju su radili TI (*Texas Instruments*) i MIT

(*Massachusetts Institute of Technology*) za potrebe DARPA (*Defense Advanced Research Projects Agency*). TIMIT je uglavnom korišćen u eksperimentima automatskog prepoznavanja fonema u normalnom govoru. Sa druge strane, wTIMIT je korpus šapata koji je sistematski uređen u parove rečenica izgovorenih u normalnom govoru i šapatu. U izradi korpusa je učestvovalo 48 govornika sa dva govorna područja engleskog jezika (28 govornika iz Singapura i 20 govornika iz severne Amerike). Snimanje je obavljeno u specijalnim kabinama namenjenim za audiometriju, pri čemu je korišćen usmereni kondenzatorski mikrofon (*MX - 2001*) na rastojanju 15cm od usta govornika. Kako bi se tokom snimanja izbegla pojave uduvavanja mikrofona, on je malo nagnut unazad (suprotno od govornika). Prilikom snimanja šapata, govornicima je rečeno da se približe mikrofonu radi postizanja boljeg dinamičkog opsega u snimcima. Svaki govornik je pročitao set od 450 rečenica, koje su fonetski balansirane i kontekstualno prilagođene svakodnevnom engleskom jeziku. Korišćene rečenice su preuzete iz TIMIT korpusa i u procesu snimanja su izgovarane u više sesija kako bi se izbegao umor govornika. Uporedo sa snimanjem je vršena kontrola kvaliteta snimaka, pri čemu su snimci sa slabom artikulacijom, pogrešnim izgovorom reči ili sa isprekidanim rečenicama odbacivani i vršeno je njihovo ponovno dosnimavanje.

Korpus wMRT je dizajniran za potrebe testiranja razumljivosti šapata [Lim, 2011]. Sastoji se iz skupa snimaka izgovora jednosložnih reči u šapatu, koje se međusobno razlikuju samo u početnom ili završnom konsonantu. Snimanje baze je obavljeno u anehoičnoj prostoriji sa usmerenim kondenzatorskim mikrofonom (*MX-2001*) na rastojanju 15cm od govornikovih usta. U snimanju je učestvovalo 27 govornika sa područija severne Amerike. Kvalitet wMRT baze je dodatno poboljšan izbacivanjem šumnih snimaka, a finalno je snimljeno 15180 reči.

UT-Vocal Effort (UT-VE) I i II [Zhang et al., 2011] predstavljaju korpuse šapata za engleski jezik koji su snimljeni na Univerzitetu Teksas u Dalasu. UT-VE I sadrži snimke 12 muških govornika u pet govorna moda: šapat, tih govor, normalan govor, glasan govor i vika. Svaki govornik je imao zadatak da u šapatu izgovori 5 različitih rečenica koje su odabrane iz TIMIT baze, a dodatno je sniman i njegov spontani šapat u trajanju od jednog minuta. UT-VE II korpus je dosta obimniji i sadrži veliku količinu snimaka neutralnog govora sa umetnutim delovima izgovorenim u šapatu (*whisper-*

island). Sadrži snimke 72 govornika. Ovaj korpus je kreiran pre svega za potrebe detekcije govora (*voice activity*) i promene u naprezanju govora (*vocal-effort*). Nije toliko pogodan za sistematičnu analizu akustičkih i fonetskih karakteristika šapata.

U literaturi se pominje i nešto stariji CHAINS korpus [Cummins et al., 2006]. Ova baza snimaka je kreirana u Irskoj za potrebe analize individualnih karakteristika govornika, a u svom sastavu pored snimaka različitih modaliteta govora sadrži i snimke rečenica u šapatu. Baza je snimljena u studijskim uslovima i sastoji se iz 33 rečenice, delom preuzetih iz TIMIT baze, koje je 36 govornika izgovaralo u 6 govorna moda. Iako je ova baza kreirana za druge namene, ona u izvesnoj meri može poslužiti i u istraživanjima šapata.

Do trenutka pisanja ove teze, u postupku izrade je i jedna audio-vizuelna baza paralelnih snimaka šapata i normalnog izgovora za engleski jezik, AV-Whisper [Tran et al., 2013], koja je namenjena za potrebe kvantitativne analize razlika produkcije govora i šapata.

Pored engleskog jezika, govorna baza šapata je snimljena i za japanski jezik. U istraživanju šapata [Ito et al., 2005] je snimljen poseban korpus, CIAR, koji se sastoji iz dva dela. Prvi deo baze čine snimci šapata sa blisko postavljenim mirkofonom, a drugi deo baze predstavljaju snimci šapata u telefonskim razgovorima. U analizi akustičkih karakteristika šapata korišćen je prvi deo baze sa snimcima 123 govornika. Govornici su u šapatu izgovarali 110 rečenica koje su delom preuzete iz ATR [Kurematsu et al., 1990] i JANAS [Itou et al., 1998] korpusa. Za potrebe eksperimenta automatskog prepoznavanja šapata [Ito et al., 2005] korišćen je samo drugi deo baze koji se sastoji iz snimaka 10 govornika koji su tokom telefonskog razgovora u šapatu čitali 30 rečenica.

Krajem 2014. godine, na međunarodnoj koferenciji *Interspeech* je predstavljena i jedna govorna baza šapata za kineski jezik, tačnije za mandarinski dijalekat kineskog jezika [Lee et al., 2014]. Ovaj korpus, nazvan *iWhisper-Mandarin*, je sličan TIMIT i wTIMIT korpusima. Sačinjen je od paralelnih snimaka normalnog govora i šapata, dobijenih čitanjem rečenica iz dnevne štampe. U snimanju je učestvovalo 80 govornika koji su čitali unapred pripremljen skup od 100 fonetski balansiranih rečenica koje u

proseku imaju 15 reči. Ovaj korpus je kreiran sa namenom obrade signala govora i šapata, a do sada je upotrebljen i u svrhe treniranja ASR sistema.

Tabelarni pregled svih navedenih korpusa i njihovih osnovnih karakteristika je pikazan u Tabeli 5.1. Spomenuti korpusi nisu javno dostupni. Neki korpusi se mogu dobiti jedino uz pismenu saglasnost autora, dok se korišćenje ostalih naplaćuje. Za potrebe analize automatskog prepoznavanja šapata u srpskom jeziku, kreirana je posebna govorna baza, pod nazivom Whi-Spe [Marković et al., 2013] i opisana je u nastavku teksta.

Tabela 5.1

Postojeći korpusi šapata koji se spominju u literaturi. Skraćenica P označava paralelni korpus normalnog govora i šapata. Skraćenica FB označava fonetski izbalansirane korpuse.

Korpus	Broj govornika	Trajanje (sati)	P	FB	Jezik
CIAIR	68M + 55Ž	≈ 15	✓	✓	Japanski
AV-Whisper	8M + 3Ž	< 10	✓	?	Engleski
CHAINS	18M + 18Ž	< 3	✓	?	Engleski
UTVE-I	12M	< 1	✓	?	Engleski
UTVE-II	37M + 35Ž	< 1	×	?	Engleski
wTIMIT	25M + 23Ž	≈ 15	✓	✓	Engleski
wMRT	12M + 15Ž	< 1	✓	×	Engleski
iWhisper-Mandarin	40M + 40Ž	≈ 15	✓	✓	Mandarinski
Whi-Spe	5M + 5Ž	< 5	✓	✓	Srpski

5.2 DIZAJN WHI-SPE KORPUSA

Whi-Spe (skraćeno od *Whispered Speech*) je korpus šapata za srpski jezik i koncipiran je tako da sadrži dva dela: prvi deo koji čine snimci izgovora izolovanih reči u šapatu i drugi deo koji se sastoji iz snimaka reči u normalnom govoru. Snimci oba govorna moda su prikupljeni od deset govornika, pet ženskih i pet muških, koji su tokom procesa snimanja čitali 50 različitih reči sa ponavljanjem od po 10 puta. Kako bi se izbegao zamor govornika, snimanje je obavljeno u više ponovljenih sesija sa pauzama od po par dana. Ukupno je bilo 10 sesija a u govornoj bazi je finalno sakupljeno 10000 audio snimaka reči, od kojih je 5000 u šapatu, a drugih 5000 u normalnom govoru. U snimanju korpusa su volontirali studenti Visoke škole tehničkih strukovnih studija iz Čačka, kojima je Srpski maternji jezik. Govornici su bili prosečne starosti od 20 do 30 godina i imali su ispravnu artikulaciju i čulo sluha. Primeri reči iz Whi-Spe baze su preuzeti iz postojećeg GEES korpusa [Jovičić et al., 2004] i radi lakše organizacije su grupisani u tri podkorpusa: boje (6 reči), brojevi (14 reči) i balansirane

reči (30 reči). Sve reči su pažljivo odabrane i zadovoljavaju osnovne lingvističke kriterijume srpskog jezika (raspodela glasova, akcenatska struktura, slogovna struktura, konsonantski skupovi, itd). Spisak svih reči zajedno sa njihovom IPA notacijom je tabelarno prikazan i može se naći u Prilogu na kraju ove disertacije. Audio snimci Whi-Spe korpusa su sistematski organizovani u veći broj direktorijuma. Normalan govor i šapat su odvojeni u posebne direktorijume u kojima su dalje snimci razvrstani u zasebne poddirektorijume za svakog od govornika. U pojedinačnim direktorijumima govornika reči su dalje klasifikovane u tri poddirektorijuma: boje, brojevi i balansirane reči. Zbog velikog broja snimaka, utvrđena je posebna notacija audio fajlova koja omogućava njihovu brzu i jednostavnu pretragu. Whi-Spe korpus je besplatan za preuzimanje i podržava mogućnost daljeg proširenja.

5.3 SNIMANJE I OBRADA WHI-SPE KORPUSA

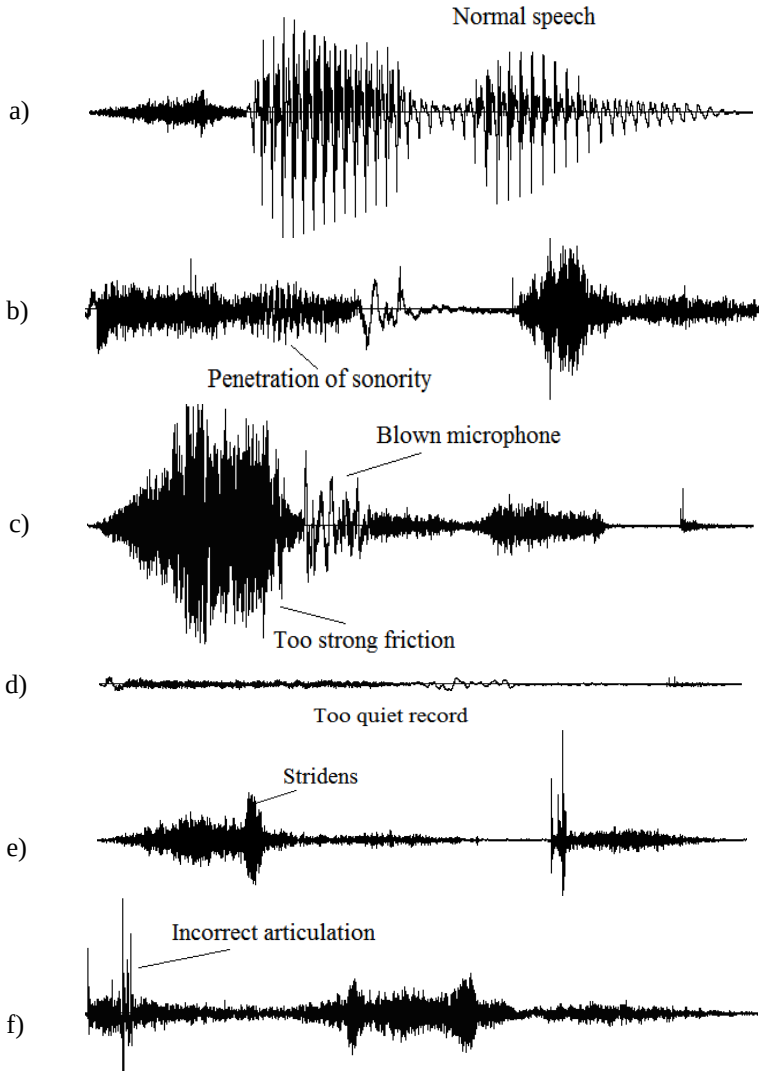
Whi-Spe baza je snimljena u tihim laboratorijskim uslovima, u specijalno konstruisanoj i akustički obrađenoj govornici, sa *Optimus* omni-direkcionim mikrofonom dobrog frekvencijskog odziva do 16kHz. Tokom snimanja normalnog govora, mikrofon je držan na rastojanju 25cm od usta govornika, dok je u šapatu bio na distanci od otprilike 5cm. Na ovaj način je pokušano da se dobiju što je moguće bolji snimci, naročito u šapatu. Govor je digitalizovan sa frekvencijom odabiranja od 22050Hz i rezolucijom 16 bita po odbirku i sačuvan je u *Windows PCM wav* formatu. Sesije snimanja su organizovane više od 10 puta kako bi se sakupio dovoljan broj dobrih snimaka. Tokom sesije, govornici su čitali u nizu 50 unapred pripremljenih reči u dva govorna moda: u šapatu i normalnom govoru. Tako snimljen set od 100 reči po govorniku je zatim manuelno segmentiran i testiran u pogledu njegovog kvaliteta. Posebna pažnja je posvećena segmentaciji reči u šapatu, u kojoj su učestvovali jedan stručnjak iz obrade govornih signala i jedan lingvista. Ukoliko je testirani snimak bio zadovoljavajućeg kvaliteta, on se obeležavao prema odgovarajućem pravilu notacije i smeštao u Whi-Spe bazu, dok se u obratnom slučaju vršila eliminacija snimka. Na ovaj način je snimljeno više od 10000 reči, ali je samo 10000 kvalitetnih snimaka zadržano u Whi-Spe bazi.

Kontrola kvaliteta snimaka je uočila razne vrste grešaka. Neke od njih su se odnosile na pogrešanu artikulaciju ili izgovor pojedinih glasova, a najveći broj grešaka

se javljao tokom snimanja šapata. (Ove greške su opisane u narednom poglavlju). Poseban problem u snimanju šapata je bio nizak SNR, koji se uglavnom pojavljivao kod ženskih osoba. U slučaju loših snimaka bila su neophodna ponovna snimanja.

5.4 SPECIFIČNE MANIFESTACIJE ŠAPATA TOKOM SNIMANJA

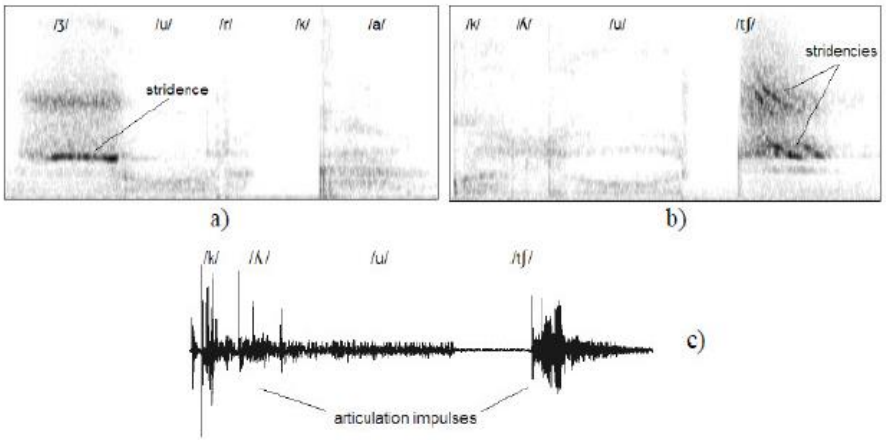
Tokom snimanja šapata pojavljivale su se dve vrste grešaka. Prvi tip su takozvane kontrolisane greške poput pogrešne artikulacije pojedinih glasova koje govornik slučajno proizvodi tokom snimanja. Ove greške se mogu ustanoviti odgovarajućom kontrolom kvaliteta snimaka i korigovati ponovnim snimanjem. Slika 5.1 ilustruje nekoliko primera pomenutih grešaka.



Slika 5.1 (a) Primer normalno izgovorene reči i pogrešnih snimaka u šapatu usled: (b) probijanja zvučnosti, (c) suviše jake frikcije i uduvanog mikrofona, (d) suviše tihog šapata, (e) pojave stridensa, (f) pogrešne artikulacije. [Marković et al., 2013 a]

Najčešće greške koje su se pojavljivale tokom snimanja šapata su probijanje zvučnosti, Slika 5.1 b) i prejaka frikcija u izgovoru frikativa i afrikata, Slika 5.1 c). Pored ovih grešaka česta je bila i pojava uduvanog mikrofona usled suviše bliskog položaja mikrofona i usta govornika. Ova pojava se prepoznaje po intenzivnom akustičkom impulsu prikazanom na Slici 5.1 c). Problematični su bili i suviše tihi šapat maskiran ambijentalnim šumom, Slika 5.1 d), omisija glasova u izgovoru pojedinih reči, pogrešan izgovor pojedinih glasova, itd. Pojava ovih grešaka tokom snimanja se može predupredi odgovarajućim treningom i blagovremenim upoznavanjem govornika sa rečima koje treba da izgovori i načinom šaputanja.

Drugi tip grešaka su takozvane nekontrolisane greške. Takve su na primer greške u artikulaciji, poput stridensa prikazanog na Slici 5.1 e) i Slici 5.2, koji se javlja kod pojedinih govornika.



Slika 5.2 Primeri različitih manifestacija u artikulaciji: (a) neželjena pojava stridensa u frikativu, (b) više stridensa u afrikatu, (c) trenutak odlepljivanja jezika od nepca. [Marković et al., 2013]

Ovakvi vidovi devijacija artikulacije mogu biti slučajni ili patološki, pri čemu su govornici sa patološkim indicijama bili eliminisani iz eksperimenta. Stridens je specifični oblik rezonanci koji se javlja tokom artikulacije frikativa [Jovičić et al., 2008 b] i prikazan je u primerima na Slici 5.2. U primeru a) je ilustrovan spektrogram šapata sa jako izraženim i stabilnim stridensom koji je uzrokovan zvučnim izgovorom frikativa /ž/, dok je na Slici 5.2 b) prikazano više jakih i nestabilanih stridensa u izgovoru afrikata /č/. Stridens nastaje usled neuobičajenog položaja jezika i nepca i prepoznaje se po jednom ili više jakih i neprijatnih zvukova zvižduka. Druga interesantna pojava u snimanju šapata se odnosi na palatalne glasove poput /lj/ i /č/ i zadnjonepčani glas /k/.

Slika 5.2 c) prikazuje niz veoma kratkih impulsa koji nastaju tokom generisanja pomenutih glasova usled kontakta jezika sa nepcima. Proces odlepljivanja jezika od nepca se ne može kontrolisati voljom i predstavlja karakteristiku pojedinih govornika. Ovakva i slične pojave predstavljaju dodatne informacije o govorniku i njegovim karakteristikama šapata i korisne su za dalju analizu. Naime, u šapatu se, kako u vremenskom tako i u spektralnom domenu, pojavljuju različita fina obeležja koja se tiču načina artikulacije pojedinca. Takve karakteristike su u normalnom govoru u velikoj meri maskirane, dok su u šapatu vidljive i mogu poslužiti u njegovom prepoznavanju, segmentaciji ili grupisanju glasova.

5.5 AKUSTIČKA ANALIZA WHI-SPE CORPUSA

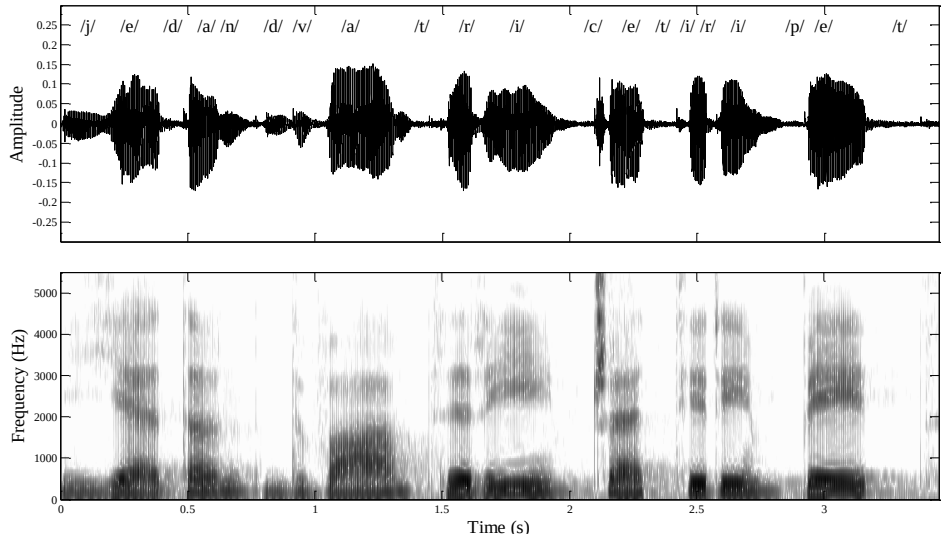
U literaturi se kao specifične karakteristike šapata najčešće spominju: smanjen spektralni nagib, duže trajanje izgovora slogova, izmenjene (podignute) pozicije formanata i šumna struktura spektrograma. U ovoj sekciji je opisan par akustičkih merenja sprovedenih na Whi-Spe korpusu koja potvrđuju ove tvrdnje.

5.5.1 TALASNI OBLICI I SPEKTROGRAMI

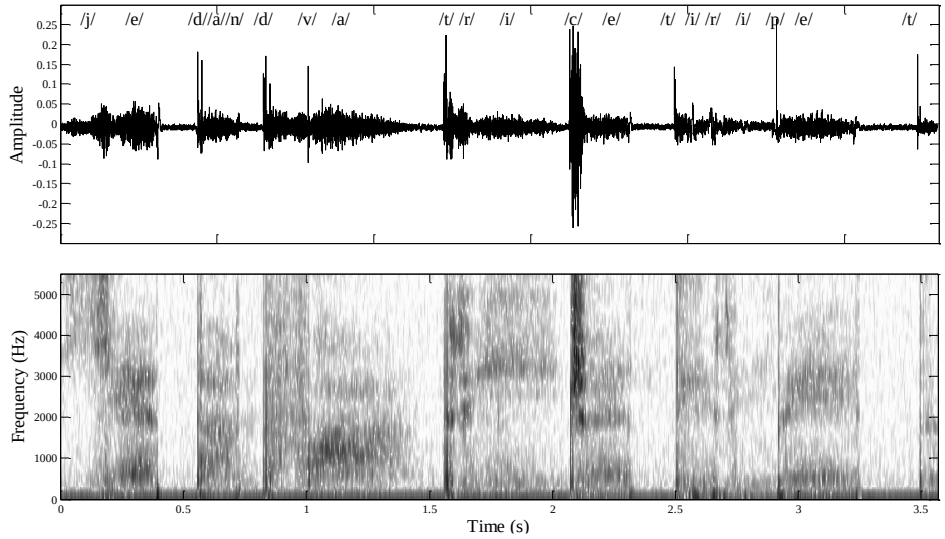
Na Slici 5.3 su prikazani talasni oblici i širokopojasni spektrogrami normalno izgovorene i prošaputane verzije istog niza reči: "jedan", "dva", "tri", "četiri" i "pet". Poredeći sa normalnim govorom, šapat se u vremenskom domenu može opisati kao talasni oblik šumne strukture znatno nižeg intenziteta, sa povremenim naglim impulsima na mestima ploziva. Nedostatak zvučnosti se odražava na amplitude zvučnih glasova koje su u šapatu primetno manje od onih u normalnom govoru, dok su kod bezvučnih fonema sličnog intenziteta kao u govoru. Do istih opservacija su došli Ito u svojoj analizi fonema japanskog jezika [Ito et al., 2005] i Jovičić u studiji šapata u srpskom jeziku [Jovičić, 1998]. Merenjem ukupnog trajanja snimaka Whi-Spe korpusa, ustanovljeno je nešto duže trajanje šapata. Naime, ukupno trajanje svih snimaka normalnog govora je 58 minuta i 53 sekunde, a šapata 1 sat i 36 sekundi. Na osnovu toga možemo zaključiti da je izgovor šapata 2,83% duži od normalnog govora, što u proseku iznosi 20,6ms duže izgovaranje jedne reči[2]. Prema tome, rezulatiti ove

[2] Ovo trajanje je znatno duže u slučaju kontinualnog šaputanja, kod kojeg je za razliku od izolovanih snimaka iz Whi-Spe baze dosta izraženija koartikulacija glasova, pauze prilikom disanja itd., što sve doprinosi trajanju šapata.

vremenske analize Whi-Spe korpusa potvrđuju tvrdnje ranijih istraživanja [Jovičić, 1998; Zhang et al., 2007; Fan et al., 2009].



(a) Normalan govor



(a) Šapat

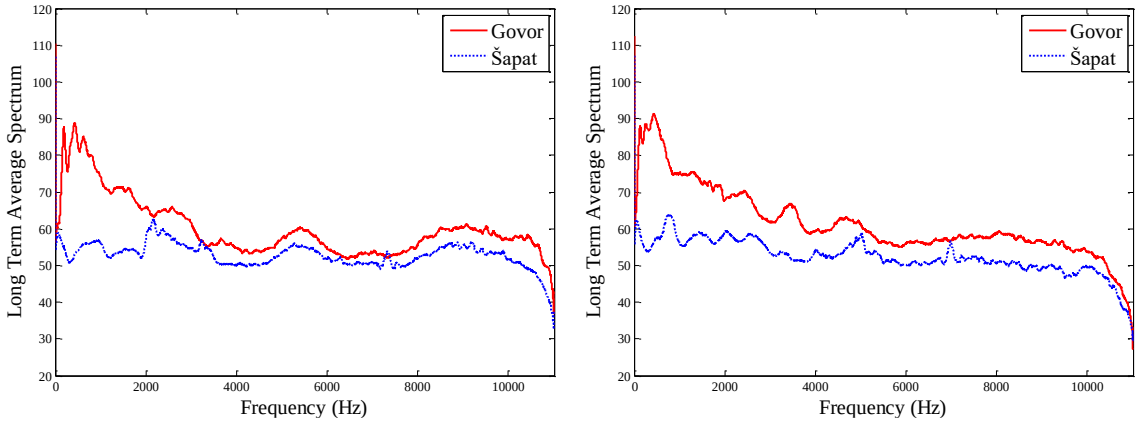
Slika 5.3 Talasni oblci i spektrogrami u (a) normalnom govoru i (b) šapatu.

Analiza širokopojasnih spektrograma ukazuje na još neke razlike između normalnog govora i šapata. Uprkos izmenama u obliku vokalnog trakta i šumnoj strukturi šapata, spektrogrami su očuvali osnovne spektralne karakteristike govora. Spektralni koncentrati, tj. formanti, su energetski oslabljeni ali i dalje vidljivi. Najveće spektralne promene su uočene kod vokala kod kojih su lokacije nižih formanata izdignute ka višim frekvencijama, što je takođe u saglasnosti sa drugim istraživanjima

[Kallail et al., 1984 a; Jovičić, 1998; Ito et al., 2005]. Za razliku od vokala, bezvučni konsonanti nisu toliko izmenjeni u spektralnom domenu [Ito et al., 2005].

5.5.2 SPEKTRALNI NAGIB

Nedostatak zvučnosti u šapatu se može primetiti i u analizi dugovremenih usrednjenih spektara (*Long Term Average Speech Spectrum - LTASS*) u kojima šapat u poređenju sa normalnim govorom ima dosta ravniji spektralni nagib [Jovičić, 1998; Ito et al., 2005; Zhang et al., 2007]. Na Slici 5.4. su upoređeni LTASS Whi-Spe baze u normalnom govoru i šapatu za muške i ženske govornike.



(a) Ženski govornici (b) Muški govornici

Slika 5.4 Dugovremeni usrednjeni spektri Whi-Spe korpusa: (a) ženski govornici (b) muški govornici. Crvena linija predstavlja LTASS normalnog govora, a plava isprekidana linija LTASS šapata. [Grozdić et al., 2017]

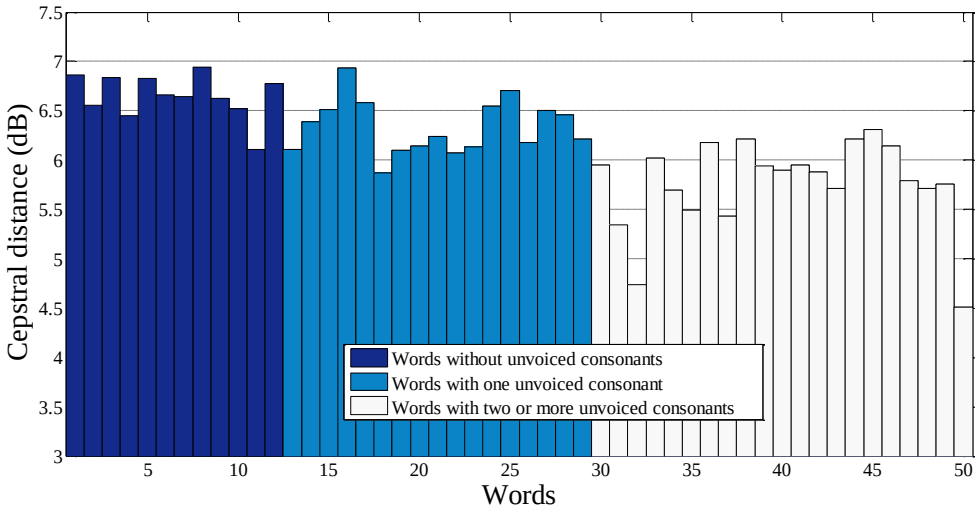
Spektri normalnog govora u poređenju sa šapatom imaju dosta višu energiju na nižim frekvencijama nego na višim, odnosno imaju strmiji spektralni nagib. Ova spektralna karakteristika govora se objašnjava snažnijom glotalnom pobudom u nižim frekvencijskim predelima nego na višim učestanostima. Sa druge strane, šapat zbog šumne i neperiodične pobude ima skoro ravan spektar. Poređenjem dugovremenih usrednjenih spektara muških i ženskih govornika, nije primećena bitnija razlika. Dobijeni rezultati spektralne analize Whi-Spe korpusa su u saglasnosti sa drugim istraživanjima šapata [Eklund et al., 1996; Konno et al., 1996; Jovičić, 1998; Ito et al., 2005; Jovičić et al., 2008].

5.5.3 *KEPSTRALNA ANALIZA*

Za potrebe kvantifikovanja razlike između reči, kao mera akustičke različitosti može se koristiti kepstralna distanca (*Cepstral Distance - CD*) [Gray et al., 1976; Tohkura, 1987; Kitawaki et al., 1988]. U tu svrhu je izvršena kepstralna analiza Whi-Spe korpusa, koja je omogućila merenje razlike između istih reči u govoru i šapatu. Upotrebljene su Mel-frekvencijske kepstralne distance (*Mel-based cepstral distance*), definisane na sledeći način [Côté, 2011]:

$$CD \cong \frac{10}{\ln(10)} \sqrt{2\sum_{i=1}^{N} (c_i^{(n)} - c_i^{(w)})^2}\ , \tag{5.1}$$

gde su $c_i^{(n)}$ i $c_i^{(w)}$ kepstralni koeficijenti (MFCC) u normalnom govoru i šapatu, respektivno, a N=132 je broj kepstralnih koeficijenata po jednoj reči (vidi poglavlje 6). Srednje vrednosti kepstralnih distanci za sve reči iz Whi-Spe baze su izračunate i prikazane na Slici 5.5.



Slika 5.5 Srednja kepstralna distanca između reči u normalnom govoru i šapatu. [Grozdić et al., 2016 a]

U zavisnosti od broja bezvučnih konsonanata, reči iz Whi-Spe korpusa su podeljene u tri grupe (reči bez bezvučnih konsonanata, reči sa jednim bezvučnim konsonantom i reči sa dva ili više bezvučnih konsonanta) i markirane različitim bojama. Sa dijagrama se može videti da najmanje vrednosti kepstralne distance imaju reči sa po dva ili više bezvučnih konsonanata, koja u proseku iznosi 5.7dB, dok reči sa jednim bezvučnim

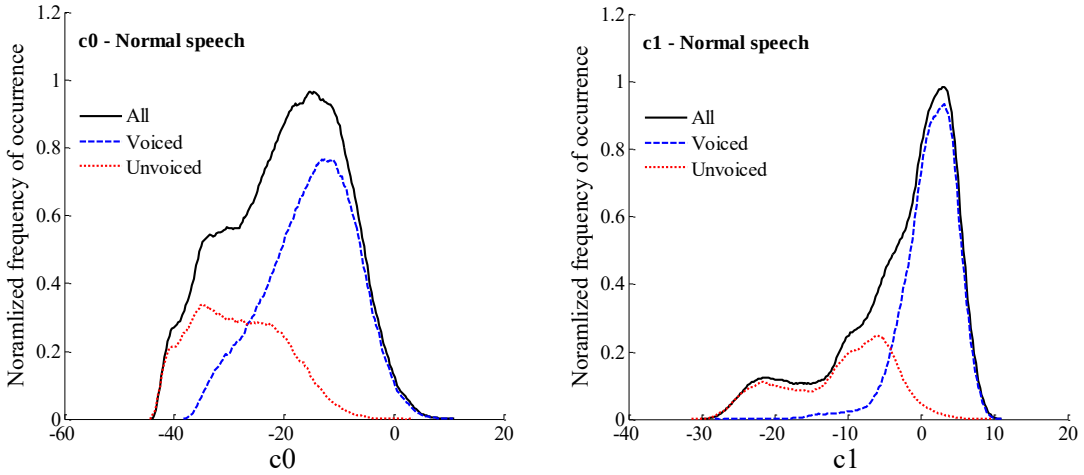
konsonantom imaju nešto veće kepstralne distance od 6.3dB. Reči bez bezvučnih konsonanata imaju najveće kepstralne distance od 6.7dB, odnosno razlike između normalnog govora i šapata su ovom slučaju najveće. Prikazani rezultati potvrđuju da su vokali i zvučni glasovi u šapatu više izmenjeni u nego što je to slučaj sa bezvučnim glasovima [Ito et al., 2005], pa je otuda i manja razlika između govora i šapata kod reči sa većim brojem bezvučnih konsonanata.

Kepstralni koeficijenti se mogu dalje analizirati i porediti u normalnom govoru i šapatu u vidu analize raspodela kepstralnih koeficijenata. Na primer, raspodele prva dva kepstralna koeficijenta, c_0 i c_1, mogu poslužiti za poređenje energije i spektralnog nagiba u rečima izgovorenim u dva različita govorna moda. Naime, c_0 koeficijenti nose informaciju o energiji signala, dok se c_1 odnosi na spektralni nagib signala [Ghaffarzadegan et al., 2014 a]. Na Slici 5.6 su prikazane raspodele c_0 i c_1 koeficijenata u Whi-Spe bazi u normalnom govoru i šapatu. Takođe su analizirane i raspodele njihovih zvučnih i bezvučnih segmenata. Odnos zastupljenosti zvučnih glasova naspram bezvučnih u Whi-Spe bazi u slučaju snimaka normalnog govora je 37.65/62.35%, dok je u šapatu taj odnos 99.2/0.8%. Prema tome, u pogledu zvučnosti snimci šapata u Whi-Spe bazi su zadovoljavajućeg kvaliteta, zahvaljujući sprovedenoj kontroli kvaliteta snimaka koja je odbacila lošije snimke sa zvučnim segmentima, odnosno snimke takozvanog mekog šaputanja (*soft whisper*).

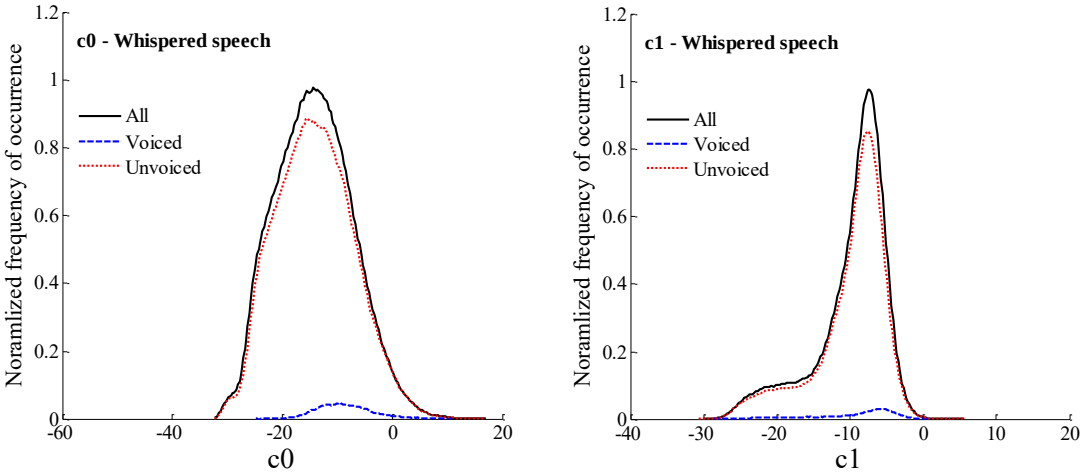
Leva strana Slike 5.6. predstavlja c_0 raspodele na osnovu kojih se mogu tumačiti nivoi energije. U normalnom govoru, c_0 raspodele pokazuju domninantnost zvučnih segmenata koji imaju višu energiju (veće c_0 vrednosti) nego bezvučni govor. U šapatu je obrnuta situacija i bezvučni glasovi su dominantni. Međutim, c_0 raspodele u normalnom govoru i šapatu su centrirane oko sličnih pozicija, što znači da snimci govora i šapata imaju sličnu energiju. Naravno, ovaj rezultat nije slučajnost već je posledica blisko primaknutog mikrofona tokom snimanja šapata.

Desna strana Slike 5.5 se odnosi na raspodele c_1 koeficijenata koje nose informacije o spektralnom nagibu. U normalnom govoru, zvučni glasovi zauzimaju više c_1 vrednosti, odnosno imaju strmije nagibe spektra. Sledeći intuiciju i položaj raspodela bezvučnih segmenta oko nižih c_1 vrednosti, zaključuje se da oni imaju manje spektralne nagibe i ravniji spektar. U šapatu su sve raspodele pomerene u levo, ka nižim c_1

vrednostima, što još jednom dokazuje prethodne tvrdnje da je spektar šapata dosta ravniji u poređenju sa normalnim govorom. Za razliku od c_0 raspodela, u slučaju c_1 raspodela postoji vidljiva razlika u njihovim položajima u normalnom govoru i šapatu.



(a) Raspodele c_0 i c_1 koeficijenata u normalnom govoru.



(b) Raspodele c_0 i c_1 koeficijenata u šapatu.

Slika 5.6 Normalizovane c_0 i c_1 raspodele u normalnom govoru i šapatu. [Grozdić et al., 2016 a]

5.6 PSEUDO-ŠAPAT

Pseudo-šapat predstavlja veštački generisan šapat dobijen pomoću inverznog filtriranja uzoraka normalnog govora i dodavanja Gausovog šuma. Ovakav tip govornog signala i inverzno filtriranje je korišćeno u dva različita eksperimenta ove doktorske teze. Prvo, inverzno filtriranje je poslužilo kao metoda smanjenja spektralnih razlika između govora i šapata koja je upotrebljena zarad dokaza hipoteze o uzroku razlike u rezultatima prepoznavanja reči u neusaglašenim obuka/test scenarijima sa MLP sistemima (Poglavlje 9.6). Drugo, pseudo-šapat je upotrebljen za obuku dubinskog

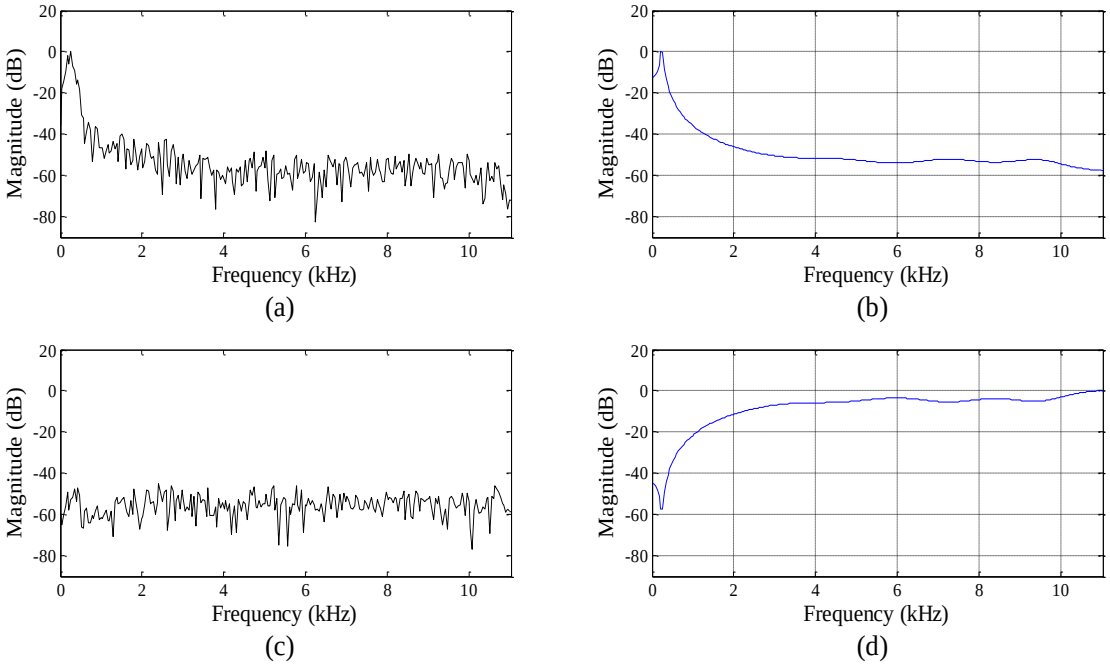
autoenkodera čiji je zadatak ekstrakcija robustnih govornih obeležja (Poglavlje 7.2). U nastavku su opisani inverzni filtar, akustička anliza inverznog filtriranja i kreiranje baze pseudo-šapata.

5.6.1 INVERZNI FILTRIRANJE

Inverzno filtriranje je primenjeno na svakom govornom stimulusu (uzorku) iz Whi-Spe baze. Inverzni filtar (IF) koji je korišćen u te svrhe se matematički može opisati sledećom formulom:

$$\mathrm{IF}(z) = \frac{1}{\mathrm{H}(z)} = 1 - \sum_{i=1}^{p} a_i z^{-i}\,, \tag{5.1}$$

gde je H(z) prenosna funkcija vokalnog trakta, a_i su LPC koeficijenti govornog stimulusa[3], a p = 10 je red LPC filtra. Na osnovu (5.1) formule, zaključuje se de je inverzni filtar, IF(z), ustvari filtar predikcije u LPC analizi [Linggard, 1985]. Zbog svog recipročnog odnosa sa *all-pole* filtrom, H(z), frekvencijski odziv inverznog filtra je obrnut anvelopi LPC spektra, kao što je ilustrovano na Slici 5.7 d).



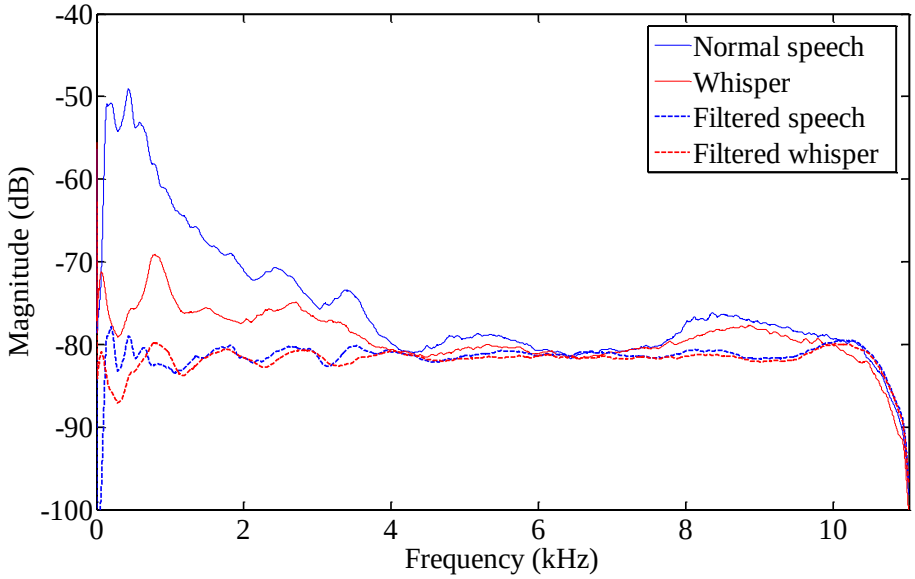
[3] U ovom slučaju pod govornim stimulusom se podrazumeva nesegmentiran govorni signal.

Slika 5.7 Primer inverznog filtriranja na reči u normalnom govoru: (a) FFT spektar reči, (b) LPC anvelopa spektra, (c) FFT spektar posle inverznog filtriranja, i (d) frekvencijski odziv inverznog filtra IF(z). [Grozdić et al., 2016 a]

Zahvaljujući ovoj karakteristici, inverzni filtar omogućava poravnanje spektra svake reči iz Whi-Spe baze, pri čemu sa porastom reda filtra raste i stepen izravnavanja spektra. Kako se ovakvim filtriranjem ne bi izgubile informacije o formantima, odnosno kako se spektar govora ne bi suviše "ispeglao", upotrebljen je filtar desetog reda koji grubo aproksimira anvelopu spektra kao što je prikazano na Slici 5.7 b).

5.6.2 AKUSTIČKA ANALIZA NAKON INVERZNOG FILTRIRANJA

Finalni rezultati inverznog filtriranja stimulusa iz Whi-Spe baze su ilustrovani u vidu dugovremenih usrednjenih spektara (LTASS) na Slici 5.8. Poredeći izglede LTASS pre i posle inverznog filtriranja, spektri govora i šapata su sada izravnati i dosta sličniji.

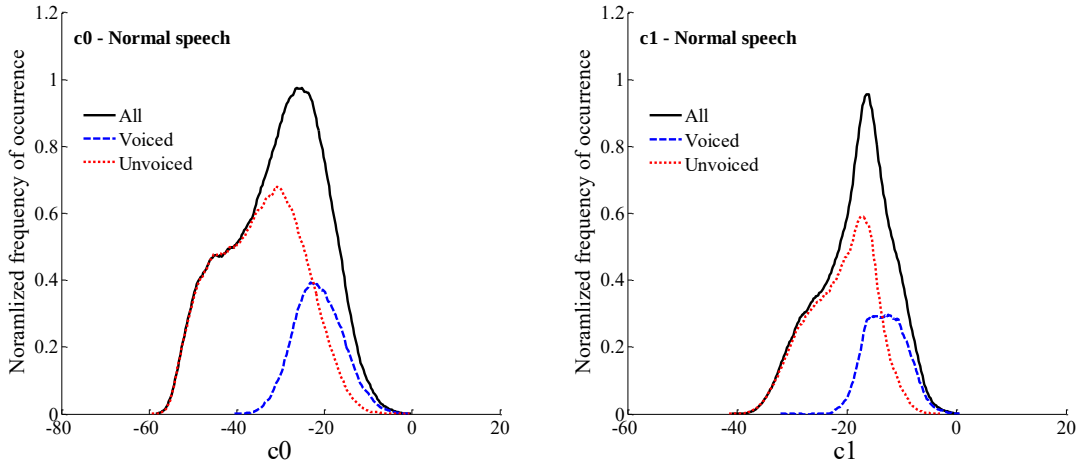


Slika 5.8 LTASS Whi-Spe baze, pre i posle inverznog filtriranja. [Grozdić et al., 2016 a]

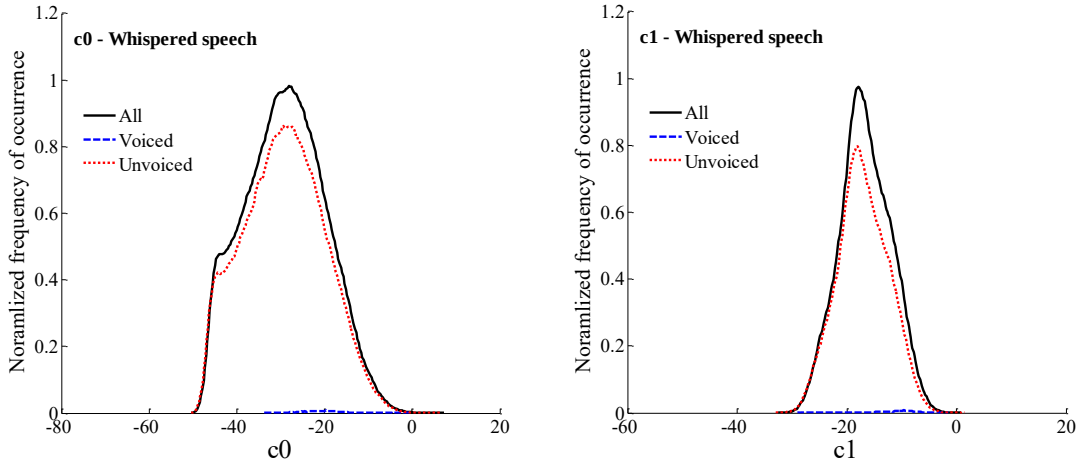
Da bi se dokazalo da se ovako postignuta spektralna sličnost govora i šapata odrazila i u kepstralnom domenu, ponovo su izračunate kepstralne distance. Srednja kepstralna distanca između reči u normlanom govoru i šapatu pre inverznog filtriranja je bila 6.19dB, dok je posle inverznog filtriranja prilično smanjena i iznosi 3.45dB.

Promene u energiji i spektralnom nagibu stimulusa posle inverznog filtriranja se mogu ustanoviti i kroz nešto detaljniju analizu raspodela c_0 i c_1 kepstralnih

koeficijenata, prikazanih na Slici 5.9. U poređenju sa raspodelama c_0 koeficijenata pre filtriranja (Slika 5.6) nove raspodele c_0 koeficijenata su pomerene ka na nešto nižim vrednostima što ukazuje na smanjene nivoa energije. Ovaj pad energije, koji je prisutan u stimulusima oba govorna moda, je posledica potiskivanja pre svega spektralnih komponenti do 5 kHz tokom inverznog filtriranja. I pored ovih izmena, raspodele c_0 koeficijenata su i dalje centrirane oko istih poziciija jedne iznad drugih, što govori da su snimci govora i šapata ostali sličnih energija.



(a) Raspodele c_0 i c_1 koeficijenata u normalnom govoru.



(b) Raspodele c_0 i c_1 koeficijenata u šapatu.

Slika 5.9 Normalizovane c_0 i c_1 raspodele u normalnom govoru i šapatu posle inverznog filtriranja. [Grozdić et al., 2016 a]

Sa desne strane Slike 5.9 prikazane su raspodele c_1 koeficijenata koje su takođe međusobno centrirane oko iste pozicije. To dokazuje da je sa inverznim filtriranjem spektralni nagib u normalnom govoru smanjen i da se sada poklapa sa nagibom spektra

šapata. Postignuti pomeraj raspodela c_1 koeficijenata u normalnom govoru ka nižim c_1 vrednostima (ravniji spektar) je očigledan prilikom poređenja Slike 5.6 i Slike 5.9. Raspodele zvučnih i bezvučnih segmenata Whi-Spe baze, takođe demonstriraju postignuto potiskivanje zvučnosti i ostvarenu dominaciju bezvučnih segmenata u normalnom govoru.

5.6.3 KREIRANJE BAZE PSEUDO-ŠAPATA

Pseudo-šapat se generiše potiskivanjem zvučnosti iz normalnog govora putem inverznog filtriranja i dodavanjem belog Gausovog šuma zarad dodatnog isticanja šumne strukture šapata. Ovako generisan pseudo-šapat iz snimaka normalnog je po svojim akustičkim karakteristikama približan prirodnom šapatu. Koristeći inverzni filtar (opisan u tački 5.6.1) uz odgovarajuće zašumljivanje sa belim šumom do 10dB SNR, od 5000 snimaka normalnog govora iz Whi-Spe korpusa je na veoma brz način kreirana baza od 5000 snimaka pseudo-šapata. Kreirana baza pseudo-šapata je korišćena u postupku obuke dubinskog *denoising* autoenkodera o čemu će biti više reči u Poglavlju 7.2.

5.7 REZIME

U ovom poglavlju je opisano kreiranje Whi-Spe korpusa, jednog od par postojećih kvalitetno snimljenih i sistematski uređenih korpusa šapata prikazanih u Tabeli 5.1. Whi-Spe baza sadrži 10000 snimaka izgovora izolovanih reči u normalnom govoru i šapatu, što je čini korpusom srednje veličine i jedinom govornom bazom tog tipa u srpskom jeziku. Whi-Spe je do sada eksploatisan u istraživanjima akustičkih karakteristika šapata kao i u eksperimentima automatskog prepoznavanja šapata, koji čine osnovu ove doktorske disertacije.

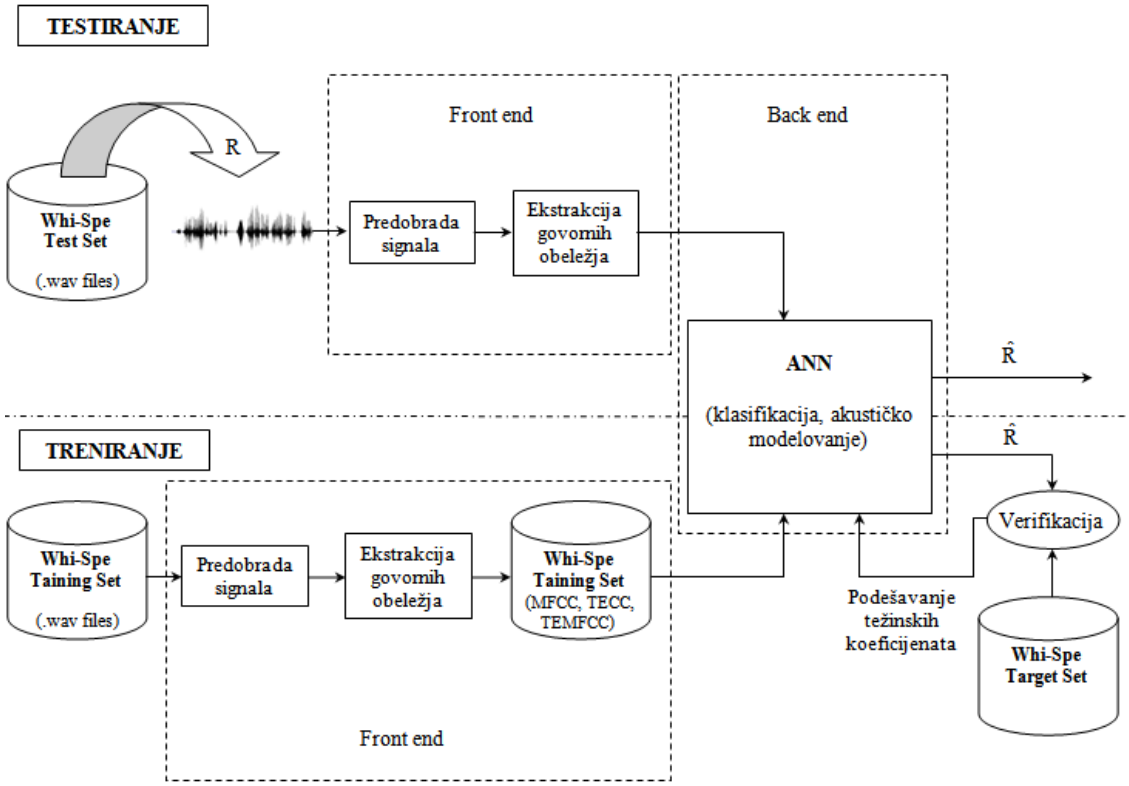
Snimanje govorne baze je obavljeno u laboratorijskim uslovima, sa korišćenjem profesionalne opreme. U snimanju su učestvovali govornici oba pola, kao i dvojica eksperta iz oblasti digitalne obrade govornih signala i lingvistike, koji su bili odgovorni za kontrolu kvaliteta snimaka. Snimanje govorne baze je vršeno u par sesija, pri čemu su sve uočene greške, najčešće u vidu specifičnih manifestacija tokom snimanja šapata, korigovane kroz nakandna dosnimavanja.

Akustička analiza Whi-Spe baze je poslužila kao provera karakteristika šapata u srpskom jeziku i demonstracija osnovnih akustičkih razlika između govora i šapata. Najistaknutije razlike šapata u odnosu na govor se ogledaju u njegovoj šumnoj i dobrim delom aperiodičnoj strukturi, niskoj energiji, drugačijem položaju formanata u vokalima i veoma blagom, skoro ravnom spektralnom nagibu. Nabrojane razlike čine šapat veoma problematičnim u pogledu atomatskog prepoznavanja pomoću tradicionalnih ASR sistema, koji su prevashodno dizajnirani za prepoznavanje normalnog govora. Neke od tih razlika se mogu ublažiti, poput razlika u nivoima energije, koji se kao što je demonstrirano u ovom poglavlju mogu smanjiti pri dobrim SNR uslovima sa približavanjem mikrofona u šaputanju. Ipak, ovako jednostavne metode ne smanjuju spektralne nejednakosti govora i šapata, koje poput razlike u spektralnom nagibu predstavljaju bitan problem, i zahtevaju njihovu odgovarajuću modifikaciju kako bi se prilagodile automatskom prepoznavanju šapata. Sa tim ciljem, u sklopu akustičke analize razlika normalnog govora i šapata, u ovom poglavlju je analiziran i postupak inverznog filtriranja kojim se pre svega utiče na smanjenje spektralnog nagiba i potiskivanje zvučnosti u normalnom govoru. Efekti inverznog filtriranja su demonstritani i kroz analizu u kepstralnom domenu. Inverzno filtriranje ujedno predstavlja i bitnu tehniku za generisanje pseudo-šapata, što je iskorišćeno za kreiranje posebne baze pseudo-šapata koja je kasnije korišćena u nekim eksperimentima ovog doktorata.

6 KREIRANJE MLP SISTEMA ZA PREPOZNAVANJE ŠAPATA

U ovom poglavlju je sistematski opisan postupak kreiranja sistema za automatsko prepoznavanje šapata baziranog na posebnom tipu veštačkih neuralnih mreža, takozvanim višeslojnim perceptronima (MLP). MLP sistem, šematski prikazan na Slici 6.1, je realizovan u MATLAB programskom jeziku i sastoji se iz ulaznog dela, takozvanog *front-end* sistema, u kome se vrši predobrada govornih signala i ekstrakcija govornih obeležja (*speech features*), i iz zadnjeg dela, poznatog kao *back-end* sistem, u kome se pomoću MLP obavlja prepoznavanje reči odnosno klasifikacija obrazaca. Procedura predobrade signala je primenjena na svakoj reči iz Whi-Spe baze i podrazumeva njihovu segmentaciju, vremensko usklađivanje, prozorovanje i filtriranje, posle čega se iz tako obrađenih govornih signala obavljala ekstrakcija tri različita tipa kepstralnih koeficijenata: MFCC, TECC (*Teager Energy Cepstral Coefficients*) i TEMFCC (*Teager Energy based Mel-Frequency Cepstral Coefficients*). Matematički postupak izračunavanja svakog od navedenih kepstralnih koeficijenata kao i njihove međusobne razlike su detaljno obrazloženi u tački 6.2.2 ovog podpoglavlja. Izdvojena govorna obeležja su dalje korišćena u postupku treniranja i simulacije MLP, zašta je bila neophodna podela Whi-Spe korpusa i formiranje posebnih baza podataka za obuku,

validaciju i testiranje neuralne mreže. Ostatak poglavlja je posvećen kreiranju ANN, pronalasku optimalne MLP arhitekture i opisu postupka obučavanja neuralne mreže.



Slika 6.1 Blok šema sistema baziranog na MLP za automatsko prepoznavanje izolovanih reči iz Whi-Spe baze.

6.1 PREDOBRADA GOVORNIH SIGNALA

Faza predobrade govornih signala je ista za sve ulazne stimuluse i uključuje: segmentaciju govornih signala na okvire, njihovo vremensko usaglašavanje, primenu preemfazisa i množenje sa Hamingovom prozorskom funkcijom.

6.1.1 SEGMENTACIJA I VREMENSKO USKLAĐIVANJE GOVORNIH SIGNALA

Ustaljeni pristup u digitalnoj obradi govornih signala je zasnovan na analizi kratkovremenih segmenata govora (*short-time analysis*), prosečne dužine 10-40 ms. Iako se smatra da je govorni signal u ovako kratkim segmentima sporopromenljiv, praksa ukazuje na pojedine izuzetke poput izgovora konsonanata na kraju reči, kod kojih se pojavljuju oštri spektralni prelazi i brzi pomeraji spektralnih pikova za čak 80 Hz/ms [Markel et al., 1976]. Brzina promene spektralnih karakteristika govora zavise i od brzine artikulacije, pa je u slučaju šapata ona nešto sporija. Ovi faktori određuju

odabir dužine a samim tim i broja vremenskih okvira koji se uzimaju u procesu segmentacije. U radu sa veštačkim neuralnim mrežama postoje dodatna ograničenja u pogledu segmentacije reči na vremenske okvire. Ulazni sloj neurona neuralne mreže sadrži fiksan broj čvorova, što znači da ona može primati samo podatke fiksne dužine. Pošto su reči u Whi-Spe bazi različitog trajanja, potrebno je da se govorni signali podele na fiksan broj vremenskih okvira (segmenata) iz kojih će se dalje obaviti ekstrakcija jednakog broja govornih obeležja. Za ovakve potrebe veštačkih neuralnih mreža postoje dve definisane metode segmentacije reči u kojima se dobija fiksan broj okvira [Pan et al., 2007]:

1) U prvoj metodi se koriste vremenski okviri promenljive dužine (*dynamic numbers of sample points*);

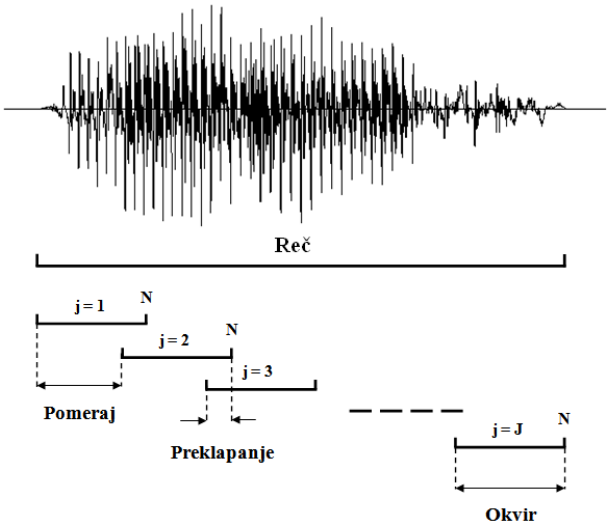
2) Druga metoda koristi vremenske okvire fiksne dužine sa promenljivim međusobnim preklapanjem (*dynamic frame overlap rates*).

U ovoj tezi su testirane obe metode segmentacije. Prvo je određen broj vremenskih okvira, J, sa kojim su reči segmentirane. Taj broj je određen na osnovu statističke analize prosečnog broja fonema u rečima. Za reči iz Whi-Spe korpusa prosečni broj fonema je između 3 i 9 (sa izuzetkom dve reči koje imaju 12 i 13 fonema). Kako bi u proseku kod dužih reči na raspolaganju postojao jedan okvir po fonemu, reči su segmentirane na J = 11 okvira, pri čemu je kod kraćih reči broj vremenskih okvira po fonemu veći. Pretpostavljeno je da ovako finija vremenska rezolucija kod kraćih reči omogućava njihovo bolje prepoznavanje, dok se kod dužih reči ovakav tip dobitka ostvaruje na račun njihovog bogatijeg fonetskog sadržaja.

U slučaju prve metode, svaka reč iz Whi-Spe korpusa je segmentirana po svojoj dužini sa J = 11 vremenskih okvira koji se međusobno preklapaju 50%, Slika 6.2, iz kojih je kasnije vršena ekstrakcija govornih obeležja. Na ovaj način je postignuta vremenska normalizacija različitog trajanja izgovora reči, pri čemu je dužina okvira srazmerna trajanju reči.

Drugi metod segmentacije podrazumeva korišćenje J = 11 vremenskih okvira iste dužine, pri čemu je njihovo preklapanje zavisno od trajanja svake pojedine reči. Dužina preklapanja susednih okvira je obrnuto srazmerna trajanju analizirane reči,

odnosno, u slučaju kratkih reči preklapanje je veće, dok je kod dugačkih reči praklapanje manje. Obe metode su pokazale slične rezultate u preliminarnim eksperimentima, a u nastavku istraživanja je korišćen samo prvi metod.



Slika 6.2 Primer segmentacija reči na okvire sa međusobnim preklapanjem.

Pored ove dve metode, za potrebe vremenskog usklađivanja govornih signala i svođenja njihovog trajanja na istu dužinu, testirane su i druge metode poput DTW-FF (*DTW-Frame Fixing*) i PCA (*Principal Component Anlaysis*). Međutim, ove metode su usled izbacivanja reudundantnih podataka iz govornih signala pokazale dosta slabije rezultate[4] u neusaglašenim obuka/test scenarijima, pa iz tog razloga nisu krišćene u kasnijim eksperimentima.

6.1.2 PREEMFAZIS I PROZOROVANJE GOVORNIH SIGNALA

Uobičajna praksa u predobradi govornih signala je primena preemfazis filtra koji poravnava spektar i vrši balansiranje energije viših i nižih spektralnih komponenti. Zvučni delovi govora po prirodi imaju veću energiju na nižim učestanostima i strmiji spektralni nagib ka višim frekvencijama, koji je obično reda 20 dB po dekadi[5] [Markel et al., 1976]. Preemfazis filtar služi da neutrališe ovaj nagib i smanji veliki dinamički opseg govornog signala pre sprovođenja spektralne analize čime se poboljšava njena efikasnost. Sa druge strane, čulo sluha je dosta osetljivije na frekvencijama iznad 1 kHz,

[4] Izbacivanje redundantnih informacija iz govornih signala nije umanjila uspeh u prepoznavanju reči u usaglašenim obuka/test scenarijima. Ipak, u neusaglašenim scenarijima primena PCA ili DTW-FF metode kompresije je rezultovala drastičnim smanjenjem uspeha prepoznavanja reči koje nije prelazilo 10%.
[5] Otprilike 6 dB po oktavi.

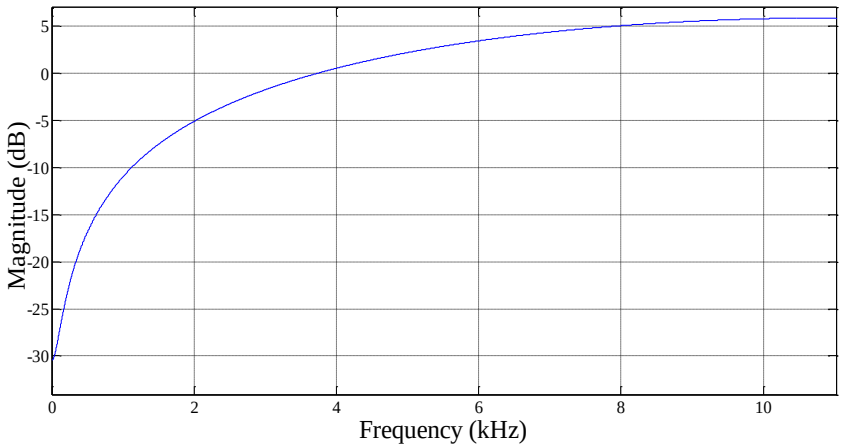
a preemfazis filtar pojačava upravo ovu oblast spektra, Slika 6.3. Mera u kojoj preemfazis pojačava energiju na višim frekvencijama se najčešće definiše u vidu koeficijenta λ = [0,9 - 1,0] FIR filtra, zapisanog na sledeći način:

$$H(z) = 1 - \lambda z^{-1}. \tag{6.1}$$

U vremenskom domenu izlaz ovakvog preemfazis filtra, $s_{out}(n)$, se može opisati sa:

$$s_{out} = s_{in}(n) - \lambda s_{in}(n-1), \tag{6.2}$$

gde su $s_{in}(n)$ odbirci ulaznog signala. U predstojećim eksperimetnima ove teze je korišćen preemfazis filtar sa koeficijentom λ = 0,97 i frekvencijskim odzivom prikazanim na Slici 6.3.

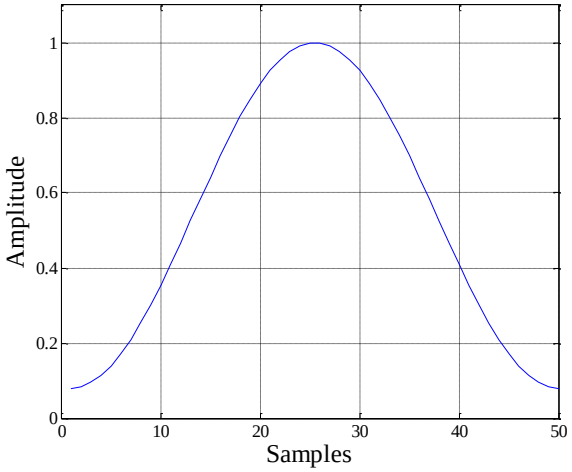


Slika 6.3 Frekvencijski odziv preemfazis filtra.

Kako bi se izbegla pojava diskontinuiteta između susednih okvira i pojava curenja spektra, vremenski okviri su posle preemfazisa pomnoženi Hamingovom prozorskom funkcijom:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \tag{6.3}$$

gde N predstavlja dužinu prozora u odbircima, Slika 6.4.



Slika 6.4 *Hamming* prozorska funkcija.

6.2 EKSTRAKCIJA GOVORNIH OBELEŽJA

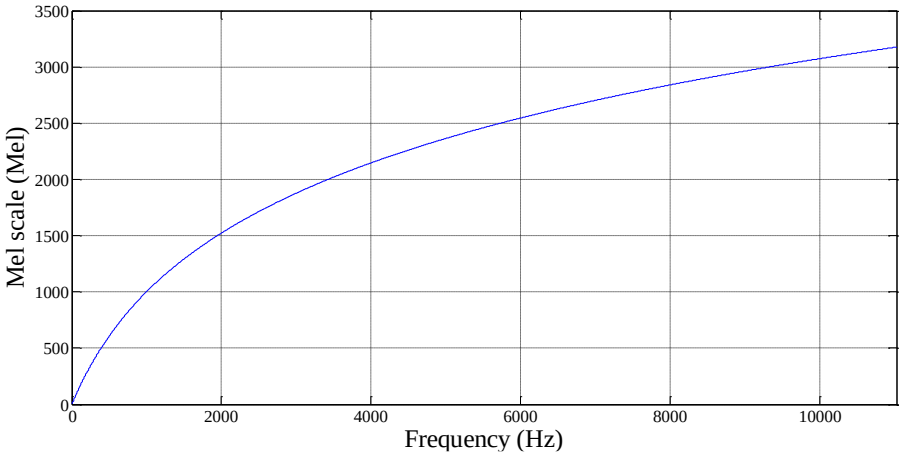
Prvi korak posle procedure predobrade govornih signala je ekstrakcija govornih obeležja. Cilj ekstrakcije govornih obeležja je transformisanje ulaznih govornih signala u sekvencu vektora akustičkih obeležja, koja nosi relevantne informacije bitne za njihovu diskriminaciju i automatsko prepoznavanje. Ekstrakcijom govornih obeležja i formiranjem odgovarajućih akustičkih obrazaca redukuju se neželjene varijabilnosti i dimenzije signala, što omogućava njihovo efikasnije procesiranje i klasifikaciju u *back-end* sistemu. Za potrebe ASR sistema kao govorna obeležja najčešće su u upotrebi kepstralni koeficijenti, od kojih su u ovom istraživanju testirana tri tipa: MFCC, TECC i TEMFCC.

6.2.1 MEL-FREKVENCIJSKI KEPSTRALNI KOEFICIJENTI (MFCC)

Mel-frekvencijski kepstralni koeficijenti (MFCC) su trenutno najdominantnija kepstralna obeležja koja se koriste u obradi govornih signala, pre svega u zadacima identifikacije, klasifikacije i prepoznavanja. MFCC su dizajnirani sa idejom da oponašaju čovekov slušni mehanizam, za koji je poznato da procesira govorne signale na nelinearan način. Melova skala [Stevens et al., 1937], Slika 6.5, je zasnovana upravo na ovoj činjenici da frekvencijski odziv čovekovog uha na pobudu zvučnog signala nije linearna funkcija frekvencije. Taj odziv se može modelovati Melovom skalom, kod koje je razmak između frekvencija iznad 1000 Hz logaritamski. Odnos između Melove skale i frekvencije u Hz se može formulisati jednačinom [O'Shaughnessy, 1987] (6.4) i ilustrovati Slikom 6.5:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700Hz}\right). \tag{6.4}$$

Zbog svojih karakteristika i mogućnosti kompaktnog reprezentovanja relevantnih informacija govornih signala, Mel-frekvencijski kepstralni koeficijenti igraju važnu ulogu u automatskom prepoznavanju govora.



Slika 6.5 Izgled Melove skale.

Postupak ekstrakcije Mel-frekvencijskih kepstralnih koeficijenta iz govornog signala je šematski prikazan na Slici 6.7.

Za svaki vremenski okvir govornog signala koji je prošao fazu predobrade, procedura izračunavanja MFCC se odvija u nekoliko koraka:

1) Izračunava se FFT (*Fast Fourier Transform*) vremenskog okvira:

$$S[n, w_k] = \sum_{m=-\infty}^{+\infty} s[m] w[n-m] e^{-jw_k m}, \tag{6.5}$$

gde je $S[n,w_k]$ spektar govornog segmenta $s[n]$, $w_k = 2\pi k/N$, a N je dužina FFT.

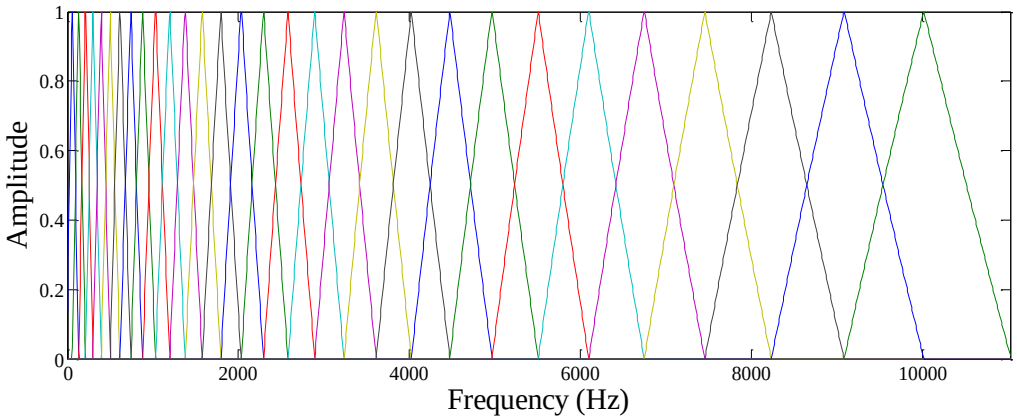
2) Određuje se spektralna gustina snage, $P[n,w_k]$, za tako izračunat spektar:

$$P[n, w_k] = \frac{1}{N} \left| S[n, w_k] \right|^2. \tag{6.6}$$

3) Primenjuje se trougaona Mel-frekvencijska banka filtara, $M_l[w_k]$, (Slika 6.6) na spektralnu gustinu snage:

$$e[n,l] = \sum_{k=L_l}^{U_l} \left| M_l[w_k] P[n, w_k] \right|, \tag{6.7}$$

gde je L = 30 ukupan broj filtara, l = 1, 2 ... L, a L_l i U_l su pojedinačne donje i gornje granične frekvencije svakog od filtra.



Slika 6.6 Melova trougaona banka filtara, predstavljena na linearnoj skali u Hz.

4) Logaritmuje se energija na izlazu svakog filtra.

5) Na logaritmovane energije se primenjuje se DCT (*Discrete Cosine Transform*) čime se dobija Mel-frekvencijski kepstar MFC (*Mel Frequency Cepstrum*), iz koga se izdvaja samo prvih 12 kepstralnih koeficijenata:

$$\mathrm{MFCC}[n, w_k] = \sum_{l=1}^{L} \log(e[n,l]) \cos\left(\frac{k\pi(l-0.5)}{L} \right), \tag{6.8}$$

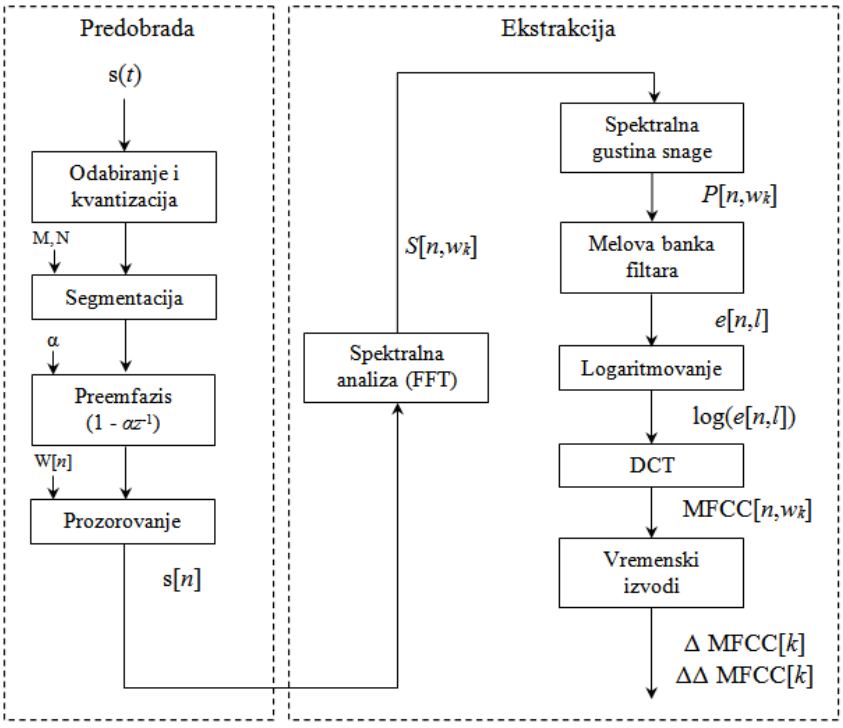
gde su indeksi koeficijenata k = 1, 2 ... N_c, a N_c = 12 maksimalni broj kepstralnih koeficijenata.

6) U cilju boljeg opisivanja dinamike signala izračunavaju se dinamička obeležja[6] u vidu prvog (*delta*) i drugog (*delta-delta*) vremenskog izvoda Mel-frekvencijskih kepstralnih koeficijenata:

$$\Delta\mathrm{MFCC_i}[k] = \frac{\sum_{m=-M}^{M} m\mathrm{MFCC_i}[k+m]}{\sum_{m=-M}^{M} m^2}, \tag{6.9}$$

$$\Delta\Delta\mathrm{MFCC_i}[k] = \frac{\sum_{m=-M}^{M} m\Delta\mathrm{MFCC_i}[k+m]}{\sum_{m=-M}^{M} m^2}, \tag{6.10}$$

gde su $\Delta\mathrm{MFCC}_i[k]$ i $\Delta\Delta\mathrm{MFCC}_i[k]$ *k*-ti *delta* i *delta-delta* koeficijenti za *i*-ti okvir, a za njihovo izračunavanje je korišćeno $M = 4$ (u slučaju $\Delta\Delta$MFCC obeležja koristi se $M = 1$). Gore opisani način izračunavanja dinamičkih obeležja predstavlja aprkosimaciju prvih i drugih vremenskih izvoda, dobijenih pomoću metode polinomskog fitovanja (*polynomial fitting*).



Slika 6.7 Procedura ekstrakcije MFCC obeležja i njenih prvih i drugih izvoda.

[6] Prema analogiji sa načinom izračunavanja brzine u kinetici, *delta* obeležja predstavljaju brzinu promene avelope spektra govornog signala. Slično, *delta-delta* obeležja predstavljaju ubrzanje, pa se zato u literaturi često javljaju pod terminom *acceleration coefficients*.

6.2.2 TEAGER-ENERGETSKA OBELEŽJA

Druga dva tipa kepstralnih koeficijenata, TEMFCC i TECC, su karakteristična po specifičnom načinu izračunavanja energije poznatom kao *Teager* energija.

6.2.2.1 TEAGER ENERGIJA

Obeležja poput LPCC (*Linear Prediction Cepstral Coeffcicients - LPCC*) i MFCC su zasnovana na modelima linearne produkcije govora u kojima je protok vazdušne struje duž vokalnog trakta predstavljen u vidu propagacije linearnog ravanskog talasa. U realnosti vazdušno strujanje nije tako jednostavno, linearno i periodično. Duž vokalnog trakta se javlja više vrtložnih struja koje međusobno interaguju i daju protoku vazduha vrtložnu i nelinearnu formu. Sa ciljem boljeg modelovanja izvora zvučne pobude u generisanju govora, osmišljen je novi model zasnovan na drugačijem izračunavanju energije vazdušnog strujanja. Predložen je jednostavan algoritam koji koristi operator za praćenje nelinearne energije, pod nazivom TEO (*Teager Energy Operator - TEO*) [Teager, 1980; Teager, 1989].

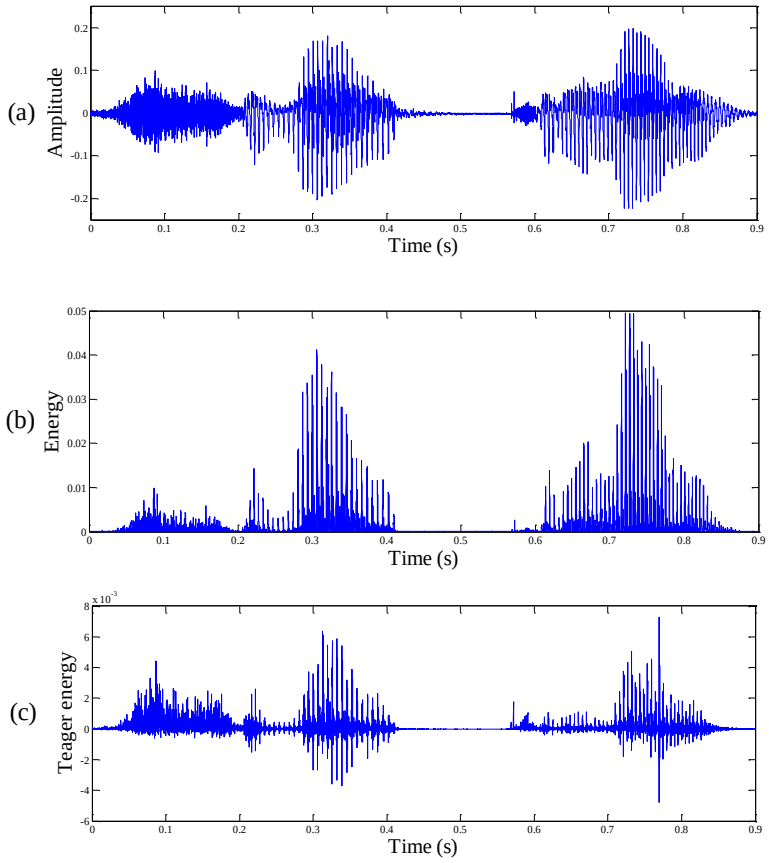
Prema konvencionalnom načinu, energija se računa kao suma kvadriranih amplituda signala (*Standard-Energy Operator,* SEO - $x(t)^2$). Vodeći se ovim načelom, može se zaključiti da su energije tonova od 10 Hz i 1000 Hz jednake. Ipak, za generisanje signala sa frekvencijom od 1000 Hz je potrebno mnogo više energije nego za generisanje signala od 10 Hz [Holambe et al., 2012]. Iz tog razloga, sa željom da se opiše trenutna energija nelinearnih vrtložnih interakcija vazdušnih struja je razvijen TEO [Teager, 1980; Kaiser, 1983]. Za razliku od SEO koji izračunava samo kinetičku komponentu energije signala, *Teager*-energetski operator izračunava "pravu" ukupnu energiju zvučnog izvora koja pored amplitudskih sadrži i frekvencijske informacije [Dimitriadis et al., 2005]. Drugim rečima, totalna energija izvora, odnosno suma potencijalne i kinetičke energije, je proprocijalna kvadratu proizvoda trenutne amplitude i frekvencije, što predstavlja ništa drugo do realnu energiju potrebnu za generisanje signala [Maragos et al., 1993]. Ove dodatne informacije mogu poboljšati vremensko-frekvencijski opis brzih energetskih promena tokom glotalnog ciklusa i reprezentaciju formantnih informacija u govornim vektorima [Kaiser, 1983]. U slučaju realnih kontinualnih vremenskih signala TEO se definiše na sledeći način:

$$\Psi\left(x(t)\right)=\left(\frac{\mathrm{d}}{\mathrm{dt}}x(t)\right)^2-x(t)\frac{d^2}{\mathrm{dt}^2}x(t)=\left(\dot{x}(t)\right)^2-x(t)\ddot{x}(t)\,. \tag{6.11}$$

Ovaj izraz se može diskretizovati, odnosno TEO se može zapisati u diskretnom obliku:

$$\Psi\left(x[n]\right)=\left(x[n]\right)^2-x[n-1]x[n+1]. \tag{6.12}$$

odakle se vidi da je njegovo izračunavanje relativno brzo i jednostavno i da zahteva samo tri susedna odbirka za izračunavanje energije u jednom vremenskom trenutku. Ova osobina dobre vremenske rezolucije omogućava da TEO "uhvati" brze energetske fuluktuacije tokom glotalnog ciklusa. Još jedna karakteristika *Teager* energije je da pored pozitivnih vrednosti ona može imati i negativne vrednosti, Slika 6.8 [Kaiser, 1983].



Slika 6.8 Prikaz: (a) talasni oblik govornog signala, (b) energija govornog signala, (c) *Teager* energija govornog signala.

U slučaju kompleksnih diskretnih signala, *Teager* energija ima sledeći oblik:

$$\Phi\left(x[n]\right)=\Psi\left(\{\Re(x[n]\}\right)+\Psi\left(\{\Im(x[n]\}\right), \tag{6.13}$$

gde je TEO predstavljen u vidu zbira realnog i imaginarnog dela energije signala, što daje uvid u realno stanje energije i odlika je dobrih operatora [Kaiser, 1983]. TEO takođe ima karakteristike filtra zbog čega se koristi u tehnikama potiskivanja šuma [Dimitriadis et al., 2005; Holambe et al., 2012].

Zbog dobrih karakteristika, *Teager* energija je našla primenu u raznim novim govornim obeležjima koja po svojim performansama i robustnosti prevazilaze tradicionalna MFCC i još starija LPCC obeležja. U ovom istraživanju su pored MFCC testirana još dva tipa kepstralnih koeficijenta, TEMFCC i TECC, koji su relativno novog datuma i bazirana na *Teager* energiji. S obzirom na mogućnosti boljeg modelovanja protoka vazdušne struje u glotalnom ciklusu i robustnosti u lošim SNR uslovima, pretpostavljeno je da ova obeležja mogu imati uspeha u automatskom prepoznavanju šapata. U narednom tekstu je opisan postupak izračunavanja ova dva obeležja.

6.2.2.2 TEAGER-ENERGETSKI ZASNOVANI MEL-FREKVENCIJSKI KEPSTRALNI KOEFICIJENTI (TEMFCC)

Postupak izračunavanja TEMFCC obeležja je veoma sličan MFCC obeležjima i razlikuje se samo u izračunavanju *Teager* energije (Slika 6.9). Ekstrakcija TEMFCC obeležja iz vremeskog okvira s[*n*] koji je prošao fazu predobrade se odvija kroz sledeće korake:

1) Primenom FFT se određuje spektar $S[n,w_k]$ govornog segmenta $s[n]$:

$$S[n,w_k]=\sum_{m=-\infty}^{+\infty} s[m]w[n-m]e^{-jw_k m}, \tag{6.14}$$

gde je $w_k = 2\pi k/N$, a N je dužina FFT.

2) Korišćenjem TEO se izračunava *Teager* energija spektra, $\Phi(S[n,w_k])$.

3) *Teager* energija se propušta kroz trougaonu Mel-frekvencijsku banku filtara $M_l[w_k]$ (Slika 6.6) i izračunavaju se energije na izlazima svakog od filtra:

$$e[n,l] = \sum_{k=L_l}^{U_l} \left| M_l[w_k] \Phi(S[n,w_k]) \right|, \tag{6.15}$$

gde je l = 1, 2 ... L, L = 30 je ukupan broj filtara, a L_l i U_l su donje i gornje granične frekvencije svakog od filtra.

4) Određuju se logaritmi energija $\log(e[n,l])$.

5) Primenom DCT se za svaki od vremenskih okvira izračunava kepstar i izdvaja prvih 12 TEMFCC:

$$\text{TEMFCC}[n,w_k] = \sum_{l=1}^{L} \log(e[n,l]) \cos\left(\frac{k\pi(l-0.5)}{L} \right), \tag{6.16}$$

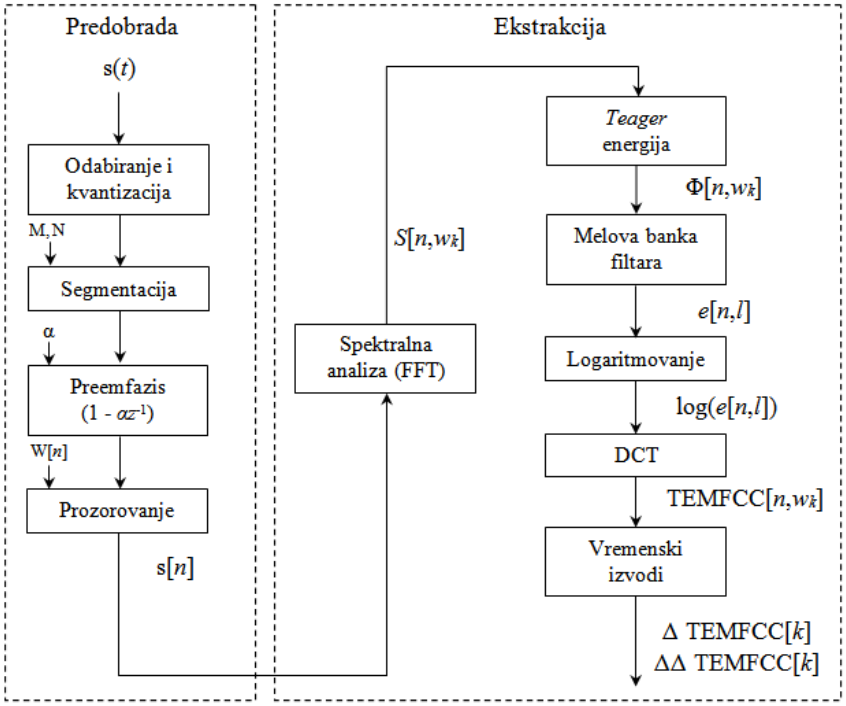
gde je N_c = 12 broj maksimalni broj kepstralnih koeficijenata, a k = 1, 2 ... N_c indeksi koeficijenata.

6) U cilju boljeg opisivanja dinamike signala izračunavaju se dinamička obeležja u vidu prvog (*delta*) i drugog (*delta-delta*) vremenskog izvoda TEMFCC obeležja. Postupak izračunavanja je isti kao i za ΔMFCC i ΔΔMFCC:

$$\Delta\text{TEMFCC}_i[k] = \frac{\sum_{m=-M}^{M} m\,\text{TEMFCC}_i[k+m]}{\sum_{m=-M}^{M} m^2}, \tag{6.17}$$

$$\Delta\Delta\text{TEMFCC}_i[k] = \frac{\sum_{m=-M}^{M} m\,\Delta\text{TEMFCC}_i[k+m]}{\sum_{m=-M}^{M} m^2}, \tag{6.18}$$

gde su $\Delta TEMFCC_i[k]$ i $\Delta\Delta TEMFCC_i[k]$ *k*-ti *delta* i *delta-delta* koeficijenti za *i*-ti okvir, a za njihovo izračunavanje je korišćeno $M = 4$ (u slučaju $\Delta\Delta$TEMFCC obeležja koristi se $M = 1$).



Slika 6.9 Procedura ekstrakcije TEMFCC obeležja i njenih prvih i drugih izvoda.

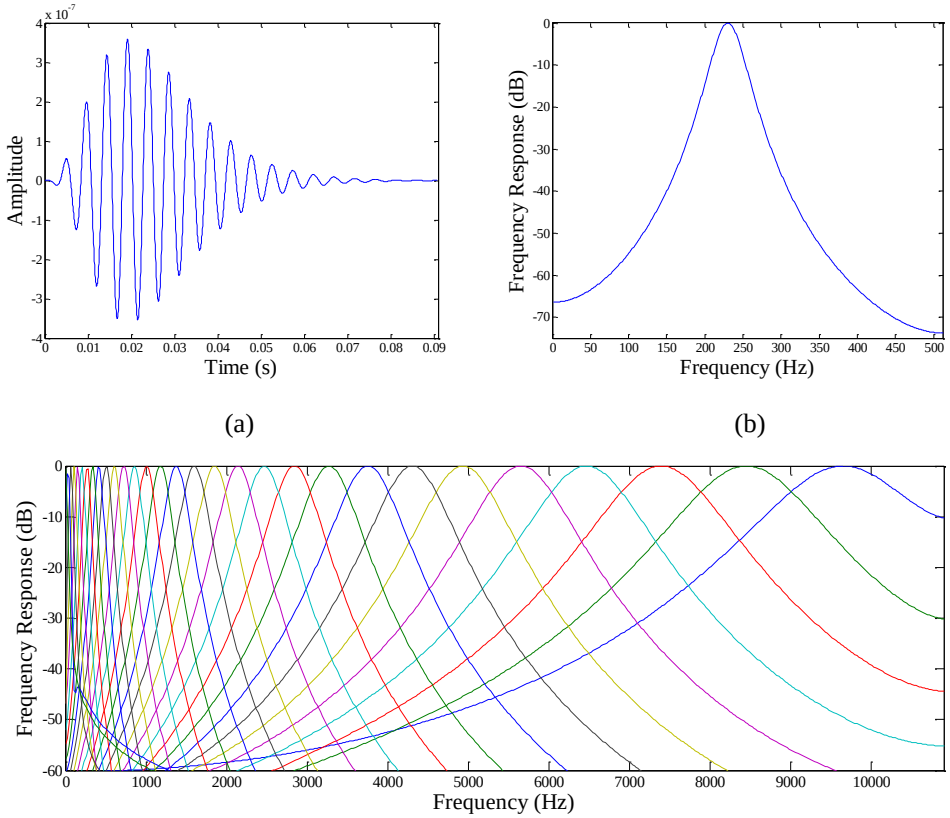
6.2.2.3 TAGER-ENERGETSKI KEPSTRALNI KOERICEIJNTI (TECC)

TECC obeležja pored upotrebe *Teager* energije imaju još jednu karakteristiku a to je implementacija *Gammmatone* banke filtara. *Gammatone* banka filtara [Patterson, 1994] je jedna od najčešće korišćenih auditivnih banki filtara koja modeluje čovekov mehanisam sluha u vidu oponašanja frekvencijske nelinearnosti bazilarne membrane i impulsnog odziva slušnog nerva. Naziv ovog filtra potiče od izgleda impulsnog odziva slušnog nerva koji podseća na čist sinusoidni signal pomnožen *Gamma* funkcijom raspodele, poznat kao *Gamma* ton, Slika 6.10 a). Još jedna odlika *Gammatone* filtara 3-5 reda je njihov frekvencijski odziv (Slika 6.10), koji po obliku liči i ima karakteristike čovekovog auditivnog filtra. On je asimetričan, zaobljen, šireg propusnog opsega, i nema konstantan Q faktor, u poređenju sa Melovim filtrima koji su asimetrični, trougaoni i sa konstantnim Q faktorom. Na značaj i popularnost *Gammatone* banke filtara u digitalnoj obradi govora su pozitivno uticali mogućnosti njenog efikasnog i

brzog izračunavanje kao i njen relativno jednostavan dizajn. Impulsni odziv jednog *Gammatone* filtra je prikazan na Slici 6.10 a) i definisan je formulom:

$$g(t) = At^{\eta-1}e^{(-2\pi f_c t)}\cos(2\pi f_c t + \phi)\,, \tag{6.19}$$

gde A predstavlja najveću amplitudu impulsnog odziva, $t^{\eta-1}$ određuje njegov početak, $e^{-2\pi bft}$ definiše širinu propusnog opsega filtra i kašnjenje impulsnog odziva, f_c je centralna frekvencija filtra, η je red filtra, a ϕ je početna faza impulsnog odziva [de Boer et al., 1978; Aertsen et al., 1980].



Slika 6.10 Karakteristika banke *Gammatone* filtara: (a) Impulsni odziv jednog *Gammatone* filtra (filtar sa centralnom frekvencijom 229Hz), (b) Frekvencijski odziv prethodnog filtra, (c) Frekvencijski odziv banke sa 30 *Gammatone* filtara prikazan na linearnoj frekvencijskoj skali.

Auditorni modeli najčešće koriste 42 opsega u banci filtara, koji pokrivaju frekvencijski opseg od 30Hz do 18kHz. Prema studiji [Glasberg et al., 1990] ustanovljena je sličnost između fiziologije čovekovog slušnog mehanizma i ERB

(*Equivalent Rectangular Bandwidth*) funkcije, koja definiše širine kritičnih opsega u zavisnosti od njihovih centralnih frekvencija:

$$ERB(f_c) = 6,23\left(\frac{f_c}{1000}\right)^2 + 93,39\left(\frac{f_c}{1000}\right) + 28,52\,, \tag{6.20}$$

gde $ERB(f_c)$ predstavlja širinu propusnog opsega jednog filtra sa centralnom frekvencijom f_c u Hz. *Gammatone* banka filtara pored svoje superiornosti u auditivnom modelovanju, za razliku od Melove banke filtara, obezbeđuje i dosta veću robustnost u situacijama narušenog SNR usled aditivnog šuma ili drugih neželjenih interferencija [Dimitriadis et al., 2005].

Postupak ekstrakcije TECC obeležja je šematski prikazan na Slici 6.11. i opisan kroz sledeće korake izračunavanja:

1) Posle završene faze predobrade, signal $s[n]$ se propušta kroz *Gammatone* banku filtara, koja se sastoji iz $L = 30$ filtara $G_l[w_k]$ (isti broj kao i u Mel-frekvencijskoj banci filtara prilikom ekstrakcije TECC i TEMFCC obeležja). Slika 6.10. c).

2) Za svaki od izlaza iz *Gammatone* banke filtara, $g_l[n]$, se primenjuje FFT i izračunava spektar $S_l[n,w_k]$:

$$S_l[n, w_k] = \sum_{m=-\infty}^{+\infty} g_l[m]w[n-m]e^{-jw_k m}\,, \tag{6.21}$$

gde je $w_k = 2\pi k/N$, a N je dužina FFT.

3) Na tako izračunate spektre $S_l[n,w_k]$ se primenjuje TEO i računaju se energije:

$$e[n,l] = \sum_{k=L_l}^{U_l} \left|\Phi(S_l[n, w_k])\right|, \tag{6.22}$$

gde l = 1, 2 ... L, određuje redni broj energije pojedinog filtra, L = 30, a L_l i U_l njegove donje i gornje granične frekvencije propusnog opsega.

4) Dobijene energije se logaritmuju.

5) Izračunava se DCT i izdvaja se prvih 12 kepstralnih koeficijenata:

$$\mathrm{TECC}[n, w_k] = \sum_{l=1}^{L} \log(e[n,l]) \cos\left(\frac{k\pi(l-0.5)}{L}\right), \tag{6.23}$$

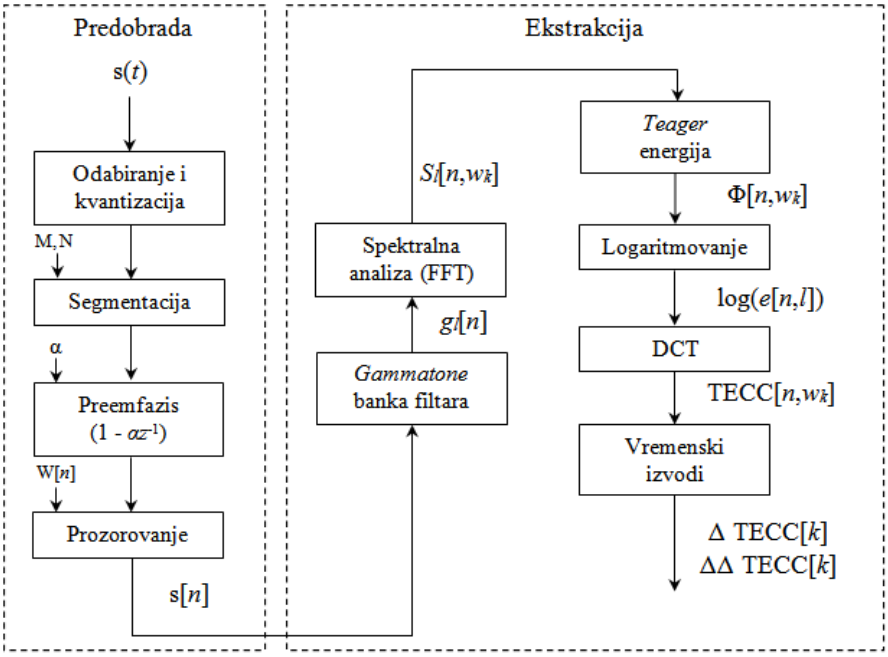
gde su k = 1, 2 ... N_c indeksi koeficijenata, a N_c = 12 maksimalni broj kepstralnih koeficijenata.

6) Kao i kod prethodnih obeležja i kod TECC se izračunavaju dinamička obeležja, *delta* i *delta-delta* koeficijenti:

$$\Delta\mathrm{TECC}_\mathrm{i}[k] = \frac{\sum_{m=-M}^{M} m\mathrm{TECC}_\mathrm{i}[k+m]}{\sum_{m=-M}^{M} m^2}, \tag{6.24}$$

$$\Delta\Delta\mathrm{TECC}_\mathrm{i}[k] = \frac{\sum_{m=-M}^{M} m\Delta\mathrm{TECC}_\mathrm{i}[k+m]}{\sum_{m=-M}^{M} m^2}, \tag{6.25}$$

gde su $\Delta\mathrm{TECC}_i[k]$ i $\Delta\Delta\mathrm{TECC}_i[k]$ k-ti *delta* i *delta-delta* koeficijenti za i-ti okvir, a za njihovo izračunavanje je u formulama korišćeno M = 4 za ΔTECC, odnosno M = 2 za ΔΔTECC koeficijenate.



Slika 6.11 Procedura ekstrakcije TECC obeležja i njenih prvih i drugih izvoda.

6.3 KREIRANJE TRENING I TEST MATRICA OBELEŽJA

Po završetku procesa ekstrakcije govornih obeležja, svaka reč iz Whi-Spe baze je predstavljena vektorom akustičkih obeležja dužine 132 koeficijenta (11 vremenskih okvira sa po 12 kepstralnih koeficijenata). U cilju boljeg opisa dinamike signala, ovim vektorima su pridodati njihovi prvi i drugi vremenski izvodi (132 *delta* koeficijenta i 132 *delta-delta* koeficijenta) pa su dužine tako proširenih vektora 264 ili 396 koeficijenata u zavisnosti od toga da li su dodavani samo *delta* ili i *delta-delta* koeficijenti. Svi ovi vektori su potom združeni u matrice dimenzija 132×500, 264×500, i 396×500 koeficijenata – za svakog govornika po jedna matrica u normalnom govoru i jedna u šapatu. Pojedinačne matrice za svakog govornika sadrže vektorske predstave svih 50 reči koje je taj govornih izgovarao deset puta u jednom od dva govorna moda. Ovakve matrice su formirane za sva tri tipa kepstralnih koeficijenata: MFCC, TEMFCC i TECC, i kasnije su korišćene u postupku obuke, validacije i testiranja MLP. Za potrebe treniranja neuralnih mreža, uz odgovarajuće opisane matrice neophodno je bilo formirati i takozvane *Target* matrice, koje definišu izlaz neuralne mreže za svaki tip stimulusa koji se može pojaviti na njenom ulaznom sloju.

Na ovaj način je kreirano ukupno 20 matrica (10 za govor i 10 za šapat) dimenzija 132×500, 264×500, i 396×500 koeficijenata i isto toliko odgovarajućih *Target* matrica.

6.4 KREIRANJE VIŠESLOJNIH PERCEPTRONA

Kao osnova *back-end* sistema za klasifikaciju reči u ovom sistemu je korišćena *feedforward* neuralna mreža, obučena sa *Back Propagation* algoritmom. Neuralna mreža je realizovana u formi višeslojnih perceptrona (*Multi Layer Perceptrons – MLP*) koristeći MATLAB Neural Network Toolbox [Demuth et al., 2008]. Dve identične mreže su kreirane – jedna za prepoznavanje reči u normalnom govoru i druga za prepoznavanje reči u šapatu. Obe mreže su imale istu strukturu, odnosno topologiju, u cilju poređenja njihovih performansi u prepoznavanju govora i šapata, respektivno.

6.4.1 ODREĐIVANJE OPTIMALNE MLP ARHITEKTURE

Višeslojni perceptroni su organizovani u tri[7] sloja: jedan ulazni, jedan skriveni i jedan izlazni sloj. U zavisnosti od dužine vektora (132, 264 ili 396 koeficijenata) koji dolaze na ulaz ANN, odnosno u zavisnosti od korišćenja govornih obeležja sa ili bez dinamičkih obeležja, kreirane su tri različite topologije neuralne mreže sa 132, 264 i 396 ulazna čvora. Izlazni slojevi mreža su imali 50 neurona koji služe za klasifikaciju 50 različitih reči. U perceptronima su kao transfer funkcije korišćene bipolarne sigmoid funkcije, Slika 3.7 f), (*hyperbolic tangent sigmoid function - tansig*).

Izbor odgovarajućeg broja skrivenih neurona predstavlja veoma važan zadatak i detaljno je analiziran. Suviše mali broj skrivenih neurona uskraćuje resurse neuralne mreže i njene mogućnosti da reši zadate probleme, dok sa druge strane suviše veliki broj neurona povećava potrebno vreme za obuku mreže, koje u izvesnim slučajevima može biti izuzetno dugo i potpuno neadekvatno. Takođe, prekomeran broj neurona povećava mogućnost ulaska mreže u *overfitting*. Iz tog razloga je u ovoj studiji pažljivo odabran optimalni broj skrivenih neurona. Testirano je više metoda njihovog izračunavanja poput: "*the rule of the thumb*", "*the geometric pyramid rule*", itd. [Blum, 1992; Masters, 1993; Boger et al., 1997; Berry et al., 1997; Karsoliya, 2012]. Međutim, preporuke ovih

[7] Pojedini autori ne računaju ulazni sloj čvorova kao zaseban sloj neurona neuralne mreže, već samo skrivene slojeve i izlazni sloj neurona.

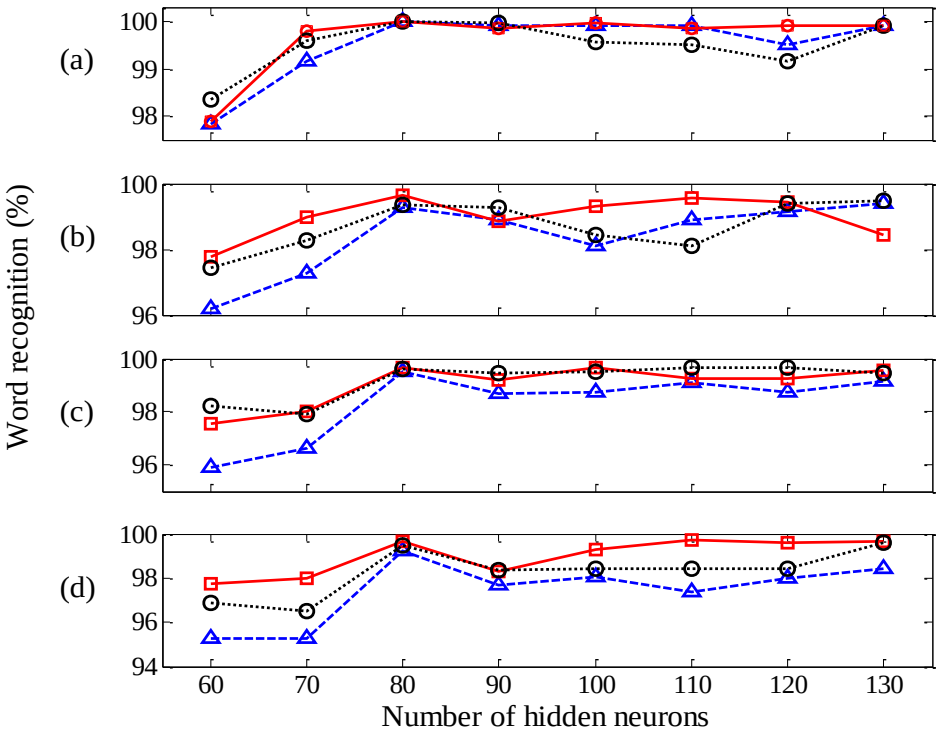
metoda nisu bile usaglašene i dale su različite rezultate u pogledu broja neurona. Na primer, metoda "*the rule of the thumb*" je prema formuli (6.26) za mreže sa 132 ulazna čvora i 50 neurona u izlaznom sloju odredila 91 neuron kao optimalni broj skrivenih neurona, dok je "*the geometric pyramid rule*" prema formuli (6.27) izračunala 81 skriveni neuron:

$$N_h = \frac{N_{in} + N_{out}}{2}, \tag{6.26}$$

$$N_h = \sqrt{N_{in} \times N_{out}}, \tag{6.27}$$

pri čemu je u formulama sa N_h obeležen broj skrivenih neurona, sa N_{in} broj ulaznih čvorova, a sa N_{out} broj neurona u izlaznom sloju mreže. Testiranjem drugih metoda, dobijene su još veće razlike u broju skrivenih neurona, pri čemu se taj broj kretao od 60 do čak 130 skrivenih neurona. Gore pomenute formule i metode predstavljaju samo grubu aproksimaciju broja neurona i ne treba ih rigorozno primenjivati. Iz tog razloga, jedina mogućnost da se odredi tačan i precizan broj neurona, je da se sprovede poseban eksperiment u kome će se broj skrivenih neurona postepeno povećavati uz uporedno praćenje performansi mreže [Grozdić et al., 2013]. U tu svrhu su formirane mreže sa startnom strukturom od po 10 skrivenih neurona, čiji je broj postepeno povećavan sa koracima od po 10 neurona uz istovremeno praćenje uspeha mreža u prepoznavanju reči. Rezultati ove analize za mreže sa 132 ulazna čvora i tri tipa govornih obeležja (MFCC, TEMFCC i TECC) su ilustrovani na Slici 6.12. Kao što se može primetiti sa dijagrama, mreže su već sa 60 neurona u skrivenim slojevima imale visok uspeh u prepoznavanju reči, dok je sa daljim povećavanjem broja neurona njihov uspeh dodatno rastao i brzo dostigao svoj maksimum. Taj maksimum je postignut sa 80 neurona u skrivenom sloju, pri čemu su performanse mreže sa daljim povećavanjem broja skrivenih neurona opadale, odnosno učestalost greške prepoznavanja reči je rasla (*Word Error Rate* - *WER*) što je pouzdani znak ulaska mreže u *overfitting*. Na ovaj način sa preciznijim dodavanjem ili oduzimanjem skrivenih neurona i uporednim testiranjem performansi mreža, su određeni optimalani brojevi skrivenih neurona, koji iznose: 81, 114 i 140 za mreže sa 132, 264 i 396 ulaznih čvorova, respektivno. Ovi brojevi se dobro

poklapaju sa teorijskom predikcijom optimalnog broja neurona, definisanom formulom (6.27).



Slika 6.12 Pronalazak optimalnog broja neurona u mrežama sa 132 ulazna čvora. Testiranje u obuci sa: MFCC (Δ), TEMFCC (O) i TECC ( ) obeležjima.

U ekperimentima menjanja arhitekture neuralne mreže, takođe je testirana upotreba mreža sa više skrivenih slojeva, ali promena broja skrivenih slojeva nije pokazala bitniji uticaj na uspeh prepoznavanja reči, zbog čega se ostalo pri korišćenju prvodbitne strukture mreže sa tri sloja (sa jednim skrivenim slojem).

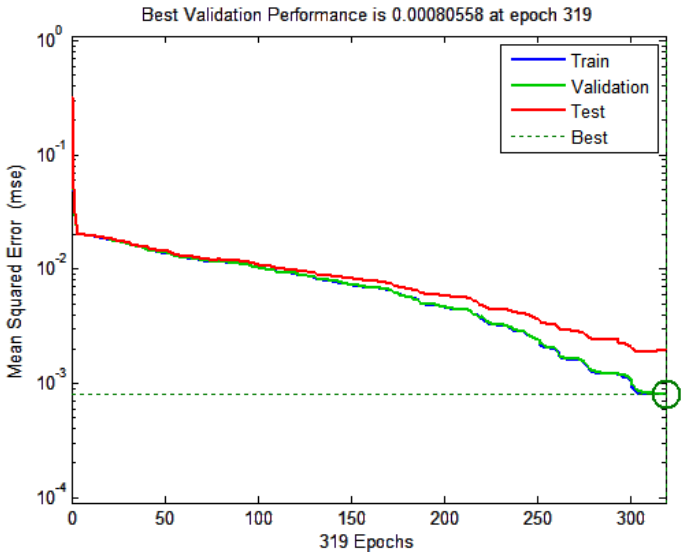
6.4.2 OBUKA VIŠESLOJNIH PERCEPTRONA

Sa stanovišta obuke neuralnih mreža, *Neural Network Toolbox* nudi već predefinisane algoritme i postavke treninga mreža među kojima je i dobro poznati *Back Propagation* algoritam. Jedna od predefinisanih postavki treninga je podela baze podataka na tri podskupa: 70% baze se koristi u treningu, 15% u validaciji, a 15% u testiranju performansi mreže. Podaci koji čine te podskupove su odabrani na statistički slučajan način, usled čega postoji rizik da neke od reči budu izostavljene iz procesa obuke mreže. Posledica ovakve obuke mreže mogu biti: (1) nemogućnost ANN da

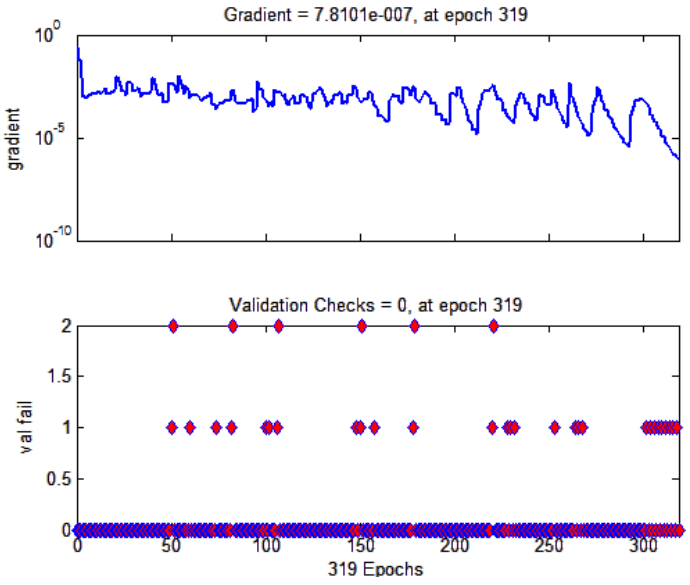
prepoznav reč koja nije bila uključena u procesu treninga, i (2) povišena varijabilnost performansi mreže tokom ponovnih treninga. Iz tog razloga, postavke ovog algoritma su dorađene u MATLAB kodu i prilagođene potrebama ovog istraživanja [Grozdić et al., 2013].

Napravljeno je par izmena. Baza podataka je podeljena na tri dela: 60% baze je korišćeno u treningu, 20% u kros-validaciji (*cross-validation*) i 20% u testiranju. Tokom podele baze podataka, vođeno je računa o uzimanju tačnog broja uzorka od 10 izgovora svake reči iz Whi-Spe baze. Šest uzoraka izgovora svake reči je slučajnim statističkim postupkom odabrano za obuku, druga dva za kros-validaciju i preostala dva za testiranje mreže. Na ovaj način je baza podeljena na tri podskupa u formi matrica govornih obeležja. Prvi podskup ima dimenzije 396×300 koeficijenata, dok drugi i treći imaju dimenzije 396×100 koeficijenata. Ovakva podela baze podataka garantuje da će ANN uspešno akustički modelovati svaku reč iz Whi-Spe korpusa, a varijabilnost performansi mreže prilikom ponovljenih treninga će biti značajno smanjena.

U procesu treninga mreže je upotrebljena *trainscg* funkcija, koja je oblik *Back Propagation* algoritma i zasniva se na principu spuštanja gradijenta (*gradient descent*). Tačnije, ovaj algoritam predstavlja kombinaciju *Levenberg-Marquardt* algoritma i principa skaliranog konjugovanog gradijenta (*scaled conjugated gradient*) [Masters, 1993]. Ovakav tip obuke neuralne mreže, zahteva nešto veći broj iteracija, ali zato značajno smanjuje broj neophodnih računskih operacija, čime se dosta skraćuje trajanje obuke mreže, što čini *trainscg* funkciju pogodnom za obuku velikih neuralnih mreža. Kako bi se izbegli *overfitting* i *overtrainging* efektri, definisani su kriterijumi za prekid obuke mreže, poput: definisanog maksimalnog broja iteracija (1000), srednje kvadratne greške (*Mean Squared Error - MSE*) u prepoznavanju reči (0.00), (Slika 6.13), maksimalnog broja uzastopnih grešaka u kros-validaciji ili takozvana *early stopping* metoda (6), (Slika 6.14), i minimalnog gradijenta (10^{-6}) (Slika 6.14) [Demuth, 2008].



Slika 6.13 Primer promene srednje kvadratne greške tokom epoha u: treningu, validaciji i testiranju.

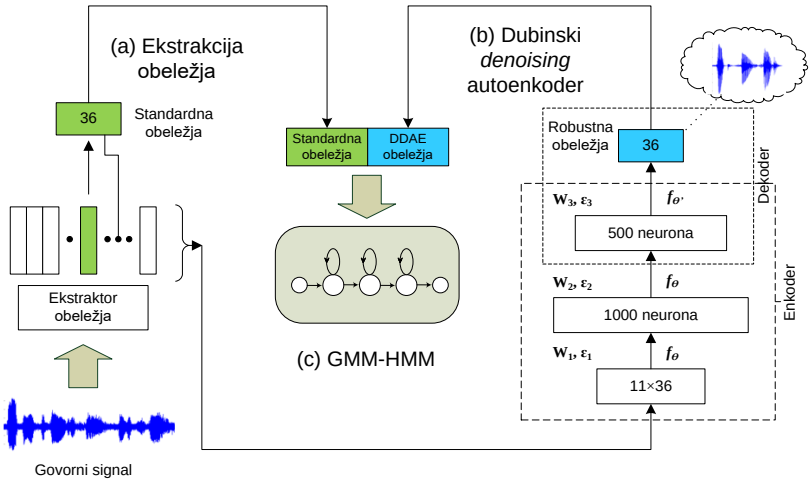


Slika 6.14 Primer spuštanja gradijenta i pojave grešaka u kros-validaciji tokom epoha.

7 KREIRANJE TANDEM DNN-HMM SISTEMA ZA PREPOZNAVANJE ŠAPATA

U ovom poglavlju je detaljno opisan postupak kreiranja još jednog sistema za efikasno prepoznavanje šapata koji je baziran na tandem DNN-HMM sistemu [Grozdić et al., 2016 b]. Predloženi sistem, šematski prikazan na Slici 7.1, je takođe koncipiran na neuralnoj mreži, tačnije na tandemu dubinskog *denoising* autoenkodera (DDAE) i HMM sistema. Osnovna karakteristika ovog sistema je njegov *front-end* deo koji čine dva odvojena dela za ekstrakciju govornih obeležja. Prvi ekstraktor služi za izdvajanje jednog od ranije opisanih standardnih govornih obeležja poput MFCC, TECC ili TEMFCC, dok drugi ekstraktor u vidu dubinske neuralne mreže (DNN), tačnije dubinskog *denoising* autoenkodera (DDAE) omogućava paralelnu ekstrakciju posebnih robustnih govornih obeležja (MFCC-DDAE, TECC-DDAE, TEMFCC-DDAE). Naime, koristeći svoju dubinsku arhitekturu DDAE omogućava rekonstrukciju govornih obeležja i karakteristika normalnog govora iz šapata, poput uklanjanja efekata šumne strukture i dodavanje karakteristika zvučnih glasova. Tako dobijena standardna i rekonstruisana govorna obeležja se potom zajedno prosleđuju u *back-end* deo koji

predstavlja tradicionalni GMM-HMM sistem u kome se vrši klasifikacija, odnosno prepoznavanje reči.



Slika 7.1 Arhitektura predloženog tandem DNN-HMM sistema za automatsko prepoznavanje šapata. Na slici su prikazane tri celine koje čine sistem: (a) Deo za ekstrakciju standardnih obeležja, (b) Dubinski *denoising* autoenkoder (DDAE) i (c) GMM-HMM *back-end* sistem.

U nastavku poglavlja biće detaljno opisan svaki deo ovog tandem sistema, uključujući predobradu govornih signala, ekstrakciju govornih obeležja, kreiranje i obuku DDAE kao i formiranje *back-end* dela tandem DNN-HMM istema.

7.1 PREDOBRADA GOVORNIH SIGNALA

S ozbzirom da *back-end* deo ovog sistema predstavlja GMM-HMM, nema potrebe za posebnim metodama za vremensko usklađivanje govornih signala i formiranje vektora govornih obeležja iste dužine kao što je to u slučaju MLP sistema. GMM-HMM ima mogućnost vremenskog modelovanja govornih signala, te se oni segmentiraju na vremenske okvire (*frames*) trajanja 25ms sa međusobnim preklapanjem od 10ms. Prema tome, svaka reč iz Whi-Spe baze se deli na odgovarajući broj vremenskih okvira u zavisnosti od dužine reči. Isto kao i u slučaju predobrade signala kod MLP sistema i u ovom slučaju segmenti se filtriraju sa istim preemfazis filtrom (koeficijent $\lambda = 0{,}97$) i množe sa Hamingovom prozorskom funkcijom, opisanim u Poglavlju 6.1.2.

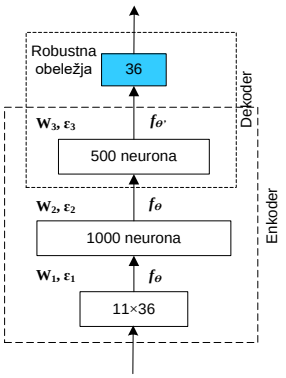
7.2 EKSTRAKCIJA GOVORNIH OBELEŽJA

Nakon predobrade i segmentacije govornih signala sledi faza ekstrakcije govornih obeležja. Kao što je prikazano na Slici 7.1, tandem DNN-HMM sistem poseduje dva

ekstraktora kepstralnih koeficijenata. Postupak ekstrakcije MFCC, TECC i TEMFCC obeležja pomoću prvog ekstraktora (prikazan na levoj strani Slike 7.1) je identičan kao kod MLP sistema i detaljno je opisan u Poglavlju 6.2. U ovoj sekciji pažnja je posvećena drugom ekstraktoru (prikazan na desnoj strani Slike 7.1) i ekstrakciji robusntih govornih obeležja pomoću dubinskog *denoising* autoenkodera.

7.2.1 KREIRANJE I OBUKA DDAE ZA EKSTRAKCIJU ROBUSTNIH OBELEŽJA

Ideja da se od govornih obeležja ekstrahovanih iz šapata dobiju rekonstruisana robustna obeležja što sličnija onim u normalnom govoru je realizovana upotrebom dubinskog *denoising* autoenkodera. Naime, korišćenjem paralelnih uzoraka šapata[8] i normalnog govora, DDAE je obučen tako da njegova dubinska struktura nauči da rekonstruiše kepstralna obeležja normalnog govora kada se na njegovom ulazu pojave kepstralni uzorci šapata. Predložena arhitektura DDAE je prikazana na Slici 7.2 i sastoji se iz dva sloja za enkodovanje sa po 1000 i 500 neurona sa *sigmoid* transfer funkcijama i jednog sloja za dekodovanje sa linearnom transfer funkcijom. Ulazni sloj ima 11×36 = 396 čvorova a izlazni sloj 36. (Slučaj kada se koriste 12 kepstralnih koeficijenata + 12Δ + 12ΔΔ koeficijenata).



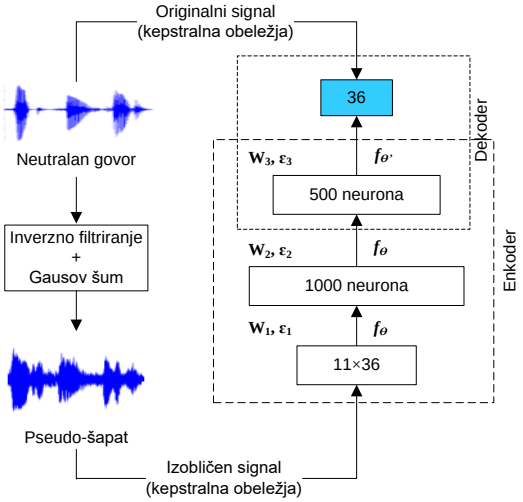
Slika 7.2 Arhitektura dubinskog denoising autoenkodera (DDAE).

Ovo znači da se na ulazu koristi 11 susednih vektora od po 36 kepstralnih koeficijenata za enkodovanje a na izlazu se dekodovanjem dobija 1 vektor od po 36 koeficijenata. Odnosno, za ekstrakciju jednog takvog vektora koriste se 5+1+5 frejmova na ulazu (1 frejm sa 5 frejmova koji mu prethode i 5 frejmova koji mu slede) pri čemu

[8] Tačnije, korišćenjem paralelnih uzoraka pseudo-šapata i normalnog govora o čemu će biti više reči u nastavku teksta.

je frejm u sredini onaj koji se autoenkoderom dekoduje na izlazu. Čitav *front-end* deo sistema, uključujući i DDAE je razvijen u MATLAB programskom sistemu.

Obuka autoenkodera je prikazana na Slici 7.3 i sastoji se iz dva dela – iz pred-obuke i završne obuke. U fazi pred-obuke, DDAE se trenira na kepstralnim koeficijentima uzoraka pseudo-šapata, a u završnoj fazi se finalno podešava sa kepstralnim koeficijentima uzoraka normalnog govora pomoću *Backpropagation* algoritma.



Slika 7.3 Dubinski *denoising* autoenkoder (DDAE) u fazi obuke.

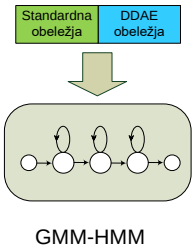
Uzorci pseudo-šapata su dobijeni iz uzoraka normalnog govora koji su inverzno filtrirani [Grozdić et al., 2014; Grozdić et al., 2017] i dodat im je Gausov beli šum, pri čemu je ostvaren SNR od 10dB. Postoji više razloga za ovakav način kreiranja pseudo-šapata. Prvo, sa dodavanjem Gausovog šuma inverzno filtriran govor po svojim karakteristikama postaje sličniji šapatu u pogledu njegovih akustičkih karakteristika i šumne strukture pa se time akustički model adaptira na slične varijacije (distorzije) signala na ulazu. Drugo, dodavanjem šuma se dodatno sprečava mogućnost DDAE da dođe do trivijalnog rešenja za zadati problem. Treće, pošto je dodavanje šuma stohastički proces, time se broj uzoraka značajno uvećava u procesu obuke čime se i smanjuje mogućnost pojave *overfitting*-a. Postoje i razlozi za korišćenje pseudo-šapata umesto pravih uzoraka šapata. Prvo, zahvaljujući tačno definisanim spektralnim karakteristikama pseudo-šapata i raspolaganjem parova takvih uzoraka pseudo-šapata sa njihovim originalima u vidu normalnog govora, DDAE dobija jasan zadatak u potrazi za

odgovarajućom transfer funkcijom. Takođe, opisani postupak generisanja pseudo-šapata omogućava relativno brzo i jednostavno generisanje velike količine uzoraka pseudo-šapata koristeći postojeće baze normalnog govora, što je od velike koristi u slučaju adaptacije *speaker-dependent* sistema kod kojih je naknadno dosnimavanje i formiranje dodatne baze šapata gotovo neizvodljivo.

Na ovakv način, bogata nelinearna struktura dubinskog *denoising* autoenkodera omogućava da DDAE efikasno nauči transfer funkciju koja će potisnuti karakteristike šapata iz govornog signala, pritom zadržavajući dovoljno fonetskih informacija za rekonstrukciju govornih obeležja normalnog govora.

7.3 KREIRANJE BACK-END SISTEMA

Kao *back-end* deo tandem DNN-HMM sistema se koristi standardni GMM-HMM sistem, koji je treniran i testiran pomoću MATLAB programskog paketa. Akustički model čine ukupno 5 stanja (od kojih su 3 emitujuća), sa *left-to-right* tipom tranzicije i 16 Gausovih mešavina. Na ulaz DNN-GMM po jednom frejmu dolazi vektor dužine 72 kepstralna koeficijenta (36 originalnih obeležja + 36 robusntih obeležja dobijenih pomoću DDAE) kao što je prikazano na slici 7.4.



Slika 7.4 *Back-end* deo tandem DNN-HMM sistema.

Broj trening ciklusa u procesu Baum-Welch re-estimacije je ograničen na 5. Parametri modela su estimirani korišćenjem *flat-start* metode. U fazi testiranja, Viterbi algoritam je korišćen kako bi se odredio model koji najviše odgovara ulaznom govornom uzorku.

7.4 OBUKA TANDEM DNN-HMM SISTEMA

Obuka tandem DNN-HMM sistema se sastoji iz dve faze – iz obuke DNN dela sistema, odnostno DDAE sistema za ekstrakciju robustnih gvornih obeležja i iz obuke *back-end* HMM sistema. DDAE se obučava sa paralelnim uzorcima pseudo-šapata i

njihovim parnjacima u vidu snimaka normalnog govora iz Whi-Spe baze. Podela uzoraka na trening (60%), validaciju (20%) i testiranje (20%) je obavljeno na isti način kao i kod MLP sistema (Poglavlje 6.4.2). Po završetku obuke DDAE, baza pseudo-šapata se više ne koristi u obuci tandem DNN-HMM sistema, a obučeni dubinski *denoising* autoenkoder je sada u funkciji ekstraktora robustnih govornih obeležja. U drugoj fazi obuke tandem DNN-HMM sistema, odnosno u obuci *back-end* HMM sistema se koristi isključivo Whi-Spe baza, koja je podeljena po istom ranije opisanom principu na deo za trening i testiranje.

8 EKSPERIMENTI SA MLP SISTEMOM

U ovom poglavlju su prezentovani rezultati eksperimenata automatskog prepoznavanja izolovanih reči u normalnom govoru i šapatu pomoću sistema baziranog na MLP tipu neuralnih mreža. Prvo je analiziran uspeh prepoznavanja reči u različitim obuka/test scenarijima, kao i performanse MLP sistema u zavisnosti od upotrebe tri različita tipa kepstralnih koeficijenata. Zatim je izvršena analiza konfuzije u prepoznavanju reči, kao i spektralno tumačenje konfuzija posebno istaknutih kritičnih parova reči. Rezultati ovih analiza su doveli do postavke hipoteze o maskiranju određenih govornih obeležja usled zvučnosti kao glavnog razloga degradiranog prepoznavanja reči u neusaglašenim obuka/test scenarijima. U nastavku poglavlja u sklopu dokaza hipoteze, opisan je eksperiment sa inverznim filtriranjem kao i ostvareni rezultati u poboljšanju uspeha prepoznavanja reči, koji su potvrdili postavljenu hipotezu. Svi prezentovani rezultati predstavljaju usrednjene vrednosti dobijene 10-fold kros validacijom, pri čemu standardna greška odstupanja od srednje vrednosti SEM[9] (*standard error of the mean*) nije bila veća od 0,23%.

[9] SEM (*standard error of the mean*), odnosno standardna greška odstupanja od srednje vrednosti, se računa na sledeći način: SEM=$\sigma/\sqrt{n}$, gde je σ standardna devijacija uzoraka a n broj uzoraka.

8.1 USAGLAŠENI OBUKA/TEST SCENARIJI

U ovoj sekciji su prikazani rezultati eksperimenta prepoznavanja izolovanih reči zavisnog od govornika u usaglašenim obuka/test scenarijima − govor/govor i šapat/šapat, sprovedenog na čitavom Whi-Spe korpusu pomoću MLP sistema. Testirana su sva tri tipa kepstralnih koeficijenata (MFCC, TEMFCC i TECC) kao i doprinos dodavanja njihovih dinamičkih obeležja. Zavisno od upotrebljene dužine vektorske reprezentacije govornih stimulusa, korišćene su neuralne mreže sa sledećim strukturama: 132/81/50, 264/114/50 i 396/140/50 ((broj ulaznih čvorova)/(broj skrivenih neurona)/(broj izlaznih neurona)). Rezultati prepoznavanja reči su prikazani u Tabeli 8.1.

Tabela 8.1

Uspeh prepoznavanja reči u usaglašenim obuka/test scenarijima (izražen u %) za muške i ženske govornike u zavisnosti od upotrebljenih govornih obeležja.

Govorni mod (obuka/test)	Govor (govor/govor)		Šapat (šapat/šapat)	
Pol govornika	Muški govornici	Ženski govornici	Muški govornici	Ženski govornici
MFCC	99,90	99,52	99,28	99,24
MFCC+Δ	99,92	99,68	99,72	99,80
MFCC+Δ+ΔΔ	100	99,70	99,70	99,80
TECC	99,90	99,68	99,64	99,68
TECC+Δ	99,88	99,68	99,66	99,64
TECC+Δ+ΔΔ	99,90	99,80	100	99,90
TEMFCC	99,90	99,60	99,36	99,48
TEMFCC+Δ	99,96	99,64	99,40	99,48
TEMFCC+Δ+ΔΔ	100	99,90	99,90	100

Kao što se vidi iz tabele, uspeh u prepoznavanju reči je kod svih govornika veoma visok u oba govorna moda, a u pojedinim slučajevima dostiže maksimalnih 100% uspeha. Iako je dostignut "plafon" u uspehu prepoznavanja reči, analizom usrednjenih vrednosti u Tabeli 8.2 se primećuju dve blage tendencije. Prvo, MFCC obeležja pokazuju nešto manji uspeh u prepoznavanju reči u oba govorna moda u poređenju sa druga dva govorna obeležja. Drugo, u šapatu TECC obeležja imaju najviše uspeha (u proseku 99,75%) što nagoveštava da TEO i *Gammatone* banka filtara bolje opisuju karakteristike šapata. Kako bi se dodatno potkrepile ove tvrdnje, sprovedena je statistička analiza dobijenih rezultata. U tu svrhu upotrebljen je dvostrani *Wilcoxon test*

označenih rangova[10] (*two-tailed Wilcoxon signed-rank test*) za analizu statističke značajnosti malih razlika u uspesima prepoznavanja reči između tradicionalnih MFCC i ostalih kepstralnih obeležja. *Z* i *p*-vredosti *Wilcoxon* testa su prikazani u Tabeli 8.3. Rezultati ovog testa kod svih govornika potvrđuju da su Teager obeležja od statističkog značaja ($p < 0.05$) u prepoznavanju šapata. TECC+Δ+ΔΔ obeležja su pokazala najveću statističku značajnost u poređenju sa ostalim obeležjima ($Z = -2.675$; $p = 0.007$). Takođe je ustanovljeno da u govor/govor scenariju izbor govornog obeležja nema bitan značaj.

Tabela 8.2

Usrednjeni rezultati prepoznavanja reči u usaglašenim obuka/test scenarijima u zavisnosti od upotrebljenih govornih obeležja. (izraženo u %).

Govorni mod (obuka/test)	Govor (govor/govor)	Srednja vr.	Šapat (šapat/šapat)	Srednja vr.
MFCC	99,71		99,26	
MFCC+Δ	99,80	99,79	99,76	99,59
MFCC+Δ+ΔΔ	99,85		99,75	
TECC	99,79		99,66	
TECC+Δ	99,78	99,81	99,65	99,75
TECC+Δ+ΔΔ	99,85		99,95	
TEMFCC	99,75		99,42	
TEMFCC+Δ	99,80	99,83	99,44	99,60
TEMFCC+Δ+ΔΔ	99,95		99,95	

Tabela 8.3

Rezultati Wilcoxon testa u poređenju uspeha prepoznavanja reči u usaglašenim obuka/test scenarijima u zavisnosti od korišćenja različitih govornih obeležja.

Govorni mod (obuka/test)	Govor (govor/govor)		Šapat (šapat/šapat)	
	Z	*p*	*Z*	*p*
MFCC	(/ ; /)		(/ ; /)	
MFCC+Δ	(*Z*=-1.604;	*p*=0.109)	(*Z*=-1.486;	*p*=0.137)
MFCC+Δ+ΔΔ	(*Z*=-1.134;	*p*=0.257)	(*Z*=-0.352;	*p*=0.725)
TECC	(*Z*=-0.604;	*p*=0.546)	(*Z*=-2.371;	***p*=0.018**)
TECC+Δ	(*Z*=-0.216;	*p*=0.829)	(*Z*=-2.151;	***p*=0.031**)
TECC+Δ+ΔΔ	(*Z*=-0.677;	*p*=0.498)	(*Z*=-2.675;	***p*=0.007**)
TEMFCC	(*Z*=-1.298;	*p*=0.194)	(*Z*=-2.207;	***p*=0.027**)
TEMFCC+Δ	(*Z*=-1.510;	*p*=0.131)	(*Z*=-2.207;	***p*=0.027**)
TEMFCC+Δ+ΔΔ	(*Z*=-1.841;	*p*=0.066)	(*Z*=-2.536;	***p*=0.011**)

(Interval poverenja = 95%)

[10] Wilcoxonov test označenih rangova je statistički test namenjen za analizu podataka dobijenih iz ponovljenih merenja, i predstavlja neparametarsku alternativu za parametarski *t-test* parova.

8.2 NEUSAGLAŠENI OBUKA/TEST SCENARIJI

Uobičajan problem ASR sistema nastaje kada govornik iz normalnog govora pređe u šapat i obratno. S obzirom da su tradicionalni ASR sistemi uglavnom obučeni za prepoznavanje normalnog govora, u ovakvim situacijama prepoznavanje šapata je u velikoj meri degradirano. Sa ciljem simulacije i analize ove pojave, ispitani su različiti oblici neusaglašenih obuka/test scenarija sa MLP sistemom. Rezultati ovog eksperimenta predstavljaju prvi korak u rešavanju navedenog problema i razumevanju konfuzije između prepoznavanja izolovanih reči u normalnom govoru i šapatu. Ispitan je uspeh MLP sistema obučenog na normalnom govoru u prepoznavanju izolovanih reči u šapatu, i obratno, prepoznavanje izolovanih reči u normalnom govoru kada je MLP sistem obučen na šapatu. Usrednjeni rezultati ovog eksperimenta za sve govornike iz Whi-Spe baze su prikazani u Tabeli 8.4.

Tabela 8.4

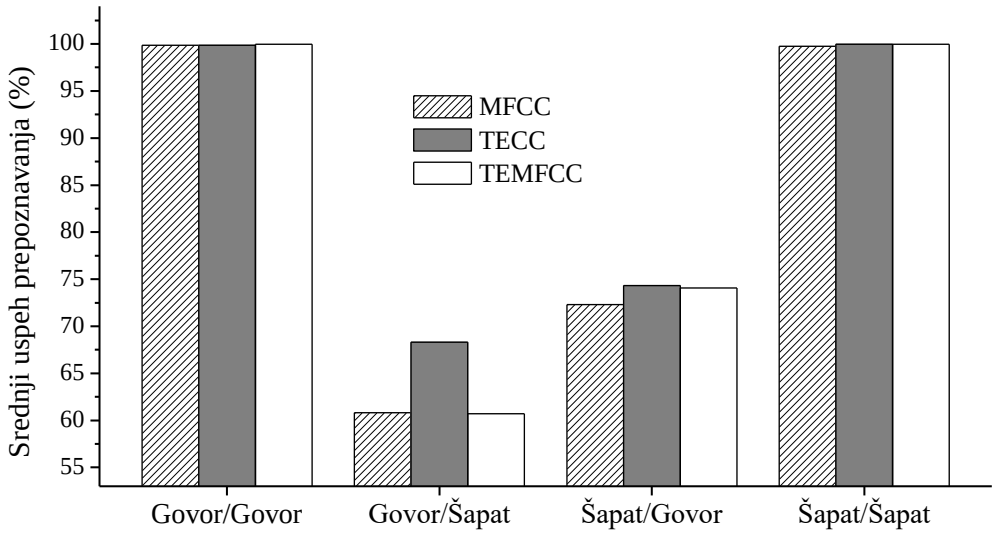
Usrednjeni rezultati prepoznavanja reči u neusaglašenim obuka/test scenarijima u zavisnosti od upotrebljenih govornih obeležja. (izraženo u %).

Govorni mod (obuka/test)	Govor (govor/šapat)	Srednja vr.	Šapat (šapat/govor)	Srednja vr.
MFCC	40,20	51,89	56,16	64,24
MFCC+Δ	54,68		64,24	
MFCC+Δ+ΔΔ	60,80		72,32	
TECC	48,44***	59,35	62,50	68,69
TECC+Δ	61,30***		69,24	
TECC+Δ+ΔΔ	68,32**		74,32	
TEMFCC	41,84*	52,49	57,58	65,69
TEMFCC+Δ	54,92*		65,42	
TEMFCC+Δ+ΔΔ	60,72		74,06	

($p < 0.05$ *; $p < 0.01$ **; $p < 0.006$ *** ; Interval poverenja = 95%)

Za razliku od usaglašenih obuka/test scenarija, neusaglašeni scenariji pokazuju značajno niži uspeh u prepoznavanju reči. Analizom rezultata iz tabele, došlo se do tri bitne opservacije. Prvo, dodavanje dinamičkih obeležja statičkim primetno poboljšava uspeh prepoznavanja reči u oba govorna moda kod svih obeležja: MFCC, TEMFCC i TECC. Drugo, u šapat/govor scenariju obeležja zasnovana na TECC imaju prilično veći uspeh u prepoznavanju reči u šapatu, u proseku 68,69% u poređenju sa 64,24% (MFCC) i 65,69% (TEMFCC). Primećene razlike su čak veće u scenariju govor/šapat, gde TECC obeležja imaju prosečni uspeh od 59,35% u poređenju sa 51,89% (MFCC) i 52,49%

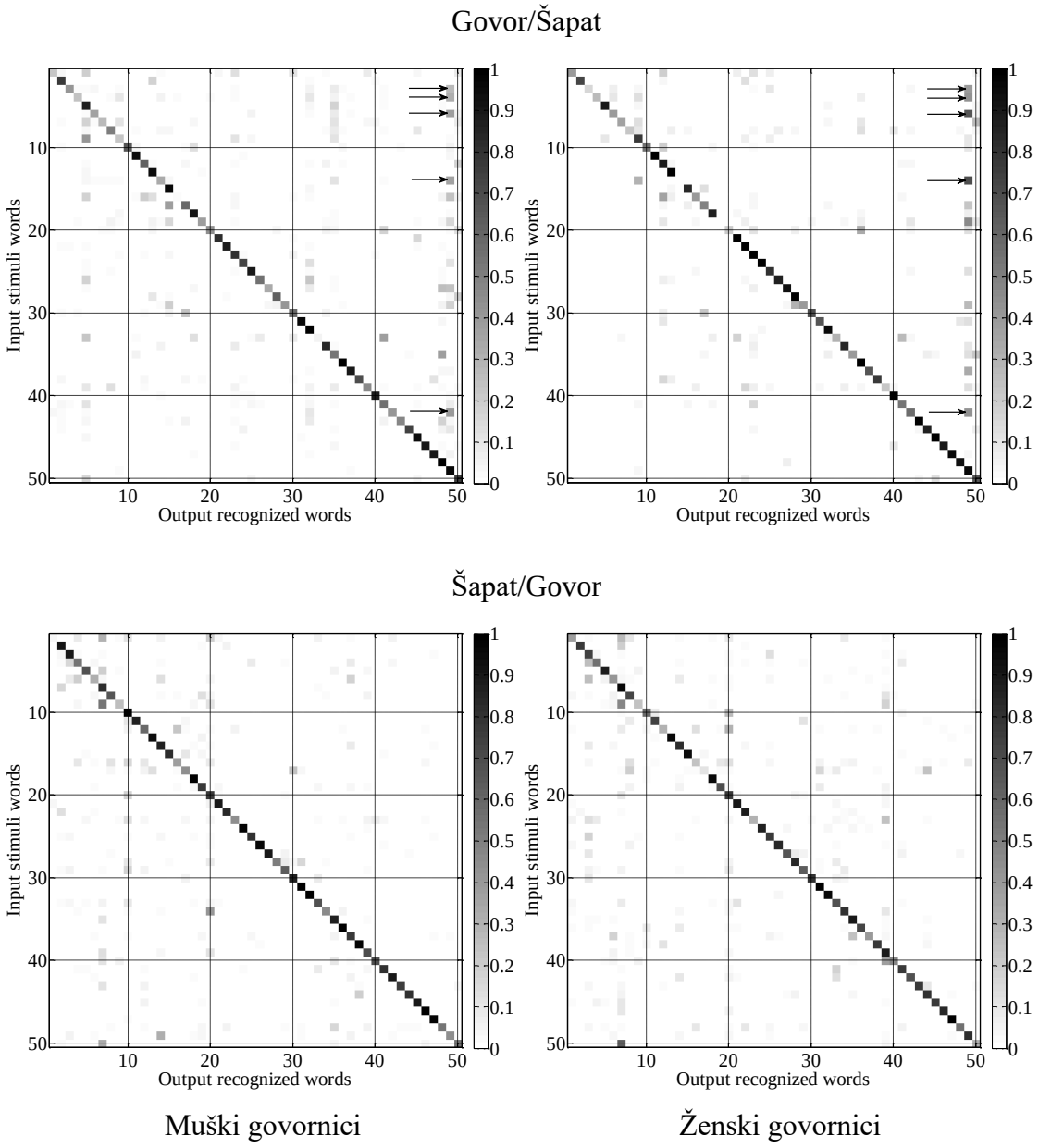
(TEMFCC). Ova tvrdnja je statistički potvrđena Wilcoxon testom (*p*-vrednosti su naznačene zvezdicama u Tabeli 8.4) čime je još jednom istaknuto da TECC obeležja najbolje modeluju karakteristike šapata. Treće, TECC obeležje pokazuje najmanju razliku između postignutih uspeha u različitim obuka/test scenarijima, posebno kada se dodaju njegova *delta* i *delta-delta* obeležja i kada ta razlika iznosi 6%, u poređenju sa 11,52% u slučaju MFCC i 13,34% u sučaju TEMFCC obeležja. Dodavanjem dinamičkih obeležja takođe se smanjuje razlika u uspehu prepoznavanja reči između MFCC, TEMFCC i TECC obeležja. Bolji uvid u odnos uspeha prepoznavanja reči u različitim obuka/test scenarijima prilikom upotrebe dinamičkih obeležja se može sagledati na Slici 8.1. Relativan odnos ovih rezultata je u saglasnosti sa rezultataima druga dva istraživanja sprovedena na HMM sistemima [Ito et al., 2005; Galić et al., 2014]. Naime, u radu [Ito et al., 2005], HMM sistem je takođe pokazao veći uspeh u prepoznavanju reči u šapat/govor scenariju (53%) u poređenju sa govor/šapat scenariom (19%). Slični zapažanja su ustanovljena i u radu [Galić et al., 2014].



Slika 8.1 Uspeh prepoznavanja reči u različitim obuka/test scenarijima prilikom korišćenja proširenog seta obeležja (obeležje+Δ+ΔΔ) za oba pola govornika.

8.3 ANALIZA MATRICA KONFUZIJE

Ova sekcija opisuje postupak analize konfuzije reči u neusaglašenim obuka/test scenarijima. Na Slici 8.2 su prikazane grafičke predstave matrica konfuzije za muške i ženske govornike u slučaju korišćenja MFCC obeležja.

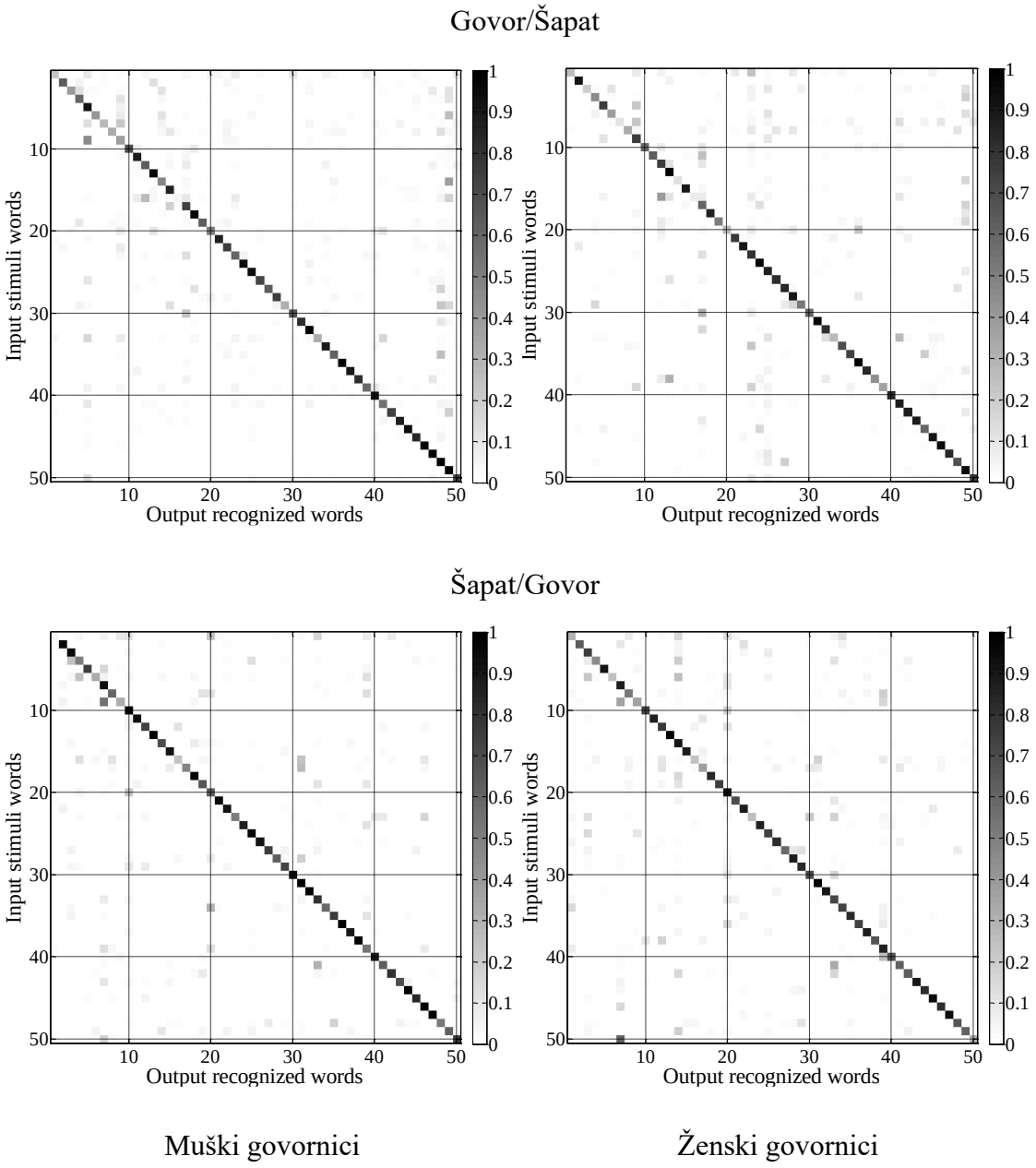


Slika 8.2 Matrice konfuzija prepoznavanja reči u dva obuka/test scenarija (govor/šapat i šapat/govor) u slučaju korišćenja MFCC obeležja. Skala sive boje sa desne strane matrica definiše opseg verovatnoća uspešnog prepoznavanja reči od 1 do 0. [Grozdić et al., 2013 b]

Svaki red matrice definiše slučaj nailaska određene stimulus reči na ulaz MLP sistema, dok kolone matrice predstavljaju očekivani izlaz neuralne mreže za zadati stimulus, odnosno prepoznatu reč. Redosled reči u matricama je formiran prema numeraciji reči u Tabeli P1.1 (videti prilog teze) koja definiše redosled reči u Whi-Spe korpusu. Za razliku od tabelarnog prikaza matrica konfuzije, grafičke predstave matrica su dosta lakše za vizuelnu analizu i olakšavaju uvid u performanse neuralnih mreža u prepoznavanju reči, pri čemu se jasno vide ispravno i pogrešno klasifikovanje reči. Indikator verovatnoće tačnog prepoznavanja reči je skala sive boje, koja sa tamnijom nijansom ilustruje veću verovatnoću ispravnog klasifikovanja reči, tako da crna boja predstavlja maksimum verovatnoće ($p = 1$), a bela minimum verovatnoće ($p = 0$).

Posmatranjem grafičkih predstava matrica konfuzije, jasno se ističu tamne dijagonale koje predstavljaju ispravno klasifikovane reći. Pored dijagonala, vide se i nešto svetlije vertikalne linije i tačke, koje ukaziju na pojedine greške i konfuzije u prepoznavanju reči i one su posebno istaknute u govor/šapat scenariju. Analiza pogrešno klasifikovanih reči u ovom scenariju otkriva 49. kolonu koja odgovara reči "sef". Detaljnija analiza je otkrila još nekoliko često pogrešno klasifikovanih stimulus reči od kojih su se najviše istakle reči: "crna" pozicionirana u 3. redu matrice, "crvena" u 4. redu, "zelena" u 6. redu, "sedam" u 14. redu i "svetlo" u 42. redu. Zajedničko za sve ove navedene reči je konfuzija sa rečju "sef" (49. red u matrici) i one su markirane strelicama na 49. koloni u matricama konfuzije. Otkriveno je da se ove konfuzije bez izuzetka javljaju kod svih govornika. Međutim, spomenuta pogrešna klasifikacija reči nije primetna u šapat/govor scenariju, što samo po sebi postavlja sledeće pitanje: Šta je uzrok ovih konfuzija i zašto se one ne pojavljuju i u šapat/govor scenariju?

Sa druge strane na Slici 8.3. su prikazane grafičke predstave matrica konfuzija u prilikom korišćenja TECC obeležja. U ovom slučaju tačnost prepoznavanja reči je veća, a broj pogrešno klasifikovanih reči je manji (videti Sliku 8.1). Ova činjenica još jednom ukazuje na to da je TECC obeležje bolje od MFCC u prepoznavanju šapata. Razlog za manju konfuziju reči prilikom korišćenja TECC obeležja je drugačiji način ekstrakcije informacija iz reči u govoru i šapatu, usled korišćenja TEO i *Gammatone* banke filtara.

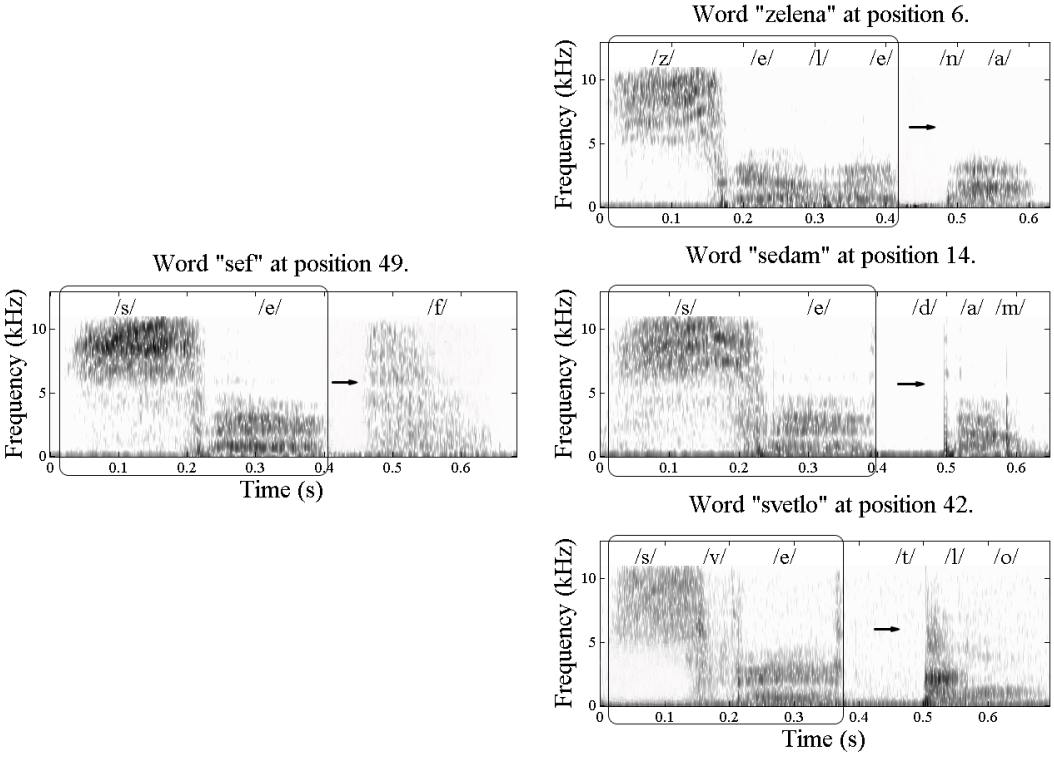


Slika 8.3 Matrice konfuzija prepoznavanja reči u dva obuka/test scenarija (govor/šapat i šapat/govor) u slučaju korišćenja TECC obeležja. Skala sive boje sa desne strane matrica definiše na opseg verovatnoća uspešnog prepoznavanja reči od 1 do 0. [Grozdić et al., 2013 b]

8.4 SPEKTRALNA ANALIZA KRITIČNIH PAROVA REČI

U cilju tumačenja posebno istaknutih grešaka u klasifikaciji reči, sprovedena je spektralna analiza za sledeće reči: “zelena”, “sedam” i “svetlo”, koje su kod svih govornika u scenariju govor/šapat prilikom prepoznavanja bile u konfuziji sa rečju “sef”. Analizu je obavio ekspert iz oblasti fonetike i digitalne obrade govornih signala

sa velikim iskustvom u forenzici govora. Na Slici 8.4 je dat primer izgleda spektrograma izgovora pomenutih reči jednog od ženskih govornika iz Whi-Spe baze.



Slika 8.4 Primer poređenja spektrograma izgovora jednog od ženskih govornika za reči: "zelena", "sedam" i "svetlo" koje su najčeće bile u konfuziji sa rečju "sef" u scenariju govor/šapat (Slika 7.2). Uokvireni delovi i strelice ukazuju na sličnosti pojedinih segmenata reči. [Grozdić et al., 2013 b]

Kao što se može videti sa spektrograma, početni segmenti ovih reči dosta liče jedni na druge i na slici su uokvireni radi lakšeg poređenja. Karakteristično za sve nabrojane reči je da počinju frikativima /s/ ili /z/ koje odlikuje veoma slična spektralna struktura. Frikativ /z/ je zvučni parnjak glasa /s/ i u šapatu je njihova konfuzija česta pojava koja nastaje kao posledica sličnog načina artikulacije. Na spektrogramima se mogu prepoznati po izraženim energetskim koncentratima na frekvencijama između 5 kHz i 10 kHz. Dalje, deo spektrograma koji odgovara vokalu /e/ u reči "sef" dosta liči na spektrograme segmenata "ele" i "ve" u rečima "zelena" i "svetlo", respektivno. U reči "zelena", likvid /l/ ima sličnu formantnu strukturu kao vokal /e/, pa u svom medijalnom položaju održava formantnu strukturu vokala /e/ koji ga okružuju. Sonant /v/, koji se nalazi između glasova /s/ i /e/ u reči "svetlo", se ponaša kao kratka pauza u spektrogramu, pa ne utiče mnogo na distinkciju segmenata "se" u reči "sef" i "sve" u reči "svetlo".

Završni segmenti reči imaju sekundarnu ulogu u pojavi konfuzije među ovim rečima i na Slici 8.4 su markirani strelicama. Naime, strukture segmenata: “na”, “dam” i “tlo”, koji predstavljaju završetke reči: “zelena”, “sedam” i “svetlo”, imaju međusobno slične spektralne strukture. One su karakteristične po slabom, opadajućem intenzitetu, koji je specifičan i sličan slučaju devokalizacije u normalnom govoru. U tom pogledu, završni segmenti ovih reči se mogu porediti u intenzitetskom pogledu sa izgovorom frikativa /f/ u reči “sef”, jer se reč “sef” u svojoj izolovanoj artikulaciji završava sa neutralnim glasom /ə/, koji ima sličnu formantnu strukturu kao vokali /a/ i /o/.

Ova analiza je pokazala veliku spektralnu sličnost reči “zelena”, “sedam” i “svetlo” sa rečju “sef” u šapatu, što uzrokuje pojavu njihove konfuzije u scnariju govor/šapat (Slika 8.2). Takođe, potrebno je imati u vidu da na ovu kofuziju u velikoj meri utiče segmentacija reči na jednak broj frejmova, što doprinosi normalizaciji reči uprkos njihovom drugačijem trajanju (videti poglavlje 6.1.1). Vizuelno, ova normalizacija je primetna sa spektrograma na Slici 8.4.

8.5 RAZLIKA U NEUSAGLAŠENIM OBUKA/TEST SCENARIJIMA

Prethodni eksperiment i analiza spektrograma su razjasnili pojavu konfuzije pojedinih parova reči, međutim pitanje odsustva ovih istaknutih grešaka u scenariju šapat/govor je ostalo i dalje nerazjašnjeno. Naime, uspeh MLP sistema u prepoznavanju reči je primetno veći u scenariju šapat/govor nego u scenariju govor/šapat, što se da lako uočiti sa Slike 8.1. Ta razlika je prisutna kod sva tri tipa govornih obeležja, približno je ista za sve govornike iz Whi-Spe baze i iznosi 11,52% za MFCC, 13,34% za TEMFCC i 6,0% za TECC. Izbacivanjem svih kritičnih parova reči kod kojih je konfuzija bila učestala i izražena, i naknadnim ponavljanjem obuke mreže i njenim testiranjem u neusaglašenim obuka/test scenarijima ova razlika je ostala nepromenjena. Na ovaj način su otklonjene sumnje da je pomenuta razlika posledica lošeg prepoznavanja pojedinih kritičnih parova reči, a u prilog tome ide i činjnica da su ponovo ustanovljene razlike u prepoznavanju reči u različitim neusaglašenim obuka/test scenarijima i dalje međusobno slične kod svih govornika. Ovaj fenomen je primećen u još dve studije [Ito et al., 2005; Galić et al., 2014], ali njegovo tumačenje i dublja analiza nisu urađeni.

8.5.1 *POSTAVKA HIPOTEZE O ZVUČNOSTI*

Za prethodno opisani fenomen u ovoj studiji je predloženo a kasnije i eksperimentalno dokazano posebno tumačenje. U scenariju šapat/govor MLP sistem je treniran na šapatu koji je u potpunosti bezvučan (videti Sliku 5.6), dok na njegov ulaz dolaze stimulusi govora sačinjeni od zvučnih i bezvučnih glasova. U ovoj situaciji, za neuralnu mrežu je dovoljno da u procesu klasifikacije prepozna obrasce bezvučnih glasova, dok su zvučne informacije suvišne. Sličnim rezonom se može zaključiti da u govor/šapat scenariju mreža očekuje da primi pored zvučnih i bezvučne glasove, ali umesto toga dobija samo bezvučne stimuluse šapata. Usled ovoga, neuralne mreže imaju poteškoće da ispravno prepoznaju reči u šapatu. Otuda MLP sistem ima više uspeha u prepoznavanju reči u šapat/govor scenariju nego u scenariju govor/šapat (pogledati Sliku 8.1), pa samim tim i manju konfuziju u prepoznavanju reči, kao što se vidi sa Slike 8.2. Vodeći se ovom idejom postavljena je sledeća hipoteza:

Neuralna mreža koja je trenirana sa šapatom ima veći uspeh u prepoznavanju izolovanih reči u šapat/govor scenariju nego neuralna mreža koja je obučena sa normalnim govorom u scenariju govor/šapat, jer je najveći deo informacija šapata (bezvučni glasovi) sadržan u normalnom govoru, što nije i obrnuti slučaj.

Prema tome uzrok ove razlike u neusaglašenim obuka/test scenarijima je zvučnost koje nema u šapatu, pa je iz tog razloga prepoznavanje šapata u govor/šapat scenariju dosta više degradirano.

8.5.2 *DOKAZ HIPOTEZE POMOĆU INVERZNOG FILTRIRANJA*

Jedan od načina da se dokaže prethodno postavljena hipoteza je da se u govornim stimulusima iz Whi-Spe korpusa umanji uticaj zvučnosti, tj. da se u akustičkom pogledu govor učini sličnijim šapatu, a zatim da se tako modifikovan govor primeni u obuci MLP sistema koji će kasnije biti testiran sa stimulusima u šapatu. Na ovaj način, zbog smanjene spektralne razlike između govora i šapata, pre svega u pogledu zvučnosti, očekuje se smanjenje konfuzije i razlike u neusaglašenim obuka/test scenarijima, što će se odraziti i na porast uspeha u prepoznavanju reči. Spektralni nagib šapata je jedna od najistaknutijih akustičkih karakteristika šapata. Sa Slike 4.5 se vidi da je on veoma blag i skoro ravan. Sa druge strane, normalni govor je karakterističan po

strmom spektralnom nagibu koji je posledica zvučnosti i dominantan je u domenu prva četiri formanta, tj. na frekvencijama ispod 5 kHz, Slika 4.4. Kako bi se spektri govora učinili sličnijim šapatu, neophodno je smanjiti ovaj spektralni nagib. U tu svrhu se može primeniti tehnika poravnanja spektra, poznata kao inverzno filtriranje, "spektralno izbeljivanje" (*spectral whitening*) ili "spektralno balansiranje" (*spectral balancing*) koja je u čestoj upotrebi u digitalnoj obradi slike i zvuka [Havelock et al, 2000; Vaseghi, 2008; Havelock et al., 2009]. U osnovi ove tehnike je inverzni filtar čije je kreiranje detaljno opisano u Poglavlju 5.6.1. Inverzni filtar, IF(z), je u proceduri predobrade signala implementiran odmah posle bloka za segmentaciju reči, na mestu preemfazis filtra, Slika 6.7. Ovakav MLP sistem sa izmenjenim delom za predobradu signala je nazvan MLP-IF sistem. Pored smanjenja spektralnog nagiba, inverznim filtriranjem se iz govornih signala uklanjaju efekti vokalnog trakta (1/H(z)) i ističu karakteristike pobudnog signala. Iz tog razloga, inverznim filtrom su pored stimulusa govora procesirani i stimulusi šapata, kako bi svi signali u istoj meri bili spektralno izmenjeni.

8.5.3 REZULATATI INVERZNOG FILTRIRANJA

Sa ovako korigovanim spektrima stimulusa iz Whi-Spe baze, eksperimenti automatskog prepoznavanja izolovanih reči su ponovljeni. Novi rezultati u usaglašenim obuka/test scenarijima su veoma slični onim iz Tabele 8.2. Na primer u govor/govor scenariju, prosečni uspeh u prepoznavanju reči za sve govornike je 99,82%, dok u slučaju šapat/šapat scenarija iznosi 99,76% (kada su TECC+Δ+ΔΔ obeležja korišćena). Prema tome, uspeh u usaglašenim obuka/test scenarijima je ostao nepromenjen.

Što se tiče rezultata u neusaglašenim obuka/test scenarijima, oni su kod MLP-IF sistema vidljivo poboljšani sa inverznim filtriranjem i prikazani su u Tabeli 8.5. Važno je zapaziti tri činjenice. Prvo, svi rezultati prepoznavanja izolovanih reči su posle inverznog filtriranja bolji. Sudeći prema Z i p-vrednostima iz *Wilcoxon* testa, ovo poboljšanje je statistički značajno samo u govor/šapat scenariju (Tabela 8.6), što je posledica potiskivanja zvučnosti tokom inverznog filtriranja. Drugo, TECC obeležja su još jednom pokazala najbolje rezultate, pri čemu je u šapat/govor scenariju ostvaren maksimalni uspeh prepoznavanja reči od 76,2%. Treće, razlika u uspesima prepoznavanja reči između različitih neusaglašenih obuka/test scenarija je dosta

smanjena. Ova razlika je najmanja u slučaju TECC obeležja, gde se uspesi prepoznavanja reči u šapat/govor i govor/šapat scenarijima razlikuju za samo 2,9%.

Tabela 8.5

Poređenje uspeha prepoznavanja reči u neusaglašenim obuka/test scenarija pre i posle inverznog filtriranja (izraženo u %).

obuka/test scenario	Govor/Šapat		Šapat/Govor		razlike u neusaglašenim scnarijima	
filtriranje	pre	posle	pre	posle	pre	posle
MFCC+Δ+ΔΔ	60,80	70,28	72,32	75,34	11,52	5,06
TEMFCC+Δ+ΔΔ	60,72	68,3	74,06	74,22	13,34	5,92
TECC+Δ+ΔΔ	68,32	73,3	74,32	76,2	6	2.9

Tabela 8.6

Rezultati Wilcoxon testa u poređenju uspeha prepoznavanja reči u neusaglašenim obuka/test scenarijima pre i posle inverznog filtriranja.

obuka/test scenario	Govor/Šapat		Šapat/Govor	
	Z	*p*	*Z*	*p*
MFCC+Δ+ΔΔ	(*Z*=-2.805;	***p*=0.005**)	(*Z*=-0.764;	*p*=0.445)
TECC+Δ+ΔΔ	(*Z*=-2.805;	***p*=0.005**)	(*Z*=-0.204;	*p*=0.838)
TEMFCC+Δ+ΔΔ	(*Z*=-2.805;	***p*=0.005**)	(*Z*=-0.255;	*p*=0.799)

(Interval poverenja = 95%)

Poboljšani uspeh prepoznavanja reči, redukovana konfuzija reči i smanjena razlika u neusaglašenim obuka/test scenarijima posle inverznog filtriranja potvrđuju hipotezu da je zvučnost u stimulusima govora glavni uzrok degradiranog prepoznavanja šapata u govor/šapat scenariju.

Druga, dosta poznatija metoda kompenzacije štetnih efekata neusaglašenosti kanala (npr. razlika u uslovima pod kojima se odvijaju obuka i testiranje ASR sistema) je normalizacija srednje kepstralne vrednosti (*Cepstral Mean Normalization - CMN* [11]). Inverzno filtriranje se može posmatrati i iz drugog ugla kao deljenje spektra sa usrednjenim spektrom (u ovom slučaju LPC spektrom). Poznato je da je ovo deljenje u frekvencijskom domenu jednako oduzimanju u log-spektralnom domenu, pa je otuda inverzno filtriranje slično oduzimanju srednje kepstralne vrednosti u kepstralnom domenu. Vodeći se ovom idejom, bilo je interesantno porediti inverzno filtriranje sa

[11] U literaturi se često sreće pod nazivom *Cepstral Mean Subtraction* (CMS).

konvencionalnom CMN tehnikom u neusaglašenim govor/šapat i šapat/govor scenarijima. Iz tog razloga je CMN implementiran kao sastavni deo predobrade govornih signala, a novi dobijeni rezultati uspeha prepoznavanja reči su prikazani u Tabeli 8.7. Dobijeni rezultati pokazuju da primena CMN tehnike poboljšava uspeh prepoznavanja reči jedino u govor/šapat scenariju i to samo u slučaju korišćenja MFCC+Δ+ΔΔ i TEMFCC+Δ+ΔΔ obeležja. Štaviše, ovo poboljšanje je manje nego ono koje se postiže sa inverznim filtriranjem. U šapat/govor scenariju, prepoznavanje reči je čak lošije nego pre upotrebe CMN. Wilcoxon statistički test, prikazan u Tabeli 8.8, ukazuje da CMN tehnika nema statistički značajan doprinos u poboljšanju uspeha prepoznavanja reči u oba neusaglašena scenarija.

Tabela 8.7

Poređenje uspeha prepoznavanja reči u neusaglašenim obuka/test scenarija pre i posle implementacije CMN (izraženo u %).

obuka/test scenario	Govor/Šapat		Šapat/Govor		razlike u neusaglašenim scnarijima	
filtriranje	pre	posle	pre	posle	pre	posle
MFCC+Δ+ΔΔ	60.80	65.35	72.32	69.13	11.52	3.78
TECC+Δ+ΔΔ	68.32	68.52	74.32	70.40	6	1.88
TEMFCC+Δ+ΔΔ	60.72	66.38	74.06	69.81	13.34	3.43

Tabela 8.8

Rezultati Wilcoxon testa u poređenju uspeha prepoznavanja reči u neusaglašenim obuka/test scenarijima pre i primene CMN.

obuka/test scenario	Govor/Šapat	Šapat/Govor
	Z-value *p*-value	*Z*-value *p*-value
MFCC+Δ+ΔΔ	(*Z*=-0.255; *p*=0.799)	(*Z*=-0.764; *p*=0.445)
TECC+Δ+ΔΔ	(*Z*=-0.357; *p*=0.721)	(*Z*=-0.968; *p*=0.333)
TEMFCC+Δ+ΔΔ	(*Z*=-0.765; *p*=0.444)	(*Z*=-0.561; *p*=0.575)

(Interval poverenja = 95%)

Postoji jedno moguće objašnjenje ovog slabijeg prepoznavanja u slučaju korišćenja CMN metode. U studiji [Boril et al., 2010] je ustanovljeno da upotreba CMN degradira uspeh u prepoznavanju reči u slučaju neusaglašenih obuka/test scenarija, ukoliko je koeficijent nagiba (*skewness*) kepstralnih raspodela podataka koji su korišćeni u procesu obuke i onih koji su upotrebljeni u postupku testiranja ASR sistema

drugačiji. Upravo se ovaj slučaj javlja u neusaglašenim scenarijima sa normalnim govorom i šapatom. Naime, kepstralne raspodele zvučnih glasova su nagnute ka desnoj strani, dok raspodele bezvučnih glasova uglavnom naginju ka levoj, na osnovu čega se ove raspodele mogu jasno diskriminisati [Boril et al., 2010]. Posebno važi da raspodele prvih kepstralnih koeficijenta u normalnom govoru imaju istaknutu karakteristiku nesimetričnosti (nagnuti su na jednu stranu) i multimodalnosti (imaju više modova koji se ogledaju u brežuljkastom izgledu raspodela) [Boril et al., 2010], što se da videti na Slici 5.6. Posle normalizacije kepstralnih koeficijenata govornih stimulusa u vidu oduzimanja njihove srednje vrednosti, raspodele kepstralnih koeficijenata u normalnom govoru i šapatu su usklađene tako da se nalaze oko iste pozicije (u ovom slučaju oko nule). Međutim, oblici ovih raspodela kao i njihovi nagibi ostaju identični onima na Slici 5.6. Sa druge strane, posle inverznog filtriranja, kepstralne raspodele postaju više simetrične i poprimaju oblik Gausove funkcije (pogledati sliku 5.9). Prema tome, nesimetričnost je smanjena, kao nagib i zvučnost, pa su kepstralne raspodele u normalnom govoru i šapatu postale dosta sličnije. Ovo objašnjava zašto je inverzno filtriranje bolje rešenje od CMN tehnike u neusaglašenim obuka/test scenarijima normalnog govora i šapata.

9 EKSPERIMENTI SA TANDEM DNN-HMM SISTEMOM

U ovom poglavlju su prezentovani rezultati eksperimenata automatskog prepoznavanja izolovanih reči u normalnom govoru i šapatu pomoću sistema baziranog na tandem DNN-HMM sistemu. Za razliku od eksperimenata sa MLP sistemom, gde je poboljšanje prepoznavanja šapata u neusaglašenim scenarijima postignuto potiskivanjem zvučnosti, odnosno inverznim filtriranjem normalnog govora, u slučaju DNN-HMM sistema se upravo radi suprotno – iz šapata se pomoću dubinskog *denoising* autoenkodera rekonstruišu karakteristike normalnog govora. Rezultati postignutog uspeha prepoznavanja reči u usaglašenim i neusaglašenim obuka/test scenarijima, kao i performanse DNN-HMM sistema u zavisnosti od upotrebe tri različita tipa kepstralnih koeficijenata (MFCC, TECC i TEMFCC) su prikazani u nastavku ovog poglavlja. Svi prezentovani rezultati predstavljaju usrednjene vrednosti dobijene 10-fold kros validacijom, pri čemu standardna greška odstupanja od srednje vrednosti SEM (*standard error of the mean*) nije bila veća od 0,18%.

9.1 USAGLAŠENI OBUKA/TEST SCENARIJI

U ovom poglavlju su prezentovani rezultati uspeha tandem DNN-HMM sistema u prepoznavanju izolovanih reči iz Whi-Spe korpusa u usaglašenim obuka/test

scenarijima – govor/govor i šapat/šapat. Usrednjeni vrednosti za muške i ženske govornike su prikazani u Tebeli 9.1.

Tabela 9.1

Uspeh prepoznavanja reči u usaglašenim obuka/test scenarijima (izražen u %) za muške i ženske govornike u zavisnosti od upotrebljenih govornih obeležja.

Govorni mod (obuka/test)	Govor (govor/govor)		Šapat (šapat/šapat)	
Pol govornika	Muški govornici	Ženski govornici	Muški govornici	Ženski govornici
MFCC	99,92	99,58	99,37	99,45
MFCC+Δ	99,80	99,78	99,76	99,52
MFCC+Δ+ΔΔ	99,81	99,85	99,70	99,86
TECC	99,84	99,90	99,59	99,67
TECC+Δ	99,90	99,90	99,71	99,63
TECC+Δ+ΔΔ	99,88	100	99,93	99,89
TEMFCC	99,72	99,80	99,37	99,61
TEMFCC+Δ	99,90	99,74	99,55	99,67
TEMFCC+Δ+ΔΔ	99,91	99,83	99,79	99,97

Kao što je očekivano, kod oba pola govornika je zabeležen visok uspeh prepoznavanja izolovanih reči u oba govorna moda. Slično kao i kod MLP sistema, analizom usrednjenih vrednosti u Tabeli 9.2 se mogu ustanoviti dve blage tendencije.

Tabela 9.2

Usrednjeni rezultati prepoznavanju reči u usaglašenim obuka/test scenarijima za MFCC, TECC i TEMFCC i njihova srodna obeležja (usrednjeno za sve govornike i izraženo u %).

Govorni mod (obuka/test)	Govor (govor/govor)	Srednja vr.	Šapat (šapat/šapat)	Srednja vr.
MFCC	99,75		99,41	
MFCC+Δ	99,79	99,79	99,64	99,61
MFCC+Δ+ΔΔ	99,83		99,78	
TECC	99,87		99,63	
TECC+Δ	99,90	99,90	99,67	99,73
TECC+Δ+ΔΔ	99,94		99,91	
TEMFCC	99,76		99,49	
TEMFCC+Δ	99,82	99,82	99,61	99,66
TEMFCC+Δ+ΔΔ	99,87		99,88	

Prvo, MFCC obeležja pokazuju nešto slabiji uspeh u prepoznavanju reči u oba govorna moda u poređenju sa druga dva govorna obeležja. Drugo, analizirajući rezultate prepoznavanja šapata, TECC obeležja pokazuju najviše uspeha (u proseku 99,73%) što

još jednom potvrđuje prednosti TECC u opisivanju karakteristike šapata. Kako bi se dobijeni rezultati statistički potvrdili primenjen je *Wilcoxon* test. *Z* i *p*-vredosti *Wilcoxon* testa su prikazani u Tabeli 9.3. Rezultati ovog testa kod svih govornika potvrđuju da su Teager obeležja u poređenju sa MFCC od statističkog značaja ($p < 0.05$) u prepoznavanju šapata. TECC+Δ+ΔΔ obeležja su pokazala najveću statističku značajnost u poređenju sa ostalim obeležjima (Z=-2.536; p=0.011). Slično kao i u analizi MLP sistema i ovde je ustanovljeno da u govor/govor scenariju izbor nekog od govornih obeležja nema statistički značaj.

Tabela 9.3

Rezultati Wilcoxon testa u poređenju uspeha prepoznavanja reči u usaglašenim obuka/test scenarijima u zavisnosti od korišćenja različitih govornih obeležja.

Govorni mod (obuka/test)	Govor (speech/speech)	Šapat (šapat/ šapat)
	Z *p*	*Z* *p*
MFCC	(/ ; /)	(/ ; /)
MFCC+Δ	(*Z*=-1.691; *p*=0.131)	(*Z*=-1.510; *p*=0.131)
MFCC+Δ+ΔΔ	(*Z*=-1.298; *p*=0.194)	(*Z*=-0.476; *p*=0.789)
TECC	(*Z*=-0.352; *p*=0.725)	(*Z*=-2.151; ***p*=0.031**)
TECC+Δ	(*Z*=-0.317; *p*=0.749)	(*Z*=-2.207; ***p*=0.027**)
TECC+Δ+ΔΔ	(*Z*=-0.610; *p*=0.474)	(*Z*=-2.536; ***p*=0.011**)
TEMFCC	(*Z*=-1.172; *p*=0.177)	(*Z*=-2.207; ***p*=0.027**)
TEMFCC+Δ	(*Z*=-1.497; *p*=0.115)	(*Z*=-2.151; ***p*=0.031**)
TEMFCC+Δ+ΔΔ	(*Z*=-1.604; *p*=0.109)	(*Z*=-2.371; ***p*=0.018**)

(Interval poverenja = 95%)

9.2 NEUSAGLAŠENI OBUKA/TEST SCENARIJI

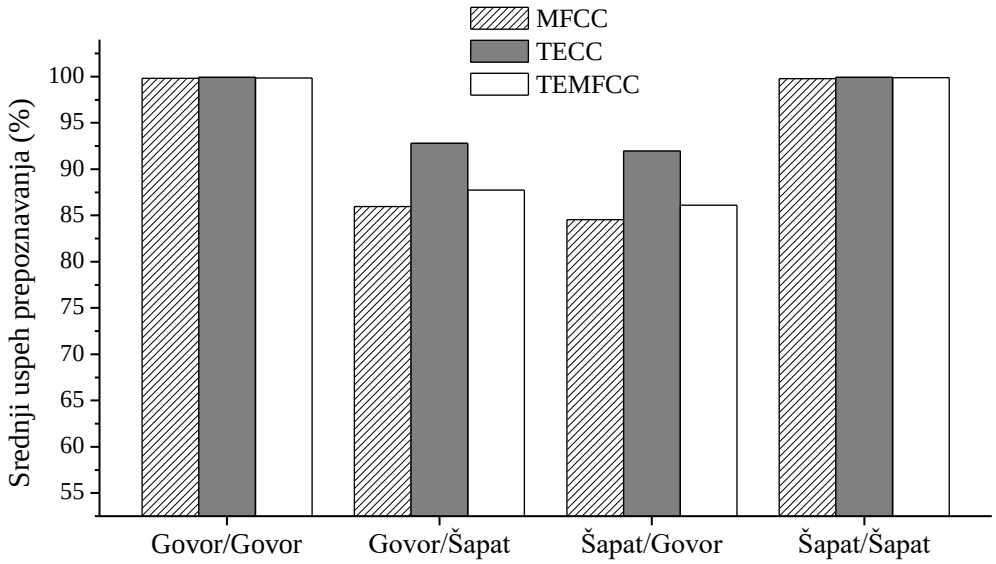
Ostvareni rezultati uspeha prepoznavanja reči iz Whi-Spe baze u neusaglašenim scenarijima govor/šapat i šapat/govor pomoću tandem DNN-HMM sistema su prikazani u Tabeli 9.4. Analizom se može doći do tri bitne opservacije. Prvo, bez obzira na upotrebljeni tip obeležja, uspeh u prepoznavanju reči je veoma sličan u oba neusaglašena scenarija. Razlika između uspeha prepoznavanja izolovanih reči u govor/šapat i šapat/govor scenariju je u proseku svedena na svega 0,67% i najmanja je u slučaju TECC obeležja (0,11%). Ovi rezultati potvrđuju benefit korišćenja dva ekstraktora, pri čemu je ekstrakcija robustnih govornih obeležja pomoću dubinskog *denoising* autoenkodera zaslužna za nivelisanje razlike u uspehu prepoznavanja reči u neusaglašenim obuka/test scenarijima. Jasniji uvid u odnos prepoznavanja reči u različitim obuka/test scenarijima je prikazan na Slici 9.1.

Tabela 9.4

Usrednjeni rezultati prepoznavanja reči u neusaglašenim obuka/test scenarijima u zavisnosti od upotrebljenih govornih obeležja. (izraženo u %).

Govorni mod (obuka/test)	Govor (govor/šapat)	Srednja vr.	Šapat (šapat/govor)	Srednja vr.
MFCC	68,15		67,45	
MFCC+Δ	75,84	76,65	75,13	75,71
MFCC+Δ+ΔΔ	85,97**		84,56	
TECC	79,12***		79,63	
TECC+Δ	86,87***	86,27	86,91**	86,16
TECC+Δ+ΔΔ	92,81**		91,95	
TEMFCC	68,54*		67,87	
TEMFCC+Δ	75,83*	77,37	75,21	76,40
TEMFCC+Δ+ΔΔ	87,73***		86,12	

($p < 0.05$ *; $p < 0.01$ **; $p < 0.006$ *** ; Interval poverenja = 95%)



Slika 9.1 Uspeh prepoznavanja reči u različitim obuka/test scenarijima sa tandem DNN-HMM sistemom prilikom korišćenja proširenog seta obeležja (obeležje+Δ+ΔΔ).

Drugo, u oba govorna moda i kod sva tri tipa kepstralnih obeležja, dodavanje dinamičkih obeležja statičkim je rezultovalo poboljšanjem uspeha prepoznavanja reči. Naime, dodavanje delta obeležja u proseku poboljšava prepoznavanje reči za 7,5% u oba govorna moda, dok dodavanje delta-delta obeležja poboljšava za još dodatnih 8,9%. Treće, pojedinačno upoređujući ostvarene rezultate MFCC, TECC i TEMFCC obeležja, TECC obeležje je ponovo demonstriralo najbolje performanse. U scenariju govor/šapat uspeh prepoznavanja reči u slučaju TECC obeležja je u proseku iznosio 86,27%, dok je kod MFCC i TEMFCC obeležja 76,65% i 77,37%. U scenariju šapat/govor je 86,16%

za TECC obeležja, naspram 75,71% i 76,40% za MFCC i TEMFCC obeležja. Ova prednost TECC obeležja je i statistički potvrđena *Wilcoxon* testom (*p*-vrednosti su naznačene zvezdicama u Tabeli 9.4).

10 KOMPARACIJA REZULTATA PREPOZNAVANJA ŠAPATA

U ovom poglavlju je predstavljena komparacija maksimalnih ostvarenih rezultata tri različita sitema za prepoznavanje izolovanih reči baziranih na veštačkim neuralnim mrežama i rezultata dobijenih pomoću konvencionalnog GMM-HMM sistema. U pitanju su MLP sistem, zatim MLP-IF sistem, odnosno MLP sistem sa predloženim izmenama u *front-end* delu u vidu inverznog filtriranja i tandem DNN-HMM sistem. Što se tiče samog GMM-HMM sistema sa kojim je vršeno poređenje, on je identičan *back-end* delu tandem DNN-HMM sistema opisanog u Poglavlju 7.3. Performanse ovih prepoznavača su upoređene u usaglašenim i neusaglašenim obuka/test scenarijima pri korišćenju TECC+Δ+ΔΔ obeležja koja su se kod svih pomenutih sistema pokazala kao najbolja u prepoznavanju šapata. Postupak obuke i treniranja kod svih ASR sistema je bio identičan i obavljen je korišćenjem uzoraka iz Whi-Spe baze. Ujedno, ovo poglavlje daje i komparaciju sa još nekim postojećim ASR sistemima iz literature, poput HTK-HMM [Galić et al., 2014] i DTW [Marković et al., 2013], koji su takođe testirani sa Whi-Spe bazom. U ovom slučaju komparacija je obavljena na osnovu rezultata koji su dobijeni pri upotrebi MFCC+Δ+ΔΔ obeležja.

10.1 REZULTATI PREPOZNAVANJA U USAGLAŠENIM OBUKA/TEST SCENARIJIMA

Maksimalni postignuti uspeh prepoznavanja reči u usaglašenim obuka/test scenarijima kod MLP, MLP-IF i DNN-HMM je postignut sa TECC+Δ+ΔΔ obeležjima. Komparacija njihovih performansi i poređenje sa GMM-HMM sistemom su prikazani u Tabeli 10.1.

Tabela 10.1

Maksimalni uspeh prepoznavanja izolovanih reči u usaglašenim obuka/test scenarijima za MLP, MLP-IF, GMM-HMM, i DNN-HMM sisteme pri korišćenju TECC+Δ+ΔΔ obeležja (izraženo u %) .

Govorni mod (obuka/test)	Govor (govor/govor)			Šapat (šapat/šapat)		
Pol govornika	Muški govornici	Ženski govornici	Svi govornici	Muški govornici	Ženski govornici	Svi govornici
MLP	99.90	99,80	99.85	100	99,9	99,95
MLP + IF	99,84	99,80	99,82	99,72	99,80	99,76
GMM-HMM	99.96	99.94	99,95	99,97	99.91	99,94
DNN-HMM	99,88	100	99,94	99,93	99,89	99,91

Kao što se vidi iz tabele, rezultati prepoznavanja reči se ne razlikuju mnogo za muške i ženske govornike. Srednja vrednost uspeha prepoznavanja reči pomoću MLP sistema u normalnom govoru za sve govornike iznosi 99,85%, dok je u šapatu 99,95%. U slučaju dodatnog inverznog filtriranja ove vrednosti su za nijansu niže. Naime, MLP-IF sistem je zabeležio tačnost od 99,82% u prepoznavanju reči u normalnom govoru i 99,76% u šapatu. Što se tiče HMM sistema, on je u proseku bio za svega 0,1% bolji u prepoznavanju govora od MLP sistema, gde je ostvario uspeh od 99,95% tačnosti dok je ta vrednost u prepoznavanju šapata 99,94%. Slične performanse ima i DNN-HMM sistem sa 99,94% uspeha u perpoznavanju govora i 99,91% u prepoznavanju šapata.
Prema tome prosečni uspeh prepoznavanja reči u usaglašenim obuka/test scenarijima za sve testirane ASR sisteme je bio sličan i očekivano visok.

U literaturi se spominju još dva ASR sistema koji su testirani sa Whi-Spe bazom. U pitanju su HMM sistem realizovan pomoću HTK softverskog paketa, takozvani HTK-HMM [Galić et al., 2014] i DTW sistem [Marković et al., 2013]. Oba sistema su testirana u usaglašenim obuka/test scenarijima sa MFCC+Δ+ΔΔ, a njihovo poređenje sa MLP, MLP-IF i DNN-HMM sistemom je prikazano u Tabeli 10.2. Analizom priloženih

rezultata došlo se do istih zapažanja – u usaglašenim scenarijima svi sistemi su demonstrirali slične performanse i visok uspeh prepoznavanja reči kako u normalnom govoru tako i u šapatu.

Tabela 10.2

Poređenje rezultata MLP, MLP-IF, DNN-HMM sistema sa rezultatima HTK-HMM i DTW sistema u usaglašenim obuka/test scenarijima pri korišćenju MFCC+Δ+ΔΔ obeležja (izraženo u %).

Govorni mod (obuka/test)	Govor (govor/govor)	Šapat (šapat/šapat)
MLP	99,85%	99,75%
MLP-IF	99,79%	99,71%
DNN-HMM	99,83%	99,78%
HTK-HMM[12]	99,66%	98,52%
DTW[13]	99,65%	94,05%

Jedina uočena razlika je konstatovana u prepoznavanju šapata sa DTW sistemom koji je demonstrirao nešto slabije prepoznavanje u odnosu na ostale ASR sisteme (lošije u proseku za 5%).

10.2 REZULTATI PREPOZNAVANJA U NEUSAGLAŠENIM OBUKA/TEST SCENARIJIMA

Najbolji rezultati perpoznavanja reči u neusaglašenim obuka/test scenarijima kod MLP, MLP-IF i DNN-HMM su postignuti sa TECC+Δ+ΔΔ obeležjima. Poređenje ovih rezultata i njihova komparacija sa rezultatima ostvarenih sa GMM-HMM sistemom su prikazani u Tabeli 10.3.

Tabela 10.3

Maksimalni uspeh prepoznavanja izolovanih reči u neusaglašenim obuka/test scenarijima za MLP, MLP-IF, GMM-HMM, i DNN-HMM sisteme pri korišćenju TECC+Δ+ΔΔ obeležja (izraženo u %) .

Govorni mod (obuka/test)	Normalan govor (govor/šapat)			Šapat (šapat/govor)		
Pol govornika	Muški govornici	Ženski govornici	Svi govornici	Muški govornici	Ženski govornici	Svi govornici
MLP	68,84	67,80	68,32	76,64	72,00	74,32
MLP + IF	75,20	71,40	73,30	77,84	74,56	76,20
GMM-HMM	71,46	71,20	71,33	78,81	75,89	77,35
DNN-HMM	93,12	92,50	92,81	92,21	91,69	91.95

[12] Rezultat HTK-HMM sistema preuzet iz rada [Galić et al., 2014].
[13] Napomena: Rezultat DTW sistema je dobijen analizom 4 govornika (2 muška i 2 ženska) iz Whi-Spe baze i preuzet je iz rada [Marković et al., 2013].

Iz priložene tabele se vidi da MLP sistem ima prosečni uspeh u govor/šapat scenariju od 68,32% a nešto veći u šapat/govor scenariju, 74,32%. Razlika u ostvarenom uspehu između ovih scenarija iznosi 6% i donekle je smanjena sa inverznim filtriranjem. Naime, MLP-IF sistem ostvaruje 73,3% uspeha u govor/šapat a 76,2% u šapat govor scenariju. Prema tome sa inverznim filtriranjem razlika između rezultata neusaglašenih scenarija je smanjena na 2,9%, dok je celokupno prepoznavanje reči poboljšano u govor/šapat scenariju za 4,98% a u šapat/govor za 1,88%. Nešto bolje rezultate u govor/šapat scenariju bez inverznog filtriranja je demonstrirao GMM-HMM sistem i to tačnost od 77,35% u prepoznavanju reči. Ipak, GMM-HMM je ostvario nešto slabije rezultate u govor/šapat scenariju od MLP-IF sistema, tačnije 71,33% uspeha u prepoznavanju reči. Interesantno je da je razlika između rezultata u neusaglašenim obuka/test scenarijima kod GMM-HMM sistema ostala ista kao i u slučaju MLP sistema bez inverznog filtriranja i iznosi 6%. DNN-HMM sistem je sa druge strane umesto potiskivanja zvučnosti iz govora, činio suprotno – vršio je rekonstruisanje karakteristika govora u šapatu i na taj način ostvario znatna poboljšanja u prepoznavanju reči, kao i smanjenje razlike u neusaglašenim obuka/test scenarijima. DNN-HMM sistem je demonstrirao najbolje performanse od svih ASR sistema u govor/šapat i šapat/govor scenarijima ostvarivši respektivno uspeh od 92,81% i 91,95% u tačnom prepoznavanju reči. Sa ovim rezultatima, DNN-HMM sistem je u poređenju sa MLP-IF sistemom pokazao u proseku za 19,51% bolje performanse u prepoznavanju reči u govor/šapat scenariju i 17,63% u šapat/govor scenariju. Takođe, DNN-HMM je nivelisao razliku u uspehu prepoznavanja reči između neusaglašenih obuka/test scenarija pri čemu je ta razlika sada svedena na svega 0,86%.

Što se tiče poređenja rezultata između muških i ženskih govornika, manje razlike su uočene kod MLP, MLP-IF i GMM-HMM sistema u šapat/govor scenariju gde je prepoznavanje reči muških govornika bilo bolje od ženskih u proseku za 3%. Slično je zabeleženo i u govor/šapat scenariju kod MLP-IF sistema. Pomenute razlike u zavisnosti od pola govornika nisu uočene kod rezultata DNN-HMM sistema.

U literaturi se pominju još dva tipa ASR sistema, HTK-HMM [Galić et al., 2014] i DTW sistem [Marković et al., 2013], koji su testirani u neusaglašenim obuka/test scenarijima sa Whi-Spe bazom i MFCC+Δ+ΔΔ obeležjima. Komparacija

rezultata ovih ASR sistema sa rezultatima MLP, MLP-IF i DNN-HMM sistema je data u Tabeli 10.4. Očekivano, DTW sistem kao predstavnih *template-based* ASR sistema, usled svojih skromnih mogućnosti u akustičkom modelovanju je imao najmanji uspeh u neusaglašenim obuka/test scenarijima i to 34,05% uspeha u govor/šapat scenariju i 26,15%. Sa druge strane HTK-HMM sistem je demonstrirao bolje rezultate. On je u scenariju šapat/govor ostvario 73,9% uspeha u prepoznavanju reči, što je slično performansama MLP sistema (72,32%) ali znatno manje u poređenju sa rezultatima MLP-IF i DNN-HMM sistema (75,34% i 84,56%). Što se tiče govor/šapat scenarija, HTK-HMM je posle DTW sistema imao namanji uspeh u prepoznavanju reči i to 57,16%.

Tabela 10.4

Poređenje rezultata MLP, MLP-IF, DNN-HMM sistema sa rezultatima HTK-HMM i DTW sistema u neusaglašenim obuka/test scenarijima pri korišćenju MFCC+Δ+ΔΔ obeležja (izraženo u %)

Govorni mod (obuka/test)	govor/šapat	šapat/govor
MLP	60,80	72,32
MLP-IF	70,28	75,34
DNN-HMM	85,97	84,56
HTK-HMM[14]	57,16	73,90
DTW[15]	34.05	26.15

[14] Rezultat HTK-HMM sistema preuzet iz rada [Galić et al., 2014].
[15] Napomena: Rezultat DTW sistema je dobijen analizom 4 govornika (2 muška i 2 ženska) iz Whi-Spe baze i preuzet je iz rada [Marković et al., 2013].

11 ZAKLJUČAK

11.1 PREGLED REZULTATA

Usled velikih razlika u produkciji šapata u odnosu na normalan govor, performanse tradicionalnih ASR sistema obučenih na normalnom govoru, značajno degradiraju prilikom prepoznavanja šapata. U cilju rešavanja navedenog problema i razvoja efikasnog sistema za prepoznavanje šapata, neophodno je suštinsko razumevanje akustičkih karakteristika i razlika između ova dva govorna moda. Iz tog razloga, ova teza je prvo prikazala analizu akustičkih karakteristika izolovanih reči u šapatu. Za potrebe ovih eksperimenata je snimljen specijalan korpus reči u šapatu za srpski jezik (Whi-Spe), koji sadrži ukupno 10000 snimaka izgovora reči. Korpus je koncipiran tako da se sastoji iz dva dela - iz snimaka normalnog govora i šapata, koji su sakupljeni od 10 govornika kojima je srpski maternji jezik i koji imaju ispravnu artikualciju i čulo sluha. Profesionalna oprema, laboratorijski uslovi snimanja, kao i preduzete mere u kontrolisanju kvaliteta i obrade snimaka su rezultovale kreiranjem jedne od malobrojnih kvalitetno snimljenih, fonetski balansiranih i sistematski uređenih baza snimaka šapata u svetu, koja ujedno predstavlja i jedinu postojeću bazu tog tipa za srpski jezik. Sprovedene akustičke analize na Whi-Spe korpusu su potvrdile istaknute karakakteristike šapata, poput dosta niže energije i SNR u poređenju sa normalnim govorom posebno u slučaju izgovora vokala i zvučnih konsonanata. U šapatu je utvrđeno da su niži formanti pomereni ka višim frekvencijama. Kepstralna analiza, pre

svega ispitivanje raspodela c_1 kepstralnih koeficijenata, i upoređivanje dugovremenih usrednjenih spektara snimaka je pokazala da je spektralni nagib u šapatu dosta ravniji nego u govoru. Takođe, analiza kepstralne distance je otkrila da se karakteristike zvučnih glasova više menjaju u šapatu od bezvučnih glasova koji ostaju gotovo nepromenjeni. Iako se problem niske energije u šapatu može rešiti približavanjem mikrofona ustima, spomenute spektralne razlike između govora i šapata ostaju nepromenjene. Pretpostavljeno je, a potom i eksperimentalno dokazano u radu sa višeslojnim perceptronima, da upravo ove razlike u zvučnosti predstavljaju ključni problem slabijeg prepoznavanja šapata.

Što se tiče samog kreiranja ASR sistema za prepoznavanje šapata, u ovoj tezi su dizajnirana dva odvojena rešenja bazirana na veštačkim neuralnim mrežama. Prvi sistem predstavlja MLP sistem koji je dodatno unapređen sa inverznim filtriranjem, dok je drugi realizovan u vidu tandem DNN-HMM sistema koji koristi poseban tip dubinskih neuralnih mreža u svom *front-end* delu, poznat kao dubinski *denoising* autoenkoder. Oba sistema su dizajnirana kao *speaker-dependent* sistemi i testirani su sa tri različita tipa kepstralnih koeficijenata: MFCC, TECC i TEMFCC. Prvi tip predstavlja tradicionalne Mel-frekvencijske kepstralne koeficijene, dok su druga dva tipa koncipirana na primeni *Teager* energetskog operatora (TEO) i do sada nisu analizirana u automatskom prepoznavanju šapata.

Obzirom da u dosadašnjim studijama neuralne mreže nisu ispitivane u zadacima automatskog prepoznavanja šapata, logičan izbor ASR sistema sa kojim su započeti eksperimenati ove teze su bili višeslojni perceptroni, odnosno MLP sistem kao jedan od najpopularnijih predstavnika neuralnih mreža. Performanse MLP sistema su analizirane u prepoznavanju izolovanih reči iz Whi-Spe baze u različitim obuka/test scenarijima, a potom su kroz dalje eksperimente poboljšane uz odgovarajući izbor govornih obeležja i izmene u *front-end* delu što je dovelo do kreiranja unapređenog MLP-IF sistema.

U usaglašenim obuka/test scenarijima, MLP sistem i sva tri tipa kepstralnih obeležja su pokazali podjednako dobre uspehe u prepoznavanju reči, u proseku 99.88% u normalnom govoru i šapatu. Prepoznavanje reči je bilo lošije u neusaglašenim scenarijima, čija analiza je od posebnog značaja sa aspekta tumačenja realnih problema testiranja ASR sistemima sa šapatom. Dodavanje *delta* i *delta-delta* kepstralnih obeležja

je poboljšalo tačnost prepoznavanja reči posebno u neusaglašenim obuka/test scenarijima. Primena TEO je potvrdila da efikasan i nelinearan način obrade govornog signala kao i mogućnost brzog praćenja modulacije energije može biti od velike koristi u prepoznavanju šapata. Zahvaljujući tome, TECC i TEMFCC obeležja su demonstrirala dosta bolji uspeh u prepoznavanju reči u neusaglašenim obuka/test scenarijima od tradicionalnih MFCC obeležja. Najviše se istaklo TECC+Δ+ΔΔ obeležje koje je obezbedilo naveći uspeh u prepoznavanju reči u govor/šapat i šapat/govor scenarijima, i to 68.32% i 74.32%, respektivno. Ovako postignute bolje performanse TECC u poređenju sa druga dva obeležja su objašnjena kao posledica korišćenja nelineranog TEO i *Gammatone* banke filtara.

Analiza pogrešno klasifikovanih reči na bazi tumačenja matrica konfuzija je tokom automatskog prepoznavanja reči otkrila bitne razlike između dva neusaglašena scenarija. Hipoteza o prisustvu zvučnosti u govoru kao glavnog uzroka degradirane tačnosti prepoznavanja u poređenju sa šapat/govor scenarijom je eksperimentalno dokazana pomoću inverznog filtriranja. Zahvaljujući ovoj analizi, predložene su izmene u *front-end* delu MLP sistema i kreiran je novi takozvani MLP-IF sistem. Naime, u cilju smanjenja spektralne razlike između dva govorna moda, inverzno filtriranje, takođe poznato i kao "spektralno izbeljivanje" (*spectral whitening*), je implementirano u procesu predobrade govornih signala. Inverzni filtar ima funkciju izravnavanja spektralnog nagiba u normalnom govoru čime snimci govora i šapata postaju međusobno sličniji u spektralnom domenu. Ova sličnost je dokazana analizama kepstralnih distanci i analizom raspodela prva dva kepstralna koeficijenta. Novi rezultati sa inverznim filtriranjem su pokazali ostvareni dobitak u uspehu prepoznavanja reči u neusaglašenim obuka/test scenarijima, koji u slučaju MFCC+Δ+ΔΔ obeležja iznosi 9,48%. Najviši uspeh u prepoznavanju reči je postignut sa TECC+Δ+ΔΔ obeležjima, 73.3% i 76.2% u govor/šapat i šapat/govor scenarijima, respektivno. Za razliku od konvencionalne CMN metode normalizacije srednje vrednosti kepstralnih koeficijenata, inverzno filtriranje se pokazalo kao bolje rešenje za neusaglašene obuka/test scenarije.

Tandem DNN-HMM sistem je takođe pristupio rešavanju problema prepoznavanja šapata u neusaglašenim obuka/test scenarijima sa aspekta smanjenja spektralnih razlika između normalnog govora i šapata kroz odgovarajuće izmene u

front-end delu i predobradi govornih signala. Međutim, za razliku od MLP-IF sistema koji je pomoću IF potiskivao zvučnost iz normalnog govora, DNN-HMM sistem radi upravo suprotno i rekonstruiše karakteristike govora. Mogućnost rekonstruisanja kepstralnih karakteristika normalnog govora iz kepstralnih uzoraka šapata je omogućeno zahvaljujući pravilno obučenoj dubinskoj strukturi *denoising* autoenkodera, koji u ulozi sekundrnog ekstraktora robustnih obeležja u *front-end* delu DNN-HMM sistema omogućava značajno poboljšanje uspeha prepoznavanja reči u neusaglašenim obuka/test scenarijima. Tako je DNN-HMM sa TECC+Δ+ΔΔ obeležjima demonstrirao najbolji uspeh u prepoznavanju reči od 92,81% tačnosti u govor/šapat scenariju i 91,95% u šapat/govor scenariju. Poredeći rezultate DNN-HMM sa ostalim ASR sistemima, on je demonstrirao poboljšanje od 24,49% u govor/šapat i 17,63% u šapat/govor scenariju u poređenju sa MLP sistemom, dok poboljšanja u odnosu na rezultate MLP-IF sistema iznose 19,51% i 15,75% respektivno. U poređenju sa tradicionalnim GMM-HMM sistemom, DNN-HMM je u proseku ostvario bolje rezultate za 21,48% u govor/šapat scenariju, a 14,6% u šapat/govor scenariju. Ukoliko uporedimo rezultate tandem DNN-HMM sistema sa TECC+Δ+ΔΔ obeležjima i rezultate GMM-HMM sistema sa MFCC+Δ+ΔΔ obeležjima (Tabela P1.2 u Prilogu), razlika u govor/šapat scenariju dostiže čak 31% u korist DNN-HMM sistema [Grozdić et al., 2016 b]. Pored očiglednog poboljšanja u neusaglašenim scenarijima, tandem DNN-HMM sistem je zadržao dobre performanse i u usaglašenim obuka/test scenarijima sa ostvarenih 99,83% uspeha u prepoznavanju govora i 99,78% u prepoznavanju šapata.

Prema tome, rezultati ove studije su potvrdili superiornist DNN-HMM sistema koji je objedinio dobre karakteristika HMM i DNN sistema, pre svega mogućnost HMM sistema u vremenskom modelovanju reči i prednosti DNN sistema u akustičkom modelovanju reči, tačnije u rekonstrukciji akustičkih karakteristika normalnog govora iz šapata.

11.2 DOPRINOS DISERTACIJE

Današnji sistemi za automatsko prepoznavanje govora su osetljivi i nepouzdani u prepoznavanju bilo kog netipičnog oblika govora. Iz tog razloga automatsko prepoznavanje šapata, samim tim i ova doktorska disertacija, predstavlja bitnu i

aktuelnu istraživačku temu sa ciljem postizanja bolje komunikacije čovek-računar. U skladu sa tim, postignuti su sledeći naučni doprinosi doktorske disertacije:

- Disertacija je omogućila suštinsko razumevanje svojstava šapata u bimodalnoj govornoj komunikaciji, pojasnila je i na odgovarajući način kvantifikovala razliku akustičkih obeležja koja karakterišu šapat i govor u bimodalnoj (govor-šapat) komunikaciji;

- Teorijski su razjašnjenji i eksperimentalno ispitani slabiji rezultati prepoznavanja u neusaglašenim obuka/test (govor/šapat i šapat/govor) scenarijima;

- Predložena su nova, robustnija obeležja od tradicionalnih MFCC, koja zahvaljujući karakteristikama *Teager* energije i *Gammatone* banke filtara mnogo bolje akustički modeluju šapat i na taj način poboljšavaju prepoznavnanje u neusaglašenim obuka/test scenarijima;

- Disertacija je po prvi put demonstrirala primenu neuralnih mreža u automatskom prepoznavanju šapata. Testirana su dva sistema bazirana na neuralnim mrežama – višeslojni perceptroni (MLP) i tandem DNN-HMM sistem. Perfomanse ovih sistema u prepoznavanju šapata su upoređene sa još nekim postojećim ASR sistemima (DTW i GMM-HMM).

- Predložena je nova metoda kompenzacije neželjenih razlika akustičkih obeležja u neusaglašenim obuka/test scenarijima, zasnovana na inverznom filtriranju, koja dodatno poboljšava prepoznavanje bimodalnog govora (govor-šapat). Inverzno filtriranje je ujedno poslužilo i kao novi metod kreiranja pseudo-šapata, čime se pružaju mogućnosti brzog i jednostavnog kreiranja velikog broja uzoraka veštačkog šapata čime se rešava problem nepostojanja odgovarajuće baze realnih snimaka šapata neophodnih za adaptaciju ASR sistema.

- Predložen je novi efikasni sistem za prepoznavanje šapata, baziran na tandem DNN-HMM sistemu koji poseduje niz prednosti u poređenju sa drugim ASR sistemima poput: (i) značajnog poboljšanja tačnosti prepoznavanja reči u neusaglašenim obuka/test scenarijima, (ii) izuzetno visokih performansi u usaglašenim obuka/test scenarijima, (iii) mogućnosti efikasne rekonstrukcije

kepstralnih karakteristika normalnog govora iz uzoraka šapata u realnom vremenu, (iv) brze i jednostavne obuke sistema pomoću uzoraka pseudo-šapata čime je otklonjena potreba za realnim uzorcima šapata.

Prisutnost i kombinovanje većeg broja modaliteta govora predstavlja veliki problem za sisteme automatskog prepoznavanja govora. Eksperimentalna istraživanja ove doktorske disertacije u prepoznavanju bimodalnog govora na relaciji neutralan govor-šapat razrešila je samo deo ovog problema, pri čemu rezultati koji su dobijeni uz pomoć neuralnih mreža treba da budu osnova za dalje analize u primeni drugih algoritama prepoznavanja.

11.3 MOGUĆNOST DALJIH ISTRAŽIVANJA

Sumirajući postignute rezultate i doprinose ove doktorske disertacije, zaključuje se da je istraživanje na ovom veoma složenom i zahtevnom problemu prepoznavanja šapata ostavilo veliki broj mogućih pravaca kojima se dalje u istraživačkom pogledu može nastaviti. Pre svega, autor ove teze ima na umu proširenje Whi-Spe baze i testiranje kompleksnijih zadataka poput prepoznavanja kontinualnog šapata pomoću tandem i hibridnih DNN—HMM sistema nezavisnih od govornika. Prostora za dalja istraživanja ima i u testiranju različitih tehnika normalizacije kepstralnih koeficijenata, u analizi pseudo-šapata i adaptaciji ASR sistema sa pseudo-šapatom, u različitom filtriranju govornih signala i sintezi zvučnih glasova iz šapata, čime bi se dodatno poboljšalo prepoznavanje šapata u neusaglašenim obuka/test scenarijima. Kao neistražena tema preostaje i analiza automatskog prepoznavanja govornika u šapatu (identifikacija/autentifikacija govronika) sa pomenutim sistemima.

LITERATURA

[Ackley et al., 1985] Ackley, D., Hinton, G., and Sejnowski, T. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science* 9, 147-169. Reprinted in Anderson and Rosenfeld (1988).

[Aertsen et al., 1980] Aertsen, A.M., Johannesma, P.I. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog: I. Characterization of tonal and neural stimuli. *Biol. Cypern.* 38: 223-234.

[Ayadi et al., 2011] Ayadi, M.E., Kamel, M.S., Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes and databases. *Pattern Recognition*, vol. 44., pp. 572-587, 2011.

[Bahl et al., 1981] Bahl, L., Bakis, R., Cohen, P., Cole, A., Jelinek, F., Lewis, B., and Mercer, R. (1981). Speech Recognition of a Natural Text Read as Isolated Words. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1981.

[Baken et al., 1991] Baken, R., Daniloff, R. (1991). *Readings in clinical spectrography of speech.* San Diego.

[Baker, 1975] Baker, J.K. (1975). The dragon system - An overview. *IEEE Trans. Acoustics, Speech,Signal Proc.*, ASSP-23 (1):24-29, February 1975.

[Bakis, 1976] Bakis, R. (1976). Continuous speech word recognition via centiseconf acoustic states. In *Proc Asa Meeting*, Whashington, DC, April 1976.

[Barnard, 1992] Barnard, E. (1992). Optimization for Training Neural Networks. *IEEE Trans. on Neural Networks*, 3(2), March 1992.

[Baum et al., 1966] Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat*, 37: 1554-1563, 1966.

[Baum et al., 1967] Baum, L.E. and Egon, J.A. (1967). An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bull. Amer. Meteorol Soc*, 73 360-363, 1967.

[Baum et al., 1968] Baum, L.E. and Sell, G.R. (1968). Growth functions for transformations on manifolds. *Pac J. Math*, 27 (2):211-227,1968.

[Baum et al., 1970] Baum, L.E., Petrie, T., Soules, G., Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat*, 41 (1): 164-171, 1970.

[Baum, 1972] Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process. *Inequalities*, 3:1-8, 1972.

[Bengio, 2009] Bengio, Y., 2009. Learning deep architectures for AI. *Found. Trends Mach. Learn*. 2, 1–127. doi:10.1561/2200000006.

[Berry et al., 1997] Berry, M.J.A., Linoff, G. (1997). *Data mining techniques for marketing, sales and customer support*. John Wiley & Sons, New York.

[Bishop, 1995] Bishop, C., (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

[Blum, 1992] Blum, A., (1992). *Neural Networks in C++*, Wiley, New York.

[Bodenhausen et al., 1993] Bodenhausen, U., and Manke, S. (1993). Connectionist Architectural Learning for High Performance Character and Speech Recognition. *In Proc. IEEE International Confer- ence on Acoustics, Speech, and Signal Processing*, 1993.

[Boger et al., 1997] Boger, Z., Guterman, H. (1997). Knowledge extraction from artificial neural network models, in: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Orlando, pp. 3030-3035.

[Bonnot et al., 1991] Bonnot, J.F.P., Chevrie-Muller, C. (1991). Some effects of shouted and whispered conditions on temporal organization. *Jouranal of Phonetics*, vol. 19, pp. 473-483, 1991.

[Boril et al., 2010] Boril, H., Hansen, J.H.L., 2010. Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments. IEEE *Trans. Audio Speech Lang. Process*. 18, 1379–1393. DOI:10.1109/TASL.2009.2034770.

[Bregler et al., 1993] Bregler, C., Hild, H., Manke, S., and Waibel, A. (1993). Improving Connected Letter Recognition by Lipreading. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993.

[Cajal, 1892] Cajal, S. (1892). A New Concept of the Histology of the Central Nervous System. In Rottenberg and Hochberg (eds.), *Neurological Classics in Modern Translation*. New York: Hafner, 1977.

[Carlin et al., 2006] Carlin, M.A., Smolenski, B.Y., Wenndt, S.J. (2006). Unsupervised Detection of Whispered Speech in the Presence of Normal Phonation. In *Proc. Interspeech*, 2006.

[Catford, 1964] Catford, J.C. (1964). Phonation types: The classification of some laryneal components of speech production. In D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, and J.L.M Trim (Eds), *In hounur of Daniel Jones* (pp. 26-37). London, Lngmans.

[Côté, 2011] Côté, N. (2011). *Integral and Diagnostic Intrusive Prediction of Speech Quality*, Springer, Berlin.

[Cummins et al., 2006] Cummins, F., Grimaldi, M., Leonard, T., and Simko, J. (2006). The CHAINS corpus: CHAracterizing INdividual Speakers. In *Proc of SPECOM'06*, pp. 431–435, St. Petersburg, RU.

[Dannenbring, 1980] Dannenbring, G.L. (1980). Perceptual discrimination of whispered phoneme pairs. *Perceptual and Motor Skills*, vol. 51, pp. 979-985, 1980.

[de Boer et al., 1978] de Boer, E., de Jough, H.R. (1978). On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *J. Acoust. Soc. Am.* 63: 115-135.

[Demuth et al., 2008] Demuth, H., Beale, M., Hagan, M. (2008). *Neural network toolbox 6: User's guide*, The Mathworks, Inc., Natick.

[Deng et al., 2014] Deng, L., Yu, D., (2014). Deep learning: methods and applications. *Foundations and Trends in Signal Processing*, 7(3-4):197-387, 2014.

[Dimitriadis et al., 2005] Dimitriadis, D., Maragos, P., Potamianos, A. (2005). Auditory teager energy cepstrum coefficients for robust speech recognition, in: *Proceedings of Interspeech 2005*, Lisbon, pp. 3013–3016.

[Eklund et al., 1996] Eklund, I., Traunmuller, H. (1996). Comparative study of male and female whispered and phoned versions of the long vowels of Swedish. *Phonetica* 54, 1-21.

[Elman, 1990] Elman, J. (1990). Finding Structure in Time. *Cognitive Science*, 14(2):179-211, 1990.

[Fan et al., 2008] Fan, X., Hansen, J.H.L. (2008). Speaker identification for whispered speech based on frequency warping and score competition, in: *Proceedings of Interspeech 2008*, Brisbane, pp. 1313-1316.

[Fan et al., 2009 a] Fan, X., Hansen, J.H.L. (2009). Speaker identification for whispered speech using modified temporal patterns and MFCCs, in: *Proceedings of Interspeech 2009*, Brisbane, pp. 896-899.

[Fan et al., 2009 b] Fan, X., Hansen, J.H.L. (2009). Speaker identification with whispered speech based on modified LFCC parameters and feature mapping, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, pp. 4553-4556.

[Fan et al., 2011] Fan, X., Hansen, J.H.L. (2011). Speaker identification within whispered speech audio stream. *IEEE Transactions on Audio, Speech and Language Processing*, 19 (5), 1408-1421.

[Galić et al., 2013 a] Galić, J., Marković, B., Grozdić, Đ., Jovičić, S. (2013). The Influence of Feature Vector Selection on Performance of Atuomatic Recognition of Whispered Speech. in *Proc. of 4th International Conference on Fundamental and Applied. Aspects of Speech and Language*., Belgrade, 2013.

[Galić et al., 2013 b] Galić, J., Popović, M., Marković, B., Grozdić, Đ.T., Jovičić, S.T. (2013). Primjena skrivenih Markovljevih modela u prepoznavanju govora u šapatu. *INFOTEH*, Jahorina, str. 387 - 390, 2013.

[Galić et al., 2014] Galić, J., Jovičić, S.T., Grozdić, Đ.T., Marković, B. (2014). HTK-based recognition of whispered speech. in: *Proceedings of the 16th International Conference Speech and Computer, SPECOM 2014*, Novi Sad, Srbija.

[Ghaffarzadegan et al., 2014 a] Ghaffarzadegan, S., Boril, H., Hansen, J.H.L. (2014). UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech, in: *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Florense, pp. 2563-2567.

[Ghaffarzadegan et al., 2014 b] Ghaffarzadegan, S., Boril, H., Hansen, J.H.L. (2014). Model and feature based compensation for whispered speech recognition, In: *Proceedings of the Annual Conference International Speech Communication Association INTERSPEECH, 2014*. Singapore, pp. 2420–2424.

[Ghaffarzadegan et al., 2015] Ghaffarzadegan, S., Boril, H., Hansen, J.H.L., 2015. Generative modeling of pseudotarget domain adaptation samples for whispered speech recognition, In: *Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. Brisbane, Australia, pp. 5024–5024.

[Glasberg et al., 1990] Glasberg, B.R., Moore, B.C.J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* 47:103–138, 1990.

[Gray et al., 1976] Gray, A.H., Markel, J.D. (1976). Distance measures for speech processing. *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-24, pp. 380-391, Oct. 1976.

[Grozdić et al., 2011 a] Grozdić Đ.T., Jovičić S.T. (2011) The role of speech quality features and accented syllables in strong emotion discrimination, In: S.T. Jovičić, M. Subotić (eds): *Verbal Comunication Quality - Interdisciplinary Research I*; LAAC, IEPSP, Belgrade, ISBN 987-86-81879-34-4, pp. 19-41, 2011.

[Grozdić et al., 2011 b] Grozdić Đ.T., Jovičić S.T., Rajković M. (2011) Uticaj kvaliteta glasa na verbalnu ekspresuju emocija, In: *Proceedings of 19th Telecommunications forum,* TELFOR2011, Belgrade 2011, pp. 663-666.

[Grozdić et al., 2012] Grozdić Đ.T., Marković B., Galić J., Jovičić S. (2012). Primena neuralnih mreža u prepoznavanju govora u šapatu. In: *Proceedings of 20th telecommunications forum TELFOR2012*. Proceedings, 2012; Belgrade, ISBN 978-1-4673-2982-8: 728-731.

[Grozdić et al., 2013 a] Grozdić, Đ.T., Galić, J., Marković, B., Jovičić, S.T. (2013). Application of neural networks in whispered speech recognition. *Telfor Journal*, 5 (2), 103-106.

[Grozdić et al., 2013 b] Grozdić, Đ.T., Jovičić, S.T., Galić, J., Marković B. (2013). Experiments in whisper recognition using neural networks. In: S.T. Jovičić, M. Subotić, M. Sovilj (eds): *Verbal Communication Quality, Interdisciplinary Research, II*; CUŽA, Belgrade, ISBN 978-86-81879-46-7, pp. 91-110, 2013

[Grozdić et al., 2013 c] Grozdić, Đ.T., Marković, B., Galić, J., Jovičić, S.T., Furundžić, D. (2013). Neural network-based recognition of whispered speech, in: *Proceedings of 4th International Conference on Fundamental and Applied Aspects of Speech and Language*, Belgrade, pp. 223-229.

[Grozdić et al., 2014] Grozdić, Đ., Jovičić, S., Galić, J., Marković, B. (2014). Application of inverse filtering in enhancement of whisper recognition. *Symposium on Neural Network Applications in Electrical Engineering NEUREL2014. Proceedings*, 2014; Belgrade, ISBN 978-1-4799-5887-0: 157-161.

[Grozdić et al., 2015] Grozdić, Đ., Jovičić, S., Šumarac Pavlović, D., Galić, J., Marković, B. (2015). Komparacija tehnika normalizacije kepstralnih koeficijenata u automatskom prepoznavanju šapata. 59. konferencija za elektroniku, telekomunikacije, računarstvo, automatiku i nuklearnu tehniku *ETRAN2015. Zbornik radova*, 2015; Srebrno jezero, Srbija, ISBN 978-86-80509-72-3: AK 1.8. 1-5.

[Grozdić et al., 2016 a] Grozdić, Đ.T., Jovičić, S.T. (2016). Application of algorithms based on neural networks in whispered speech recognition. *Specific Applications of Information Technology and Signal Processing in Speech Disorder Diagnosis and Therapy*, Editors: Jovičić S., Šarić Z., Subotić M., ISBN 978-86-89431-14-8, LAAC&IEPSP, Belgrade, pp. 125-176.

[Grozdić et al., 2016 b] Grozdić, Đ.T., Jovičić, S.T., Subotić, M. (2016). Whispered Speech Recognition Using Deep Denoising Autoencoder. *Engineering Applications of Artificial Intelligence*, vol. 59 (1), 15-22. DOI: 10.1016/j.engappai.2016.12.012

[Grozdić et al., 2017] Grozdić, Đ.T., Jovičić, S.T., Šumarac Pavlović, D., Galić, J., Marković, D. (2017). Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition, *Advances in Electrical and Computer Engineering,* 2017; vol. 17 (1), 21-26. DOI: 10.4316/AECE.2017.01004

[Havelock et al., 2000] Havelock, D., Kuwano, S., Vorländer, M. (2000). Handbook of Signal Processing in Acoustics. Springer Science & Business Media, 2000.

[Havelock et al., 2009] Havelock, D., Kuwano, S., Vorländer M. (Eds.). (2009). *Handbook of Signal Processing in Acoustics*, Springer, New York.

[Hebb, 1949] Hebb, D. (1949). The Organization of Behavior. New York: Wiley. Partially reprinted in Anderson and Rosenfeld (1988).

[Heracleous,2009] Heracleous, P., (2009). Using Teager Energy Cepstrum and HMM distances in Automatic Speech Recognition and Analysis of Unvoiced Speech, (2009) 31–37.

[Higashikawa et al., 1996] Higashikawa, M., Nakai, K., Sakakura, A., Takahashi, H. (1996). Perceived pitch of whispered vowels. *Journal of Voice*, vol. 10, no. 2, pp. 155-158, 1996.

[Higashikawa et al., 1999] Higashikawa, M., Minifie, F.D. (1999). Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels. *Journal of Speech, Language and Hearing Research*, vol. 42, pp. 583-591, 1999.

[Higashikawa et al., 2003] Higashikawa, M., Green, J.R., Moore, C.A., Minifie, F.D. (2003). Lip kinematics for /p/ and /b/ production during whispered and voiced speech. *Foli Phoniatrica*, 2003.

[Hild et al., 1993] Hild, H. and Waibel, A. (1993). Connected Letter Recognition with a Multi-State Time Delay Neural Network. In *Advances in Neural Information Processing Systems 5*, Hanson, S., Cowan, J., and Giles, C.L. (eds), Margan Kaufmann Publishers.

[Hinton et al., 2006] Hinton, G., Osindero, S., Teh, Y., 2006. A fast learning algorithm for deep belief nets, *Neural Comput.* 18(7), 1527–1554. DOI: 10.1162/neco.2006.18.7.1527

[Hinton, 1989] Hinton, G. (1989). Connectionist Learning Procedures. *Artificial Intelligence* 40:1(3), 185-235.

[Hinton, 2002] Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 1771–1800. doi:10.1162/089976602760128018

[Holambe et al., 2012] Holambe, R.S., Deshpande, M.S. (2012). Noise Speaker Identification Using Nonlinear Modeling Techniques. In: Neustein, A., Patil, H.A. (Eds.): *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism,* Springer Science & Business Media.

[Holmes et al., 1983] Holmes, J.N., Stephens, A.P., 1983. Acoustic correlates of intonation in whispered speech. *J. Acoust. Soc. Am.* 73, S87.

[Hopfield, 1982] Hopfield, J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. National Academy of Sciences USA*, 79:2554-58, April 1982. Reprinted in Anderson and Rosenfeld (1988).

[Hultsch et al., 1992] Hultsch, H., Todt, D., Zuhlke, K. (1992). Einsatz und soziale Interpretation geflusterter Signale. In: Pawlik, K., Stapf,K.H. (Eds.), *Umwelt und Verhalten.* Bern, Gottingen, Toronto: Huber Berlag, pp. 391-406.

[Hwang et al., 1993] Hwang, M.Y., Huang, X.D., and Alleva, F. (1993). Predicting Unseen Triphones with Senones. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993.

[Idan et al., 1992] Idan, Y., Auger, J., Darbel, N., Sales, M., Chevallier, R., Dorizzi, B., and Cazuguel, G. (1992). Comparative Study of Neural Networks and Non-Parametric Statistical Methods for Off-Line Handwritten Character Recognition. In *Proc. International Conference on Arti- ficial Neural Networks*, 1992.

[Itakura, 1975] Itakura, F. (1975). Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 23(1): 67-72, February 1975. Reprinted in Waibel and Lee (1990).

[Ito et al., 2005] Ito, T., Takeda, K., Itakura, F. (2005). Analysis and recognition of whispered Speech. *Speech Communication* 45, 129-152.

[Itoh et al., 2002] Itoh, T., Takeda, K., Itakura, F. (2002). Acoustics analysis and recognition of whispered speech. *Acoustic Speech and Signal Processing*, vol. 1, pp. 389-392, 2002.

[Itou et al., 1998] Itou, K., Takeda, K., Takezawa, T., Matsuoka, T., Shikano, K., Kobayashi, T., Itahashi, S. (1998). Design and development of Japanese speech corpus for large vocabulary continuous speech recognition. In: *Proceedings of Oriental COCOSDA*, May 1998, Tsukuba.

[Jelinek et al., 1975] Jelinek, F., Bahl, L.R., Mercer, R.L. (1975) Design of a linquistic statistical decoder for the recognition of continous speech. *IEEE Trans. Information Theory*. IT-21: 250-256, 1975.

[Jin et al., 2007] Jin, Q., Jou, S.S., Schultz, T. (2007). Whispering speaker identification, in *IEEE Intl. Conf. on Multimedia and Expo*, Beijing, China, pp. 1027–1030, Jul. 2007

[Jovičić et al., 2004] Jovičić, S.T., Kašić, Z., Djordjević, M., Rajković , M. (2004). Serbian emotional speech database: design, processing and evaluation. In: *SPECOM-2004*, pp. 77–81. St. Petersburg, Russia (2004).

[Jovičić et al., 2008 a] Jovičić, S.T., Šarić, Z.M. (2008). Acoustic analysis of consonants in whispered speech. *Journal of Voice*, vol. 22, no. 3, pp. 263-274, 2008.

[Jovičić et al., 2008 b] Jovičić, S.T., Punišić, S., Šarić, Z. (2008). Time-frequency detection of stridence in fricatives and affricates. In: *Int. Conf. Acoustics 2008*, Paris, pp. 5137–5141 (2008).

[Jovičić, 1998] Jovičić, S.T. (1998). Formant feature differences between whispered and voiced sustained vowels. *Acustica-Acta* 84 (4), 739-743.

[Jovičić, 1999] Jovičić, S.T. (1999). *Govorna komunikacija - fiziologija, psihoakustika i percepcija*. Izdavačko preduzeće Nauka, Beograd, 1999.

[Kaiser, 1983] Kaiser, J.F. (1983). Some observations on vocal tract operation from a fluid flow point of view, in: Titze, I.R., Scherer, R.C. (Eds.), *Vocal Fold Physiology: Biomechanics,acoustics, and phonatory control*, Denver Center for the Performing Arts, Denver, pp. 358-386.

[Kallail et al., 1984 a] Kallail, K.J., Emanuel, F.W. (1984). An acoustic comparison od isolated whisper and phonated vowel samples produsedby adult male subjects. *Journal of Phonetics*, vol. 12, pp. 175-186, 1984.

[Kallail et al., 1984 b] Kallail, K.J., Emanuel, F.W. (1984). Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *J. Speech Hearing Res.* 27, 245-251.

[Karsoliya, 2012] Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *International Journal of Engineering Trends and Technology* 3 (6), 714-717.

[Kimura, 1990] Kimura, S. (1990). 100,000-Word Recognition Using Acoustic-Segment Networks. In P*roc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990.*

[Kitawaki et al., 1998] Kitawaki, N., Nagabuchi, H., Itoh, K. (1988). Objective quality evaluation for low-bit-rate speech coding systems, *IEEE J. Select. Areas Commun.*, vol. 6, pp.242-248, Feb. 1988.

[Klich, 1982] Klich, R.J. (1982). Effects of speech level and vowel context on intraoral air pressure in vocal and whispered speech. *Folia Phoniatrica,* vol. 34, pp. 33-40, 1982.

[Kohonen, 1989] Kohonen, T. (1989). *Self-Organization and Associative Memory* (3rd edition). Berlin: Springer-Verlag.

[Konno et al., 1996] Konno, H., Toyama, J., Shimbo, M., Murata, K. (1996). The effect of formant frequency and spectral tilt of unvoiced vowels on their perceived pitch and phonemic quality, in: *IEICE Technical Report*, SP95-140, pp. 39–45.

[Kurematsu et al., 1990] Kurematsu, A., Takeda, K., Kuwabara, H., Shikano, K., Sagisaka, Y., Katagiri, S. (1990). ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9 (4), 357–363.

[Lang, 1989] Lang, K., (1989). *A Time-Delay Neural Network Architecture for Speech Recognition*. PhD Thesis, Carnegie Mellon University.

[Lashley et al., 1980] Lashley, C., Hicks, D.M. (1980). Vibratory action of the vocal folds during whisper. *The IASCP Bulletin*, vol. 2, pp. 32-35, 1980.

[Lass et al., 1976] Lass, N.J., Hughes, K.R., Bowyer, M.F., Waters, L.T., Bourne, V.T. (1976). Speaker sex identification from voiced, whispered and filtered isolated vowels. *Journal of Acoustical Society of America*, vol. 59, pp. 675-678, March 1976.

[Laver, 1980] Laver, J. (1980). *The phonetic description of voice quality*. Cambridge, England: Cambridge University Press.

[Lee et al., 2014] Lee, P.X., Wee, D., Si, H., Toh, Y., Lim, B.P., Chen, N., Ma, B., College, V.J., 2014. A whispered Mandarin corpus for speech technology applications, In: *Proceedings of the Annual Conference International Speech Communication Association INTERSPEECH*, 2014. Singapore, pp. 1598–1602.

[Lee, 1988] Lee, K.F. (1988). *Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. PhD Thesis, Carnegie Mellon University.

[Leggetter et al., 1995] Leggetter, C.J., Woodland, P.C., 1995. Flexible speaker adaptation using maximum likelihood linear regression. In: *Proceedings of the ARPA Spoken Language Technology Workshop*, 1995, Barton Creek.

[Lim, 2011] Lim, B.P. (2011). *Computational Differences between Whispered and Non-whispered Speech*, Ph.D. Thesis, University of Illinois, 2011.

[Linggard, 1985] Linggard, R. (1985). *Electronic Synthesis of Speech*, Cambridge University Press, New York.

[Maragos et al., 1993] Maragos, P., Kaiser, J.F., Quatieri, T.F., 1993. Energy Separation in Signal Modulations with Applications to Speech Analysis. *IEEE Transaction on Signal Processing* 41(10), pp. 3024-3051.

[Markel et al., 1976] Markel, J., Gray A.H. (1976). *Linear Prediction of Speech*. Springer-Verlag, New York, USA, ISBN: 0-13-007444-6, 1976.

[Marković et al., 2013 a] Marković, B., Jovicić, S.T., Galić, J., Grozdić, Đ.T. (2013). Whispered Speech Database: Design, Processing and Application. In: *Proceedings of 16th International Conference TSD 2013*, pp. 591- 598. Springer, Pilsen, Czech Republic (2013)

[Marković et al., 2013 b] Marković, B., Galić, J., Grozdić, Đ., Jovičić, S. (2013). Application of DTW method for Whispered Speech Recognition. In *Proc. of 4th*

International Conference on Fundamental and Applied. Aspects of Speech and Language, Belgrade, 2013.

[Marković et al., 2014] Marković, B.R., Grozdić, Đ.T. (2014). The LPCC-DTW analysis for whispered speech recognition. In: *Proceedings of 1. International Conference on Electrical, Electronic and Computing Engineering IcETRAN 2014*, Vrnjačka Banja, Serbia, 2014.

[Marković et al., 2015] Marković, B., Jovičić S., Galić J., Grozdić, Đ. (2015). Recognition of the Multimodal Speech Based on the GFCC Features. in: *Proceedings of 2. International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2015*, Vrnjačka Banja, Serbia, 2015.

[Marković et al., 2016] Marković, B., Jovičić, S., Miomir M., Galić J., Grozdić, Đ. (2016). Recognition of Whispered Speech Based on PLP Features and DTW Algorithm. in: *Proceedings of 3. International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2016*, Vrnjačka Banja, Serbia, 2016, pp. AK1.2.1-4.

[Masters, 1993] Masters, T. (1993). *Practical neural network recipies in C++*, Academic Press, New York.

[Mathur et al., 2012] Mathur, A., Reddy, S.M., Hegde, R.M. (2012) Significance of parametric spectral ratio methods in detection and recognition of whispered speech, *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–20, 2012.

[Matsuda et al., 1999] Matsuda, M. and Kasuya, H., 1999. Acoustic nature of the whisper. In *Proc. of Eurospeech*, vol 1., pp. 137–140.

[Meyer-Eppler, 1957] Meyer-Eppler, W., 1957. Realisation of prosodic features in whispered speech. *J. Acoust. Soc. Am.* 29, 104–106.

[Mills, 2009] Mills, T.I.P. (2009). *Speech motor control variables in the production of voicing contrasts and emphatic accent*. Ph.D. dissertation, University of Edinburgh, August 2009.

[Minsky et al., 1969] Minsky, M. and Papert, S. (1969). *Perceptrons.* Cambridge: MIT Press. Partially reprinted in Anderson and Rosenfeld (1988).

[Minsky, 1967] Minsky, M. (1967). *Computation: Finite and Infinite Machines*. Englewood Cliffs: Prentice-Hall.

[Miyatake et al., 1990] Miyatake, M., Sawai, H., and Shikano, K. (1990). Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990.*

[Monoson et al., 1984] Monoson, P., Zemlin, W.R., (1984). Quantitative study of whisper. *Folia Phoniatrica*, vol. 36, pp. 53-65, 1984.

[Morris, 2003] Morris, R.W. (2003). *Enhancement and recognition of whispered speech*. Ph.D. dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, August 2003.

[Osfar, 2011] Osfar, M.J.O (2011). *Articulation of whispered alveolar consonants*. Master thesis, Unicersity of Illinois at Urbana-Chamaign, 2011.

[O'Shaughnessy, 1987] O'Shaughnessy, D., (1987). *Speech communication: human and machine*. Addison-Wesley. p. 150. ISBN 978-0-201-16520-3.

[Pan et al., 2007] Pan, S.T., Lai, C.C. (2007). Using genetic algorithm to improve the performance of speech recognition based on artificial neural network. In: Grimm, M., Kroschel, K.(Eds.): *Robust Speech Recognition and Understanding*, I-Tech Education and Publishing, Vienna, Austria.

[Patterson, 1944] Patterson, R.D. (1944). The sound of a sinusoid: Spectral models. *J. Acoust. Soc. Am.* 96, 1409-1418.

[Quatieri, 2002] Quatieri T.F. (2002). *Discrete-Time Speech Signal Processing*, Prentice Hall, NJ.

[Rabiner et al., 1993] Rabiner, L.R., Juang, B.H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[Rabiner et al., 2007] Rabiner, L.R., Schafer, R.W. (2007). Introduction to Digital Speech Processing. In: *Foundations and Trends in Signal Processing 1*.1-2, pp. 1–194.

[Rosenblatt, 1962] Rosenblatt, F. (1962). *Principles of Neurodynamics*. New York: Spartan.

[Rubin et al., 2006] Rubin, Adam D; Praneetvatakul, Veeraphol; Gherson, Shirley; Moyer, Cheryl A; Sataloff, Robert T., (2006). Laryngeal hyperfunction during whispering: reality or myth?. *Journal of voice*. 2006 Mar; 20(1): 121-127.

[Rumelhart, 1986] Rumelhart, D., McClelland, J., and the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press.

[Sakoe et al., 1978] Sakoe, H. and Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 26(1):43-49, February 1978. Reprinted in Waibel and Lee (1990).

[Scherer et al., 1998] Scherer, K.R., Kappas, A. (1998). Primate vocal expression of affective state. In D. Todt, P. Goedeking, D. Symmes (Eds.), *Primate vocal vommunication*, pp. 171-194. Berlin: Springer.

[Schwartz, 1970] Schwartz, M.F. (1970). Power spectral density measurements of oral and whispered speech. *Journal of Speech and Hearing Research*, pp. 438-448, 1970.

[Sharifzadeh et al., 2009] Sharifzadeh, H.R., McLoughlin, I.V., Ahamdi, F. (2009). Voiced speech from whispers for post-laryngectomised patients. *IAENG International Journal of Computer Science*, 36 (4), 367-377.

[Siniscalchi et al., 2013] Siniscalchi, S.M., Yu, D., Deng, L., Lee, C. (2013). Exploiting deep neural networks for detection-based speech recognition. *Neurocomputing* 106, 148-157.

[Solomon et al., 1989] Solomon, N.P., McCall, G.N., Trosset, M.W., Gray, W.C. (1989). Laryngeal configuration and constriction during two types of whispering. *Journal of Speech and Hearing Research*, vol. 32, pp. 161-174, march 1989.

[Stathopoulos et al., 1991] Stathopoulos, E.T., Hoit, J.D., Hixon, T.J., Watson, P.J., Solomon, N.P. (1991). Respiratory and laryngeal function during whispering. *Journal of Speech and Hearing Research*, vol. 34, pp. 764-767, August 1991.

[Stevens et al., 1937] Stevens, S.S., Volkmann, J., Edwin, B., (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*. 8 (3): 185–190. doi:10.1121/1.1915893.

[Stevens et al., 1999] Stevens, H.E., Wickesberg, R.E. (1999). Ensemble responses of the auditory nerve to normal and whispered stop consonants. *Hearing Research*, vol. 131, pp. 47-62, May 1999.

[Sugito et al., 1991] Sugito, M., Higasikawa, M., Sakakura, A., Takahashi, H., 1991. Perceptual, acoustical, and physiological study of Japanese word accent in whispered speech. *IEICE Technical Report*, SP91-1, May 1991, pp. 1–8.

[Sundberg et al., 2009] Sundberg, J., Scherer, R., Hess, M., Muller, F. (2009). Whispering a single-subject study of glottal configuration and aerodynamics. *Journal of Voice*, vol. 24, pp. 574-584, 2009.

[Tao et al., 2014] Tao, F., Busso, C., 2014. Lipreading approach for isolated digits recognition under whisper and neutral speech. In: *Proceedings of the Annual Conference International Speech Communication Association INTERSPEECH, 2014*. Singapore, pp. 1154–1158.

[Tartter, 1989] Tartter, V.C. (1989). What's in a whisper? *Journal of the Acoustical Society of America*, vol. 86, pp. 1678-1683. November 1989.

[Tartter, 1991] Tartter, V.C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perceptrion and Psychophysics*, vol. 49, no. 4, pp. 365-372, 1991.

[Tartter, 1994] Tartter, V.C., Brun, D. (1994). Hearing smiles and frowns in normal and whispered register. *Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101-2107, October 1994.

[Teager et al., 1989] Teager, H.M., Teager, S.M. (1989). Evidence for nonlinear sound production mechanisms in the vocal tract. In: Hardcastle W, Marchal A (eds) *Speech production and speech modeling*, vol 55. NATO Advanced Study Institute Series D, Bonas, France.

[Teager, 1980] Teager, H.M. (1980). Some observations on oral air flow during phonation. *IEEE Trans Speech Audio Process* 28(5):599–601.

[Tebelskis, 1995] Tebelskis, J. (1995). *Speech Recognition using Neural Networks*. PhD thesis, School of Computer Science, Pittsburgh, PA (1995).

[Thomas, 1969] Thomas, I.B. (1969). Perceived pitch of whispered vowels. *Journal of the Acoustical Society of America*, vol. 46, pp. 468-470, August 1969.

[Tohkura, 1987] Tohkura, Y. (1987). A weighted cepstral distance measure for speech recognition. *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-35, pp. 1414-1422, Oct. 1987.

[Tran et al., 2013] Tran, T., Mariooryad, S., Busso, C. (2013). Audiovisual corpus to analyze whisper speech, in *ICASSP*, 2013

[Tsunoda et al., 2012] Tsunoda, K., Sekimoto, S. and Baer, T. (2012) Brain Activity in Aphonia after a Coughing Episode: Different Brain Activity in Healthy Whispering and Pathological Aphonic Conditions. *Journal of Voice*, 26, 668.e11-668.e13. doi:/10.1016/j.jvoice.2011.11.004

[Vaseghi, 2008] Vaseghi, S.V. (2008). *Advanced Digital Signal Processing and Noice Reduction*. Johh Wiley & Sons, 2008.

[Vincent et al., 2008] Vincent, P., Larochelle, H., Bengio, Y., Manzagol PA., (2008). Extracting and composing robust features with denoising autoencoders, In: *25th Interntaionsal Conference on Machine Learning, ICML 2008*. Helsinki, Finland. pp. 1096-1103.

[Vintsyuk, 1971] Vintsyuk, T. (1971). Element-Wise Recognition of Continuous Speech Composed of Words from a Specified Dictionary. *Kibernetika* 7:133-143, March-April 1971.

[Waibel et al., 1989] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37(3), March 1989. Originally appeared as Technical Report TR-1-0006, ATR Interpreting Telephony Research Laboratories, Japan, 1987. Reprinted in Waibel and Lee (1990).

[Wenndt et al., 2002] Wenndt, S.J., Cupples, E.J., Floyd, R.M., 2002. A study on the classification of whispered and normally phonated speech. In: *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 16–20 September 2002, Denver, pp. 649–652.

[Yang et al., 2012] Yang, C.Y., Brown, G., Lu, L., Yamagishi, J., King, S. (2012). Noise-robust whispered speech recognition using a non- audible-murmur microphone with VTS compensation, in: *Proc. of 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2012, pp. 220–223.

[Young et al., 2002] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., (2002). *The HTK Book (for HTK Version 3.2)*, CambridgeUniversity Engineering Department, 2002.

[Zhang et al., 2007] Zhang, C., Hansen, J.H.L. (2007). Analysis and classification of speech mode: Whisper through shouted. In: *Proceedings of Interspeech*, Antwerp, pp. 2289-2292.

[Zhang et al., 2010] Zhang, C., Hansen, J.H.L., 2010. Advancements in whisper-island detection using the linear predictive residual. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, USA, pp. 5170–5173. DOI: 10.1109/ICASSP.2010.5495022.

[Zhang et al., 2011] Zhang, C., Hansen, J.H.L. (2011). Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. *IEEE Trans. Audio Speech Lang. Process*. 19(4), 883–894.

[Zhou et al., 1998] Zhou, G., Hansen, J., Kaiser, J., (1998). Classification of speech under stress based on features derived from the nonlinear Teager energy operator, *Acoust. Speech Signal*, (1998) 2–5. doi:10.1109/ICASSP.1998.674489.

[Zue et al., 1990] Zue, V., Seneff, S., Glass, J. (1990). Speech database development at mit: Timit and beyond. *Speech Communication*, no. 9, pp. 351-356, 1990.

PRILOZI

SPISAK REČI U WHI-SPE KORPUSU

Tabela P1.1

Spisak reči u Whi-Spe korpusu*.

	Serbian	*English*		*Serbian*	*English*
1.	bela	white	26.	pijatsa	market place
2.	ʒuta	yellow	27.	padavine	drops
3.	tsrna	black	28.	ponedeʌak	Monday
4.	tsrvena	red	29.	godina	year
5.	plava	blue	30.	predstava	play
6.	zelena	green	31.	kompjuteri	computers
7.	nula	zero	32.	inostranstvo	abroad
8.	jedan	one	33.	drvo	tree
9.	dva	two	34.	Mirjana	Mirjana (name)
10.	tri	three	35.	more	sea
11.	tʃetiri	four	36.	kiʃa	rain
12.	pet	five	37.	zgrade	buildings
13.	ʃest	six	38.	klintsi	kids
14.	sedam	seven	39.	Milan	Milan (name)
15.	osam	eight	40.	rezultati	results
16.	devet	nine	41.	telefon	telephone
17.	deset	ten	42.	svetlo	light
18.	sto	hundred	43.	prozor	window
19.	hiʌadu	thousand	44.	ruke	hands
20.	milion	million	45.	lokal	locale
21.	Mirko	Mirko (name)	46.	kʌutʃ	key
22.	ʒurka	party	47.	suntse	sun
23.	Petar	Petar (name)	48.	pare	money
24.	demonstratsije	demonstration	49.	sef	safe
25.	standard	standard	50.	blok	block

* Srpska i IPA notacija za konsonante i vokale su iste izuzev za sledeće konsonanate: ʃ(š), h(x), ʒ(ž), ts(c), tɕ(ć), tʃ(č), dʑ(đ), dʒ(dž), ɲ(nj), ʌ(lj).

REZULTATI GMM-HMM SISTEMA

Tabela P1.2

Uspeh prepoznavanja reči (%) u neusaglašenim obuka/test scenarijima ostvaren sa GMM-HMM sistemom i različitim kepstralnim obeležjima. [Grozdić et al., 2016 b]

Govorni mod (obuka/test)	Govor/Šapat	σ
MFCC	45,19	5,35
MFCC+Δ	59,67	4,14
MFCC+Δ+ΔΔ	61,81	5,32
TECC	54,42**	4,46
TECC+Δ	65,29***	3,56
TECC+Δ+ΔΔ	71,33**	2,13
TEMFCC	47,85	3,95
TEMFCC+Δ	59,94*	3,33
TEMFCC+Δ+ΔΔ	63,71	2,20

BIOGRAFIJA AUTORA

Đorđe (Tomislav) Grozdić je rođen 20.01.1987. godine u Beogradu, gde je završio osnovnu školu i Treću beogradsku gimnaziju sa Vukovom diplomom. Diplomirao je 2010. godine na Elektrotehničkom fakultetu u Beogradu na Odseku za telekomunikacije i informacione tehnologije. Tema diplomskog rada je bila „Multidimenzionalna analiza akustičkih obeležja u govornom signalu", pod mentorstvom prof. dr Slobodana Jovičića. Master studije na Elektrotehničkom fakultetu u Beogradu (Smer za audio i video tehnologije) je završio 2011. godine. Tema master rada je bila „Analiza varijacija akustičkoh obeležja u govornom signalu", takođe pod mentorstvom prof. Slobodana Jovičića.

Radnu karijeru je započeo 2011. godine u firmi Telefonkabl a.d. u sektoru za projektovanje i razvoj telekomunikacionih sistema. Decembra 2011. godine, nastavlja studije na Elektrotehničkom fakultetu i upisuje doktorske studije, modul Telekomunikacije, sa rukovodiocem naučno-istraživačkog rada prof. Slobodanom Jovičićem. Početkom 2012. godine stiče naučno zvanje istraživač-saradnik i započinje radni odnos u Centru za unapređenje životnih aktivnosti u Laboratoriji za forenzičku akustiku i fonetiku, gde je angažovan na tehnološkom projektu Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, pod brojem TP32032 i nazivom „e-logoped". Odlaskom prof. Jovičića u penziju 2014. godine, mentorstvo u izradi doktorske teze preuzima prof. dr Dragana Šumarac Pavlović. Područje istraživačkog rada Đorđa Grozdića obuhvata multidisciplinarno istraživanje govora i govorne komunikacije i to pre svega oblasti: digitalne obrade govornih signala, komunikacije čovek-računar, veštačke inteligencije, automatskog prepoznavanja govora i netipičnih oblika govora (poput emotivnog govora, šapata…), kao i automatskog prepoznavanja govornika. Od novembra 2016. godine, Đorđe Grozdić je zaposlen na poziciji softverskog inženjera u kompaniji Fincore Ltd.

Kao rezultat dosadašnjeg naučno-istraživačkog rada, Đorđe Grozdić ima preko 40 objavljenih stručnih i naučnih radova, od toga: 3 rada u međunarodnim časopisima sa SCI liste, 1 rad u čaopisu nacionalnog značaja, 6 poglavlja u monografijama međunarodnog značaja, 19 radova na međunarodnim konferencijama, 9 radova na nacionalnim konferencijama i 5 tehničkih rešenja.

Đorđe Grozdić je član Udruženja studenata elektrotehnike Evrope. Dobitnik je nagrade za najbolji rad mladog autora na međunarodnoj konferenciji TELFOR 2012. godine, a 2016. godine je nagrađen za najbolji rad mladog istraživača na konferenciji ETRAN.

Прилог 1.

I. Изјава о ауторству

Потписани-а Ђорђе Гроздић

Број индекса 5013/2011

Изјављујем

да је докторска дисертација под насловом

„Примена неуралних мрежа у препознавању шапата"

- резултат сопственог истраживачког рада,
- да дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 9.6.2017.

Ђорђе Гроздић

Прилог 2.

II. Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Ђорђе Гроздић

Број индекса 5013/2011

Студијски програм телекомуникације

Наслов рада „Примена неуралних мрежа у препознавању шапата"

Ментор др Драгана Шумарац Павловић

Потписани Ђорђе Гроздић

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања 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 докторанда

У Београду, 9.6.2017.

Ђорђе Гроздић

Прилог 3.

III. Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

„Примена неуралних мрежа у препознавању шапата”

_____

_____

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 докторанда

У Београду, 9.6.2017.

Ђорђе Грозданић

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.