УНИВЕРЗИТЕТ У НОВОМ САДУ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА
КАТЕДРА ЗА ТЕЛЕКОМУНИКАЦИЈЕ
И ОБРАДУ СИГНАЛА

АНАЛИЗА МЕЛ-ФРЕКВЕНЦИЈСКИХ КЕПСТРАЛНИХ КОЕФИЦИЈЕНАТА КАО ОБЕЛЕЖЈА КОРИШЋЕНИХ ПРИ АУТОМАТСКОМ ПРЕПОЗНАВАЊУ ГОВОРНИКА

ДОКТОРСКА ДИСЕРТАЦИЈА

Ментор:
Проф. др Владо Делић

Кандидат:
мр Иван Јокић

Нови Сад, 2014.

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

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, **РБР**:		
Идентификациони број, **ИБР**:		
Тип документације, **ТД**:		Монографска документација
Тип записа, **ТЗ**:		Текстуални штампани материјал
Врста рада, **ВР**:		Докторска дисертација
Аутор, **АУ**:		мр Иван Јокић
Ментор, **МН**:		Проф. др Владо Делић
Наслов рада, **НР**:		Анализа мел-фреквенцијских кепстралних коефицијената као обележја коришћених при аутоматском препознавању говорника
Језик публикације, **ЈП**:		Српски
Језик извода, **ЈИ**:		Српски / енглески
Земља публиковања, **ЗП**:		Србија
Уже географско подручје, **УГП**:		Војводина
Година, **ГО**:		2014.
Издавач, **ИЗ**:		Ауторски репринт
Место и адреса, **МА**:		Нови Сад, Трг Доситеја Обрадовића 6
Физички опис рада, **ФО**: (поглавља/страна/ цитата/табела/слика/графика/прилога)		6 / 72 / 36 / 23 / 21 / 14 / 0
Научна област, **НО**:		Електротехничко и рачунарско инжењерство
Научна дисциплина, **НД**:		Телекомуникације и обрада сигнала
Предметна одредница/Кључне речи, **ПО**:		Аутоматско препознавање говорника, чујни критични опсези, мел-фреквенцијски кепстрални коефицијенти, експоненцијални чујни критични опсези, вишедимензионална Гаусова расподела, коваријансна матрица.
УДК		
Чува се, **ЧУ**:		Библиотека Факултета техничких наука у Новом Саду.
Важна напомена, **ВН**:		
Извод, **ИЗ**:		Рад је окренут ка анализи мел-фреквенцијских кепстралних коефицијената као обележја говорника која се користе при аутоматском препознавању говорника. Испитан је утицај промене облика чујних критичних опсега као и модификације енергије у њима на тачност препознавања говорника. Такође испитане су и неке трансформације ради умањења временске променљивости модела истих говорника.
Датум прихватања теме, **ДП**:		29.01.2014.
Датум одбране, **ДО**:		
Чланови комисије, **КО**:	Председник:	Др Зоран Перић (редовни професор)
	Члан:	Др Срђан Крчо (доцент)
	Члан:	Др Милан Гњатовић (ванредни професор)
	Члан:	Др Милан Сечујски (доцент)
	Члан, ментор:	Др Владо Делић (редовни професор)

Потпис ментора

UNIVERSITY OF NOVI SAD ● **FACULTY OF TECHNICAL SCIENCES**
21000 NOVI SAD, Trg Dositeja Obradovića 6

KEY WORDS DOCUMENTATION

Accession number, **ANO**:	
Identification number, **INO**:	
Document type, **DT**:	Monograph documentation
Type of record, **TR**:	Textual printed material
Contents code, **CC**:	PhD thesis (Doctoral dissertation)
Author, **AU**:	mr Ivan Jokić
Mentor, **MN**:	PhD Vlado Delić
Title, **TI**:	Analysis of mel-frequency cepstral coefficients as features used for automatic speaker recognition
Language of text, **LT**:	Serbian
Language of abstract, **LA**:	Serbian / English
Country of publication, **CP**:	Serbia
Locality of publication, **LP**:	Vojvodina
Publication year, **PY**:	2014
Publisher, **PB**:	Author's reprint
Publication place, **PP**:	Novi Sad, Trg Dositeja Obradovića 6
Physical description, **PD**: (chapters/pages/ref./tables/pictures/graphs/appendixes)	6 / 72 / 36 / 23 / 21 / 14 / 0
Scientific field, **SF**:	Electrical and Computer Engineering
Scientific discipline, **SD**:	Telecommunications and Signal Processign
Subject/Key words, **S/KW**:	Automatic speaker recognition, auditory critical bands, mel – frequency cepstral coefficients, exponential auditory critical bands, multidimensional Gaussian distribution, covariance matrix.
UC	
Holding data, **HD**:	Library of Faculty of Technical Sciences in Novi Sad.
Note, **N**:	
Abstract, **AB**:	The work is oriented towards the analysis of mel-frequency cepstral coefficients as speaker features used in automatic speaker recognition. The influence of the shape of auditory critical bands as well as the proposed energy modification inside them is tested. Also, some transformations for reducing of time variability of models of the same speakers are proposed.
Accepted by the Scientific Board on, **ASB**:	29.01.2014.
Defended on, **DE**:	

Defended Board, **DB**:		
President:	PhD Zoran Perić (professor)	
Member:	PhD Srdjan Krčo (assistant professor)	
Member:	PhD Milan Gnjatović (associate professor)	Menthor's sign
Member:	PhD Milan Sečujski (assistant professor)	
Member, Mentor:	PhD Vlado Delić (professor)	

ЗАДАТАК ДОКТОРСКОГ РАДА

ТЕКСТ ЗАДАТКА :

Аутоматско препознавање говорника као одређено програмско решење извршиво на рачунарима опште или специјализоване намене одликује се извесним степеном сложености те стога и својом модуларношћу. Ради могућности контроле сваког параметра примењених модула потребно је извршити њихову практичну реализацију. Стога се један од постављених задатака огледа у формирању програмског решења аутоматског препознавача говорника. Особине аутоматског препознавача говорника проистичу из начина на који он врши представу разматраног говорног сигнала: коришћена обележја говора и примењени модели. Имајући у виду широку примену мел-фреквенцијских кепстралних коефицијената као обележја која се користе при аутоматском препознавању говорника као и њихову зависност од енергије присутне у говору неизбежна је констатација да ова обележја поред тога што зависе од самог говорника зависе и од текстуалне садржине говора као и осећања при којима је та текстуална садржина изказана. Стога је пред овај рад постављен задатак анализе могућности утицања како на саме мел-фреквенцијске коефицијенте тако и на моделе којима су они моделовани ради побољшања резултата аутоматског препознавања говорника. У циљу остварења очекиваног доприноса резултујуће програмске имплементације аутоматског препознавача говорника потребно је спровести следеће кораке:

1) Извршити потребан и довољан увид у досадашње резултате у области аутоматског препознавања говорника,
2) Сходно очекиваном доприносу дисертације извршити реализацију аутоматског препознавача говорника коришћењем програмског C++ језика,
3) Испитати утицај различитих параметара аутоматског препознавача говорника који директно утичу како на квалитет израчунатих мел-фреквенцијских кепстралних коефицијената, која ће се користити као обележја говорника, тако и на квалитет процењених модела говорника.

САЖЕТАК

Рад је окренут ка анализи мел-фреквенцијских кепстралних коефицијената као обележја говорника која се користе при аутоматском препознавању говорника. Посредно, рад је окренут ка практичној реализацији аутоматског препознавача говорника. Сви модули су реализовани у C++ програмском језику. Реализовани аутоматски препознавач говорника је независан од изговореног текста. Вектори обележја садрже мел-фреквенцијске кепстралне коефицијенте. Претпостављено је да је расподела вектора обележја говорника гаусовског типа. Стога је моделовање расподеле вектора обележја посматраних говорника извршено проценом облика одговарајућих вишедимензионалних Гаусових расподела кроз одређивање одговарајућих коваријансних матрица.

Испитивање тачности препознавања препознавача спроведено је над говорним базама раније развијеним у оквиру АлфаНум пројекта. Једна говорна база садржи изговоре 121 говорника док друга садржи изговоре 44 говорника од којих су снимци 37 говорника коришћени при тестирању рада аутоматског препознавача говорника. Различит квалитет снимака у овим говорним базама искоришћен је као тест за понашање препознавача над говорним снимцима различитог квалитета.

Приликом анализе варирани су различити параметри везани за препознавач и окружење у ком се примењује, као што су: део говорне базе над којом се врши тестирање препознавања, облик аудиторних критичних опсега на основу којих се одређују мел-фреквенцијски кепстрални коефицијенти као и утицај увођења енергијске поправке унутар примењених чујних критичних опсега. Такође је праћено понашање модела за различите снимке истих говорника, тренинг односно тест снимке, и на основу тога испитан утицај примене неких трансформација које би омогућиле мање разликовање модела који одговарају различитим снимцима истих говорника и самим тим допринеле тачнијем препознавању говорника.

Кључне речи - аутоматско препознавање говорника, чујни критични опсези, мел-фреквенцијски кепстрални коефицијенти, експоненцијални чујни критични опсези, вишедимензионална Гаусова расподела, коваријансна матрица.

ABSTRACT

The work is oriented towards the analysis of mel-frequency cepstral coefficients as speaker features used in automatic speaker recognition. Indirectly, the work is also oriented towards the practical realization of automatic speaker recognizer. All modules are implemented in C++ programming language. Implemented automatic speaker recognizer is text-independent. Feature vectors contain the mel-frequency cepstral coefficients. Multidimensional Gaussian distribution of feature vectors which belongs to a speaker is assumed. The shape of assumed Gaussian distribution is determined by appropriate covariance matrix. Therefore speaker modeling was performed throw determination of covariance matrices of feaure vectors which belongs to a given speaker.

Testing of recognition accuracy was conducted over speech databases earlier developed within the AlfaNum project. One speech database contains 121 speakers while another contains utterances of 44 speakers, 37 speakers are used for tests in this work. The different quality of recordings in these speech databases was used for testing of recognizer on speech recordings of different quality.

Different parameters of recognizer and applied environment are varied: part of used speech database for testing of recognition, the shape of auditory critical bands, the influence of the proposed energy correction inside auditory critical bands applied. Also the difference between models of training and test speech of the same speakers is monitored. In order to reduce this difference and to improve the recognition accuracy some transformations are applied.

Keywords – automatic speaker recognition, auditory critical bands, mel – frequency cepstral coefficients, exponential auditory critical bands, multidimensional Gaussian distribution, covariance matrix.

САДРЖАЈ:

1. МЕТОДОЛОШКЕ ОСНОВЕ РАДА 1
1.1. ПРОБЛЕМ ИСТРАЖИВАЊА 1
1.2. ПРЕДМЕТ ИСТРАЖИВАЊА 1
1.3. ЦИЉ ИСТРАЖИВАЊА 2
1.4. ПРЕТПОСТАВКЕ ИСТРАЖИВАЊА 2
1.5. НАЧИН ИСТРАЖИВАЊА 3
1.6. ДРУШТВЕНА И НАУЧНО-СТРУЧНА ОПРАВДАНОСТ ИСТРАЖИВАЊА 3
2. ОПШТЕ ПОСТАВКЕ 5
2.1. КЛАСИФИКАЦИЈА СИСТЕМА ЗА АУТОМАТСКО ПРЕПОЗНАВАЊЕ ГОВОРНИКА 5
2.2. ОБЕЛЕЖЈА ГОВОРА 8
2.2.1. Кепструм 10
2.2.2. Мел-фреквенцијски кепстрални коефицијенти (MFCCs) 10
2.2.3. Динамичка обележја 13
2.2.4. Висина гласа 14
2.3. МОДЕЛИ ГОВОРНИКА 14
2.3.1. Мешавина Гаусових расподела (GMM) 16
2.3.2. Скривени Марковљев модел (HMM) 16
2.3.3. Вештачке неуронске мреже (ANN) 18
3. СМАЊЕЊЕ СТАНДАРДНО КОРИШЋЕНОГ СКУПА ОБЕЛЕЖЈА УЗИМАЊЕМ У ОБЗИР ГЛАВНИХ ПРАВАЦА ЕНЕРГИЈЕ У ИЗВОРНОМ ПРОСТОРУ ОБЕЛЕЖЈА 20
3.1. ЕКСПЕРИМЕНТАЛНИ УСЛОВИ АНАЛИЗЕ УТИЦАЈА PCA 22
3.2. РЕЗУЛТАТИ ПРИМЕНЕ PCA 23
4. РЕАЛИЗАЦИЈА ЈЕДНОГ РЕШЕЊА АУТОМАТСКОГ ПРЕПОЗНАВАЊА ГОВОРНИКА 25
4.1. РЕАЛИЗАЦИЈА ИЗДВАЈАЊА ОБЕЛЕЖЈА 25
4.1.1. Канонички wav формат дигиталног записа 26
4.1.2. Формирање вредности одбирака на основу расположивог "wav" дигиталног записа 27
4.1.3. Израчунавање MFCCs 29
4.1.3.1. Израчунавање брзе Фуријеове трансформације 32
А) Реализација преуређења временског низа 35
Б) Примењени начин израчунавања FFT 36
4.1.3.2. Израчунавање MFCCs посматрањем енергије унутар аудиторних критичних опсега 37
4.2. МОДЕЛОВАЊЕ ГОВОРНИКА И НАЧИН ОДЛУЧИВАЊА 37
5. РЕЗУЛТАТИ ПРЕПОЗНАВАЊА 39
5.1. ТЕСТОВИ ПРОМЕНЕ ТЕСТ ФАЈЛА 40
5.2. ПОБОЉШАЊЕ ТАЧНОСТИ ПРЕПОЗНАВАЊА ПРОМЕНОМ ОБЛИКА КРИТИЧНОГ ОПСЕГА 46
5.2.1. Примена троугаоних чујних критичних опсега 47
5.2.2. Примена експоненцијалних чујних критичних опсега 54
5.3. ПРАЋЕЊЕ ПРОМЕНЉИВОСТИ ЕЛЕМЕНАТА МОДЕЛА И ЊИХОВ УТИЦАЈ НА ТАЧНОСТ АУТОМАТСКОГ ПРЕПОЗНАВАЊА ГОВОРНИКА 64
6. ЗАКЉУЧАК 69
ЛИТЕРАТУРА 71

Хвала ментору проф. др Влади Делићу на подршци, помоћи, усмеравању, идејама, саветима, при раду на докторату, писању неопходних радова и комплетном последипломском раду.
Хвала председнику комисије проф. др Зорану Перићу на идејама и саветима које су помогле објављивању потребних радова и употпуниле докторат.
Хвала члановима комисије,
доц. др Крчо Срђану, проф. др Милану Гњатовићу, доц. др Милану Сечујском, на саветима, помоћи, подршци, како током различитих периода последипломског усавршавања тако и током самог састављања текста доктората.
Хвала проф. др Драгани Бајић као и проф. др Жељену Трповском који су ми као ментор и председник комисије при дипломском раду отворили врата последипломских студија и даљег усавршавања.
Хвала проф. др Војину Шенку и проф. др Дејану Вукобратовићу на помоћи и сарадњи у току магистарских студија.
Хвала родитељима Милосави и Драгутину и брату Стевану на подршци, помоћи, саветима, идејама, које су ми пружали током свих досадашњих година.

1. МЕТОДОЛОШКЕ ОСНОВЕ РАДА

1.1. ПРОБЛЕМ ИСТРАЖИВАЊА

Развој технологије и технике олакшава пружање различитих услуга. Многе од тих услуга подразумевају контролу приступа или препознавање идентитета. Дакле потребно је обезбедити приступ одређеној особи или извршити њено препознавање. Омогућавање поменутог у општем случају подразумева примену неког од поступака препознавања облика. Примена ових поступака или неког од њих, зависно од потребе, на рачунарима опште или специјализоване намене, често се декларише појмом аутоматско препознавање. У оквиру ове широкообухватне проблематике као један од њених видова истиче се и аутоматско препознавање говорника.

Аутоматско препознавање говорника се бави проблемом препознавања идентитета кроз посматрани глас или говор. Посматран у светлу узрок – последица глас би се могао окарактерисати као резултат воље да се нешто изговори. При том се над изворном ваздушном струјом дешава низ трансформација: лингвистичких, семантичких, артикулаторних и акустичких, распоређених на различитим нивоима (Campbell P. Josef, Jr., 1997; Делић Д. Владо et al., 2008). Ниво и начин дејства сваке трансформације својствен је посматраном говорнику. Стога се глас може сматрати једним од биометријских обележја.

Имајући у виду начин настанка говора и посматрајући га у домену сигнала, може се рећи да је он сложен сигнал. Стога је потребно дефинисати одређени скуп обележја говора која ће имати потребну и довољну репрезентативност за сваког посматраног говорника. На тај начин за сваки посматрани узорак говора израчунава се одговарајући скуп репрезентативних обележја односно вектор обележја. Постављајући проблем препознавања говорника у оквир великог броја говорних узорака, како једног тако и осталих говорника, води ка тежњи за њиховим обједињавањем у јасну целину. Ове целине се праве за сваког говорника и често се за њих каже да су то модели одговарајућих говорника. Коначно, разрешавање проблема аутоматског препознавања говорника своди се на поређење говорних узорака непознатих идентитета, или претпостављених зависно да ли се има или не икаква информација о њиховим идентитетима, са раније направљеним моделима и утврђивање највеће сличности.

1.2. ПРЕДМЕТ ИСТРАЖИВАЊА

Практично решавање проблема аутоматског препознавања говорника води ка дефинисању предмета истраживања. Посредством њега могу се испитивати разни поступци препознавања односно тачност која се њима може постићи, те као такав, предмет истраживања се у овом случају поистовећује са аутоматским препознавачем говорника. Дакле потребно је на одређени начин направити техничко решење аутоматског препознавача говорника.

Поступак аутоматског препознавања говорника одликује се извесним степеном сложености која зависи од примењених поступака унутар препознавача. Ова сложеност се огледа у спровођењу издвајања обележја, прављењу модела говорника и начину на који се врши само препознавање и доноси одлука о извршеном препознавању. Из постојања корака при реализацији препознавача следи утицај њихове реализације на успешност препознавања. Формирање аутоматског препознавача говорника на овај начин омогућује испитивање утицаја сваког параметра на тачност препознавања.

Боја гласа представља једну од особина гласа на основу које је могуће извршити разликовање говорника. Посматрајући спектар говорног сигнала, боја гласа је последица хармонијске специфичности конкретног гласа и она се огледа у обвојници спектра конретног говорног сигнала. Стога су као посматрана обележја говора коришћени мел – фреквенцијски кепстрални коефицијенти – MFCCs (енг. Mel – Frequency Cepstral Coefficients) чије добијање

зависи како од примењеног прозора у временском домену и учесталости његове примене на разматраном говорном сегменту, тако и од примењених филтарских секција те учесталости њихове примене приликом симулације аудиторних критичних опсега у разматраном спектралном опсегу. Ради моделовања коришћене су вишедимензионалне Гаусове расподеле. Обзиром на одређеност њиховог облика елементима у припадајућој коваријансној матрици, одлука о препознавању представља последицу мере сличности између коваријансних матрица расположивих тренинг модела и модела посматраног тест говора. Вишедимензионална расподела која описује говор једног говорника последица је вредности у коришћеним MFCCs векторима обележја те на основу претходне кратке представе поступака унутар самог аутоматског препознавача говорника уочава се да на предмет истраживања тј. аутоматски препознавач говорника у највећој мери посредно утичу како избор броја MFCCs обележја тако и параметри који се дефинишу ради њиховог израчунавања. Вредност MFCCs директно је зависна од енергије унутар чујних критичних опсега, што повлачи њихову директну условљеност обликом чујних критичних опсега. Такође, тачност примењеног поступка одлучивања условљена је елементима израчунатих модела говорника тј. међусобним разликовањем модела тренинг и тест говора истих говорника.

1.3. ЦИЉ ИСТРАЖИВАЊА

Поступак аутоматског препознавања говорника одликује се низом корака као што су: издвајање обележја говора, моделовање расподеле посматраних обележја као и поређење ипитиваних говорних узорака и модела добијених приликом обуке. Реализација ових корака подразумева сложено програмско решење чије управљање ће се вршити постављањем одговарајућих параметара. Аутоматски препознавач говорника се одликује мноштвом параметара који утичу на тачност препознавања која се може постићи. Приступ сваком параметру препознавача понаособ омогућава детаљно испитивање утицаја сваког параметра на тачност препознавања у циљу прављења што ефикаснијег препознавача. Формирање програмског решења аутоматског препознавача говорника омогућиће утицај на произвољан број параметара, у зависности од примењене реализације.

Обзиром да избор MFCCs као обележја говорника декларише одговарајући приступ аутоматског препознавача говорника посматраним говорним сигналима у наставку ће се као кључни параметри аутоматског препознавача говорника посматрати облици чујних критичних опсега, енергија садржана у њима и могуће умањење разликовања тест и тренинг модела истих говорника. Циљ је наравно реализацију подредити оптималности аутоматског препознавача говорника обзиром на тачност препознавања и параметре коришћене при препознавању. Тим ће се створити адекватно окружење за формирање модуларног решења аутоматског препознавача говорника.

1.4. ПРЕТПОСТАВКЕ ИСТРАЖИВАЊА

Аутоматско препознавање говорника се са становишта проблематике може посматрати као независан технички проблем али се често и налази у склопу неког општијег система. Имајући у виду ове различитости његове примене следе претпоставке које су утицале на покретање истраживања на пољу аутоматског препознавања говорника у циљу анализе MFCCs као обележја коришћених при аутоматском препознавању говорника:

- аутоматско препознавање говорника позитивно утиче на повећање природности двосмерне говорне комуникације између човека и машине,
- као такво оно представља део дијалошких система,
- програмски језик C++ може се сматрати адекватним окружењем за развој рачунарског програма који ће вршити аутоматско препознавање говорника у реалном времену,

- могуће је унапредити аутоматско препознавање говорника применом одређених модификација на уобичајени поступак израчунавања MFCCs. Чујни критични опсези описују појаву маскирања присутну приликом људског доживљаја интензитета звука, стога је уведена претпоставка да ће чујни критични опсези веће стрмине позитивно допринети постигнутој тачности аутоматског препознавања говорника. Повећању тачности препознавања говорника ће такође допринети тежња ка конструкцији модела који ће показивати малу временску променљивост односно мало разликовање између тест и тренинг модела истих говорника.

1.5. НАЧИН ИСТРАЖИВАЊА

Спроведено истраживање представља својеврстан наставак раније започетог истраживања у оквиру магистарског рада (Јокић Иван, 2010). Особеност истраживања које ће бити описано истиче се у тежњи за конструкцијом аутоматског препознавача говорника коришћењем C++ програмског окружења. Стога је поред избора начина реализације, посматрајући постављену проблематику са становишта теорије препознавања облика, такође било потребно ускладити реализацију са могућностима која пружа C++ окружење. Имајући то у виду током обраде предвиђене теме истраживање је пролазило кроз различите како теоријске тако и експерименталне фазе:

- консултовање литературе како ради утицаја на сам поступак препознавања тако и ради остварења што ефикаснијег решења у C++ окружењу,
- имплементација аутоматског препознавача говорника у изабраном окружењу,
- извођење експеримената препознавања говорника над расположивим говорним базама у циљу:
    - утврђивања постигнуте тачности и извођења одговарајућих закључака за различите облике чујних критичних опсега и количине енергије садржане у њима,
    - испитивања утицаја разних трансформација које би требале својим утицајем на смањење разликовања тест и тренинг модела истих говорника да утичу на повећање тачности аутоматског препознавања говорника.

1.6. ДРУШТВЕНА И НАУЧНО-СТРУЧНА ОПРАВДАНОСТ ИСТРАЖИВАЊА

По принципу свог функционисања техничка решења подразумевају интеракцију са корисником. У тежњи за њеним што природнијим остварењем потребно је у оквиру техничког система имплементирати могућност препознавања корисника. Ради што комплетнијег препознавања потребно је да систем буде у могућности да препозна присуство корисника посматрано са различитих аспеката, као што су изглед корисника или његов глас. Усмереност система на препознавање корисника кроз његов глас подразумева развој дијалошких система[1] који би били задужени за обезбеђивање двосмерне комуникације између корисника и посматраног техничког система.

Развој телекомуникационих система и информационих технологија мотивише стварање нових идеја за имплементације услуга које ће побољшати квалитет већ постојећих. На овај начин се тежи што природнијој комуникацији између корисника и система. Један сегмент који би ову комуникацију учинио природнијом јесте аутоматско препознавање говорника. Поред тога што би овакав систем био у могућности да из гласа сваког корисника додатно

[1] Овај рад представља резултат истраживања на пројекту "Развој дијалошких система за српски и друге јужнословенске језике" (ТР-32035), Министарства просвете, науке и технолошког развоја Републике Србије.

препозна неког од претходних саговорника или потврди да комуницира са новим корисником и самом кориснику овакав систем би пружао безбеднији и ауторизованији приступ. Наравно ова повећања у нивоу сигурности система у погледу идентитета корисника последица су примене разних начина препознавања на страни система, при чему је аутоматско препознавање говорника само један од њих. Промене облика чујних критичних опсега, од правоугаоних преко троугаоних ка експоненцијалним, модификације енергије садржане у њима и смањење променљивости елемената модела истих говорника биће спроведене у циљу повећања тачности аутоматског препознавања говорника.

У наредном делу рада биће приказане неке опште поставке у области аутоматског препознавања говорника које су блиско повезане са истраживањима описаним у овом раду. Након тога дат је опис експеримента који је имао за циљ испитивање могућности смањења броја потребних обележја говора приликом аутоматског препознавања говорника применом поступка анализе главних компонената (енг. Principal Component Analisys – PCA). У наставку је описан реализовани аутоматски препознавач говорника и приказани резултати тестова промене облика чујних критичних опсега, примењене енергијске корекције у посматраним чујним критичним опсезима и примене одређених трансформација на елементе модела говорника.

2. ОПШТЕ ПОСТАВКЕ

Аутоматско препознавање говорника представља један сегмент области препознавања облика. У овом случају предмет препознавања представља глас односно говор посматраног говорника. Стога аутоматско препознавање говорника као поступак препознавања имплементиран на одређеном систему има за циљ препознавање личности на основу разматрања њеног гласа односно говора. Као и у раду свих система за препознавање облика и код система за аутоматско препознавање говорника разликују се две фазе рада: обука и тест. При обуци се формирају модели за сваког посматраног говорника на основу расположивих говорних узорака. Приликом тестирања препознавач анализира непознати, тест, говорни узорак и на основу имплементираног начина препознавања доноси одлуку о идентитету говорника који је изговорио разматрани говорни узорак.

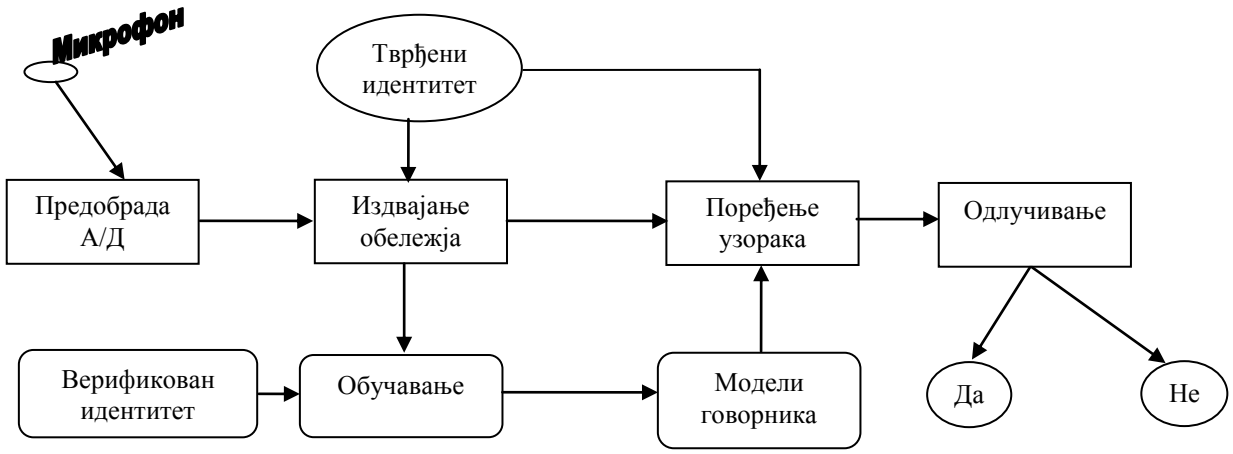
Гледано са становишта аутоматског препознавача говорника који представља имплементацију одређеног алгоритма на рачунару опште односно посебне намене или процесору за обраду сигнала, посматрани говор као акустички сигнал потребно је превести у одређени еквивалентан сигнал који ће бити разумљив препознавачу. На основу тога, поступак аутоматског препознавања говорника подразумева следеће кораке:

- предобраду говорног сигнала након које следи А/Д конверзија,
- издвајање релевантних обележја,
- обучавање ради генерисања референтних модела говорника,
- поређење узорака,
- одлучивање.

Предобрада подразумева филтрирање улазног говорног сигнала у складу са учестаношћу одабирања која ће се користити приликом аналогно – дигиталне конверзије. А/Д конвертор обично има резолуцију од 12 до 16 bit при учестаности одабирања од 8 до 20 kHz. Након дигитализације посматраног говорног сигнала потребно је на неки начин извршити његово моделовање а први корак при томе представља издвајање репрезентативних обележја. На основу издвојених обележја следи фаза обуке система за препознавање. Овај процес се огледа у формирању репрезентативних модела (Wildermoth Richard Breth, 2001) за сваког од посматраних говорника. Након моделовања говорника, препознавач је могуће подвргнути тестовима препознавања приликом којих врши упоређивање модела добијених током обуке са узорцима говора говорника чије се препознавање врши, на основу чега доноси одлуку о препознаваном говорнику.

2.1. КЛАСИФИКАЦИЈА СИСТЕМА ЗА АУТОМАТСКО ПРЕПОЗНАВАЊЕ ГОВОРНИКА

Аутоматско препознавање говорника у својој основи подразумева два основна вида препознавања: верификацију и идентификацију говорника. Ова подела последица је саме поставке аутоматског препознавања говорника односно мере у којој је потребно извршити препознавање говорника. При верификацији потребно је потврдити или не истинитост тврдње о идентитету посматраног говорника док је при идентификацији потребно утврдити идентитет говорника. Наиме, системи који врше верификацију говорника за сваки тест говорни узорак располажу и тврдњом о идентитету посматраног говорника коме припада разматрани говорни узорак. На основу мере сличности између тест говорног узорка и модела који припада тврђеном идентитету системи који врше верификацију доносе одлуку да ли је то тврђење тачно или не (слика 2.1). Дакле одлука система за аутоматску верификацију говорника је стриктно бинарна (Wildermoth R. B., 2001) у облику прихватити ("*Да*") односно одбацити ("*Не*") тврђени идентитет. Препознавач доноси одлуку на основу мере слагања говорног узорка препознаваног говорника и одговарајућег модела добијеног приликом обуке.



Слика 2.1. Основна представа система за аутоматску верификацију говорника.

Мера слагања се може посматрати као одређени вид растојања између тест говорног узорка и модела који одговара тврђеном идентитету. Ако се са $d_i(x)$ означи мера сличности између посматраног говорног узорка x и модела i*-тог* говорника, тада се задатак поменутог система своди на испитивање услова:

$$\begin{array}{ll} \textit{ако је } d_i(x) < d_{ПРАГА} & \textbf{\textit{ПРИХВАТИТИ}} \\ \textit{у супротном} & \textbf{\textit{ОДБАЦИТИ.}} \end{array}$$

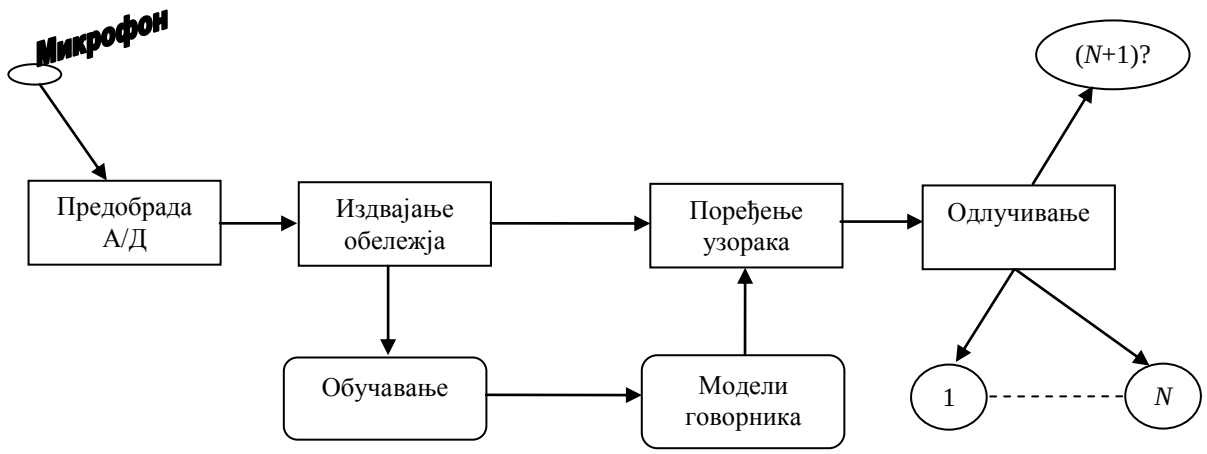
Дакле чим је растојање између модела и препознаваног говорног узорка веће од прага, верификатор одбацује тврдњу о идентитету односно препознавану личност сматра уљезом. Растојање $d_{ПРАГА}$ се експериментално одређује као мера степена дозвољеног одступања између узорка чија се верификација врши и одговарајућег модела.

Идентификација говорника подразумева да систем који врши препознавање доноси одлуку о томе ком од говорника из говорне базе, којом систем за препознавање располаже, припада посматрани говор. Према томе да би систем извршио идентификовање посматраног говора он мора проћи кроз своју говорну базу говорника и утврдити ком од говорника овај говор припада. Опште посматрано тада се могу десити два случаја (слика 2.2):

- препознавач препознаје посматрани говор и идентификује га неким од говорника из базе, идентификација на коначном скупу, тзв. случај *1* од *N,* где опет могу настати два догађаја:
  - препознати говорник је и стварни говорник или
  - систем је у посматраном говору препознао погрешног говорника;
- препознавач не може посматрани говор приписати нити једном од говорника из говорне базе којом располаже, идентификација на отвореном скупу, тзв. случај *1* од *N+1*.

Имајући у виду да се од система за идентификацију говорника очекује да на основу садржаја базе од *N* модела говорника добијених приликом тренирања препознавача одреди ком од њих приписати разматрани исказ следи да је поступак одлучивања система за аутоматску идентификацију говорника доста сложенији у поређењу са верификацијом говорника. Према томе, полазећи од захтева за проналажењем модела који показује минимално одступање у односу на препознавани тест говор, препознавач у случају идентификације доноси одлуку да исказ припада ***i****-том* говорнику ако је испуњен услов да важи $d_i(x) < d_j(x), \forall j = 1,2,...,N \wedge j \neq i$. Други наведени случај који се односи на идентификацију на

отвореном скупу говорника захтевао би додатну проверу саме мере утврђеног минималног одступања одговарајућег модела из говорне базе и тест говорног узорка. Слично као и при верификацији било би потребно поставити одређену вредност прага за посматрану говорну базу. Уколико би при поређењу тест говора и модела из расположиве базе минимално растојање било веће у односу на вредност прага посматрани тест говорни узорак би се приписивао непознатом говорнику, (N+1)? на слици 2.2.



Слика 2.2. Основна представа система за аутоматску идентификацију говорника.

Поређењем верификације и идентификације са становишта начина одлучивања долази се до закључка да препознавач приликом верификације врши једну проверу за сваки разматрани тест говорни узорак док идентификација говорника подразумева N тестова по сваком разматраном говорном исказу. Као последица тога следи да грешка аутоматске идентификације говорника расте увећавањем броја говорника N, док проценат погрешног одлучивања аутоматске верификације говорника не зависи од истог.

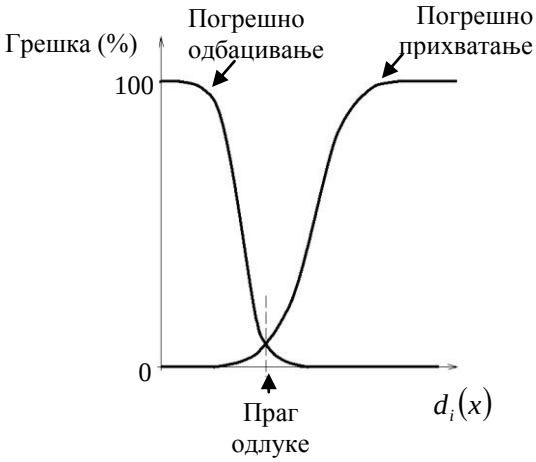
Грешка система за аутоматску идентификацију огледа се у погрешном идентификовању посматраног говорника док се грешке при верификацији говорника могу сврстати у две групе (Imperl Bojan):

- погрешно прихватање, које се огледа у прихватању идентитета особе која се лажно представља, тзв. уљеза,
- погрешно одбацивање, подразумева неприхватање идентитета особе која се тачно представља.

Вероватноће појављивања ових грешака условљене су вредношћу постављеног прага одлучивања (слика 2.3), тако да се на основу захтеваног понашања препознавача вредност прага поставља у циљу минимизације критичније од ове две грешке. Код система са финим одлучивањем, ради заштите од неауторизованог приступа, поставља се што нижа вредност прага. То значи да је дозвољено мало одступање између модела и узорка чија се верификација врши, да би он био прихваћен. Дакле потребан је висок степен усклађености између препознаваних узорака и њима одговарајућих модела, што повлачи као врло вероватан догађај да систем изврши одбацивање стварног идентитета особе, чија се верификација врши, уколико њен глас поприми некарактеристична својства (као последице варијације боје гласа током дана, неких здравствених проблема и сл.) у односу на модел који се налази у бази. Систем са оваквим видом заштите одбија приступ неауторизованим личностима, док као последицу тога врши одбијање идентитета оних који се покушају пријавити на неуобичајен начин. Према томе овакви системи имају низак степен погрешног прихватања док им је висок ступањ погрешног одбацивања.

Често се приликом пројектовања система за верификацију тежи постизању смањења укупне грешке. У том случају потребно је обезбедити да обадве претходно поменуте грешке

карактеристичне за овај систем буду минималне. Обзиром на приближно инверзни облик зависности грешке погрешног прихватања односно одбацивања у односу на вредност прага одлуке, минимизација укупне грешке одлучивања се постиже постављањем вредности прага на ниво за који су обадве грешке међусобно једнаке (слика 2.3).



Слика 2.3. Утицај прага одлучивања на погрешно одбијање и прихватање.

У многим случајевима примене аутоматског препознавања говорника систем који врши препознавање очекује да говорник каже неку унапред дефинисану фразу или опште посматрано неки унапред одређен текст. Тада се ради о препознавању говорника зависном од изговореног текста. У општијем случају, када се обука препознавача врши независно од садржине говора објекта препознавања, реч је о препознавању говорника независном од изговореног текста.

2.2. ОБЕЛЕЖЈА ГОВОРА

Коришћење говорних сигнала у системима за аутоматско препознавање говорника подразумева њихову употребу у дигитализованом формату. Свака даља обрада говорног сигнала у циљу његовог прилагођења аутоматском препознавачу говорника подразумева његову обраду у дискретном временском домену. Стога ако се претпостави да су на самом улазу система за аутоматско препознавање говорника говорни сигнали аналогни тада је на самом улазу, пре дела који врши издвајање обележја говора, потребно извршити њихову дискретизацију тј. одабирање потребном учестаношћу одабирања а затим и дигитализацију. Поменути улазни процеси по својој спектралној природи су нискофреквентни тако да утичу на слабљење виших спектралних компонената у говорном сигналу. Такође и степени у даљој дигиталној обради говорног сигнала одликују се коначном прецизношћу што такође има особине нискофреквентног филтрирања. Из тог разлога након извршене дигитализације а пре самог издвајања обележја потребно је на одговарајући начин извршити предобраду претходно снимљених говорних сигнала. Пошто систем за обраду говора редукује више спектралне компоненте процесираног говорног сигнала, ради укупног поништавања тог ефекта на улазу се врши његово пропуштање кроз филтар који додатно истиче, критичне, више спектралне компоненте. Ово у суштини представља високофреквентно филтрирање говорног сигнала, које има за циљ поравнање нивоа спектралних компонената. Поменуто се остварује дигиталним високофрекфентним филтром првог реда чија је преносна функција:

$$H(z)=1-az^{-1}, \tag{2.1}$$

при чему се параметар a у општем случају налази у интервалу [0.95, 0.98] и представља степен издизања виших спектралних компонената.

Говор посматраног говорника представља начин његовог изражавања путем гласа. Три основне особине гласа: јачина (интензитет), висина – мера реакције гласница на дејство побудне ваздушне струје и боја на одређени начин пресликавају се у говор посматраног говорника. Говор представља сложен акустички сигнал и може се сматрати као резултат заједничког деловања: семантичких, лингвистичких, артикулаторних и акустичких дејстава (Campbell P. Joseph, Jr., (1997.); Делић Д. Владо и др. (2008.)). Оваква сложеност формирања говора одређује начин пресликавања три основне особине гласа у говор. Може се истаћи да су акустичка дејства првенствено последица анатомских особина говорног тракта посматраног говорника, затим да су артикулаторна дејства првенствено узрокована наученим правилима изговора као и да су семантичка и лингвистичка дејства последица мисаоних процеса у човековом бићу. Сложеност говорног сигнала узрокована низом дејстава одговорних за његов настанак усложњава поступак избора њему одговарајућих обележја на основу којих би се могло вршити успешно разликовање говорника. Стога је потребно дефинисати одговарајућа правила на основу којих би се вршио избор одговарајућих обележја говора. Ради што боље представе говорника кроз обележја говора као и ради што успешнијег рада система за аутоматско препознавање говорника пред сама обележја се постављају следећи захтеви (Mølgaard L Lasse et al., (2005.); Kinnunen Tomi et al., (2010.)):

- испољавају велике међуговорничке разлике и мале варијације за једног посматраног говорника,
- једноставно се могу измерити на основу разматраног говорног сигнала,
- стабилна су током времена тј. не нарушава их евентуална болест говорника или дуготрајне променљивости у његовом гласу,
- дешавају се природно и често у говору,
- мало се мењају приликом промене говорног окружења тј. показују отпорност на дејство шума и изобличења,
- да их је тешко имитирати.

Од три особине гласа: интензитет, висина и боја, боја најбоље испуњава претходно наведене услове постављене пред обележја говорника. Боја гласа је последица његовог хармонијског састава. Гласови различитих говорника се могу одликовати истом јачином односно интензитетом затим и истом висином тј. истом учестаношћу основног хармоника али распоред и величина виших хармоника је својство које одређује боју гласа и разликује се од говорника до говорника. Сазнања о боји гласа говорника су садржана у спектру његовог говорног сигнала.

Природа настанка гласа као процеса проласка побудне ваздушне струје кроз органе говорног тракта указује на конволутивну природу говорног сигнала. Стога посматрајући појаву настанка гласа у домену дискретних сигнала може се извршити њен опис конволуционом једначином:

$$s(n) = e(n) * \theta(n) \tag{2.2}$$

при чему $s(n)$ представља резултантни говорни сигнал тј. еквивалент гласа, $e(n)$ представља побудну ваздушну струју и $\theta(n)$ означава импулсни одзив органа говорног тракта. Овим проласком првобитно формирана ваздушна струја из говорникових плућа на свом путу преко говорног, назалног тракта и усана са којих се емитује у околни простор, доживљава модификације које је у спектралном домену најједноставније пратити у променама на њеној спектралној обвојници. До сазнања о боји гласа потребно је доћи на основу познавања обвојнице спектра говорног сигнала. Обзиром на конволутивну природу говорног сигнала следи да његов спектар има мултипликативну природу, дакле спектар побудне ваздушне струје је амплитудски модулисан преносном карактеристиком органа говорног тракта:

$$S(\omega) = E(\omega) \cdot \Theta(\omega). \tag{2.3}$$

Да би се могао узети у обзир утицај ових компонената на боју гласа, првенствено преносне карактеристике органа говорног тракта $\Theta(\omega)$, потребно је извршити раздвајање компонената

које представљају еквиваленте поменутих делова спектра. Суштина решења је у логаритмовању спектра што је дефинисано као кепструм.

2.2.1. Кепструм

Појам кепструма у разматраној литератури (Mølgaard L Lasse et al., 2005) је дефинисан на следећи начин:

$$c_S(n) = F^{-1}\{\log|F\{s(n)\}|\}, \tag{2.4}$$

при чему F одговара дискретној Фуријеовој трансформацији (*DFT*) посматраног говорног сигнала $s(n)$, док F^{-1} представља инверзну *DFT*. Решење за одвајање компонената које одговарају побудном сигналу односно преносној карактеристици говорних органа уследило је из чињенице да логаритам производа две компоненте представља збир логаритмованих компонената, те применом логаритма на амплитудски спектар множење се преводи у сабирање и добија се адитивна комбинација логаритмованих амплитудских спектралних компонената:

$$\log|S(\omega)| = \log|E(\omega)\cdot\Theta(\omega)| = \log|E(\omega)| + \log|\Theta(\omega)| = C_e(\omega) + C_\theta(\omega). \tag{2.5}$$

Коначно, применом инверзне Фуријеове трансформације добија се кепстрална представа говорног сигнала:

$$c_S(n) = F^{-1}\{C_e(\omega) + C_\theta(\omega)\} = F^{-1}\{C_e(\omega)\} + F^{-1}\{C_\theta(\omega)\} = c_e(n) + c_\theta(n). \tag{2.6}$$

У облику сигнала $c_S(n)$ уочљиве су области у којима доминирају еквиваленти импулсне ваздушне побуде $c_e(n)$ односно импулсног одзива говорних органа $c_\theta(n)$. Утицај импулсне ваздушне побуде доминантнији је при већим вредностима аргумента, односно виши кепстрални коефицијенти углавном носе информацију о побудном сигналу тј. висини гласа. Из тога следи да је ради издвајања дела који носи обележја везана за боју гласа посматраног говора из кепстралне представе потребно издвојити део за ниже вредности аргумента *n*. На овај начин могуће је применом одговарајуће прозорске функције извршити издвајање обележја која носе информацију о боји гласа посматраног говорника.

2.2.2. Мел-фреквенцијски кепстрални коефицијенти (MFCCs)

Претходно описани кепстрални коефицијенти се често користе у применама везаним за аутоматско препознавање говорника. У циљу опонашања људског начина доживљаја различитих учестаности унутар самог аутоматског препознавача говорника, учестаности се посматрају унутар мел скале. На овај начин долази се до појма мел-кепструма, који се рачуна на исти начин као и изворно дефинисани кепструм описан у одељку 2.2.1, с тим што је фреквентна скала трансформисана у мел скалу.

Мел скала је усклађена са људским осећајем висине гласа односно његове учестаности. Њено добијање се врши експериментално, слушаоцу се репродукује тон учестаности 1000 Hz и као његово запажање о висини овог тона се бележи вредност од 1000 mel, ова вредност се користи као мера упоређивања за даље добијање мел скале. Затим се учестаност тона повећава све док слушалац не примети да тон који слуша има дупло већу висину од упоредне вредности и та висина се означава вредношћу 2000 mel. Овај принцип се понавља даље ради добијања осталих вредности мел скале, како за вредности веће од 1000 mel тако и за мање вредности, када се учестаност смањује све док нпр. слушалац не запази да је тон дупло нижи од референтног и та вредност се означава као 500 mel. Експерименти су ипак показали да је међусобна зависност мела и херца линеарна до 500 Hz (Јовичић Т. Слободан, (1999.)) док изнад ове учестаности једнаким променама мела одговара све већа промена у херцима. Дакле зависност осећаја висине звука у мелима и његове учестаности у херцима линеарна је до 500 Hz или приближно линеарна до 1000 Hz, док изнад ових учестаности ова зависност се може

сматрати приближно логаритамском и често се апроксимира једнакошћу (Wildermoth R. B., (2001.); Mølgaard L L et al., (2005.)):

$$f[mel] = 2595 \cdot \log_{10}\left(1 + \frac{f[Hz]}{700}\right). \tag{2.7}$$

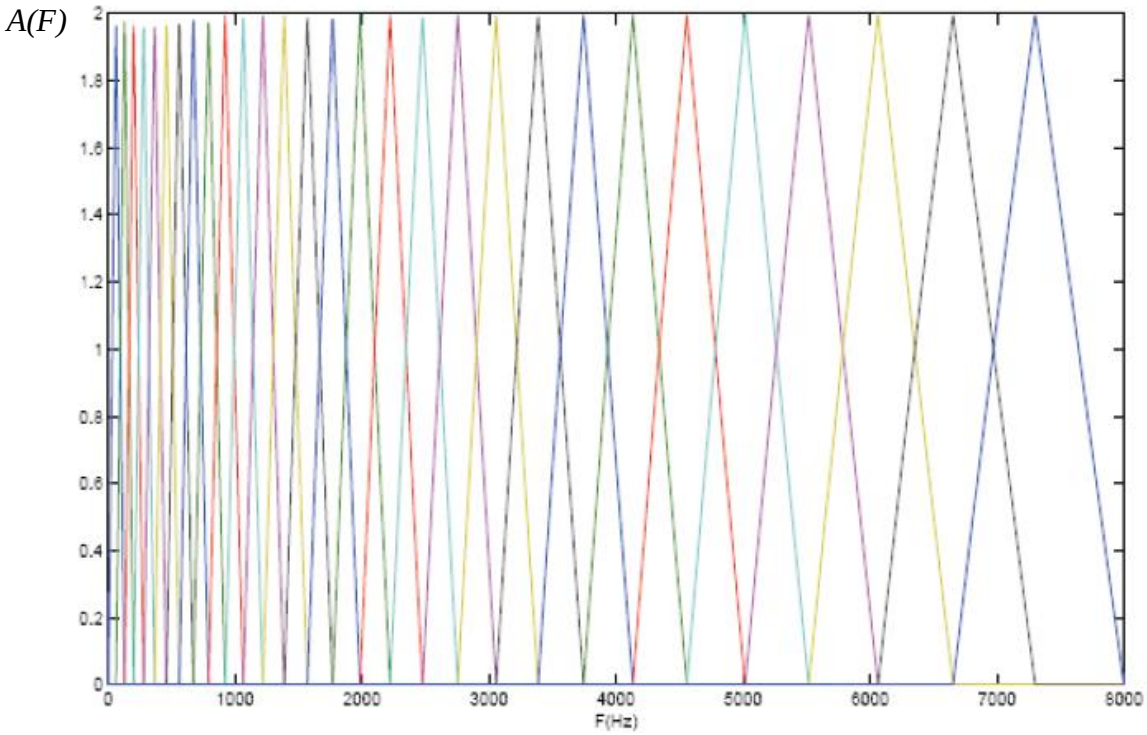
На основу дефиниције кепструма, једнакост 2.4, израчунавање MFCCs може се вршити применом дискретне косинусне трансформације на логаритме енергија унутар целих посматраних рамова говорног сигнала (Molau Sirko et al. (2001.)). Чешће се фреквентни опсег од интереса дели аудиторним критичним опсезима (Lee Chulhee et al. (2003.); Lyon Richard F. et al. (2010.); Siafarikas Mihalis et al. (2007.)). Постојање аудиторних критичних опсега је такође особина која условљава људски доживљај различитих учестаности. Ова појава је везана за чујни доживљај слушаоца приликом слушања два различита тона на различитим учестаностима. Дакле уколико се слушаоцу пусти тон учестаности f_1 која се налази у чујном опсегу тада слушалац има доживљај висине тона у складу са мел скалом и једнакошћу 2.7. Уколико се поред тог тона пусти други тон учестаности f_2 тада звучни доживљај слушаоца зависи од међусобне фреквентне блискости ова два тона. Наиме уколико је фреквентни размак тонова довољно мали тако да се налазе унутар истог критичног чујног опсега долази до појаве маскирања и слушалац чује тон учестаности f_1 повећане гласности у односу на његову изворну гласност у складу са постојањем тона на учестаности f_2, дакле тон на учестаности f_2 у овом случају само повећава гласност тона на учестаности f_1. Уколико је фреквентни размак ових тонова већи од ширине чујног критичног опсега тада слушалац чује два тона на одговарајућим претходно поменутим учестаностима. Дакле чујни доживљај слушаоца као последица ове појаве, који се огледа у томе да ли ће слушалац чути тон повећане гласности или два тона различитих висина зависи од тона који је први наступио као и од међусобног положаја ових тонова на фреквентној скали. Према томе за постојећи тон у одређеном чујном критичном опсегу оно што је својствено доживљају слушаоца огледа се у гласности коју он може чути, обзиром на особине органа чула слуха. Чујни критични опсези се према томе могу описати и енергијом коју садрже. У (Wildermoth R. B., (2001.)) ради израчунавања *M* мел-фреквентних кепстралних коефицијената (енг. Mel-Frequency Cepstral Coefficients – MFCCs) примењена је косинусна трансформација на логаритме енергија у одговарајућим аудиторним критичним опсезима:

$$c_n = \sum_{k=1}^{L} X_k \cos\left[n\left(k - \frac{1}{2}\right)\right], \quad n = 1,2,...,M, \tag{2.8}$$

при чему *L* представља број коришћених филтарских секција и X_k логаритам енергије унутар *k*-те филтарске секције. Филтарске секције су међусобно исте ширине у мелима и симулирају аудиторне критичне опсеге.

Обзиром да се аудиторни критични опсези опонашају низом филтарских секција тј. мел филтар-банком, за израчунате коефицијенте помоћу једнакости 2.8 се може наћи у литератури и назив мел филтар-банка кепстрални коефијенти (енг. Mel Filter-bank Cepstral Coefficients – MFCCs). У различитим радовима може се наићи на различите вредности ширине филтарских секција које се користе. Обично је њихов број већи или једнак 20, тако да је у (Wildermoth R. B., (2001.)) коришћено 20 филтарских секција приближне ширине по 300 mel при чему су суседне филтарске секције биле међусобно померене за по 150 mel, у (Mølgaard L L et al., (2005.)) је примењено 29 филтарских функција на опсегу од 8 kHz (слика 2.4), док је у (Quarteri F. Thomas, (2002.)) дат предлог употребе 24 филтарске секције на опсегу од 4 kHz. Како је приказано на слици 2.4 амплитудски спектар посматраног говорног сигнала се филтрира низом филтарских функција троуганих амплитудских карактеристика чије су централне учестаности и пропусни опсези у складу са аудиторним критичним опсезима претпостављеним за примену у датом случају. Поред троугаоне преносне амплитудске карактеристике може се користити и филтар-банка чији филтри имају правоугаоне амплитудске преносне карактеристике (Lee Chulhee et al., (2003.)). Нулти

кепстрални коефицијент c_0, представља еквивалент средње снаге посматраног рама. Често се приликом дефинисања скупа мел-кепстралних коефицијената који се користе као обележја експлицитно наводи да ли се овај коефицијент узима у вектор обележја или не. У случају да се он не користи у оквиру вектора обележја онда се његова вредност може искористи и као нормализациони фактор за остале MFCCs. Коефицијент c_1 рефлектује прерасподелу енергије између виших и нижих учестаности док остали коефицијенти приказују финије спектралне детаље рама посматраног говорног сигнала.



Слика 2.4. Пример распореда 29 мел филтарских секција (Mølgaard L L et al., (2005.)).

Имајући у виду да су нижи MFCCs нарушени постојањем споропроменљивог адитивног шума као и да су MFCCs вишег реда мање битни за аутоматско препознавање говорника, у циљу постизања бољих могућности израчунатих M MFCCs може се извршити њихово пондерисање тзв. лифтер прозорском функцијом (Wildermoth R. B., (2001.)):

$$\omega_m = 1 + \frac{M}{2}\sin\left(\frac{\pi m}{M}\right), \quad m = 1,2,...,M, \tag{2.9}$$

чиме се врши умањивање утицаја нижих као и виших MFCCs.

Често се MFCCs карактеришу као обележја ниског нивоа (Bimbot Frédéric et al., (2004.)) која су првенствено последица анатомских особености говорника. У односу на такву категоризацију обележја, поред обележја ниског нивоа дефинишу се и прозодијска и обележја високог нивоа. Особеност обележја ниског нивоа огледа се у томе да се она добијају на основу спектралне анализе краткотрајних временских сегмената у посматраним говорним сигналима. Краткотрајни временски сегменти посматраних говорних сигнала издвајају се из изворних говорних сигнала узастопном применом изабране прозорске функције над њима. Имајући у виду временску споропроменљивост говорног сигнала може се претпоставити његова квазистационарност унутар кратких временских интервала. Квазистационарност у овом случају подразумева да су посматране особине тј. обележја говора која је потребно израчунати, у складу са претходним излагањем то би били MFCCs, приближно константне унутар тих временских интервала. Углавном се у радовима примењује трајање ових сегмената у интервалу од 20 до 30 ms (Kinnunen T., Haizhou L., (2010.)). То значи да је потребно извршити прозорирања улазног говорног сигнала

одговарајућом прозорском функцијом чије је трајање у претходно поменутом временском интервалу. Ова функција треба да испољава што мање спектрално цурење, односно потребно је да је оно, односно бочни лобови примењене прозорске функције, испод динамике говорног сигнала. Овај услов задовољавају Ханингова односно Хамингова прозорска функција:

$$\omega(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), \ 0 \le n \le N-1, \tag{2.10}$$

те се као таква она врло често користи приликом дигиталне обраде аудио сигнала.

Иако примене прозорске функције кроз примену поступака брзе Фуријеове трансформације (енг. Fast Fourier Transform – FFT) значајно убрзавају поступак издвајања обележја, ипак у извесној мери и нарушавају сама обележја говора. Ради смањивања тог утицаја прозорирања на обележја која треба издвојити, врши се међусобно делимично преклапање суседних прозора на 30-50% свог трајања. Ово се постиже тако што се почетак сваке наредне прозорске функције у односу на претходну помера унапред за 10 ms до 20 ms. На овај начин сваки говорни сегмент (рам) тј. прозор посматраних говорних сигнала се представља одговарајућим вектором мел-фреквентних кепстралних коефицијената.

Као последица научених и мисаоних процеса, могу се користити и тзв. обележја вишег нивоа (Bimbot Frédéric et al., (2004.)), која у обзир узимају како речи које посматрани говорник користи тако и прозодијске особености његовог говора. У односу на дужине сегмената говора који се анализирају ради израчунавања MFCCs, да би се добиле информације о прозодији посматраног говора потребно је извршити анализу дужих временских сегмената као што су: слогови, речи и искази и на основу њихове анализе проценити начин усменог изражавања посматраног говорника. Обзиром да обележја вишег нивоа у обзир узимају речник посматраног говорника следи да је ради њиховог обезбеђивања потребно извршити спрегу између система за аутоматско препознавање говорника и препознавање самог говора.

Приликом експеримената аутоматског препознавања говорника у овом раду су као обележја која описују обвојницу спектра говорног сигнала коришћени MFCCs. Као такви MFCCs зависе и од боје гласа посматраног говорника. MFCCs рачунати на основу једнакости 2.8 зависе од енергије унутар посматраних аудиторних критичних опсега што указује на чињеницу да ова обележја нису зависна само од боје гласа посматраног говорника. Енергија коју носи говорни сигнал зависи од распореда фонема у њему, тј. од текстуалне садржине посматраног говора. Може се очекивати да текстуална зависност мел-фреквентних кепстралних коефицијената утиче на повећање разликовања ових обележја за истог говорника, чиме се у извесној мери одступа од првог наведеног захтева који се поставља пред обележја говорника. Следи да MFCCs нису идеална обележја на основу којих се може извршити разликовање говорника. Стога се ради добијања успешнијег скупа обележја могу спровести додатне поправке на изворно израчунате MFCCs помоћу једнакости 2.8. Један од начина се огледа у примени одређених трансформација на претходно израчунате MFCCs (Wu Dalei et al., (2008.)). У следећем поглављу приказан је утицај примене поступка анализе главних компонената (енг. Principal Component Analisys – PCA). Други поступак би могао бити заснован на детаљнијем разматрању утицаја појединих чиниоца на резултујуће MFCCs, као што су облик аудиторних критичних опсега и енергија обухваћена у њима.

2.2.3. Динамичка обележја

Да би се описала промена MFCCs у времену уведена су динамичка обележја. На овај начин коришћена обележја директно носе и информацију о променама између суседних рамова говорног сигнала који се анализирају. Ова информација је садржана унутар самих MFCCs тако да се може рећи и да су динамичка обележја сувишна тј. редундантна. Ради израчунавања ових обележја користе се полиномијалне апроксимације првог и другог извода кепстралних коефицијената (Bimbot Frédéric et al., (2004.)):

$$\Delta c_n = \frac{\sum_{k=-l}^{l} k \cdot c_{n+k}}{\sum_{k=-l}^{l} |k|}, \quad \Delta\Delta c_n = \frac{\sum_{k=-l}^{l} k^2 \cdot c_{n+k}}{\sum_{k=-l}^{l} k^2}. \tag{2.11}$$

На овај начин динамичка обележја су представљена линеарном комбинацијом $2 \cdot l + 1$ MFCCs, при чему l представља ред примењене линеарне предикције.

У решењима аутоматског препознавања говорника која користе векторе обележја засноване на MFCCs и њиховим изводима често се у вектору обележја говора користи 12 MFCCs и по 12 њихових првих и других извода, тако да коначни вектор обележја има 36 елемената. Узимањем у обзир и нултог кепстралног коефицијента заједно са његовим првим и другим изводом вектор обележја има 39 елемената.

2.2.4. Висина гласа

Кепстрални коефицијенти су често коришћена обележја у циљу препознавања говорника, но и поред тога ради даљег побољшања могућности препознавача могу се користити и особине говора које су везане за обележја побуде говорног тракта као нпр. висина гласа и облик глоталних импулса. Висина гласа, такође позната и као основна учестаност (f_0) представља врло важну особину говора и дефинише периодичност говорног сигнала. Када је говорни сигнал звучан односно када се одликује периодичношћу висина гласа одговара учестаности осциловања гласних жица. Мушки гласови се одликују нижом док женски и дечји имају вишу основну учестаност гласа. Сматра се да висина гласа представља једно од важнијих обележја говора које људи користе ради разликовања говорника. Осим ње а као логична последица, важну селективну особину представља и мера количине звучности односно периодичности у говорном сигналу. Висина гласа у поређењу са претходно описаним спектралним обележјима (кепстрални коефицијенти) има битну предност која се огледа у њеној независности од фреквентних особина преносног система.

Почеци имплементације поступака процене висине гласа ослањају се на нискофреквентно филтрирање говорног сигнала чиме би се из говорног сигнала уклањали сви виши хармоници сем првог. Овакви поступци суочавају се са бројним проблемима од којих је један од основних могућност непостојања првог хармоника у посматраном говорном сигналу нпр. као последица његовог претходног проласка кроз телефонски канал чији је пропусни опсег $200 - 3400Hz$.

Висина гласа се врло ретко директно користи као обележје говора зато што није увек присутна у рамовима говора, приближно у 50% рамова који су иначе безвучни она се не може одредити, а такође и методи који се користе за њену процену нису сасвим поуздани.

2.3. МОДЕЛИ ГОВОРНИКА

Имајући у виду да аутоматско препознавање говорника у основи представља један вид препознавања облика потребно је да систем који врши препознавање говорника створи одговарајућу представу о говорницима као објектима препознавања. На основу расположивих вектора обележја за сваког од посматраних говорника аутоматски препознавач говорника треба да на одређени начин изврши њихову компактну представу у виду одговарајућих модела. Приликом одлучивања секвенце вектора обележја које представљају тест говор пореде се са сваким од постојећих модела говорника при чему препознавач доноси одлуку на основу највеће мере сличности између посматраног тест говора и расположивих модела. Квалитет примењеног модела и сам начин одлучивања значајно утичу на ефикасност система за препознавање говорника.

У актуелним применама поступака аутоматског препознавања тежи се остварењу вештачке интелигенције која ће бити способна да извршава препознавања у задатом

окружењу. Стога је пројектовање система за препознавање потребно извршити на тај начин да буде функционалан у окружењу велике количине података интересантних за његов рад. У случају аутоматског препознавања говорника интересантно окружење препознавача представљају вектори MFCCs који представљају посматране говорне сигнале. Вредности мел-фреквенцијских кепстралних коефицијената се могу апроксимативно описати одговарајућим случајним променљивама тј. њима припадајућим функцијама густине расподеле. Случајне променљиве у том случају представљају моделе говора посматраних говорника. Овакав начин моделовања се декларише као стохастички начин моделовања говорника.

Изговорена мисао се може схватити као низ изговорених фонема. У текстуалном смислу она се може представити низом знакова. У зависности од наметнутих ограничења која се односе на текстуалну садржину говора на основу кога аутоматски препознавач говорника врши препознавање, проблем аутоматског препознавања говорника може бити независан или зависан од текста. Очекивани тип овог проблема узрокује коришћење одговарајућег стохастичког модела ради моделовања говорника. У случају аутоматског препознавања говорника независног од текста ради моделовања говорника примењују се модели засновани на Гаусовим вишедимензионалним расподелама. Уколико се у оквиру таквих модела користи више Гаусових вишедимензионалних расподела такав модел се назива модел мешавине Гаусових расподела (енг. Gaussian Mixture Model - GMM). Када је говор посматраних говорника унапред познате текстуалне садржине или је његова текстуална садржина последица пермутације унапред познатог низа речи (Che Wei Chi, et al., 1996) примењује се моделовање говорника скривеним Марковљевим моделима (енг. Hidden Markov Model - HMM).

Пошто системи за аутоматско препознавање говорника граде тј. генеришу GMM или HMM на основу расположивих вектора обележја за посматраног говорника, често се ови модели декларишу као генеративни модели. С друге стране проблем моделовања се може тако поставити да циљ буде формирање параметарских модела који ће максимизовати међусобна разликовања. Такви модели се често називају и дискриминативним моделима и један од модела са таквим особинама је и вештачка неуронска мрежа (енг. Artificial Neural Network – ANN).

Одлучивачи засновани на стохастичким моделима одлуку о препознавању доносе на основу израчунате мере веродостојности тест говорних узорака у односу на познате моделе говорника.

Нека је λ^S стохастички модел добијен на основу говорних тренинг секвенци *s*-тог говорника. Препознавач у својој бази има N_S стохастичких модела за сваког од N_S говорника на које је претходно обучен. Означивши са $Y=\left(y_1,y_2,...,y_L\right)$ секвенцу вектора обележја тест говорног узорка од L говорних рамова, циљ препознавања јесте одређивање ком од N_S говорника посматрана тест секвенца припада. Овај проблем своди се на одређивање условне вероватноће:

$$p\left(Y\middle|\lambda^S\right)=p\left(y_1,y_2,...,y_L\middle|\lambda^S\right), \tag{2.12}$$

за $S=1,2,...,N_S$, те одређивање идентитета говорника на основу њене максимизације:

$$S^*=\arg\max_{1\le S\le N_S} p\left(Y\middle|\lambda^S\right). \tag{2.13}$$

Уколико суседни вектори обележја нису међусобно корелисани односно уколико су независни једнакост (2.12) се може написати у облику производа:

$$p\left(Y\middle|\lambda^S\right)=\prod_{i=1}^{L} p\left(y_i\middle|\lambda^S\right). \tag{2.14}$$

Обзиром да су вектори обележја еквиваленти посматраног говорног сигнала који представља смислено изговорени низ фонема, без обзира да ли контекст има или нема значењски смисао,

следи неизбежна међусобна зависност суседних вектора обележја. Апроксимативно ова међузависност се може занемарити тако да се према томе проблем препознавања своди на одређивање вероватноће тест вектора обележја за конкретни модел говорника, $p\left(y_i \middle| \lambda^S\right)$. Имајући у виду да ова вероватноћа зависи од примењеног модела говорника следи да се она може одредити кроз претпоставку GMM или HMM као одговарајућих генеративних модела говорника или ANN као дискриминативног начина моделовања говорника.

2.3.1. Мешавина Гаусових расподела (GMM)

Мотивација увођења GMM лежи у тежњи за моделовањем акустичког простора говорника коришћењем одређеног броја акустичких класа при чему обично свака од њих апроксимативно моделује по један фонем. Овим начином се вероватноћа припадности вектора обележја, посматраног *n*-тог говорног рама, *s*-том говорнику, моделује линеарном комбинацијом *M* пондерисаних мултидимензионалних Гаусових функција густине расподеле вероватноћа:

$$p\left(y_n \middle| \lambda^S\right) = \sum_{i=1}^{M} p_i^S \cdot b_i^S\left(y_n\right), \tag{2.15}$$

при чему $b_i^S\left(y_n\right)$ представља Гаусову вишедимензионалну функцију густине расподеле вероватноћа *i*-те компоненте мешавине односно *i*-те акустичке класе, вектора средњих вредности μ_i^S и коваријансне матрице Σ_i^S:

$$b_i^S\left(y_n\right) = \frac{1}{\sqrt{\left(2\cdot\pi\right)^D \left|\Sigma_i^S\right|}} \cdot e^{-\frac{1}{2}\cdot\left(y_n-\mu_i^S\right)^T\cdot\left(\Sigma_i^S\right)^{-1}\cdot\left(y_n-\mu_i^S\right)}. \tag{2.16}$$

D представља димензионалност простора обележја. Тежински коефицијенти појединих Гаусових компонената, p_i^S, i=1,2,…,M, (једнакост 2.15), испуњавају ограничење $\sum_{i=1}^{M} p_i^S = 1$. Пошто кепстрална обележја која се уобичајено користе у системима за препознавање говорника показују висок степен међусобне некорелисаности могуће је искористити ову особину у циљу смањења количине података неопходне за израчунавање у (2.16) и (2.15), кроз претпоставку да је коваријансна матрица коришћена у (2.16) дијагонална. На основу претходних разматрања, Гаусов параметарски модел *S*-тог говорника се може представити уређеном тројком:

$$\lambda^S = \left\{p_i^S, \mu_i^S, \Sigma_i^S\right\}, \qquad 1 \le i \le M. \tag{2.17}$$

Пошто многи говорни језици имају око 30 до 40 гласова, обично се за вредност *M* у Гаусовим моделима узима вредност 32. Овакав начин моделовања примењује се у системима за препознавање говорника независним од изговореног текста.

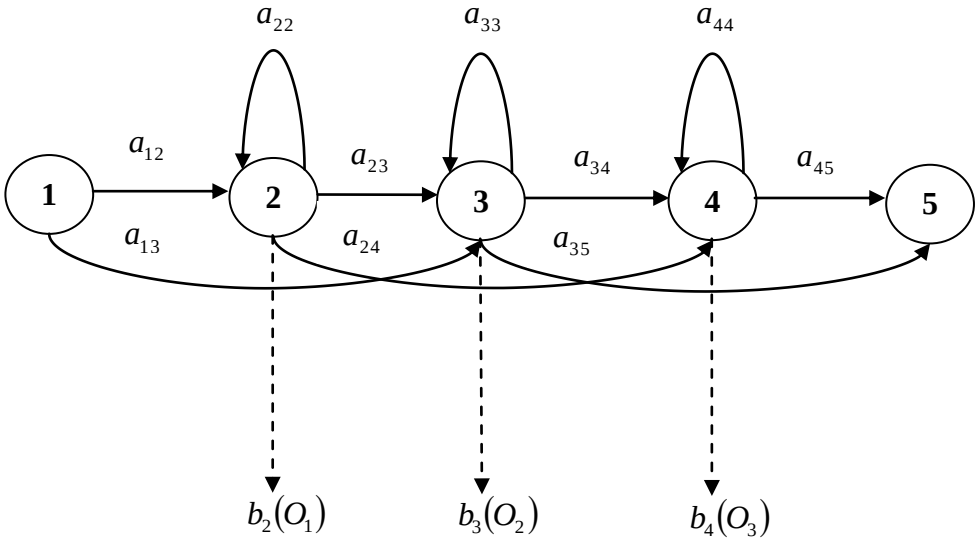
2.3.2. Скривени Марковљев модел (HMM)

Скривени Марковљев модел представља обично коришћени модел за моделовање говорних секвенци. Стога су системи за препознавање говорника засновани на оваквом моделу подесни за аутоматско препознавање говорника зависно од изговореног текста. У светлу акустичких класа поменутих у претходном делу посвећеног опису GMM, које се најчешће везују за фонеме или евентуално делове фонема, говорне секвенце представљају низове овако дефинисаних акустичких класа.

Сходно сврси примене овакав модел се одликује секвенцом стања, *i*=1,2,…,*N*, чији број варира у зависности од дужине акустичке целине која се моделује њиме. У случају моделовања фонема и евентуално и субфонема HMM обично има пет стања, као на слици 2.5, док у случају када се моделују дуже изговорене целине као што су речи или дуже изговорене форме, број стања се сразмерно увећава. Пошто се на овај начин моделују

секвенце вектора обележја, опсервација, које одговарају различитим изговорима посматране текстуалне садржине, следи потреба да се појава сваког од вектора обележја из те секвенце опише адекватном густином расподеле. Приликом моделовања вектори обележја који одговарају говорним секвенцама су познати тако да се на основу њихових вредности могу проценити и густине расподела које описују њихово појављивање. Свака од ових расподела се додељује преласку у одређено стање скривеног Марковљевог модела, тако да ради употпуњења модела потребно је одредити и вероватноће прелаза између стања. Стога се често каже да су стања односно вероватноће прелаза између њих скривене иза густина расподела у стањима на основу чега је модел и добио назив скривени Марковљев модел.

Уводи се додатно ограничење по ком се првом и последњем стању не придружују вектори обележја тако да се она често карактеришу као неемитујућа стања. Такође, она се одликују и нултом сопственом вероватноћом прелаза, тако да се она могу посматрати само као битна ради задовољења саме форме Марковљевог ланца, док се суштина скривеног Марковљевог модела крије иза емитујућих стања, стања која се одликују одговарајућим густинама расподеле вектора обележја. Из првог стања HMM може само прећи у неко друго стање, док по уласку у последње стање прелази више нису могући, те се за њега може рећи да је апсорбујуће стање. Према томе HMM се како секвенцом стања и вероватноћама прелаза између њих, $a_{ij} = p\left(q_i \middle| q_j\right)$, карактерише и густином расподеле вероватноћа вектора обележја по емитујућим стањима, $b_i(O_n)$, при чему индекс *n* означава редни број вектора обележја у оквиру секвенце вектора обележја која се посматра.



Слика 2.5. Скривени Марковљев модел са три емитујућа стања.

За свако стање осим последњег, из кога се врши прелаз у неко друго, важи да је сума вероватноћа прелаза једнака јединици:

$$\sum_{j=1}^{N} a_{ij} = 1. \tag{2.18}$$

HMM приказан на слици 2.5 погодан је за моделовање фонема као посматраног говорног узорка. У конкретном случају на слици 2.5 моделоване секвенце би садржале три вектора обележја чије појаве се описују густинама расподела вектора обележја $b_2(O_1)$, $b_3(O_2)$, $b_4(O_3)$. За већ израчунате вероватноће прелаза модел приликом уласка у одређено емитујуће стање емитује односно генерише одговарајући вектор обележја са вероватноћом која је дефинисана придруженом густином расподеле. За модел фонема који би био направљен по узору на модел са слике 2.5 уласци у стања, редом, *i*=2, *i*=3, *i*=4, резултују генерисањем

одговарајућих вектора обележја са вероватноћама генерисања која су одређена припадајућим густинама расподела вектора обележја, $b_2(O_1)$, $b_3(O_2)$ и $b_4(O_3)$. Најчешће је расподела разматраних вектора обележја по стањима HMM-а моделована мешавином Гаусових расподела. На тај начин HMM представља Марковљев модел чије је свако емитујуће стање, у зависности од путање која је до њега довела, моделовано одговарајућим GMM-ом, тј. HMM представља низ међусобно повезаних, на одговарајући начин вероватноћама прелаза, модела заснованих на Гаусовим расподелама.

Уколико се за пример модела на слици 2.5 претпостави да су секвенцама прелаза 1-2, 1-2-3, 1-2-3-4, придружене појаве вектора обележја, O_1, O_2, O_3, тада вероватноћа појаве ове секвенце вектора обележја с обзиром на примењени модел износи:

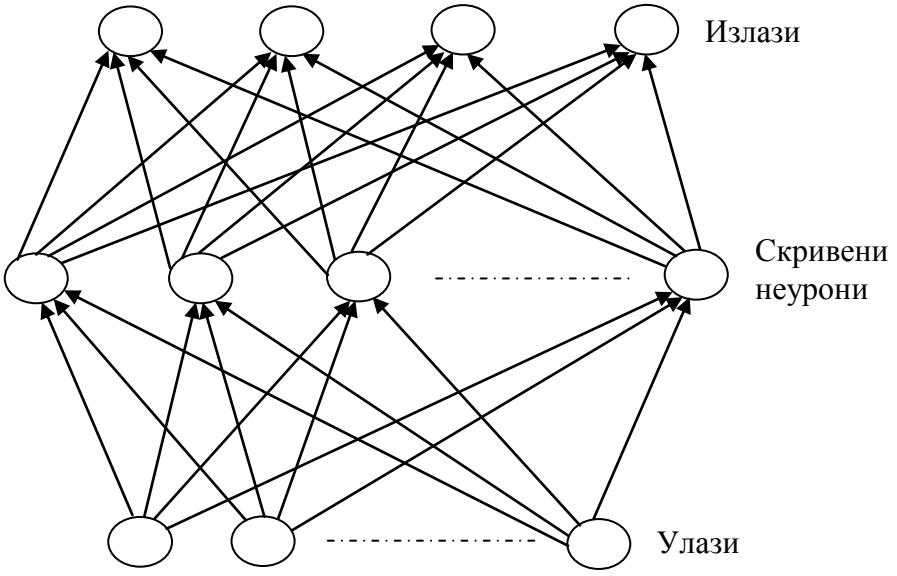
$$p(O_1O_2O_3|\textit{модел}) = a_{12}\cdot b_2(O_1) + a_{12}\cdot a_{23}\cdot b_3(O_2) + a_{12}\cdot a_{23}\cdot a_{34}\cdot b_4(O_3). \qquad (2.19)$$

Максимизација мере веродостојности посматраног низа рамова тј. вектора обележја у односу на примењени модел може се извршити Витербијевим алгоритмом.

2.3.3. Вештачке неуронске мреже (ANN)

Одлучивачи засновани на вештачким неуронским мрежама примењују се у системима за аутоматско препознавање говорника како у случајевима зависним тако и независним од изговореног текста. Оне се могу изузетно ефикасно обучити у циљу извршења сложених пресликавања између улаза и излаза те као такве могу апроксимирати вероватноће класа на које су трениране. ANN се састоје од малих функционалних јединица (неурона) које су међусобно повезане ради добијања потребне преносне функције. Најчешће коришћена форма неуронске мреже је тзв. одлучивач у више нивоа (енг. Multi-Layer Perceptron – MLP).

MLP садржи улазни ниво, одређен број скривених нивоа и излазни ниво (слика 2.6). Улазни ниво је нефункционалан и одговоран је за преусмеравање улазних информација ка неуронима у скривеном нивоу. Преостали нивои су фукционални и као такви описани припадајућим пондерисаним улазима и нелинеарним активационим функцијама. Сваки неурон у излазном нивоу директно се односи на одговарајућу класу препознавања. Преко улазних неурона улазна информација се похрањује у MLP и сваки излазни неурон садржи резултујућу вероватноћу за ту одређену класу. Након тога информација са улаза се класификује у класу чији одговарајући излазни неурон има највећи резултат.



Слика 2.6. Вештачка неуронска мрежа у два нивоа која има 4 излазна неурона.

У применама за аутоматско препознавање говорника ANN се најчешће јављају у следеће две опште форме:

- једна MLP је тренирана са N_S излазних неурона, при чему је N_S број трениpаних говорника,
- по једна MLP је тренирана за сваког од N_S говорника, при чему је сада број мрежа једнак броју говорника, те свака од њих садржи два излазна неурона који се односе на трениpаног говорника и на остале говорнике.

Тренирање односно обука MLP-а обавља се тзв. агоритмом пропагационе грешке уназад (енг. error back-propagation algorithm). Овај алгоритам итеративно проналази вредности тежинских коефицијената чворова MLP-а, које се иницијално постављају на случајан начин у опсегу од -0.5 до 0.5. Алгоритам се извршава проласком у напред и у назад.

Приликом проласка у напред сви тренинг вектори и њима одговарајуће лабеле се доводе на улаз MLP-а, на основу чега се процењује грешка добијене вредности на излазу. Излаз *k*-тог излазног неурона за дати *n*-ти улазни тренинг вектор y_n, може се представити у форми:

$$O_k(n) = f\left(\sum_i \omega_{ki} \cdot f\left(\sum_j \omega_{ij} \cdot y_n(j)\right)\right), \tag{2.20}$$

при чему индекси *j* односно *i* означавају одговарајуће улазне односно скривене неуроне, док пондери ω_{ij} односно ω_{ki} одговарају тежинским коефицијентима везе између неурона у посматраном и претходном нивоу. Активациона функција има облик:

$$f(x) = \frac{1}{1+e^{-x}}. \tag{2.21}$$

Уколико $d_k(n)$ представља идеалну вредност *k-тог* излазног неурона за посматрани *n-ти* тренинг вектор, укупна квадратна грешка коришћеног *MLP-а* износи:

$$E = \sum_n \sum_k \left(O_k(n) - d_k(n)\right)^2. \tag{2.22}$$

Приликом проласка у назад, од излазног ка скривеном нивоу, пондери се модификују у циљу минимизације ове грешке. Процес итерација израчунавања грешке проласком у напред односно кориговање пондера проласком у назад траје све док тежински коефицијенти MLP-а конвергирају локалном минимуму квадратне грешке *E*.

Током тестирања, сваки од *L* тест вектора обележја $Y = (y_1, y_2, ..., y_L)$ доводе се на улаз MLP-а. Препознавач доноси одлуку на основу посматрања вредности излазних неурона као мере веродостојности посматраног модела у односу на препознавани вектор обележја.

3. СМАЊЕЊЕ СТАНДАРДНО КОРИШЋЕНОГ СКУПА ОБЕЛЕЖЈА УЗИМАЊЕМ У ОБЗИР ГЛАВНИХ ПРАВАЦА ЕНЕРГИЈЕ У ИЗВОРНОМ ПРОСТОРУ ОБЕЛЕЖЈА

Приликом коришћења MFCCs као обележја ради имплементације аутоматског препознавања говорника често се користи првих 12 MFCCs заједно са нултим мел-фреквенцијским кепстралним коефицијентом и њихови први и други изводи. На овај начин сваки рам посматраног говорног сигнала се представља одговарајућим 39-то димензионалним вектором обележја. Да би се изградио што комплетнији модел говора посматраног говорника потребно је располагати довољним бројем вектора обележја како би се што већи број могућих варијабилности у његовом гласу узео у обзир. Број коришћених обележја говора условљава и величину потребног тренинг скупа односно број потребних тренинг вектора обележја ради конструкције адекватних модела говорника. Имајући у виду експоненцијалну зависност између величине потребног тренинг скупа и димензионалности коришћеног вектора обележја следи потреба избора потребног и довољног скупа обележја ради тачног препознавања аутоматског препознавача говорника.

Дакле потребно је одредити минималан број обележја говора на основу којих је могуће извршити разликовање говорника. Минималност скупа обележја последица је избора међусобно независних обележја говора. Посматрајући овај захтев постављен пред обележја говора са становишта теорије информација следи потреба за проналажењем таквих обележја која ће показивати минималну могућу редундансу међу собом. У претходно поменутом често коришћеном 39-то димензионалном вектору обележја први и други изводи MFCCs се најчешће апроксимативно изражавају линеарном комбинацијом одговарајућих MFCCs, као што је то исказано једнакостима 2.11. На основу постојања линеарне зависности може се говорити о информационој сувишности тј. редунданси обележја која представљају први односно други извод у односу на обичне MFCCs тј. у односу на првих 13 обележја у коришћеном 39-то димензионалном вектору обележја.

Битна особина посматраног скупа података огледа се у концетрисаности вредности односно у одређивању домена у простору посматраног скупа података који одговарају најизраженијим правцима постојања посматраног скупа података. Ова особина се може проценити на основу разматрања енергетског профила посматраног скупа података. Скуп података од интереса приликом имплементације аутоматског препознавања говорника представљају вектори обележја добијени на основу анализе говора посматраних говорника. Представа енергетске прерасподеле тј. енергетска слика посматраног скупа вектора обележја огледа се у коваријансној матрици.

Уколико се *d*-димензионални вектори обележја посматраног говора поређају у матрицу тако да сваки од *n*-вектора обележја попуњава одговарајућу колону посматране матрице добија се матрица вектора обележја $X = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}_{d \times n}$. Вектори обележја се групишу у матрицу ради дефинисања начина одређивања коваријансне матрице. Претходно је потребно дефинисати вектор средњих вредности μ који одговара посматраној матрици *X*. Претпостављајући да се елементи у некој од *d* врста посматране матрице могу сматрати елементима једне од *d* одговарајућих случајних променљивих, елементи вектора средњих вредности μ се дефинишу као очекивања сваке од *d* поменутих случајних променљивих, или посматрајући целокупни скуп вектора обележја, као очекивање матрице вектора обележја *X* по њеним врстама: $\mu = E[X]$. За коначан скуп од *n* вектора обележја вектор средњих вредности се може израчунати коришћењем матричне форме као:

$$\mu = \frac{1}{n} \cdot \left[\sum_{i=1}^{n} x_{1,i} \quad \sum_{i=1}^{n} x_{2,i} \quad \dots \quad \sum_{i=1}^{n} x_{d,i} \right]^T . \tag{3.1}$$

Коваријансна матрица посматраног скупа вектора обележја тј. матрице вектора обележја се дефинише као очекивање производа нормализоване матрице вектора обележја

вектором средњих вредности и транспонованог облика овако добијене матрице: $\Sigma = E[(X-\mu)\cdot(X-\mu)^T]$. Такође у случају реалних проблема над коначним скупом од n вектора обележја обично се израчунавање коваријансне матрице врши на основу једнакости:

$$\Sigma_{d\times d} = \frac{1}{n-1}\cdot(X-\mu)\cdot(X-\mu)^T. \tag{3.2}$$

Елементи коваријансне матрице:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1d} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{d1} & \Sigma_{d2} & \cdots & \Sigma_{dd} \end{bmatrix}, \tag{3.3}$$

по својој природи представљају одговарајућа енергетска поља у посматраном скупу обележја. Дијагонални елементи представљају варијансе односно енергије дуж сваке од d димензија, док вандијагонални елементи репрезентују меру међусобне повезаности димензија, тако да на овај начин коваријансна матрица даје слику прерасподеле енергије посматраног скупа вектора обележја дуж унапред дефинисаних димензија. Такође евидентна је симетричност вандијагоналних елемената у односу на главну дијагоналу коваријансне матрице. Постојање информационе сувишности између коришћених d димензија указује да коришћене димензије не показују главне правце простирања енергије посматраних говорних сигнала. Поступак којим је могуће одредити међусобну истакнутост енергетских праваца у посматраном скупу вектора обележја заснива се на израчунавању својствених вектора коваријансне матрице Σ посматране матрице вектора обележја X и сходно томе може се именовати као анализа главних компонената (енг. Principal Component Analisys – PCA). Својствени вектори одговарају новим димензијама дуж којих се сада врши анализа посматраног скупа вектора обележја X.

PCA се заснива на вези између коваријансне матрице, њеним својственим векторима и њима одговарајућим својственим вредностима:

$$\Sigma\cdot EVect = EVect\cdot EVal. \tag{3.4}$$

Посматраном скупу d-димензионалних вектора обележја одговара коваријансна матрица која има d врста и d колона. Својство коваријансне матрице Σ јесте да за њу постоји d својствених вектора и њима одговарајућих d својствених вредности повезаних једнакошћу 3.4. Примена PCA ради смањења димензионалности посматраних вектора обележја при аутоматском препознавању говорника заснива се на примени матричне трансформације чија форма зависи од својствених вектора који се узимају у обзир. Имајући у виду да својствени вектори репрезентују правце нових димензија у посматраном скупу вектора обележја следи да број својствених вектора узетих у разматрање дефинише димензионалност пресликаног тј. трансформисаног скупа вектора обележја. Сваком својственом вектору одговара својствена вредност. Величина својствене вредности $EVal_i$, i=1,2,...,d, описује истакнутост i-те димензије коју репрезентује одговарајући својствени вектор $EVect_i$ у посматраном скупу вектора обележја дефинисаних матрицом X. Трансформациона матрица C_{tr} се формира на основу избора одговарајућих својствених вектора. У тежњи да се овом трансформацијом истакну правци највеће варијансе тј. енергије унутар изворног скупа података трансформациона матрица се формира својственим векторима којима одговарају највеће својствене вредности. Ради формирања трансформационе матрице која може да изврши пресликавање из изворног d-димензионалног простора обележја у мање димензионалан простор димензионалности реда m потребно је формирати трансформациону матрицу коришћењем m својствених вектора највећих својствених вредности поређаних по опадајућем редоследу одговарајућих својствених вредности. Трансформисани вектор обележја се израчунава као:

$$x_{tr} = C_{tr}^{\ T}\cdot x. \tag{3.5}$$

Обзиром на чињеницу да PCA резултује димензијама које указују на главне правце присуства поматраних говорних сигнала као и да су полазне димензије показивале извесну количину међусобне редундансе следи могућност примене PCA ради смањења димензионалности изворног скупа вектора обележја. Стога је у наставку анализиран утицај смањења димензионалности 39-то димензионалног често коришћеног вектора обележја на тачност препознавања говорника применом PCA.

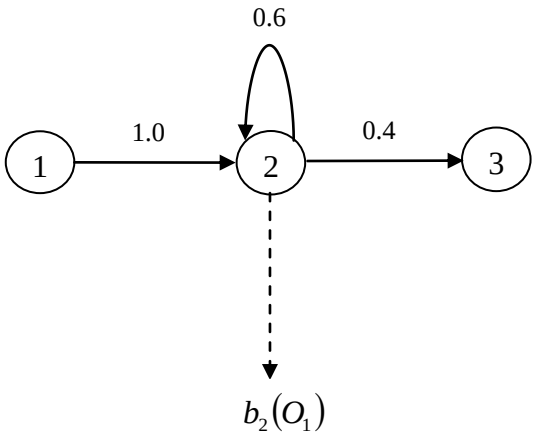
3.1. ЕКСПЕРИМЕНТАЛНИ УСЛОВИ АНАЛИЗЕ УТИЦАЈА PCA

Ради експеримената коришћен је део говорне базе преузет од стране АлфаНум групе потекле са Факултета техничких наука у Новом Саду. Ради експеримената употребљен је део говорне базе који садржи изговоре 5 мушких и 5 женских говорника. Говорници су изговорили исте текстуалне садржине при чему је сваки од говорника једну текстуалну садржину изговорио једанпут. Снимци се одликују тематско-значењском садржином тако да су сходно томе подељени у три групе:

- цифре – за сваког говорника постоје два снимка изговора низа цифара 1 – 2 – 3 – 4 – 5 и 6 – 7 – 8 – 9 – 0,
- речи – за сваког говорника постоји 11 снимака унапред дефинисаног низа речи при чему су низови речи исти за све говорнике,
- имена – по један снимак за сваког говорника који садржи идентификациони број посматраног говорника и његово презиме и име.

Коришћени снимци су у wav (енг. Waveform Audio File Format – WAVE или WAV) формату учестаности одабирања 22050Hz квантизационе резолуције 16 бита по одбирку. Имајући у виду текстуалну садржину расположивих говорних снимака, снимци из група "Цифре" и "Речи" су коришћени за обуку модела говорника док је тестирање препознавања вршено над снимцима из групе "Имена".

Издвајање обележја говора, моделовање говорника и тестирање препознавања вршено је употребом програмског алата HTK (енг. Hidden Markov Models Toolkit – HTK) (Young Steve et al., 2009). Ради анализе посматраних говорних сигнала разматрани су рамови говора настали применама Хамингове прозорске функције трајања 25ms на сваких 10ms посматраног говорног сигнала. Сваки рам говора анализираних говорних сигнала представљен је 39-то димензионалним вектором обележја тј. помоћу нултог и првих 12 MFCCs и њима одговарајућим првим и другим изводима. Обзиром да је сваки говорник једну текстуалну садржину изговорио једанпут примењен је стохастички начин моделовања независан од изговореног текста. На овај начин говор сваког говорника описан је одговарућим HMM који има једно емитујуће стање при чему је густина расподеле обележја у емитујућем стању описана помоћу GMM (слика 3.1, једнакости 3.6 и 3.7).



Слика 3.1. Примењени модел говорника.

$$b_2(O_1) = \sum_{k=1}^{64} \frac{1}{64} \cdot \mathrm{N}(O_1, \mu_k, \Sigma_k) \tag{3.6}$$

$$\mathrm{N}(O, \mu, \Sigma) = \frac{1}{\sqrt{(2 \cdot \pi)^d \cdot |\Sigma|}} \cdot e^{-\frac{1}{2} \cdot (O-\mu)^T \cdot \Sigma \cdot (O-\mu)} \tag{3.7}$$

Ради реализације јединствене PCA трансформације за све посматране говорнике у циљу израчунавања коваријансне матрице посматран је комплетан скуп за обуку обзиром на коришћену говорну базу. Матрицу X су чинили сви вектори обележја коришћени за обуку модела говорника. Применом програмског пакета MATLAB на основу једнакости 3.2 израчуната је коваријансна матрица Σ а затим и њени својствени вектори и њима одговарајуће својствене вредности. Ради обраде тј. трансформације MFCCs потребно је запис унутар бинарних mfc HTK фајлова који садрже векторе MFCCs прилагодити формату који је погодан за примену програмског пакета MATLAB. Стога су mfc фајлови конвертовани у текстуалне txt фајлове из којих су вектори MFCCs учитавани у MATLAB ради примене PCA над њима.

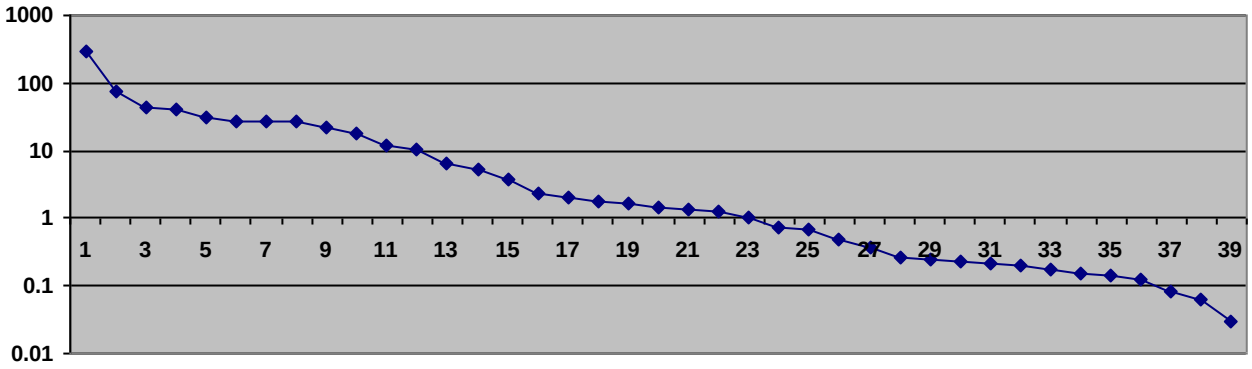
Бинарни HTK формат записа унутар mfc фајла организован ради чувања вектора обележја састоји се из заглавља и дела који садржи податке тј. на одговарајући начин кодоване вредности у посматраним векторима обележјима. Заглавље чини 12 бајта по следећем редоследу:

- број узорака у фајлу тј. број вектора обележја – 4-бајтна целобројна вредност (4-byte integer),
- период посматрања вектора обележја исказан у односу на јединицу мере коју користи HTK и износи 100ns – 4-бајтна целобројна вредност,
- број бајта по вектору обележја – 2-бајтна целобројна вредност,
- ознака података који су у фајлу – 2-бајтна целобројна вредност.

Дефинисањем корисничких података у четвртом пољу заглавља могуће је користити векторе обележја добијене применом PCA на MFCCs и на основу њих формирати моделе говорника.

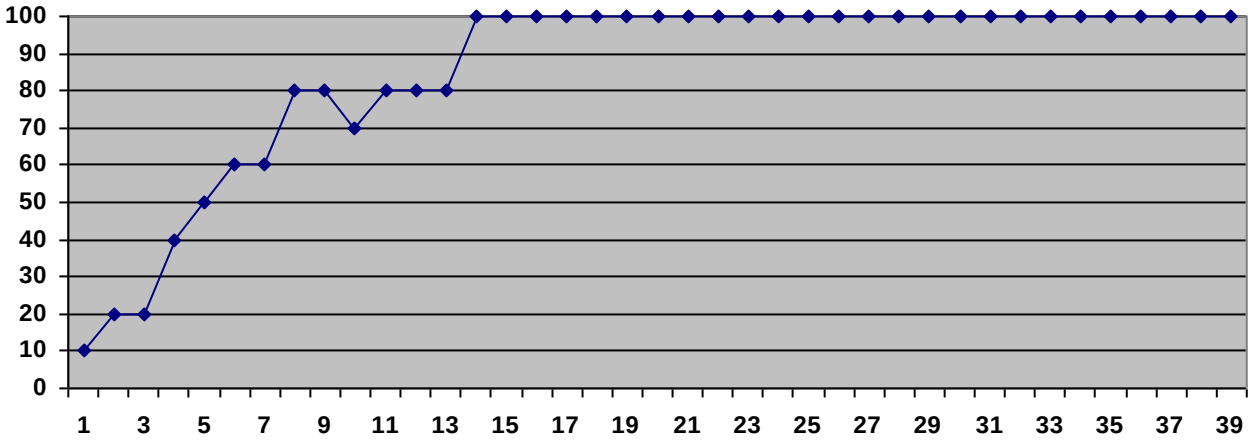
3.2. РЕЗУЛТАТИ ПРИМЕНЕ PCA

Експерименти су били усмерени ка испитивању могућности у погледу тачности аутоматског препознавања говорника које пружа PCA у циљу смањења димензионалности полазних 39-то димензионалних MFCCs вектора обележја. Својствене вредности су поређане по опадајућем редоследу (слика 3.2). Трансформациона матрица за постизање d-димензионалног вектора обележја формирана је употребом d својствених вектора којима одговарају највеће својствене вредности. На овај начин димензионалност трансформисаног вектора обележја је постепено повећавана у корацима за један и праћена тачност препознавања говорника.



Слика 3.2. Својствене вредности у опадајућем редоследу.

Као што је евидентно са приказа на слици 3.3 повећање димензионалности углавном је резултовало неопадајућом тачношћу препознавања. Изузетак је случај 10-то димензионалног вектора када је тачност препознавања опала са 80 на 70%. Обзиром да је циљ ових експеримената био одређивање потребне димензионалности трансформисаних вектора обележја за коју неће доћи до погоршања тачности препознавања од 100% добијене у 39-то димензионалном простору MFCCs вектора обележја (Јокић Иван, 2010), постојање поменутог изолованог случаја опадања тачности препознавања није детаљније испитивано. Експерименти су показали да се изворна тачност од 100% постиже при 14-то димензионалном трансформисаном вектору облежја.



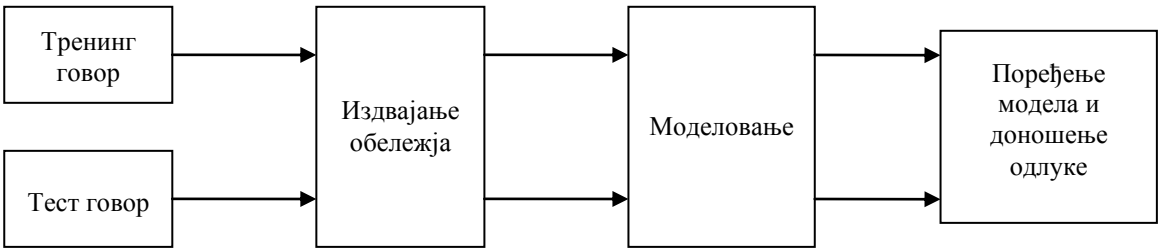
Слика 3.3. Тачност препознавања обзиром на димензионалност трансформисаног вектора обележја.

У случају конкретне говорне базе примена PCA је показала постизање значајног смањења димензионалности изворних вектора обележја при чему се не нарушава тачност препознавања у односу на стандардно коришћени простор вектора обележја. Сматрајући да својствене вредности указују на важности димензија одређених одговарајућим својственим векторима на основу разматрања приказа на слици 3.2 евидентно је да је 16 својствених вредности мање од 1 тј. више од 300 пута мање у односу на највећу својствену вредност што указује на прилично малу значајност ових димензија при репрезентацији обележја говора. На овај начин PCA пондерише димензије чиме избегава гомилање редундансе дуж њих. Ова трансформација рационалније распоређује димензије дуж којих треба посматрати векторе обележја говора. Димензије које одговарају изразито малим својственим вредностима потенцијално се могу занемарити при изграђивању модела говорника, што смањењем димензионалности простора у коме се одлучивање врши смањује и његову сложеност.

Претходно описани експерименти указују на потребу детаљног испитивања свих елемената који се налазе у саставу аутоматског препознавача говорника, што води ка реализацији комплетног програмског решења.

4. РЕАЛИЗАЦИЈА ЈЕДНОГ РЕШЕЊА АУТОМАТСКОГ ПРЕПОЗНАВАЊА ГОВОРНИКА

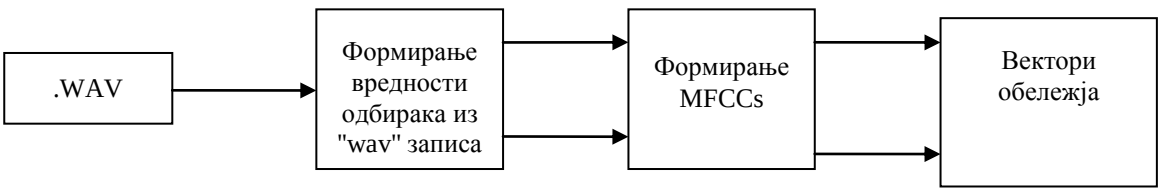
Извршена је имплементација аутоматског препознавача говорника (слика 4.1) у виду програма написаног у C++ програмском језику. Препознавач је реализован на тај начин да на основу расположивог тренинг и тест говора направи одговарајуће моделе и на основу њихове сличности донесе одлуку препознавања о идентитету посматраног говорника. Као и приликом сваког препознавања у раду реализованог аутоматског препознавача говорника постоје два режима рада: обука и тестирање. Приликом обуке врши се моделовање говора посматраних говорника на основу расположивих говорних снимака тренинг говора. У првом кораку врши се издвајање обележја односно формирање вектора обележја за сваки рам посматраних говорних сигнала. Након тога у наредном модулу се врши формирање модела говорника на основу њима припадајућих вектора обележја. Имајући у виду тежњу да се направи аутоматски препознавач говорника који ће бити независан од текстуалне садржине посматраног говора примењено је стохастичко моделовање коришћењем вишедимензионалних Гаусових расподела. У тест фази препознавач за расположиве говорне снимке такође пролази кроз издвајање обележја и затим моделовање вектора обележја тест говора. Након тога следи поступак утврђивања идентитета говорника за посматрани тест говор. Суштина овог поступка огледа се у утврђивању мере сличности између модела тест говора и модела добијених приликом обуке.



Слика 4.1. Блок дијаграм реализованог аутоматског препознавача говорника.

4.1. РЕАЛИЗАЦИЈА ИЗДВАЈАЊА ОБЕЛЕЖЈА

Реализовани аутоматски препознавач говорника је прилагођен раду над звучним записима у основном, каноничком "WAVE" формату. Овај формат представља један од Microsoft-ових "RIFF" (енг. Resource Interchange File Format – RIFF) дигиталних записа намењених за чување дигиталних звучних записа. Заједничка општа особина свих "RIFF" (http://en.wikipedia.org/wiki/Resource_Interchange_File_Format) записа јесте подељеност на означене делове. Сваки од делова садржи одговарајуће информације које описују садржани звучни запис.



Слика 4.2. Кораци при формирању вектора обележја.

Ради израчунавања MFCCs потребно је располагати вредностима одбирака које представаљају њихов напонски еквивалент. Стога је било потребно извршити конверзију вредности података из "WAVE" записа у вредности одбирака припадајућег звучног сигнала.

Имајући у виду резултате препознавања остварене применом PCA, информациону сувишност првих и других извода MFCCs у односу на саме MFCCs, у раду се пошло од коришћења само мел-фреквенцијских кепстралних коефицијената као обележја, без њихових извода. Излазни вектори обележја представљају векторе MFCCs.

4.1.1. Канонички wav формат дигиталног записа

Канонички "WAVE" запис као целина одликује се информацијама везаним за сам формат и количину садржаних података као и постојањем два поддела у којима су дефинисане подробније особине података као и сами подаци. Запис унутар каноничког wav формата је дефинисан по бајтима тако да је распоред записа по редним бројевима бајта следећи (https://ccrma.stanford.edu/courses/422/projects/WaveFormat/):

- 0 – 3, четворобајтна ознака генеричког формата, "RIFF", који је основа многим осталим форматима као што су "AVI" (Windows audiovisual), "ANI" (Animated Windows cursors) и "WAVE" (Windows audio). Ознака "RIFF" је у прва четири бајта уписана у ASCII формату у "big-endian" бајт распореду, 52 49 46 46h, при чему по две хексадецималне цифре (h) дефинишу одговарајући бајт. Преузет назив из енглеске литературе, "big-endian", за бајт распоред дефинише да значајност уписаних бајта опада с лева у десно. За разлику од њега "little-endian" бајт распоред, који је својствен запису података намењеним обради на "Intel" процесорима који раде под "Windows" оперативним системима, подразумева да значајност бајта расте с лева у десно тј. прво је уписан бајт најмањег тежинског фактора а затим бајтови веће важности.
- 4 – 7, четворобајтна вредност која означава величину преосталог дела фајла уписана у "little-endian" бајт поретку.
- 8 – 11, четворобајтна ознака "WAVE" у "big-endian" форми, 57 41 56 45h.
- 12 – 15, четворобајтна ознака "fmt " у "big-endian" форми, 66 6d 74 20h, која означава почетак поддела у коме су дефинисани параметри дигиталног звучног записа као што су: врста кодног записа, учестаност одабирања, број бита по одбирку.
- 16 – 19, величина поддела "fmt " која следи иза овог четворобајтног записа. У случају да су одбирци кодовани импулсном кодном модулацијом (енг. Pulse Code Modulation – PCM), што и јесте код каноничког wav формата, уписана је декадна вредност 16 у "little-endian" бајт поретку. Дакле у оквиру наредних 16 бајта уписане су особине посматраног каноничког "WAVE" записа.
- 20 – 21, двобајтна бројна ознака формата аудио записа у "little-endian" бајт поретку. У случају звучног записа коришћењем линеарне квантизације (PCM) што је карактеристика каноничког wave формата у ова два бајта уписана је декадна вредност 1.
- 22 – 23, двобајтна "little-endian" ознака броја канала у којима је аудио запис снимљен, нпр. моно снимак је означен бројем 1, стерео бројем 2, итд.
- 24 – 27, у оквиру ова четири бајта у "little-endian" формату уписана је вредност учестаности одабирања, нпр. у случају учестаности 22050 Hz редом по бајтима су уписане хексадецималне цифре 22 56 00 00.
- 28 – 31, уписана је "little-endian" вредност бајт брзине која је једнака производу: учестаност одабирања * број канала * број бита по одбирку /8.
- 32 – 33, број бајта којима је кодован један одбирак звучног сигнала укључујући у обзир све канале посматраног записа (моно, стерео или нека друга врста

груписања канала звучног записа) рачунат као производ: број канала * број бита по одбирку/8 и уписан у "little-endian" бајт поретку.

- 34 – 35, у оквиру ова два бајта уписана је вредност броја бита који се користе за кодовање сваког одбирка, нпр. за квантизацију помоћу 16 бита уписана је декадна вредност 16 тј. у "little endian" поретку редом по бајтима 10 00. Након ових података у случају да је кодовање одбирака извршено на неки други начин у односу на PCM у оквиру поддела "fmt " постоје још неки додатни параметри уписани у одговарајуће додатне бајте. За канонички wav формат у следећем бајту започиње означавање поддела који се односи на податке које посматрани "WAVE" запис садржи.
- 36 – 39, у оквиру ова четири бајта садржана је текстуална ознака "data" записана у "big-endian" бајт поретку тако да бајти редом садрже хексадецималне записе: 64 61 74 61.
- 40 – 43, број бајта у којима су садржане кодоване вредности одбирака посматраног звучног сигнала тј. број бајта који следе након бајта 43. Број бајта података једнак је производу: број одбирака * број канала * број бита по одбирку/8 и уписан је у "little-endian" формату.
- 44 – (43+број бајта података), бајти који садрже кодоване одбирке тј. податке. Вредности одбирака су уписане у "little-endian" бајт поретку.

Дакле прва 44 бајта у оквиру каноничког "WAVE" формата представљају заглавље посматраног "WAVE" записа. Све знаковне тј. словне ознаке у оквиру заглавља које представљају имена одређених делова унутар каноничког wav формата као што су: "RIFF", "WAVE", "fmt " и "data" су кодоване у "big-endian" бајт поретку док пошто је "WAVE" формат прилагођен "Windows" оперативном систему и "Intel" процесорима (http://www.sonicspot.com/guide/wavefiles.html) све бројне вредности како у самом заглављу тако и вредности одбирака које представљају податке представљене су у "little-endian" форми. Кодовање одбирака унутар каноничког "WAVE" формата извршено је у некомпримованом PCM запису.

4.1.2. Формирање вредности одбирака на основу расположивог ''wav'' дигиталног записа

Сви коришћени звучни фајлови су снимљени у PCM моно формату при учестаности одабирања од 22050 Hz при чему је сваки одбирак квантован помоћу 16 бита. Приликом анализе звучних фајлова ради одређивања вредности одбирака у сваком посматраном "WAVE" фајлу прескочена су прва 44 бајта која одговарају заглављу након чега су сваком одбирку придружена по два одговарајућа бајта, водећи рачуна да се приликом њиховог читања прво наилази на бајт нижег тежинског фактора и да су 16-то битни одбирци представљени као целобројне означене вредности у комплементу двојке, у опсегу од -32768 до 32767.

Комплемент двојке је погодан начин за представу означених бинарних бројева. На овај начин врши се представа негативних целобројних вредности у бинарном облику. За посматрани N-битни позитивни бинарни број B одговарајући комплемент двојке C_2, тј. представа броја $-B$ се дефинише као допуна тј. комплемент у односу на вредност 2^N:

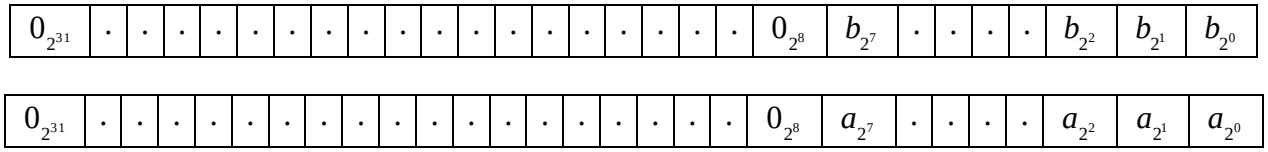
$$C_2(B) = 2^N - B. \tag{4.1}$$

Имајући у виду да се представа бинарних бројева у комплементу јединице врши инвертовањем вредности одговарајућих бита као и последичне чињенице да је збир посматраног бинарног броја B и његовог комплемента јединице C_1 бинарни број који садржи само једнице (http://en.wikipedia.org/wiki/Two's_complement) следи да се комплемент двојке бинарног броја B може израчунати и као збир његовог комплемента јединице и броја 1:

$$C_2(B) = C_1(B) + 1. \tag{4.2}$$

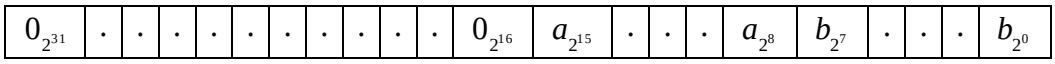
Уколико се претпостави да је *B* одређени позитиван цео број чија се представа у бинарном облику може извршити помоћу *N* бита тада се комплемент двојке посматраног броја *B* користи за представу негативног целог броја *–B*. Посматрајући 16-то битне бинарне бројеве којима се врши приказ одбирака у посматраним каноничким "WAVE" фајловима следи да се позитивни бројеви из опсега од 0 до 32767 пресликавају у опсег од 0000h до 7FFFh док се негативне целобројне вредности из опсега од –32768 до –1 пресликавају у опсег хексадецималних бројева од 8000h до FFFFh. Из чињенице да је 0 водећи бит при бинарној четворобитној представи хексадецималних цифара у опсегу од 0 до 7 а 1 бит највећег тежинског фактора при истој таквој представи хексадецималних цифара у опсегу од 8 до F следи примењено правило за означавање знака вредности у посматраним каноничким "WAVE" 16-то битним записима, позитивне вредности на највећем тежинском месту имају бит 0 док за негативне вредности тај бит има вредност 1.

На основу ове чињенице вршено је одређивање алгебарске вредности, вредности са предзнаком, посматраног кодованог одбирка. Пошто је кодована PCM вредност сваког одбирка представљена у оквиру двобајтне кодне речи у "little-endian" бајт поретку у првом кораку је било потребно извршити померање бајта већег тежинског фактора на почетак двобајтне кодне речи која одговара сваком посматраном одбирку. Вредност сваког бајта је посебно читана из дела који одговара подацима у посматраном "WAVE" фајлу. За посматрани одбирак говорног сигнала одговарајући бајти су чувани у две посебне целобројне променљиве (слика 4.3).



Слика 4.3. Изглед заузећа 32-о битних меморијских локација у којима су привремено чувани бајти који одговарају посматраном одбирку, *b* – нижи бајт, *a* – виши бајт.

Програмски језик C++ пружа могућност извршења одређених операција над битима као што су: померања у лево или десно као и логичке операције. Имајући у виду положај бајта који одговарају посматраном одбирку у оквиру 32-о битних меморијских локација (слика 4.3), коришћењем одговарајућих претходно поменутих операција над битима извршено је одређивање вредности посматраног одбирка. У првом кораку било је потребно извршити операцију над битима бајта већег тежинског фактора, бајт *a* на слици 4.3, како би се омогућило добијање праве кодне речи чија су тежинска места распоређена с лева у десно (слика 4.4). Ради тога потребно је извршити померање бајта већег тежинског фактора за 8 бита у лево.

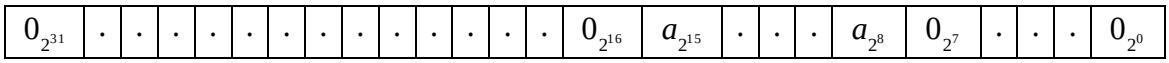


Слика 4.4. Прави облик кодне речи која одговара посматраном одбирку унутар 32-о битне меморијске локације.

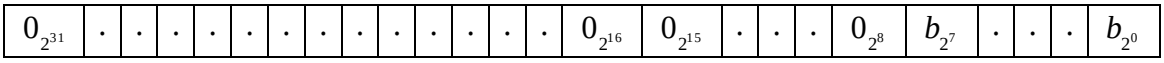
За тако формирану кодну реч са правилно распоређеним тежинским позицијама потребно је одредити бројну вредност коју представља. Поступак одређивања еквивалентне бројне вредности из опсега од –32768 до 32767 разликовао је два случаја у зависности да ли је посматрани одбирак имао позитивну или негативну вредност односно да ли је бит a_{2^7} имао вредност 0 или 1. Обзиром да је сваком бајту података вршен приступ односно читање његове децималне вредности, посматрајући претходно поменути услов у светлу вредности бајта *a* одговарајућег посматраном одбирку он се може изразити кроз следећа два случаја:

- вредност одбирка је позитивна уколико вредност одговарајућег бајта a испуњава услов $a \geq 0 \wedge a \leq 127$,
- у супротном тј. када је $a \geq 128 \wedge a \leq 255$ вредност одбирка је негативна.

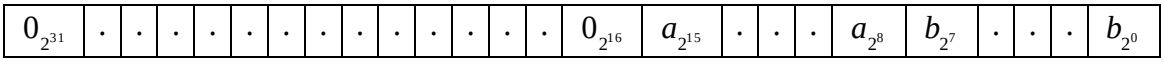
Уколико је вредност одбирка позитивна тада кодна реч представљена у исправном бајт поретку (слика 4.4) директно изражава бројну вредност одбирка. Померањем кодне речи a за 8 бита у лево, првих 8 бита меморијске локације остаје без садржаја односно на тим местима се налазе бинарне нуле (слика 4.5а). За добијање исправне кодне речи (слика 4.5в) потребно је извршити уметање бајта b у првих 8 бита меморијске локације која садржи померени бајт a за 8 бита улево. Обзиром да су вредности ових првих 8 бита бинарне нуле резултантна кодна реч која представља вредност позитивног одбирка је добијена као резултат логичке "или" операције над одговарајућим битима садржаним у меморијској локацији која садржи померени бајт a (слика 4.5а) и меморијској локацији која садржи бајт b (слика 4.5б).



а)



б)



в)

Слика 4.5. Изглед попуњености меморијских локација: а) након померања бајта a за 8 места у лево, б) након читања бајта b, в) након израчунавања бројне вредности позитивног одбирка.

Приликом израчунавања вредности негативног одбирка пошло се од чињенице да су негативни цели бројеви у "WAVE" запису кодовани у комплементу двојке. Исправна кодна реч је добијена на исти начин као и при одређивању вредности позитивног одбирка само што се сада водило рачуна да је прочитана вредност записана у комплементу двојке. Имајући то у виду ради израчунавања вредности негативног одбирка искоришћена је дефиниционa једнакост 4.1 за одређивање комплемента двојке неког позитивног целог броја B. Пошто $C_2(B)$ представља еквиваленту представу у меморији рачунара негативног целог броја $-B$ израчунавање вредности негативног одбирка $-B$ извршено је на основу једнакости 4.1 за 16-то битне кодне речи:

$$-B = C_2(B) - 2^{16} = C_2(B) - 65536. \tag{4.3}$$

На претходно описане начине добијене су целобројне вредности одбирака у опсегу $V_1 \in [-32768, 32767]$. На тако добијене вредности одбирака примењена је нормализација тако да се изврши њихово пресликавање у опсег $V_2 \in [-1.0, 1.0]$. Сходно томе коначне вредности одбирака су израчунате као:

$$V_2 = \begin{cases} \dfrac{V_1}{32767}, \forall V_1 \geq 0. \\ \dfrac{V_1}{32768}, \forall V_1 < 0. \end{cases} \tag{4.4}$$

4.1.3. Израчунавање MFCCs

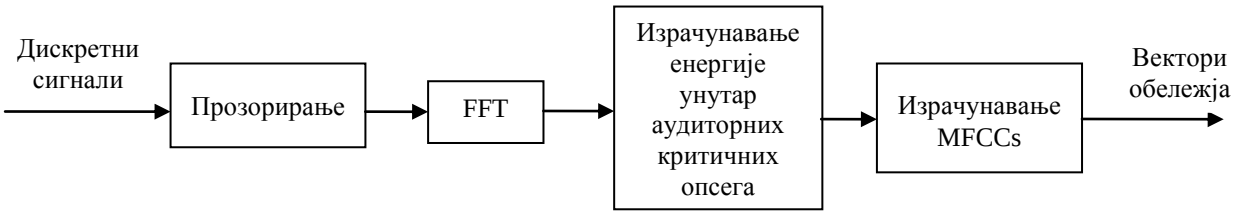
Након одређивања вредности одбирака звучних сигнала у "WAVE" формату који припадају посматраном говорнику формиран је њему одговарајући вектор одбирака који представља еквивалент њему припадајућих дискретних сигнала. Над тако добијеним еквивалентним дискретним сигналом који представља низ дискретних сигнала припадајућих

посматраном говорнику извршен је низ трансформација ради израчунавања вредности у одговарајућим векторима MFCCs (слика 4.6). Одређивање MFCCs извршено је применом једнакости (Wildermoth Richard Brett, 2001.):

$$c_n = \sum_{i=1}^{N} X_i \cdot \cos\left[n \cdot \left(i - \frac{1}{2} \right) \right], \tag{4.5}$$

при чему n=1,2,…,M, представља редни број мел-фреквенцијског кепстралног коефицијента који се израчунава у односу на укупан број M MFCCs који се израчунавају. Суштина једнакости 4.5 се огледа у примени косинусне трансформационе функције на логаритме енергија, X_i, у посматраним аудиторним критичним опсезима, при чему i представља редни број аудиторног критичног опсега. Ради израчунавања претходно поменутих енергија потребно је претходно располагати дискретном Фуријеовом трансформацијом (енг. Discrete Fourier Transform – DFT) посматраног дискретног сигнала. Ово условљава примену одговарајуће прозорске функције на посматрани дискретни сигнал. Ради ефикасног израчунавања DFT примењен је радикс 2 поступак брзе Фуријеове трансформације (енг. Fast Fourier Transform – FFT). За примену радикс 2 поступка је потребно да дужина, l, временског низа над којим се трансформација примењује буде:

$$l = 2^n, \text{ при чему } n \in \mathrm{N}. \tag{4.6}$$



Слика 4.6. Блок приказ формирања вектора обележја за посматраног говорника.

При анализи говорних сигнала, било за потребе аутоматског препознавања говора или говорника, уобичајено се усваја да величина примењене прозорске функције буде у интервалу 20 – 30ms (Kinnunen Tomi et al., 2010). Над сваким сегментом, издвојеним из посматраног дискретног сигнала примењеном прозорском функцијом, извршен је поступак FFT. Стога је било потребно дужину прозорске функције ускладити и са условом исказаним једнакошћу 4.6. Имајући у виду учестаност одабирања посматраних дискретних сигнала, $f_s = 22050\,\mathrm{Hz}$, установљено је да дужини прозорске функције од $l = 2^9 = 512$ тачака одговара временско трајање $\tau = \frac{l}{f_s} \approx 23.2\,\mathrm{ms}$. Обзиром да први мањи број тачака који испуњава услов 4.6 износи $l_1 = 2^8 = 256$ при временском трајању $\tau_1 \approx 11.6\,\mathrm{ms}$ као и да први већи износи $l_2 = 2^{10} = 1024$ при $\tau_2 \approx 46.4\,\mathrm{ms}$ усвојена је величина прозорске функције $l = 512$ тачака. Периоди примене прозорске функције унутар посматраног дискретног сигнала, $\Delta\tau$, се уобичајено додељује вредност 10 ms. У експериментима препознавања ова вредност је варирана у односу на претходно поменуту уобичајену, ради праћења њеног утицаја на тачност аутоматског препознавања говорника.

Поступак примене прозорске функције, $\omega(n)$, на посматрани дискретни сигнал, $s(n)$, подразумева множење посматраног дискретног сигнала и прозорске функције тј. померених верзија посматране прозорске функције при чему је период померања $\Delta\tau$. Најједноставнија примена прозорске функције подразумевала би да се као резултати прозорирања посматраног дискретног сигнала добијају његови неизобличени делови трајања τ, евентуално скалирани у зависности од амплитуде примењене прозорске функције. То значи да је прозорирање извршено применом правоугаоне прозорске функције. Примена прозорске

функције у спектралном домену еквивалентна је конволуцији спектра посматраног дискретног сигнала и спектра примењене прозорске функције. Да би процес прозорирања дискретног сигнала резултовао што мањим изобличењем потребно је да спектар прозорске функције буде присутан у што ужем фреквентном опсегу. Спектар правоугаоне прозорске функције се одликује ниским потискивањем првог бочног лоба те се стога ради обраде звучних сигнала примењују временски заобљеније прозорске функције. Ове прозорске функције се одликују мањом стрмином временских промена те стога и нижом амплитудом виших спектралних компонената што директно за последицу има смањење амплитуде бочних лобова. Постојање бочних лобова у спектралној карактеристици прозорске функције резултује спектралним проширењем прозорираног сигнала у односу на прави спектар посматраног дискретног сигнала, тако да се ова појава често назива и спектрално цурење. Примењена прозорска функција треба да пружа што мање спектрално цурење тј. што веће потискивање бочних лобова. Имајући у виду ове тежње обично се при анализи говорних сигнала користе прозорске функције чији је импулсни одзив заснован на функцији подигнутог косинуса: Ханова и Хамингова прозорска функције.

Ханова прозорска функција се у дискретном временском домену описује једнакошћу:

$$\omega(n) = 0.5 \cdot \left(1 - \cos\frac{2 \cdot \pi \cdot n}{N-1}\right), \text{ за } 0 \le n \le N-1. \tag{4.7}$$

Уопштено посматрајући прозорска функција која је заснована на функцији подигнутог косинуса може се описати једнакошћу:

$$\omega(n) = a + b \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{N-1}\right), \text{ за } 0 \le n \le N-1. \tag{4.8}$$

За вредности параметара a=0.54 и b= - 0.46, добија се Хамингова прозорска функција:

$$\omega(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{N-1}\right), \text{ за } 0 \le n \le N-1. \tag{4.9}$$

Ова прозорска функција се врло често примењује при анализи говорних сигнала пошто она минимизује вредност максималног тј. првог бочног лоба на око једну петину вредности првог бочног лоба код Ханове прозорске функције. Посматрајући ову особину у светлу нивоа сигнала следи да Хамингова прозорска функција обезбеђује за око 10 dB веће слабљење првог бочног лоба у односу на Ханову прозорску функцију. Спектралне карактеристике ове две прозорске функције такође се разликују и на вишим спектралним компонентама, фреквенцијама изнад фреквенције на којој се појављује први бочни максимум тј. лоб у спектру, тако што амлитуда спектра Ханове прозорске функције на тим спектралним компонентама монотоно опада и ниже је вредности у односу на амлитудски спектар Хамингове прозорске функције која у том фреквентном подручју показује много блаже опадање. Стога је у поступку развоја аутоматског препознавача говорника вршено упоредно испитивање утицаја обе ове прозорске функције на постигнуту тачност препознавања.

У анализи утицаја примењених прозорских функција на остварену тачност препознавања испитиван је и утицај Блекманове прозорске функције:

$$\omega(n) = 0.42659 - 0.49656 \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{N-1}\right) + 0.076849 \cdot \cos\left(\frac{4 \cdot \pi \cdot n}{N-1}\right), \text{ за } 0 \le n \le N-1. \tag{4.10}$$

Повећано слабљење првог бочног лоба од око 26 dB у односу на Ханову и око 16 dB у односу на Хамингову прозорску функцију препоручило је коришћење и ове прозорске функције приликом развојног испитивања тачности препознавања пројектованог аутоматског препознавача говорника (Јокић Д. Иван и др., (ТЕЛФОР 2012.)).

4.1.3.1. Израчунавање брзе Фуријеове трансформације

Дискретна Фуријеова трансформација (DFT) $X(k)$, дискретног сигнала $x(n)$ коначног трајања у N тачака:

$$X(k)=\sum_{n=0}^{N-1} x(n)\cdot W_N^{kn} \text{ за } k=0,1,...,N-1 \text{ и } W_N=e^{-j\cdot\frac{2\cdot\pi}{N}}, \tag{4.11}$$

представља начин израчунавања спектра на појединим међусобно еквидистантним учестаностима чији је број одређен бројем тачака у којима се посматра дискретни сигнал $x(n)$. Дискретна природа ове трансформације тј. резултата њене примене који постоји у коначном броју тачака чини је погодном за израчунавање спектралних особина посматраних сигнала применом дигиталних кола за обраду сигнала. Директна примена једнакости 4.11 резултује полиномијалном сложеношћу другог степена, тј. поступак захтева: $N\cdot(N-1)$ комплексних сабирања, N^2 комплексних множења и $2\cdot N^2$ вредности синуса и косинуса, што чини да је сложеност поступка $O(N)\sim N^2$. Уочавајући ротациону природу језгра DFT, W_N, тј. његову периодичност, као и на основу симетрије елемената трансформационе матрице:

$$W=\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N & W_N^2 & \cdots & W_N^{(N-1)} \\ 1 & W_N^2 & W_N^4 & \cdots & W_N^{2\cdot(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{N-1} & W_N^{(N-1)2} & \cdots & W_N^{(N-1)(N-1)} \end{bmatrix}, \tag{4.12}$$

обзиром на идентичне границе параметара n и k, следи да се велики број рачунских операција при директном израчунавању DFT на основу израза 4.11 понавља. У складу с тим могуће је уочити скупове операција које се понављају приликом израчунавања DFT и на основу тога извршити њихово груписање у одговарајуће кораке поступка израчунавања трансформације. На овај начин свака група операција формира одговарајућу итерацију израчунавања DFT при чему су операције садржане у различитим итерацијама међусобно различите. Свака итерација уноси нова израчунавања користећи резултате претходних итерација што значајно убрзава поступак израчунавања DFT смањујући његову сложеност на ниво $O(N)\sim N\cdot ldN$ и чини поступак брзе Фуријеове трансформације (FFT).

Један од поступака FFT, који је коришћен у овом раду, заснива се на раздвајањима парних и непарних чланова посматраног временског низа:

$$X(k)=\sum_{n=0}^{N-1} x(n)\cdot W_N^{n\cdot k}=\sum_{n=2\cdot i} x(n)\cdot W_N^{n\cdot k}+\sum_{n=2\cdot i+1} x(n)\cdot W_N^{n\cdot k}, \quad k=0,1,...,N-1. \tag{4.13}$$

На тај начин се добијају две DFT у $N/2$ тачака:

$$X(k)=\sum_{i=0}^{N/2-1} x(2\cdot i)\cdot W_{N/2}^{k\cdot i}+W_N^k\cdot\sum_{i=0}^{N/2-1} x(2\cdot i+1)\cdot W_{N/2}^{k\cdot i}=X_{10}(k)+W_N^k\cdot X_{11}(k), \quad k=0,1,...,N/2-1, \tag{4.14}$$

које су представљене међурезултатима $X_{10}(k)$ и $X_{11}(k)$ (слика 4.7). Брзина тј. ефикасност рачунског поступка огледа се у начину израчунавања DFT у преосталих, других $N/2$ тачака:

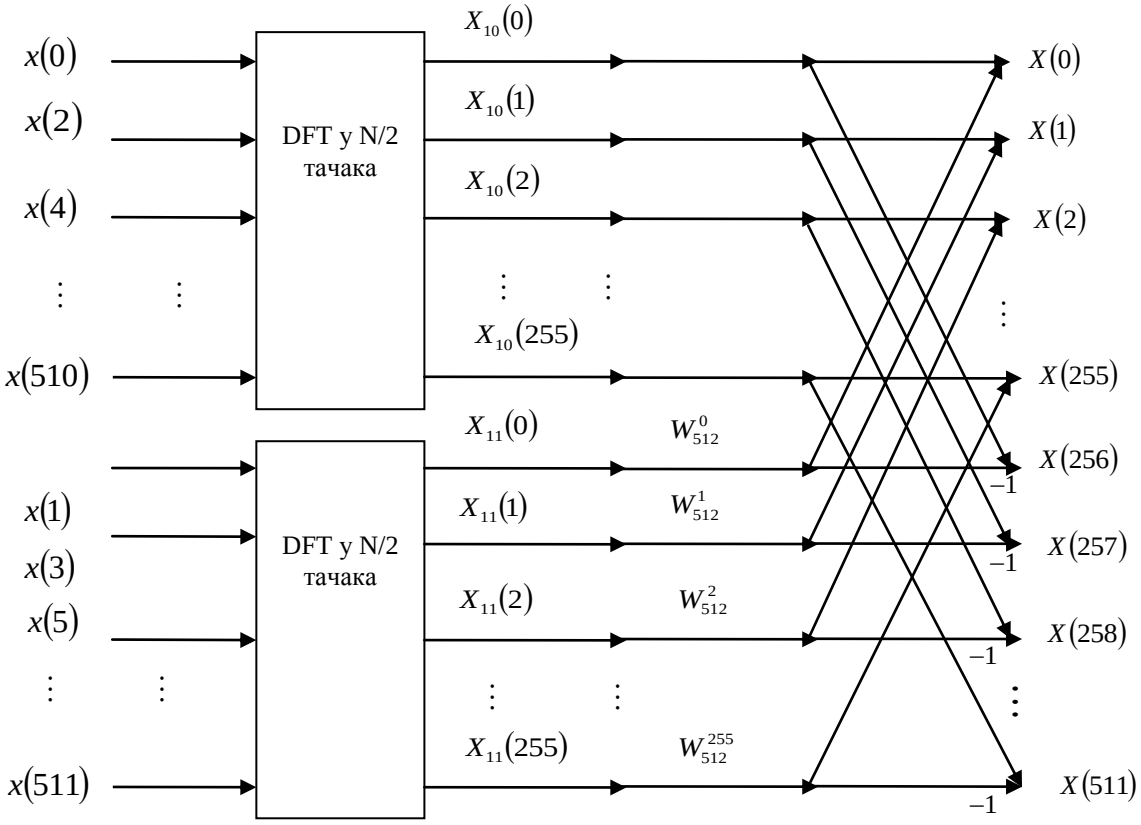
$$X(k+N/2)=X_{10}(k+N/2)+W_N^{k+N/2}\cdot X_{11}(k+N/2)=X_{10}(k)-W_N^k\cdot X_{11}(k), \quad k=0,1,...,N/2-1, \tag{4.15}$$

пошто је $W_N^{k+N/2}=W_N^k\cdot W_N^{N/2}=-W_N^k$. Дакле други сабирак у 4.15 само мења знак у односу на други сабирак у једнакости 4.14. Након овог корака полазна DFT у N тачака растављена је у две DFT. Следећи корак FFT, корак $k=2$, подразумева посматрање сваке од две новонастале DFT у $N/2$ тачака на начин како је разматрана полазна DFT у N тачака, тј. за сваку од две новонастале DFT се врши раздвајање на две DFT, сада у $N/4=N/2^k=N/2^2$ тачака које

одговарају парним односно непарним члановима одговарајућих временских низова. Поступак раздвајања сваке од новонасталих DFT у $N/2^{k-1}$ тачака у одређеном *k*-том кораку FFT се итеративно наставља све док се као улазни временски низови не појаве временски низови у две тачке. Стога је потребно да дужина полазног временског низа представља природни степен броја 2 (услов исказан једакошћу 4.6) и често се овај поступак FFT означава као један од радикс 2 поступака брзе Фуријеове трансформације. Број потребних корака декомпозиције ради извршења оваквог поступка FFT износи:

$$K = ldN - 1, \tag{4.16}$$

тако да за посматране временске низове дужине $N = 512$ тачака, што је дужина прозорираних делова сигнала посматраних у овом раду за које је рачуната FFT, потребно је извршити $K = 8$ корака декомпозиције.



Слика 4.7. Први корак декомпозиције израчунавања DFT у 512 тачака.

Свака новонастала DFT након сваког корака FFT подразумева као улазне временске низове парне односно непарне чланове временских низова који одговарају дискретним Фуријеовим трансформацијама из претходног корака. Следи да сваки корак примењеног поступка FFT врши одређено преуређење елемената полазног временског низа датог у *N* тачака, сходно итерацији у којој се поступак тренутно налази, те се стога често овај поступак брзе Фуријеове трансформације класификује и као радикс 2 поступак FFT са преуређењем у временском домену. Преуређење чланова улазног низа након *K*-тог корака је у складу са битским читањем посматраних изворних вредности индекса у супротном смеру. Наиме, за посматране индексе полазног временског низа у бинарном облику, индекси преуређеног низа у бинарном облику се добијају читањем полазних бинарних индекса с десна у лево, од бита најмањег тежинског фактора ка биту највећег тежинског фактора (табеле: 4.1а, 4.1б). Стога ради имплементације FFT потребно је поступак њеног одређивања посматрати с лева у десно

и прво имплементирати преуређење посматраног временског низа да би се као резултат добиле вредности DFT у исправном редоследу.

Индекс у изворном временском низу	Бинарни запис индекса у изворном временском низу	Бинарни запис индекса у преуређеном временском низу	Индекс у преуређеном временском низу
0	000	000	0
1	001	100	4
2	010	010	2
3	011	110	6
4	100	001	1
5	101	101	5
6	110	011	3
7	111	111	7

Табела 4.1а. Преуређење индекса временског низа дужине 8 бита.

Индекс у изворном временском низу	Бинарни запис индекса у изворном временском низу	Бинарни запис индекса у преуређеном временском низу	Индекс у преуређеном временском низу
0	0000	0000	0
1	0001	1000	8
2	0010	0100	4
3	0011	1100	12
4	0100	0010	2
5	0101	1010	10
6	0110	0110	6
7	0111	1110	14
8	1000	0001	1
9	1001	1001	9
10	1010	0101	5
11	1011	1101	13
12	1100	0011	3
13	1101	1011	11
14	1110	0111	7
15	1111	1111	15

Табела 4.1б. Преуређење индекса временског низа дужине 16 бита.

Да би се претходно поменуто читање бинарне кодне речи у супротном смеру могло реализовати применом рачунарског програма писаног у C++ програмском језику било је потребно искористити могуће операције над битима које програмски језик C++ пружа:

- померања посматраног низа бита који одговара посматраној кодној речи за одређени број бита у лево или десно,
- логичке операције над битима.

Имајући у виду да је читање бинарног записа у супротном смеру еквивалентно симетричном пресликавању тј. ротацији бита у односу на централну позицију у посматраном бинарном запису, преуређење чланова улазног временског низа извршено је применом одговарајућих

претходно описаних операција над битима посматране кодне речи које програмски језик C++ пружа.

A) Реализација преуређења временског низа

Израчунавање дискретне Фуријеове трансформације вршено је над сегментима дужине 512 тачака посматраних дискретних сигнала. Дакле приликом реализације FFT индексе изворних улазних временских низова из интервала $[0,511]$ потребно је преуредити уопштавањем претходно приказаног поступка за временске низове дужине 8 и 16 бита. Ради бинарног тежинског кодовања целих бројева из интервала $[0,511]$ минимално је потребно 9 бита. Приликом обраде одређеног сегмента посматраног дискретног сигнала индекси сваког од 512 одбирака налазе се у одговарајућим меморијским локацијама при чему је њихов значајни део записа смештен у првих девет бита (слика 4.8). Централно место у односу на кога је потребно извшити симетрично пресликавање тј. ротацију бита, ради преуређења индекса временског низа, јесте позиција i_{2^4}.

$0_{2^{31}}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0_{2^9}	i_{2^8}	i_{2^7}	i_{2^6}	i_{2^5}	i_{2^4}	i_{2^3}	i_{2^2}	i_{2^1}	i_{2^0}

Слика 4.8. Бинарни запис индекса изворног временског низа.

Посматрајући индекс временског низа у бинарном облику симетрично пресликавање бита у односу на позицију i_{2^4} се може извршити адекватним избором померања за одговарајући број бита у лево ("<<") односно десно (">>") и применом логичке "И" ("&") операције над битима са одговарајућим бројем (табела 4.2).

Померање бита које је потребно извршити	Начин извођења потребног померања
$i_{2^8} \to i_{2^0}$	$i >> 8$
$i_{2^7} \to i_{2^1}$	$i >> 6 \ \ \& \ \ 2_{10} \ (010)_2$
$i_{2^6} \to i_{2^2}$	$i >> 4 \ \ \& \ \ 4_{10} \ (0100)_2$
$i_{2^5} \to i_{2^3}$	$i >> 2 \ \ \& \ \ 8_{10} \ (01000)_2$
$i_{2^4} \to i_{2^4}$	$i \ \ \& \ \ 16_{10} \ (010000)_2$
$i_{2^3} \to i_{2^5}$	$i << 2 \ \ \& \ \ 32_{10} \ (0100000)_2$
$i_{2^2} \to i_{2^6}$	$i << 4 \ \ \& \ \ 64_{10} \ (01000000)_2$
$i_{2^1} \to i_{2^7}$	$i << 6 \ \ \& \ \ 128_{10} \ (010000000)_2$
$i_{2^0} \to i_{2^8}$	$i << 8 \ \ \& \ \ 256_{10} \ (0100000000)_2$

Табела 4.2. Приказ потребних померања.

У табели је коришћено означавање потребних операција над битима у складу са правилима програмског језика C++:

- "$i >> n$" – померање битског низа *i* за *n* бита у десно,
- "$i << n$" – померање битског низа *i* за *n* бита у лево,
- "*a*&*b*" – логичка "И" операција над битима у бинарној представи бројева *a* и *b*.

Да би се остварило пресликавање одређеног бита на симетричну локацију у односу на позицију i_{2^4} потребно је извршти померање целокупне секвенце бита за утврђени број бита у потребну страну а након тога да би се издвојио само пресликани бит за свако померање

урађена је и одговарајућа "И" логичка операција по битима са бројем који у бинарној представи логичку јединицу има на позицији пресликаног бита. Приликом пресликавања бита највећег тежинског фактора са позиције i_{2^8} довољно је само урадити померање битске секвенце за 8 бита у десно пошто је остатак секвенце након померања изашао ван опсега меморијске локације. Такође бит на централној позицији, i_{2^4}, остаје на својој позицији тако да је урађено само његово издвајање из битске секвенце применом логичке "И" операције са бројем 16. На овај начин извршено је пресликавање сваког од бита из бинарне представе посматраног индекса и након сваког пресликавања у одговарајућој меморијској локацији рачунара налази се бинарни број који на пресликаној позицији има логичку јединицу а на осталим позицијама логичке нуле. Стога је ради добијања вредности индекса након пресликавања свих бита потребно на одговарајући начин сабрати све претходно добијене бинарне бројеве. Имајући у виду да посматраних 9 бинарних бројева имају по једну логичку јединицу при чему сваки на различитој позицији у односу на остале, ово је постигнуто применом логичке "ИЛИ" операције над битима претходно добијених бинарних бројева.

Б) Примењени начин израчунавања FFT

На основу итеративне природе радикс 2 поступка FFT са преуређењем временског низа и разматрања декомпоноване DFT у N=8 тачака уочена итеративна правила су имплементирана у једнакост која се користила ради израчунавања FFT у N=512 тачака:

$$X_{k-1}\left(2^k \cdot i + j\right) \pm W_{2^k}^{j} \cdot X_{k-1}\left(2^k \cdot i + j + 2^{k-1}\right) = \begin{cases} X_k\left(2^k \cdot i + j\right) \\ X_k\left(2^k \cdot i + j + 2^{k-1}\right) \end{cases}, \tag{4.17}$$

при чему:

- $1 \le k \le ldN$ – редни број корака итеративног поступка израчунавања,
- $0 \le i \le \frac{N}{2^k} - 1$ – редни број посматране DFT k-ог корака итерације,
- $0 \le j \le 2^{k-1} - 1$ – помоћна променљива за позиционирање у оквиру посматране DFT у $(j+1)\cdot 2$ тачака.

Улазни временски низ, у ознаци X_0 складно са једнакошћу 4.17, представља преуређен реални временски низ добијен након примене прозорске функције на посматрани говорни сигнал. Имајући у виду комплексну природу резултата и међурезултата при израчунавању DFT уведена су два вектора у којима су привремено чувани реални и имагинарни делови међурезултата, као и четири помоћне променљиве ради израчунавања реалног и имагинарног дела сваког од сабирака из 4.17. Такође у два вектора, *sintab* и *costab* су смештени реални и имагинарни делови који одговарају могућим вредностима ротационог фактора $W_{N=512}^{j}$ у првих $N/2 = 256$ тачака. Обзиром да је ротациони фактор у 4.17 исказан у односу на број тачака елементарне DFT за k-ти корак итеративног поступка, потребне вредности тригонометријских функција су добијене на основу приступа одговарајућим елементима вектора *sintab* и *costab*.

Ради израчунавања енергије унутар посматраних аудиторних критичних опсега у следећем кораку израчунавања MFCCs, која се у општем случају за дискретни сигнал $x(n)$ унутар опсега дискретних учестаности $[k_1, k_2]$ и расположиве коефицијенте DFT X_i израчунава као:

$$E = 2 \cdot \sum_{i=k_1}^{k_2} \left|X_i\right|^2, \tag{4.18}$$

резултати који одговарају последњем кораку итерације FFT су директно прерачунати у вредности квадрата модула резултујћих коефицијената DFT посматраног прозорираног дела сигнала.

4.1.3.2. Израчунавање MFCCs посматрањем енергије унутар аудиторних критичних опсега

Израчунавање MFCCs вршено је применом трансформације на логаритме енергија, $E_{\log(k)}$, у посматраним критичним аудиторним опсезима:

$$c_n = \sum_{k=1}^{L} E_{\log(k)} \cdot \cos\left[n \cdot \left(k - \frac{1}{2} \right) \right], \quad n = 0,1,2,...,d-1. \tag{4.19}$$

Слушалац не примећује разлику у висини два различита тона која се фреквентно посматрано налазе довољно близу тј. који се налазе у оквиру истог аудиторног критичног опсега. Он има осећај повећане гласности тона који је прво чуо. Сходно томе следи да је енергија унутар посматраног чујног критичног опсега једна од његових особина. Посматрајући аудиторне критичне опсеге у фреквентном опсегу може се претпоставити њихова правоугаона односно троугаона преносна карактеристика (Lee Chulhee et al., 2003.). У раду су почетно примењени правоугаони критични опсези при чему је стартно постављено да њихова ширина буде по 300 mel и да је почетак наредног критичног опсега у односу на посматрани критични опсег померен за 150 mel. За почетно примењене правоугаоне критичне опсеге потребно израчунавање енергија унутар посматраних аудиторних критичних опсега извршено је директним коришћењем једнакости 4.18 што убрзава поступак израчунавања MFCCs у односу да су примењени критични опсези троугаоне преносне карактеристике. Потребно прерачунавање вредности граничних учестаности извршено је посредством једнакости 2.7 и применом правила за прелаз у домен дискретних учестаности:

$$k = N \cdot \frac{f}{f_s}, \qquad k = 0,1,...,N-1, \tag{4.20}$$

при чему је учестаност f прерачуната коришћењем једнакости 2.7.

4.2. МОДЕЛОВАЊЕ ГОВОРНИКА И НАЧИН ОДЛУЧИВАЊА

При прављењу аутоматског препознавача говорника пошло се од претпоставке да се на основу разликовања у боји гласа посматраних говорника може извршити потребно разликовање говорника. Сходно томе као обележја су коришћени MFCCs. Боја гласа утиче на облик спектралне обвојнице посматраног говорног сигнала тј. утиче на појаву односно расподелу присутности одређених вектора MFCCs у гласу посматраног говорника. Претпостављајући да је појава вектора MFCCs односно вектора обележја својствених гласу посматраног говорника у складу са одговарајућом Гаусовом вишедимензионалном расподелом (једнакост 2.16) следи да ова усклађеност повезује тј. доводи у последичну везу облик спектралне обвојнице посматраног говорног сигнала и облик Гаусове вишедимензионалне расподеле која описује расподелу одговарајућих вектора обележја. Облик Гаусове вишедимензионалне расподеле огледа се у њеној коваријансној матрици те је стога моделовање говорника извршено одговарајућим коваријансним матрицама (једнакост 3.3) рачунатим како је то приказано једнакошћу 3.2.

Уколико се препознавање врши над скупом од N говорника тада су модели говорника дефинисани одговарајућим коваријансним матрицама из скупа $\{\Sigma_1, \Sigma_2, ..., \Sigma_N\}$. Приликом препознавања односно одређивања идентитета говорника на основу расположивог говорног сигнала или скупа говорних сигнала такође се у првом кораку врши моделовање посматраног говорног узорка одговарајућом коваријансном матрицом Σ. Ради одлучивања ком од моделованих идентитета из скупа $\{1,2,...,N\}$ приписати посматрани говор израчунава се мера разликовања између модела посматраног говора и модела раније моделованих N говорника. Мера разликовања између посматраног тест модела и неког *i*-тог модела из скупа од N говорника дефинисана је једнакошћу:

$$m(i,test)=\frac{1}{d^2}\cdot\sum_{j=1}^{d}\sum_{k=1}^{d}\left|\Sigma_i(j,k)-\Sigma_{test}(j,k)\right|, \tag{4.21}$$

при чему d представља димензионалност коришћених вектора MFCCs. Говору чији се идентитет препознаје приписује се идентитет из скупа $\{1,2,...,N\}$ који минимизује претходно дефинисану меру растојања, односно тест говор (*test*) припада *i*-том говорнику ако важи:

$$m(i,test)<m(j,test) \qquad \forall j\in\{1,2,...,N\}\setminus\{i\}. \tag{4.22}$$

5. РЕЗУЛТАТИ ПРЕПОЗНАВАЊА

Испитивање рада аутоматског препознавача говорника вршено је над две говорне базе. Резултат препознавања је директна последица особина посматраних обележја говора. Посматрајући MFCCs као обележја говора може се рећи да су њихове особине квантитативне природе тј. огледају се у њиховој величини. Израчунавање MFCCs своди се на примену потребних трансформација над кратким деловима, реда 25 ms, посматраних говорних сигнала. Дакле сама природа процене вектора MFCCs је локална. Локалност процене за последицу има временску променљивост сваког елемента унутар посматраног вектора обележја. На локалном нивоу, нивоу анализираног рама говора уобичајеног трајања реда 25 ms, MFCCs се могу сматрати обележјима која описују обвојницу спектра посматраног рама говора. Посматрајући их у односу на цео сигнал ова обвојница спектра је локална.

Полазећи од тога да различити сигнали у временском домену имају различит спектрални садржај следи да је свака новост или новина коју посматрани говорни сигнал носи одсликана на његов спектрални садржај или у зависности од спектралног садржаја посматраног говорног сигнала слушалац сазнаје новости које га интересују. Следи да је ово пресликавање између спектралног садржаја говора и његовог информативног садржаја обострано једнозначно односно један на један пресликавање. Сходно интересу из посматраног говорног сигнала могу се сазнати односно препознати различите особине говора као што су: текстуална садржина говора, личност која говори, како личност говори и стање њених осећања. Све ове особине говора одсликавају се на његов спектрални садржај, обвојницу спектра тј. на његову боју. MFCCs као обележја која описују спектралну обвојницу посматраног рама говорног сигнала погодна су за примену у различитим видовима аутоматског препознавања везаним за говорни сигнал, с тим што се сам препознавач унапред кроз изабране ентитете моделовања, било да су то гласови, говорници, начин говора или осећања, прилагођава предмету препознавања односно информацији коју би хтео да извуче из посматраног говора.

Вредности обележја која се израчунавају на основу анализе краткотрајних временских делова сигнала зависна су од околине анализираних сегмената сигнала односно од њихових контекста. Такође и вредности MFCCs посматраног говорног рама зависе од говорних рамова који су му претходили односно од рамова који му следе, тј. посредно од текстуалне садржине посматраног говора. Стога је ради потпуније анализе резултата аутоматског препознавања говорника потребно узети у обзир и садржине говорних снимака на основу којих су вршени обука модела говорника и тестирање препознавања.

Прва говорна база, у даљем тексту Говорна база 1, садржи снимке говора 121-ог говорника, 61-ог женског и 60 мушких говорника. Садржински снимци су подељени у три групе: Имена, Цифре и Речи. Свака од ове три групе садржи по један снимак једне текстуалне садржине за једног посматраног говорника и то:

- Имена – један снимак изговора личних података одговарајућег говорника који га јединствено представљају у оквиру Говорне базе 1. Текстуална садржина снимка зависи од говорника, при чему се она у општем облику може представити као: <идентификациони број говорника у оквиру Говорне базе 1>, <име и презиме говорника> ;
- Цифре – два снимка изговора низа цифара:
    1. један, два, три четири, пет;
    2. шест, седам, осам, девет, нула;
- Речи – једанаест снимака следећих низова речи:
    1. командант, пук, одељење, батаљон, чета, вод;
    2. батерија, операција, штаб, десант, покрет, обезбеђење;
    3. уређај, напад, веза, одбрана;
    4. дан, час, минут, војник, ракета, авион;
    5. непријатељ, возило, брод, извештај, стоп, одмах;

6. сутра, окончано, замени, упути, реон;
7. напред, је, са, да, не, од;
8. назад, на, из, до, изврши, готовост;
9. правац, пријем, циљ, уништен, задатак, понови;
10. група, разумео, тачка, опасност, јединице, аеродром;
11. метео, сигнал, објекат.

Снимци говора унутар друге говорне базе, у даљем тексту Говорна база 2, садрже изговоре одређене четири цифре. За разлику од прве говорне базе текстуална садржина снимака у оквиру друге говорне базе своди се само на цифре, 0 – 9, при чему су изговори исте текстуалне садржине снимани у оквиру више сесија (Јокић Д. И., Добријевић Н. Т., et al., 2009). Говорна база садржи снимке говора 44 говорника. У већини случајева једна сесија снимака за посматраног говорника садржи 12 снимака. За већину говорника у оквиру ове говорне базе постоје снимци у десет сесија. Известан број говорника је снимљен у оквиру мањег броја сесија тако да је обука модела и тестирање препознавања вршено над 37 говорника.

Говорне базе 1 и 2 су снимане у различитим условима. Говорна база 1 снимана је у студијским условима док је за разлику од ње Говорна база 2 снимана у канцеларијским условима коришћењем рачунарског микрофона, звучне картице и рачунара (Јокић Д. И., Добријевић Н. Т., et al., 2009).

5.1. ТЕСТОВИ ПРОМЕНЕ ТЕСТ ФАЈЛА

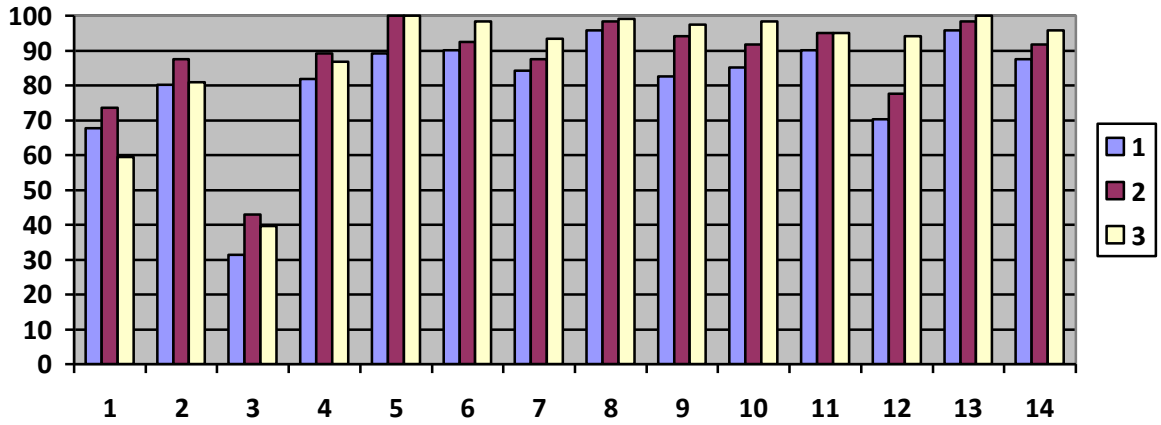
Тестирање препознавања у оквиру Говорне базе 1 вршено је над сваким говорним снимком у оквиру ње. Редослед тестова је одређен списком фајлова у оквиру посматраних говорних база који направи аутоматски препознавач говорника приликом његовог покретања. Аутоматски генерисан списак говорних снимака условљен је знаковним садржајем њихових назива. Пошто Говорна база 1 садржи по 14 снимака за сваког говорника, тестови препознавања су разврстани у 14 група, како то приказују графици који следе и који се односе на тачност препознавања говорника остварену над Говорном базом 1. Тестирање тачности препознавања аутоматског препознавача говорника на Говорном базом 1 вршено је кроз 14 тестова редоследом приказаним у табели 5.1. Тестирање аутоматског препознавача говорника над Говорном базом 2 извршено је над снимцима из прве сесије у оквиру 12 тестова.

Редни број теста	Текстуална садржина
1.	један, два, три, четири, пет
2.	шест, седам, осам, девет, нула
3.	<идентификациони број, име и презиме>
4.	командант, пук, одељење, батаљон, чета, вод
5.	батерија, операција, штаб, десант, покрет, обезбеђење
6.	метео, сигнал, објекат
7.	уређај, напад, веза, одбрана
8.	дан, час, минут, војник, ракета, авион
9.	непријатељ, возило, брод, извештај, стоп, одмах
10.	сутра, окончано, замени, упути, реон
11.	напред, је, са, да, не, од
12.	назад, на, из, до, изврши, готовост
13.	правац, пријем, циљ, уништен, задатак, понови
14.	група, разумео, тачка, опасност, јединице, аеродром

Табела 5.1. Редослед тестова при тестирању тачности на Говорном базом 1.

Имајући у виду да је учестаност одабирања говорних сигнала у говорној бази 22050Hz следи да у складу са теоремом одабирања фреквентни опсег од интереса у посматраним говорним сигналима износи [0 – 11025Hz], што пружа могућност распореда највише 20 критичних опсега ширине 300 мела који су међусобно померани за 150 мела као што је то приказано на слици 5.1. Ради постизања највеће тачности израчунавања вектора MFCCs потребно је у обзир узети све расположиве податке који су доступни у оквиру посматраног опсега учестаности. У овом случају то значи узимање свих 20 критичних опсега у обзир при израчунавању MFCCs. Прецизност односно резолуција представе посматраног скупа података расте како расте и број параметара или обележја која описују посматрани скуп података. Зато се у тестирањима пошло од вектора обележја који заједно са нултим садржи 20 MFCCs. Према томе у првом случају тестирања (слике 5.2 (1) и 5.3 (1)) коришћени су вектори обележја који садрже и нулти MFCC, тако да је тада вектор обележја садржао нулти и првих 19 MFCCs рачунатих на основу логаритама енергије у 20 правоугаоних критичних опсега у складу са једнакошћу 4.19. Модели говорника као и модели тест говорних снимака представљени су одговарајућим коваријансним матрицама димензија 20×20.

Слика 5.1. Распоред примењених правоугаоних критичних опсега.



Слика 5.2. Тачност препознавања над Говорном базом 1 у зависности од посматраног тест фајла, нулти+19 MFCCs, 20 правоугаоних критичних опсега.
1 – пуна коваријансна матрица као модел,
2 – пуна коваријансна матрица изузев елемента $\Sigma_{0,0}$,
3 – без нултог MFCC, 19MFCC и 20 правоугаоних критичних опсега.

У случају Говорне базе 1 евидентна је изразито најмања остварена тачност препознавања при тестирању препознавања над снимцима говора из групе Имена. Снимци из ове групе одликују се најкраћим трајањем у односу на снимке у оквиру преостале две групе: Цифре и Речи, тј. они имају најмање гласовног звучног садржаја у односу на остале снимке. Наиме, ради успешног препознавања говорника потребно је располагати довољном

количином информација о његовој боји гласа односно о спектралној садржини његовог гласа на основу које је могуће извршити његово разликовање у односу на остале познате говорнике. Дакле, посматрајући глас у спектралном домену учестаности које су његов саставни део у складу са Фуријеовом трансформацијом, ради међусобног разликовања гласова потребно је познавати дискретне скупове учестаности посматраних говорних сигнала који су различити за различите говорнике. Према томе елементи говорног сигнала који се одликују дискретним спектралним компонентама тј. имају одређену временску периодичност која се доживљава као звучност, доприносе разликовању различитих гласова. Богатство говора елементима који му дају звучност тј. најчешће вокалима чини говор једног говорника препознатљивијим у односу на друге говорнике. Са становишта пауза у говору изговори из групе Имена се углавном могу посматрати као да садрже три речи а понекад и две. Наиме изговор сложених бројева, нпр. као изговор броја 123 – сто двадесет три, често са становишта слушаоца изгледа као једна реч. Наиме, речи које се налазе у саставу сложеног броја се изговарају једна за другом без изражене паузе. Ово често резултује бржим изговором од уобичајеног или стандардног када би се свака реч у саставу сложеног броја изговарала посебно, нпр. за претходно поменути изговор броја 123 то би изгледало као: сто, двадесет, три, као што је то случај у групи снимака цифре. Такође и изговори презимена и имена често звуче као једна реч што све доводи до тога да се снимци из групе Имена често одликују звучним сегментима скраћеног трајања у односу на преостали део базе који је коришћен ради обуке модела говорника. Из тог разлога препознавач је доста чешће доносио погрешну одлуку при препознавању говорника над снимцима из групе Имена у односу на препознавања вршена када су се као тест снимци говора користили снимци из група Цифре или Речи.

Снимци из групе Цифре садрже изговоре по пет одређених речи што повећава гласовну разноликост ових снимака у односу на снимке из групе Имена. Ово је резултовало повећањем тачности препознавња. Изговори низа цифара 1 – 5, односно 6 – 0, такође се одликују различитом гласовном разноликошћу што је резултовало већом тачношћу препознавања за низ цифара 6 – 0. У изговору низа цифара: шест, седам, осам, девет, нула постоји 9 вокала док се изговор низа цифара: један, два, три, четири, пет, одликује постојањем 8 вокала. Према томе изговор низа цифара 6 – 0 поседује већу количину звучности на основу чега је могуће тачније извршити препознавање говорника.

Препознавање говорника над говорним снимцима из групе Речи резултовало је повећањем тачности у односу на тестове над снимцима из група Имена и Цифре. Тачност препознавања је сада у 2 теста била реда 90% док је у 2 теста тачност достизала ниво од 95%. Евидентно је да је знаковна тј. текстуална дужина говора у тестовима 6 и 11, снимци изговора скупа речи: метео, сигнал, објекат, и скупа речи: напред, је, са, да, не, од, краћа у односу на изговор коришћен у тесту 1. Ипак тачност у тестовима 6 и 11 је значајно већа у односу на тачност остварену у тесту 1. Снимци говора из групе Речи, над којима је вршено тестирање у оквиру тестова 4, 5, 6, ... , 14, одликују се неповезаношћу изречених речи. Свака реч у оквиру посматраног снимка говора који се налази у оквиру групе Речи изговорена је засебно тј. независно у односу на остале које се налазе у посматраном снимку говора. То значи да су речи изговорене чистије односно гласови у речи су јасније изговорени у односу на изговоре из групе Имена. Осим тога снимци говора над којима је вршено тестирање у оквиру тестова 6 и 11 обилују вокалима тј. имају велику количину звучности у себи која у комбинацији са јасношћу изговора резултује значајним повећањем тачности аутоматског препознавања говорника у односу на тачност препознавања остварену у тесту 1.

За конфигурацију аутоматског препознавача говорника – 20 правоугаоних критичних опсега на основу којих су рачунати вектори обележја који садрже нулти и првих 19 MFCCs, над Говорном базом 1 постигнута је највећа тачност препознавања реда 95% у оквиру тестова 8 и 13. Снимци изговора речи (табела 5.1) који се користе при овим тестовима садрже значајан број вокала у односу на изговоре из групе Речи коришћене у осталим тестовима (табела 5.2). Ипак приметно је да изговори у оквиру тестова 5 и 14 садрже највише

вокала али тачност у тим тестовима није достигла највеће вредности. Наравно, на основу тога следи да постојање већег броја вокала не значи да ће се и тачност препознавања говорника повећати. Утицај звучности на разликовање међу говорницима у општем случају је позитивна али стварна количина звучности поред тога што зависи од броја вокала такође зависи и од њихових позиција у речима односно од њиховог окружења сугласницима који су такође присутни у речима и који у суштини, више или мање у зависности од природе своје творбе, пригушују звучност присутних вокала.

Редни број теста	1	2	4	5	6	7	8	9	10	11	12	13	14
Број вокала	8	9	14	19	8	10	12	14	14	7	10	13	19

Табела 5.2. Број места на којима се појављују вокали у тестовима из група Цифре и Речи.

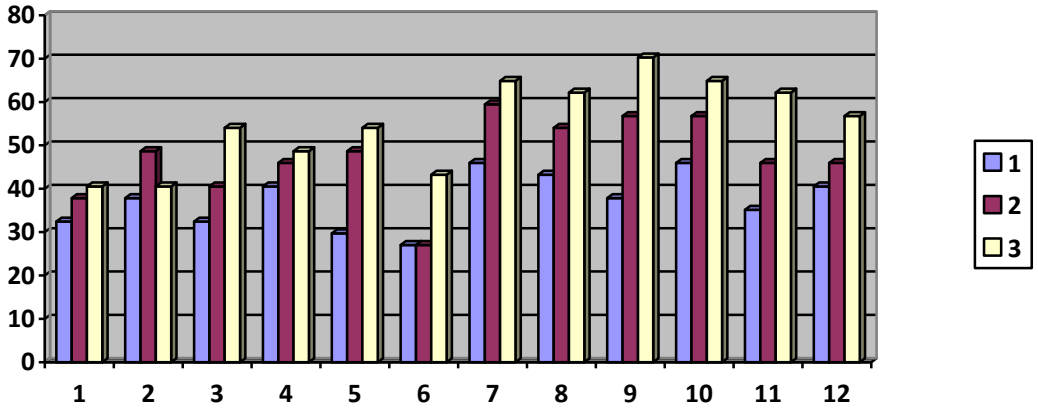
За разлику од Говорне базе 1, Говорна база 2 се са становишта текстуалне садржине изговорене од стране једног тачно одређеног говорника одликује прилично мањом фонетском разноврсношћу. Сваки њен снимак садржи низ од четири цифре. Гласовна разноврсност једног говорника обухваћена овом говорном базом огледа се у постојању снимака изговора једне исте текстуалне садржине у различитим сесијама снимања односно снимања су вршена у временским размацима. За препознавање говорника битни су звучни делови у посматраном говору односно делови говорног сигнала који садрже вокале: а, е, и, о, у. Подаци који се посматрају при аутоматском препознавању говорника су *d*-димензионални вектори MFCCs. Процена расподеле која представља модел посматраног говорног снимка или посматраног скупа говорних снимака извршена је израчунавањем просечне вредности збира производа вредности MFCCs у оквиру одговарајућих димензија расположивих вектора обележја као што је то приказано једнакошћу 3.2. На овај начин сваки скуп снимака говора који су се користили за обуку или сваки снимак говора који је коришћен ради тестирања представљен је скупом вредности варијанси у оквиру посматраних мел-фреквенцијских кепстралних димензија као и коваријанси између посматраних димензија. Дакле бројне вредности варијанси и коваријанси за сваког говорника као и за сваки тест снимак говора представљају очекиване односно средње вредности производа расположивих скупова бројних вредности у одговарајућим димензијама. Са становишта присутности шума у говорним снимцима квалитет Говорне базе 2 је лошији у односу на квалитет Говорне базе1. Постојање шума у говорним снимцима утиче на ове очекиване вредности варијанси односно коваријанси померајући њихове вредности у односу на стварне вредности када шум не би био присутан. У зависности од квалитета односно природе шума присутног у снимцима коришћеним при обуци и тестирању може доћи до значајнијег повећања разлике између модела посматраног говорника добијеног при обуци и тест модела његовог говора, што ће резултовати нетачним препознавањем.

Процена претпостављеног типа вишедимензионалне расподеле боља је уколико је већи скуп вектора над којим се та процена врши. У случају Говорне базе 1 постоји шири скуп расположивих вредности посматраних вектора MFCCs, вокали битни за препознавање говорника налазе се у ширем спектру могућих позиција у коришћеним речима. У Говорној бази 2 то су само цифре и промена вредности вектора обележја условљена је малим бројем различитости појединих вокала у речима у говорној бази. Ово нарочито долази до изражаја при моделовању снимака говора који су намењени тестирању препознавања. Изречени гласови у речи међусобно утичу једни на друге тј. утичу на квалитет вокала у посматраној речи. Ова зависност се може описати узимањем у обзир најближег окружења вокала у посматраној речи тј. декларисањем појаве одређеног вокала на одређеном месту у речи кроз

трознак у коме знак посматраног вокала заузима централну позицију. На овај начин појава пет могућих вокала у Говорној бази 2 може се описати кроз следеће трознаке:

- а: ла_, дан, ва_, дам;
- е: јед, чет, пет, шес, сед;
- и: ри_, тир, ри_;
- о: _ос;
- у: нул.

Види се да су вокали "о" и "у" присутни у само по једној речи, осам и нула, и може се десити да у тест снимку ове речи буду присутне само једанпут. Постојање шума у овој говорној бази утиче на померање вектора обележја који одговарају вокалима и самим тим може довести до погрешног препознавања. Као пример оваквог случаја може се издвојити препознавање говорника ID05 при тесту 1 и тест исказу "нула седам осам један". Дакле у овом случају у тест снимку постоје само по један узорак за гласове "у" и "о" и у сва три случаја препознавања приказаног на слици 5.3 говорник ID05 је погрешно препознат.



Слика 5.3. Тачност препознавања над Говорном базом 2 у зависности од посматраног тест фајла, нулти+19 MFCCs, 20 правоугаоних критичних опсега.
1 – пуна коваријансна матрица као модел,
2 – пуна коваријансна матрица изузев елемента $\Sigma_{0,0}$,
3 – без нултог MFCC, 19MFCC и 20 правоугаоних критичних опсега.

За разлику од тачности остварене над Говорном базом 1 тачност препознавања над Говорном базом 2 у првом скупу тестова (слика 5.2 – 1) показивала је доста мању вредност, у просеку више него двоструко мању тачност (табела 5.3). Пошто снимци у овој говорној бази намењени обуци и тесту текстуално гледано имају исте садржине следи да и мали случајни спољашњи утицај може довести до разликовања између модела добијеног при обуци и тест модела. Стога се може закључити да је ова говорна база у односу на Говорну базу 1 осетљивија на присуство спољашњег шума. Узимајући у обзир већи ниво шума присутан у Говорној бази 2, на основу претходног разматрања следи да је мања тачност аутоматског препознавања говорника над Говорном базом 2 очекивана.

Тип теста	Просечна тачност препознавања [%]	
	Говорна база 1	Говорна база 2
1	80.87	37.387
2	87.190	46.88
3	88.488	55.18

Табела 5.3. Просечна тачност препознавања при рачунању MFCCs коришћењем 20 правоугаоних критичних опсега у зависност од типа теста (слике 5.2 и 5.3).

Релативни однос просечних тачности препознавања остварених у првом скупу тестова над говорним базама 1 и 2 приближно се задржао и за односе максималних тачности препознавања за тестове извршене над ове две говорне базе. Наиме, тачност препознавања у првом скупу тестова над делом Говорне базе 2 у оквиру снимака из прве сесије достизала је највећу вредност реда 45%. Овакво приметно разликовање у оствареним тачностима препознавања над ове две говорне базе указивало је на потребу детаљнијег разматрања разликовања модела добијених при обуци и модела тест говорних снимака.

Упоређујући елементе коваријансних матрица за истог говорника добијене при обуци и тестирању уочене су приметне разлике у вредностима првог, $\Sigma_{0,0}$, елемента у одговарајућим коваријансним матрицама у односу на разликовања осталих елемената. Овај елемент представља варијансу нултог MFCC односно средњу вредност збира квадрата нултих MFCCs посматраних рамова говора који су моделовани посматраном коваријансном матрицом. Дакле он представља упросечену енергијску слику нултог MFCC и као такав веома је зависан како од гласовне садржине посматраног говора тако и од присуства евентуалног шума у говорним снимцима. Последично, вредност енергије нултог MFCC може се доста разликовати у коваријансној матрици израчунатој на основу скупа говорних снимака намењених обуци у односу на вредност на истом месту у коваријансној матрици прорачунатој на основу тест снимка за посматраног говорника. Модел говорника као модел његовог гласа требао би да у себи садржи информације о његовом гласу које нису значајно зависне од текстуалне садржине, пошто у реалним ситуацијама аутоматског препознавања говорника одлука аутоматског препознавача говорника не би требала бити, бар не значајно, условљена садржином текста говорниковог изговора.

Стога је у следећем кораку тестирања извршено испитивање тачности препознавања када се за исту димензионалност вектора обележја, нулти и првих 19 MFCCs, први елемент у одговарајућим коваријансним матрицама постави на нулту вредност односно изврши занемаривање његовог утицаја. Ова преправка у коваријансним матрицама добијеним при обуци односно тестирању резултовала је повећањем тачности препознавања над обе говорне базе, што се уочава у постигнутим тачностима у појединачним тестовима (слике 5.2 – 2 и 5.3 – 2) као и у оствареној просечној тачности у скупу тестова 2 (табела 5.3). Тест 5 над говорном базом 1 резултовао је тачношћу 100%, такође тестови 8 и 13 резултовали су тачношћу преко 95%. Тест 5 на известан начин потврђује и значајност броја вокала у говорниковом исказу, пошто тест снимак из теста 5 уз тест снимак из теста 14 садржи највећи број вокала (табела 5.2). Пошто је у Говорној бази 2 присутан израженији шум, позитивна последица изузимања првог члана у коваријансној матрици више се осећа у тестовима над Говорном базом 2.

Утицај нултог MFCC одражава се и у осталим елементима коваријансних матрица који представљају коваријансу између нултог и осталих MFCCs. Ово се такође да приметити у израженијој мери разликовања елемената коваријансних матрица добијених при обуци и тесту за истог говорника на овим местима ($\Sigma_{0,1},\Sigma_{0,2},...,\Sigma_{0,d-1}$ односно $\Sigma_{1,0},\Sigma_{2,0},...,\Sigma_{d-1,0}$) у односу на одговарајуће елементе на осталим местима у којима нема утицаја нултог MFCC-а. Из изложеног следи да нулти MFCC значајно доприноси одступању између модела говора посматраног говорника добијеног при обуци и при тесту тако да је извршено тестирање препознавања без нултог MFCC да би се у одређеној мери избегли директни утицаји текстуалне садржине посматраног говора као и присутног шума при препознавању говорника.

Посматрајући коваријансне матрице добијене при обуци и тесту за истог говорника када је из вектора обележја изузет нулти MFCC и задржано преосталих 19 MFCCs уочено је да су вредности елемената на истим местима међусобно уједначеније. Избегнут је утицај нултог MFCC-а који је резултовао највећим вредностима на првим местима у коваријансним матрицама модела и самим тим узроковао веће одступање између модела добијених при обуци и тест модела. Тачност препознавања се повећала над обе говорне базе, што је сада било доста израженије над Говорном базом 2 (слике: 5.2 – 3, 5.3 – 3, табела 5.3).

Наиме, у прва четири теста над Говорном базом 1 тачност се у извесној мери смањила у односу на случај када је у моделима занемарен само први дијагонални елемент који је директна последица нултог MFCC. Смањење је најизраженије у тесту препознавања говорника над скупом цифара 1 – 5 и тада је тачност нешто мања и у односу на случај када су као модели коришћене пуне коваријансне матрице. Тачност препознавања у тестовима 2, 3 и 4 била је већа у односу на тачност остварену када су као модели коришћене пуне коваријансне матрице. Остали тестови над Говорном базом 1 резултовали су тачношћу која је била на нивоу или је превазилазила тачност постигнуту када су као модели коришћене пуне коваријансне матрице без првог дијагоналног елемента.

Тестови над Говорном базом 2 показују само један случај, тест 2 (слика 5.3 – 3), када је изузимање нултог MFCC погоршало тачност препознавања у односу на случај када је из модела изузет само елемент који је директна његова последица. Сви остали тестови резултовали су побољшањем тачности у односу на претходни скуп тестова. Тест 9 је показивао вршну вредност тачности препознавања реда 70% што је представљало релативно побољшање за више од 30% у односу на исти тест када је у векторима обележја коришћен нулти MFCC и као модели коришћене пуне коваријансне матрице.

При анализи вредности елемената коваријансних матрица добијених на основу вектора обележја који су садржали и нулти MFCC уочено је да поред значајног разликовања првог дијагоналног елемента у коваријансним матрицама обуке и теста за исте говорнике постоји и истакнутије разликовање последњег дијагоналног елемента, елемента $\Sigma_{19,19}$, који је директна последица 19-тог MFCC-а. Међусобна разликовања ових елемената коваријансних матрица за исте говорнике су приближно неколико пута мања у односу на разликовања првог дијагоналног елемента. Имајући у виду тежњу за што прецизнијом представом говорника у расположивом скупу обележја као и чињеницу да MFCC-и реда изнад нултог одсликавају финије логаритамске енергетско-спектралне детаље односно логаритамске енергетске детаље на вишим учестаностима у посматраном раму говора а самим тим и целокупном говорном сигналу, за разлику од нултог MFCC у даљим тестовима, усмереним ка побољшању ефикасности коришћених обележја говора ради тачнијег аутоматског препознавања говорника, задржан је 19-ти MFCC, тако да је вектор обележја чинило првих 19 MFCC.

5.2. ПОБОЉШАЊЕ ТАЧНОСТИ ПРЕПОЗНАВАЊА ПРОМЕНОМ ОБЛИКА КРИТИЧНОГ ОПСЕГА

У поступку аутоматског препознавања облика део за издвајање обележја директно утиче на остварену тачност препознавања. Његови утицаји се примећују у сваком наредном сегменту који чини целину аутоматског препознавача. У случају аутоматског препознавања говорника овде су као обележја говора говорника посматрани MFCCs. Из претходног разматрања евидентно је да уклањање делова модела из посматраних модела говорника који нису директна последица гласа посматраног говорника, него зависе и од текста који је изговорен као и од присутног шума, води ка побољшању тачности аутоматског препознавања говорника.

Коришћена једнакост 4.19 за израчунавање MFCCs указује да њихове вредности директно зависе од логаритама енергија у посматраним чујним критичним опсезима. За унапред дефинисан распоред чујних критичних опсега који се огледа у 20 опсега ширине 300 мела при чему је сваки наредни у односу на претходни опсег на фреквентној оси померен унапред за 150 мела, садржани логаритам енергије у оквиру тако распоређених опсега директна је последица облика постављених опсега.

Својство чујног критичног опсега које се огледа у немогућности детектовања два тона различитих фреквенција а који се налазе унутар истог критичног опсега са становишта детекције звука одликује се селективношћу у одређеном опсегу учестаности. Ова појава везана за осећај висине тонова блиских учестаности такође је повезана са појавом маскирања два блиска тона. Особеност појаве маскирања огледа се у томе да слушалац од два тона на

блиским учестаностима чује тон чија је амплитуда већа при чему је резултантна гласност коју слушалац чује последица међусобног слагања тј. интерференције примењених тонова. Бројчано посматрано као еквивалент гласности може се посматрати логаритам енергије. Појам блискости тонова у смислу могућности разликовања њихових учестаности тј. висина узет је у обзир кроз концепт аудиторних критичних опсега. Наиме, слушалац висину два или више тонова који се налазе у оквиру истог чујног критичног опсега чује као тон на висини тона највеће гласности при чему је укупна гласност коју слушалац чује резултат слагања присутних тонова. Дакле, за већ постојећи говорни сигнал својственост одређеног чујног критичног опсега огледа се у гласности коју би могући слушалац чуо на висини спектралне компоненте која у посматраном чујном критичном опсегу има највећу амплитуду. Енергетска својственост чујног критичног опсега уведена је у концепт израчунавања MFCCs кроз логаритме енергија у оквиру унапред распоређених чујних критичних опсега. Израчунавањем MFCCs као вектора обележја говора посматраног говорника на овај начин, у векторима обележја се покушава сачувати онаква информација о говорнику како би је доживео и могући слушалац који је такође на основу тога у стању да препозна говорника или да га класификује као непознатог.

Облик чујног критичног опсега дефинише начин слагања односно интерферирања енергија појединих компонената унутар њега. Правоугаони облик квадрата амплитудске карактеристике чујног критичног опсега подразумева подједнаку важност свих спектралних компонената унутар њега. На овај начин није узета у обзир способност слушаоца да се усредсреди на одређени тон чију висину на основу енергетске доминантности чује. На тај начин остале спектралне компоненте унутар посматраног чујног критичног опсега дају мањи енергетски допринос у односу на доминантни тон, укупном осећају гласности слушаоца. Опсег чујног критичног опсега везан је за тон доминантне енергије и његову најближу околину што значи да посматрајући га у фреквентном домену учестаности изражених у Hz опсег чујног критичног опсега је симетричан у односу на централну позицију на којој се налази тон доминантне енергије. Уколико се фреквентна скала посматра у мелима, симетричност у мелима у односу на централну позицију чујног критичног опсега би се претворила у антисиметричност на фреквентној скали израженој у Hz. Мањи допринос недоминантних спектралних компонената у посматраном чујном критичном опсегу у односу на доминантни тон који заузима централну позицију у Hz опсегу учестаности указује да су амплитудске особине чујног критичног опсега опадајуће у односу на централну позицију. На овај начин укупна гласност унутар чујног критичног опсега је смањена у односу на случај када је претпостављено да су чујни критични опсези правоугаоног облика. Тиме се у одређеном степену смањује и ниво шума присутан у посматраном чујном критичном опсегу. Израженије присуство шума у снимцима говора из друге говорне базе као и доста мања тачност аутоматског препознавања говорника остварена над говорним снимцима из те говорне базе дали су подстрек увођењу претпоставке опадајућег чујног критичног опсега у односу на његово средиште.

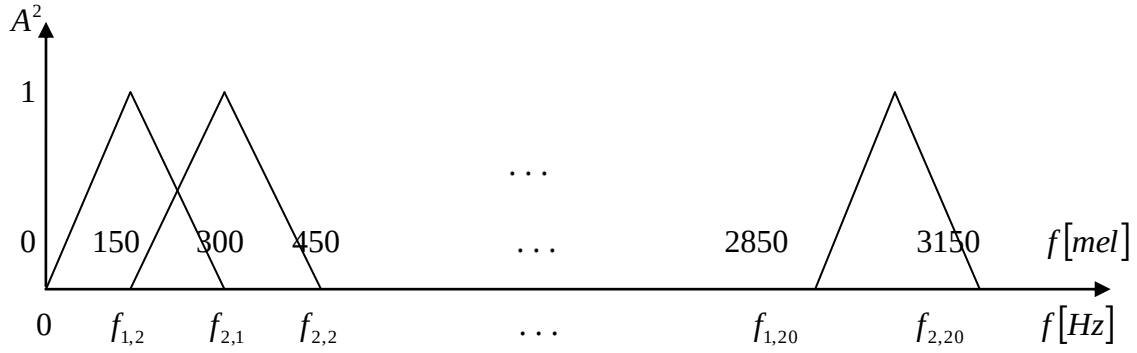
5.2.1. Примена троугаоних чујних критичних опсега

Као први начин апроксимације опадајуће природе амплитуде чујног критичног опсега намеће се примена линеарних нагиба у односу на средиште посматраног чујног критичног опсега. На овај начин апроксимација чујних критичних опсега изведена је функцијама троугаоног облика, што се у литератури често користи.

При испитивању утицаја примене троугаоних чујних критичних опсега на тачност аутоматског препознавања говорника задржан је исти распоред троугаоних филтарских функција (слика 5.4) као и при тестовима када су претпостављени правоугаони чујни критични опсези. Као чујни критични опсези примењене су симетричне троугаоне функције квадрата амплитудске карактеристике:

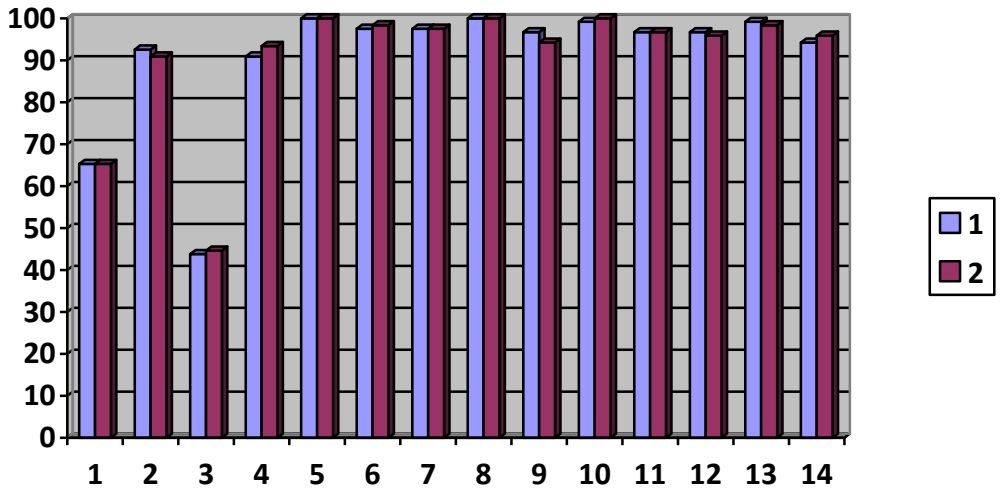
$$A^2(k)=\begin{cases}\dfrac{2}{k_{2,n}-k_{1,n}}\cdot(k-k_{1,n}), & k_{1,n}\le k\le\dfrac{k_{1,n}+k_{2,n}}{2},\\ \dfrac{2}{k_{1,n}-k_{2,n}}\cdot(k-k_{2,n}), & \dfrac{k_{1,n}+k_{2,n}}{2}<k\le k_{2,n},\end{cases} \tag{5.1}$$

при чему *n* означава редни број посматраног чујног критичног опсега из опсега $n=\{1,2,...,20\}$ и *k* означава дискретну учестаност која је на основу учестаности *f* у Hz рачуната у складу са једнакошћу 4.20.



Слика 5.4. Распоред примењених троугаоних критичних опсега.

Имајући у виду резултате аутоматског препознавања говорника у појединим тестовима када су претпостављени чујни критични опсези правоугаоног облика, примена троугаоних критичних опсега у највећем броју случајева резултовала је повећањем тачности препознавања над говорним снимцима из обе говорне базе.



Слика 5.5. Тачност препознавања над Говорном базом 1 у зависности од посматраног тест фајла, 19 MFCCs, 20 троугаоних критичних опсега.
1 – немодификовани критични опсези,
2 – вредности енергија у критичном опсегу кориговане енергијом најмањег критичног опсега.

Тачност препознавања у првом и трећем тесту над говорном базом 1 била је изразито мања у односу на тачност у осталим тестовима. Остали тестови резултовали су тачношћу реда 90% и изнад тог нивоа. Тестови 5 и 8 резултовали су тачношћу 100% док у тестовима 10 и 13 на скупу од 121 говорника 120 говорника је тачно препознато. Ради лакшег поређења тачности препознавања остварене применом правоугаоних односно троугаоних чујних

критичних опсега табеларно је дат приказ разликовања тачности препознавања говорника у ова два случаја. У табели су поређени случајеви, односно дата је разлика у постигнутој тачности аутоматског препознавања говорника, када је у вектору обележја коришћено 19 MFCCs рачунатих на основу 20 троугаоних (слика 5.5 – 1) односно 20 правоугаоних чујних критичних опсега (слика 5.2 – 3).

Редни број теста	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Број препознатих говорника при примени правоугаоних чујних критичних опсега	72	98	48	105	121	119	113	120	118	119	115	114	121	116
Разликовање тачности [%]	5.7	11.5	4.1	4.1	0	-0.8	4.1	0.8	-0.8	0.8	1.6	2.4	-0.8	-1.6

Табела 5.4. Разликовање тачности препознавања при примени троугаоних и правоугаоних чујних критичних опсега у тестовима над Говорном базом 1 када су као вектори обележја коришћени 19 MFCCs, поређење приказаних резултата на слици 5.5 – 1 и 5.2 – 3.

Највеће побољшање постигнуто је у другом тесту када је за вектор обележја који садржи 19 MFCCs уместо 98 препознатих говорника при примени правоугаоних критичних опсега, при примени истог броја тј. 20 троугаоних чујних критичних опсега, препознато 112 од укупног броја посматраног 121-ог говорника. Процентуална поправка тачности као што се види у табели 5.4 за тај случај износи око 11.5%. Ипак не може се рећи да је ово апсолутна величина поправке за тај тест пошто у датом тесту при примени правоугаоних чујних критичних опсега бољи резултати су постигути када је из модела избачен само елемент који представља варијансу нултог MFCC. Тест 2 је и у поређењу са тим случајем резултовао већом тачношћу препознавања када су MFCCs рачунати на основу троугаоних чујних критичних опсега. Следеће значајно побољшање остварено је у тесту 1 али само у односу на случај примене правоугаоних чујних критичних опсега када су унутар вектора обележја коришћени 19 MFCCs. Наиме у случају коришћења и нултог MFCC а нарочито у случају када је у оквиру посматраних модела говорника занемарен елемент који представља варијансу нултог MFCC, при коришћењу правоугаоних чујних критичних опсега постигнути су бољи резултати (табела 5.5).

Опис упоредног теста 1 при коришћењу 20 правоугаоних чујних критичних опсега	Нулти MFCC + 19 MFCCs	Нулти + 19 MFCCs, при чему је занемарен елемент који у моделу одговара нултом MFCC	19 MFCCs
Број тачно препознатих говорника за примењене правоугаоне чујне критичне опсеге	82	89	72
Разликовање тачности [%]	-2.4793	-8.2644	5,7852

Табела 5.5. Разликовање тачности остварене у тесту 1 над Говорном базом 1 у случају употребе троуганих чујних критичних опсега у односу на случај коришћења правоугаоних чујних критичних опсега.

Разликовања тачности приказана у табели 5.4 такође не одсликавају прави добитак у тачности препознавања у тестовима 3 и 4. У тим тестовима при примени правоугаоних чујних критичних опсега најбоља тачност препознавања постигнута је у случају када је из модела искључен елемент који одговара варијанси нултог MFCC. Тачност је у тим случајевима била за око 2 до 3% боља у односу када је занемарен нулти MFCC из вектора обележја. Ипак примена троугаоних чујних критичних опсега при чему су вектори обележја садржали 19 MFCCs у ова два теста резултовала је већом тачношћу и у односу на случај примене правоугаоних чујних критичних опсега када је у моделима занемарен члан који одговара варијанси нултог MFCC. Побољшање тачности је у ова два теста износило око 1%.

Тачност препознавања у тестовима 5 – 14 изузев теста 7 не показује значајнија одступања у односу на тачност постигнуту при примени правоугаоних чујних критичних опсега. Побољшања тачности препознавања у тестовима 7 и 12 допринела су да у тестовима 5 – 13 тачност буде изнад 95%. Евидентна мала погоршања у тачности препознавања реда 0.8% односно у тесту 14 реда 1.6%, што се у броју препознатих говорника одсликава у умањењу за једног односно два говорника, показују да у извесним случајевима примена овако конципираних троугаоних чујних критичних опсега може и деградирати тачност препознавања односно указује на апсолутну неидеалност примењених троугаоних филтарских функција (једнакост 5.1).

Примењени распоред чујних критичних опсега који за посматрани *i*-ти чујни критични опсег, $i=\{2,3,\ldots,20\}$, подразумева преклапање прве половине посматраног опсега и друге половине његовог претходника (слике 5.1 и 5.4), има за последицу пресликавање одређеног дела енергије из претходног у посматрани чујни критични опсег. Мера овог пресликавања зависи од претпостављеног облика чујних критичних опсега и од вредности амплитуда спектра у деловима чујних критичних опсега који се преклапају. У зависности од квалитета говорног снимка односно од односа сигнал – шум у њему може се говорити и о одређеној деградирајућој мери коју уноси преклапање чујних критичних опсега а која се одражава кроз израчунате MFCCs за посматрани рам говорног сигнала. Наиме, узастопним пресликавањима одређеног дела енергије из претходног чујног критичног опсега у наредни пресликавају се постојеће сметње тј. шум присутан у посматраном фреквентном опсегу анализираног говорног рама. На тај начин стиче се нестварна слика тј. представа о вредностима MFCCs који се рачунају за посматрани говорни рам.

Имајући у виду израженије присуство шума у Говорној бази 2 као и чињеницу да сама појава шума у говорном снимку не мора бити последица околине или несавршености уређаја за снимање него и могућих појава нерегуларности у гласу говорника, након сваког скупа тестова испитивања тачности аутоматског препознавања говорника за изабрани облик чујних критичних опсега извршен је и скуп тестова када је примењена корекција на енергије посматраних чујних критичних опсега за анализирани говорни рам.

Примењена корекција састојала се од три корака:

- израчунавање логаритама енергија, $E_{\log(k)}$, за сваки чујни критични опсег посматраног говорног рама, $k=\{1,2,\ldots,20\}$,
- одређивање најмањег логаритма енергије, $E_{\min\log(k)}=\min\left\{E_{\log(k)}\right\}$, $k=\{1,2,\ldots,20\}$,
- кориговање логаритама енергија чујних критичних опсега одузимањем најмањег логаритма енергије посматраних чујних критичних опсега од првобитно израчунатог логаритма енергије дотичног чујног критичног опсега,

$$E_{\log(k)\textit{ново}}=E_{\log(k)}-\min\left\{E_{\log(k)}\right\}. \tag{5.2}$$

Даље израчунавање MFCCs у складу са једнакошћу 4.19 вршено је коришћењем вредности $E_{\log(k)\textit{ново}}$ као вредности логаритма енергије посматраног *k*-тог чујног критичног опсега.

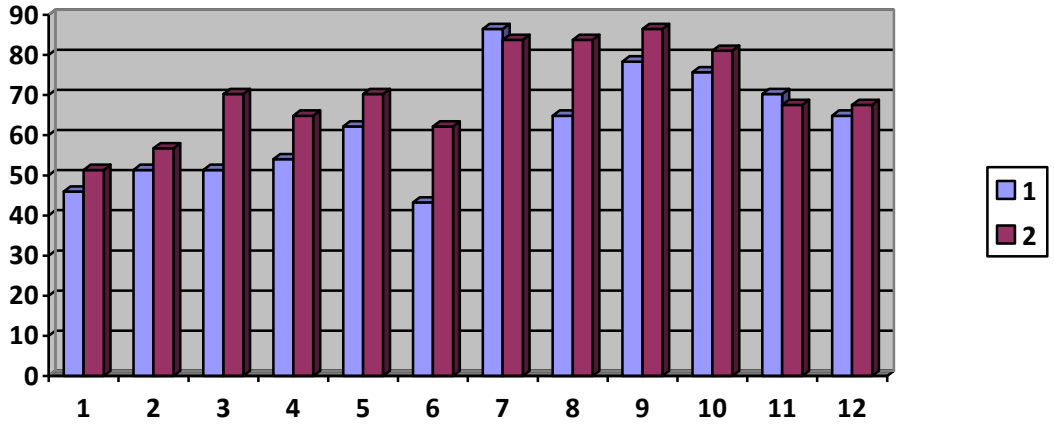
Као што приказује хистограм на слици 5.5 примена корекције логаритама енергије на претходни описани начин није значајно утицала на промену тачности аутоматског препознавања говорника у односу на случај када на логаритме енергије у чујним критичним опсезима није примењена корекција.

Редни број теста	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Број тачно препознатих говорника без примењене корекције енергија	79	112	53	110	121	118	118	121	117	120	117	117	120	114
Разликовање тачности [%]	0	-1.6	0.8	2.4	0	0.8	0	0	-2.4	0.8	0	-0.8	-0.8	1.6

Табела 5.6. Разликовања постигнутих тачности за тестове над Говорном базом 1 при примени троугаоних чујних критичних опсега са примењеном корекцијом енергије у односу на случај када корекција није примењена.

Из табеле 5.6 може се закључити да је средња промена тачности препознавања једнака нули. Имајући у виду да разликовање тачности од 0.8% одговара порасту броја тачно препознатих говорника за 1, највеће повећање броја тачно препознатих говорника износило је 3 у оквиру теста 4. За исти број се смањио број тачно препознатих говорника у оквиру теста 9. Тачност у оквиру тестова 5 и 8 од 100% није нарушена применом корекције на логаритме енергија у оквиру чујних критичних опсега. Поред ова два теста, у скупу тестова у којима је примењена корекција логаритама енергија, у оквиру теста 10 тачно је препознат још један говорник тако да је и у оквиру овог теста постигнута тачност аутоматског препознавања говорника 100%.

Поступак којим је спроведена корекција логаритама енергије у оквиру чујних критичних опсега првенствено је уведен ради побољшања тачности аутоматског препознавања говорника у тестовима над Говорном базом 2 која се одликује постојањем израженијег шума, шума који се чује при слушању снимака говора. Стога овај поступак не доноси значајнија побољшања у постигнутој тачности препознавања над Говорном базом 1 која се одликује доста мањим нивоом шума у односу на Говорну базу 2. За Говорну базу 1 примена ове енергијске корекције у извесним случајевима је сувише груба, сувише је груба процена присутног шума, тако да је резултат смањење тачности препознавања.



Слика 5.6. Тачност препознавања над Говорном базом 2 у зависности од посматраног тест фајла, 19 MFCCs, 20 троугаоних критичних опсега.
1 – немодификовани критични опсези,
2 – вредности енергија у критичном опсегу кориговане енергијом најмањег критичног опсега.

Редни број теста	1	2	3	4	5	6	7	8	9	10	11	12
Број тачно препознатих говорника при примени правоугаоних чујних критичних опсега	15	15	20	18	20	16	24	23	26	24	23	21
Разлика броја препознатих говорника при примени троугаоних и правоугаоних чујних критичних опсега	2	4	-1	2	3	0	8	1	3	4	3	3

Табела 5.7. Приказ разликовања броја препознатих говорника при примени троугаоних и правоугаоних чујних критичних опсега у тестовима над Говорном базом 2 када су у оквиру вектора обележја коришћени првих 19 MFCCs, поређење приказаних резултата на сликама 5.6 – 1 и 5.3 – 3.

Примена троугаоних критичних опсега у оквиру тестова над Говорном базом 2 у већини случајева резултовала је повећањем броја препознатих говорника у односу случај када су претпостављени правоугаони критични опсези (табела 5.7). Највећа поправка препознавања остварена је у тесту 7, где је након примене троугаоних критичних опсега тачно препознато 32 говорника од укупног броја 37 говорника. Процентуално посматрано тачност у овом тесту је поправљена за око 21.6% што је скоро двоструко већа процентуална промена у односу на највеће увећање тачности у истим тестовима над Говорном базом 1 (табела 5.4 – тест 2). Велико процентуално увећање представља последицу више него двоструко мањег броја говорника у Говорној бази 2. Ипак највећа остварена тачност у овим тестовима мања је у односу на тестове над Говорном базом 1. Примена троугаоних чујних критичних опсега утицала је на поправку просечне тачности препознавања над обе говорне базе (табела 5.8). Процентуална поправка тачности над Говорном базом 1 је прилично мала и износи око 2%, што говори да је у просеку број препознатих говорника повећан за 2 – 3 говорника. Релативна поправка тачности над Говорном базом 2 је нешто већа, приближно око 7% што такође одговара повећању броја препознатих говорника за 2 до 3 говорника.

Просечна тачност у односу на тип чујног критичног опсега [%]	Говорна база 1	Говорна база 2
правоугаони	88.488	55.18
троугаони	90.732	62.387

Табела 5.8. Просечна тачност аутоматског препознавања говорника у зависности од говорне базе и типа примењеног чујног критичног опсега.

Примена корекције на енергије у чујним критичним опсезима допринела је повећању броја препознатих говорника у већини тестова, изузев тестова 7 и 11 (табела 5.9). Пратећи промене броја препознатих говорника у табелама 5.7 и 5.9 уочљиви су случајеви тестова 5 и 9. Тада је повећање броја препознатих говорника за 3 при прелазу са правоугоних на троугаоне чујне критичне опсеге праћено још једним повећањем за такође 3 говорника при извршеној енергијској корекцији на троугаоне чујне критичне опсеге.

Редни број теста	1	2	3	4	5	6	7	8	9	10	11	12
Број тачно препознатих говорника без примењене корекције енергија	17	19	19	20	23	16	32	24	29	28	26	24
Разлика броја препознатих говорника при примени корекције енергије и без њене примене	2	2	7	4	3	7	-1	7	3	2	-1	1

Табела 5.9. Промена броја препознатих говорника за тестове над Говорном базом 2 при примени троугаоних чујних критичних опсега за случај са примењеном корекцијом енергије у односу на случај када корекција није примењена.

Примена корекције логаритама енергија у посматраним троугаоним чујним критичним опсезима највише је повећала број препознатих говорника у тестовима 3, 6 и 8 када је број препознатих говорника увећан за 7 говорника. Из приказаних резултата очигледно је да се примена енергијске корекције различито испољава на постигнуту тачност препознавања говорника у различитим тестовима. Дакле, тачност препознавања зависи од теста тј. од садржине говора на основу које се врши препознавање говорника.

Тип теста	Просечна тачност препознавања [%]	
	Говорна база 1	Говорна база 2
1	90.732	62.387
2	90.791	70.27

Табела 5.10. Просечна тачност препознавања при коришћењу 20 троугаоних критичних опсега (1 – без примене, 2 – са применом, корекције на логаритме енергија чујних критичних опсега).

Постигнуте тачности препознавања над Говорном базом 2 су мање у односу на тачност препознавања над Говорном базом 1. Стога се примене одговарајућих корекција, прво троугаоних чујних критичних опсега а затим и енергијске корекције унутар њих, просечно посматрано више одсликавају на повећање тачности препознавања над Говорном базом 2 (табела 5.10). Посматрајући повећања тачности препознавања у односу на почетне тестове када је уз првих 19 MFCCs коришћен и нулти MFCC и када је просечна тачност над Говорном базом 1 износила 80.87% а над Говорном базом 2 37.387% уочава се изразито већи утицај примене извршених корекција над Говорном базом 2. Упоређујући резултате из табела 5.10 и 5.3 уочава се да се тачност препознавања над Говорном базом 2 повећала за приближно 32.883% док је повећање над Говорном базом 1 износило приближно 9.921%.

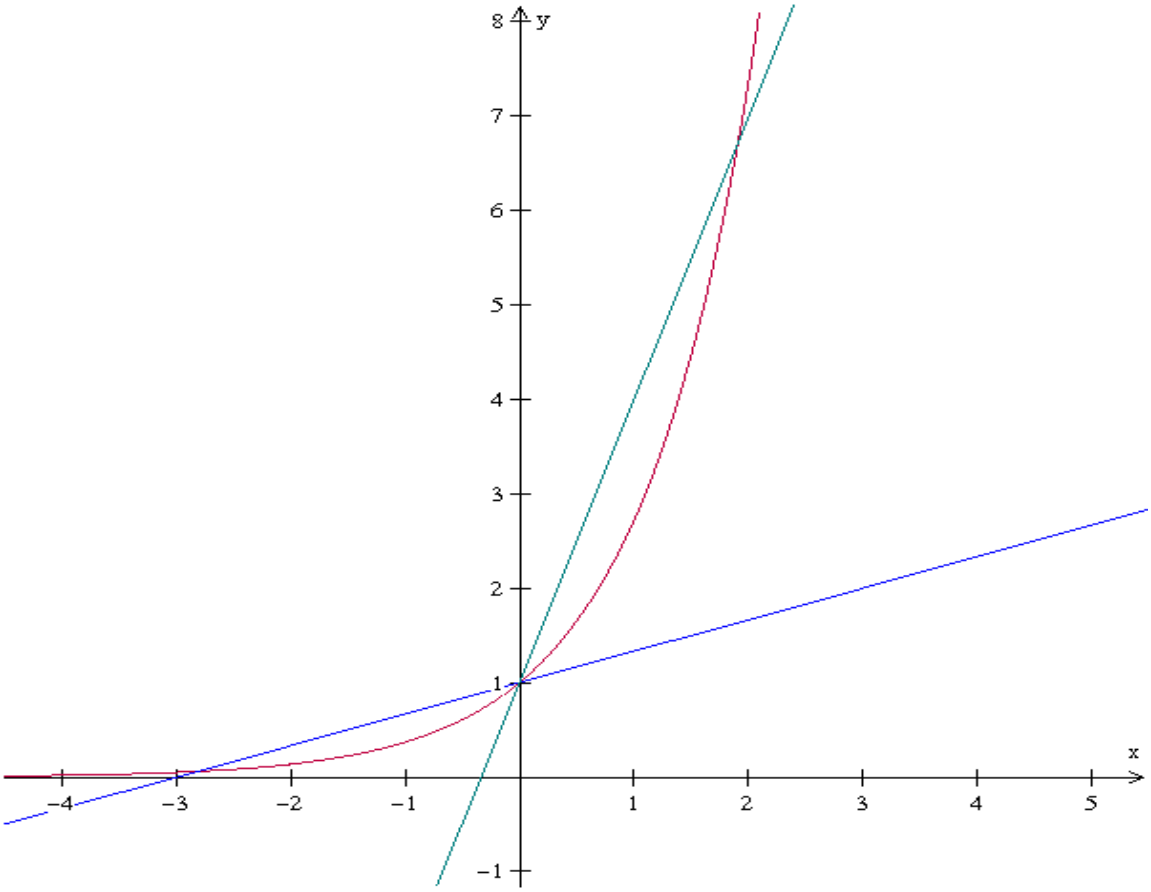
Чињеница да је примена троугаоних чујних критичних опсега допринела побољшању тачности препознавања говорника над обе говорне базе указује да фреквентна селективност унесена применом оваквих чујних критичних опсега доприноси повећању тачности препознавања. Троугаона карактеристика критичног опсега умањује утицај спектралних компонената у односу на компоненту у његовом средишту. Ради повећања селективности критичних опсега потребно је додатно смањити амплитуду компонената у околини средишта опсега. На основу овога следи да је апроксимацију стрмина критичног опсега потребно извршити применом нелинеарних функција. Као функција која има брз пораст у следећем

скупу експеримената ради повећања селективности посматраних критичних опсега примењена је експоненцијална функција.

5.2.2. Примена експоненцијалних чујних критичних опсега

Експоненцијална функција општег облика $y = e^x$ за вредности независно промењљиве $x \in (-\infty, +\infty)$ може се посматрати у два дела (слика 5.7):

1. део (доњи део) за $x \in (-\infty, 0]$ при чему функција има вредности из опсега $y \in (0,1]$ и
2. део (горњи део) за $x \in (0, \infty)$ при чему фунција има вредности из опсега $y \in (1, \infty)$.



Слика 5.7. Експоненцијална функција као погодна за апроксимацију чујних критичних опсега у односу на линеарну функцију.

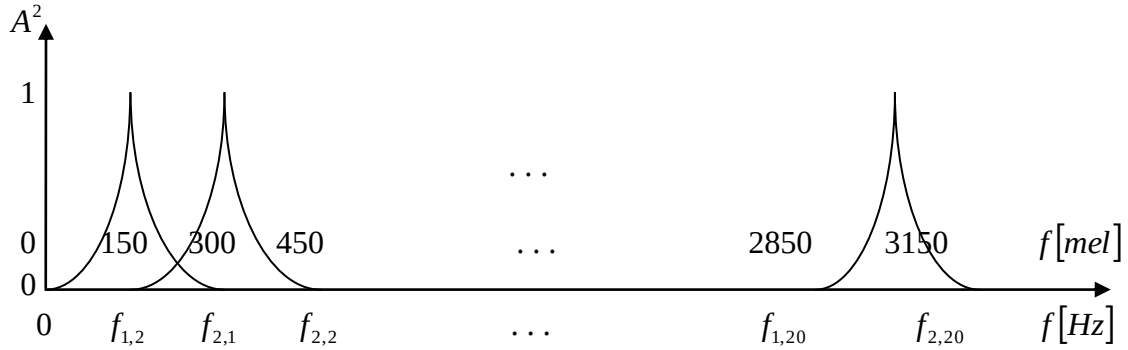
На слици 5.7 је дат упоредни приказ експоненцијалне функције $y = e^x$ (црбена боја) и две линеарне функције: $y_1 = \frac{1}{3} \cdot x + 1$ (плава боја) односно $y_2 = 3 \cdot x + 1$ (зелена боја). Посматрајући график на слици 5.7 уочава се да са становишта опсега $x \in (-\infty, 0]$ доњи део експоненцијалне криве показује бољу селективност у односу на линеарну функцију y_1 односно на скуп линеарних функција $y_{1,A} = \frac{1}{A} \cdot x + 1$, A={2, 3, 4, …}, при чему A одговара вредности опсега који се посматра односно при апроксимацији чујних критичних опсега троугаоним функцијама A одговара половини његове ширине. Други, горњи део експоненцијалне функције има већу стрмину пораста у односу на њен први део. Такође овај део експоненцијалне криве има већу селективност у односу на скуп линеарних функција $y_{1,B} = B \cdot x + 1$, B={1, 2, 3, …}, при чему би сада коефицијент B одговарао мери потребне

истакнутости средишње компоненте у посматраном критичном опсегу. Пораст параметра *B* означава пораст стрмине линеарне функције, повећање посматране средишње али и компонената у њеној околини, док експоненцијална крива мање утиче на повећање компонената у изабраној околини средишта односно показује већу селективност. У зависности од интервала у ком се налази аргумент експоненцијалне функције следи и део који ће се користити. Уколико се средишња вредност преносне карактеристике чујног критичног опсега веже за тачку у којој експоненцијална функција има вредност 1 тада се апроксимација чујних критичних опсега врши доњим делом експоненцијалне функције. Када се крајње тачке преносних карактеристика чујних критичних опсега вежу за тачке у којима експоненцијална функција има вредност 1 тада се апроксимација чујних критичних опсега врши горњим деловима одговарајућих експоненцијалних функција.

Уколико крајње тачке посматраног *n*-тог чујног критичног опсега одговарају нормализованим учестаностима $k_{1,n}$ и $k_{2,n}$ тада за позицију средишта чујног критичног опсега у тачки $k_{c,n}=\dfrac{k_{1,n}+k_{2,n}}{2}$ коришћење доњих делова одговарајућих експоненцијалних функција ради његове апроксимације вршено је на следећи начин:

$$A_{\exp}^2(k)=\begin{cases} e^{(k-k_{c,n})a}, & k_{1,n}\le k\le k_{c,n}, \\ e^{-(k-k_{c,n})a}, & k_{c,n}<k\le k_{2,n}, \end{cases} \tag{5.3}$$

при чему највећа вредност преносне карактеристике чујних критичних опсега има вредност једнаку 1 (слика 5.8). Ради подешавања селективности примењеног чујног критичног опсега уведен је фактор стрмине посматране експоненцијалне функције, *a* у једнакости 5.3. Повећање фактора стрмине резултује удаљавањем фреквентних компонената. У случају коришћења доњег дела експоненцијалне криве, компоненте у околини средишта се умањују, док се у случају примене горњег дела експоненцијалне функције централна компонента највише додатно истиче у односу на остале компоненте. У тестовима, како при апроксимацији чујних критичних опсега доњим деловима одговарајућих експоненцијалних фукнција тако и при коришћењу њихових горњих делова вршено је упоредно тестирање утицаја фактора стрмине за вредности $a=1$ и $a=2$ на тачност препознавања говорника.



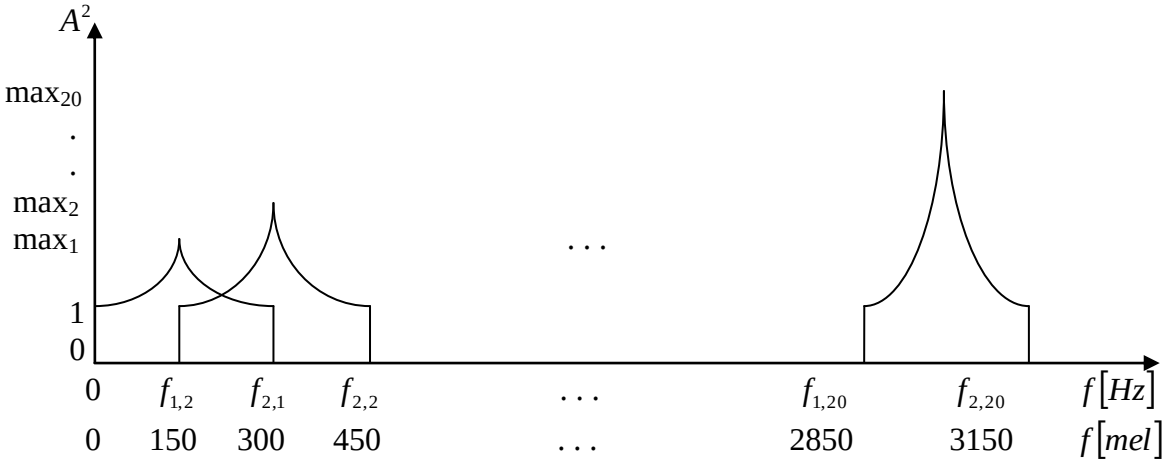
Слика 5.8. Распоред примењених експоненцијалних чујних критичних опсега при коришћењу доњег дела експонецијалне функције.

Горњи део експоненцијалне криве $y=e^{x}$ започиње вредношћу функције $y(0)=e^{0}=1$ и ова експоненцијална крива је погодна за апроксимацију левог дела критичног опсега. Имајући у виду да је претпостављена симетричност чујних критичних опсега у односу на средишњу спектралну компоненту десну ивицу чујног критичног опсега погодно је апроксимирати експоненцијалном кривом $y=e^{-x}$ тако да завршна тачка чујног критичног опсега, као и почетна, такође има вредност једнаку 1. Крајње тачке овако апроксимираног чујног критичног опсега увећане су за 1 у односу на троугаону апроксимацију чујних критичних опсега (слика 5.4). Претпостављајући да је дискретна целобројна учестаност *k*

израчуната на основу једнакости 4.20 у тестовима је коришћен следећи облик квадрата амплитуде преносне карактеристике чујних критичних опсега:

$$A^2_{\exp}(k)=\begin{cases} e^{(k-k_{1,n})\cdot a}, & k_{1,n}\le k\le k_{c,n},\\ e^{-(k-k_{2,n})\cdot a}, & k_{c,n}<k\le k_{2,n}, \end{cases} \tag{5.4}$$

при чему n означава редни број посматраног чујног критичног опсега, $n=\{1,2,...,20\}$. У односу на изворну експоненцијалну функцију у коришћеној експоненцијалној функцији за апроксимацију чујних критичних опсега аргумент је помножен фактором $a\in\{1,2\}$ ради упоредног испитивања утицаја фактора селективности на тачност препознавања. При фактору селективности $a=2$ апроксимирани чујни критични опсег постиже већу селективност већим издизањем средишње компоненте у односу на остале у посматраном чујном критичном опсегу. Најмања вредност квадрата амплитуде овако дефинисаних чујних критичних опсега износи 1 (слика 5.9).

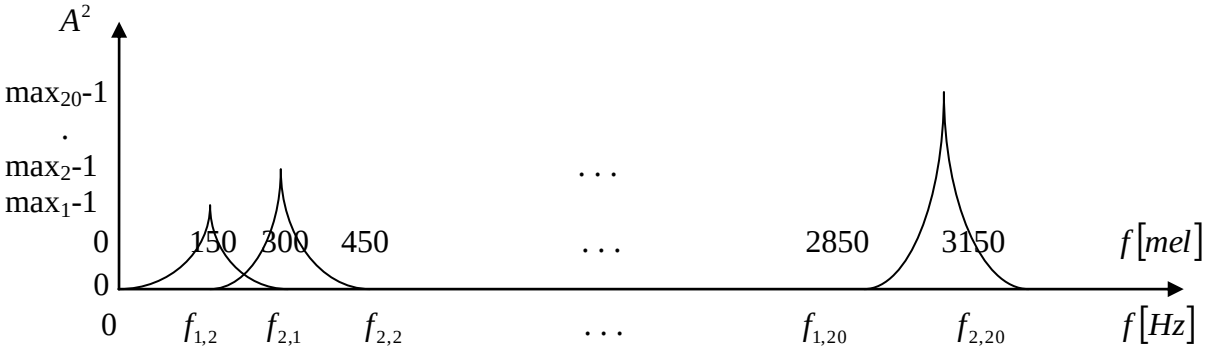


Слика 5.9. Распоред примењених експоненцијалних критичних опсега заснованих на горњем делу експоненцијалне функције.

Ради испитивања утицаја критичних опсега чија ће вредност апроксимације у крајњим тачкама опадати на нулту вредност, као што је то било при примени троугаоних критичних опсега (једнакост 5.1, слика 5.4), а који ће користити горње делове експоненцијалне функције, коришћени су чујни критични опсези следеће апроксимације (слика 5.10):

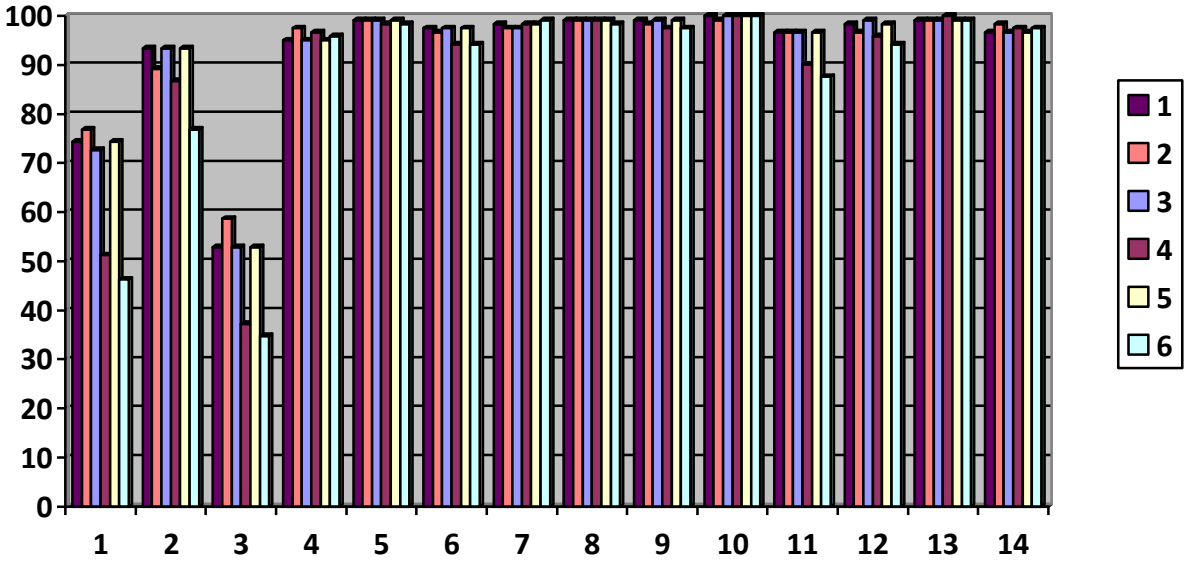
$$A^2_{\exp-1}(k)=\begin{cases} e^{(k-k_{1,n})\cdot a}-1, & k_{1,n}\le k\le k_{c,n},\\ e^{-(k-k_{2,n})\cdot a}-1, & k_{c,n}<k\le k_{2,n}, \end{cases} \tag{5.5}$$

такође за $n=\{1,2,...,20\}$.

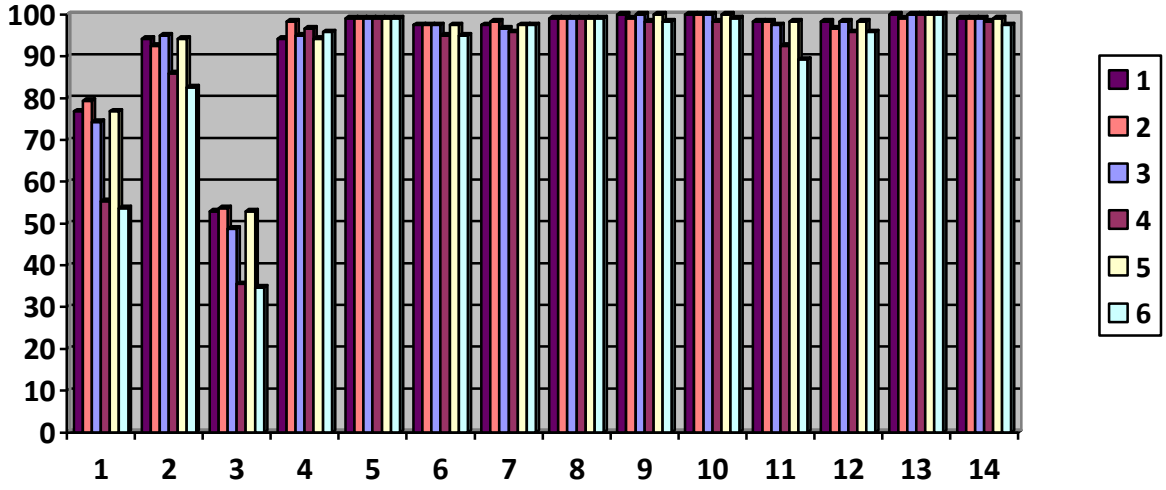


Слика 5.10. Распоред примењених спуштених експоненцијалних критичних опсега заснованих на горњем делу експоненцијалне функције.

Као и у претходним тестовима када су претпостављени троугаони критични опсези, и при тестирању над експоненцијалним критичним опсезима вршена су и испитивања утицаја истоветне корекције логаритма енергије унутар посматраних чујних критичних опсега на тачност препознавања.



Слика 5.11. Тачност препознавања над Говорном базом 1 у зависности од посматраног тест фајла, 19 MFCCs, 20 експоненцијалних критичних опсега, фактор стрмине $a=1$, (1 и 2 доњи део експоненцијалне функције, 3 и 4 – горњи део експоненцијалне функције спуштен за један; 5 и 6 – горњи део експоненцијалне функције).
1, 3, 5 – немодификовани критични опсези,
2, 4, 6 – примењена корекција енергије.



Слика 5.12. Тачност препознавања над Говорном базом 1 у зависности од посматраног тест фајла, 19 MFCCs, 20 експоненцијалних критичних опсега, фактор стрмине $a=2$, (1 и 2 доњи део експоненцијалне функције, 3 и 4 – горњи део експоненцијалне функције спуштен за један; 5 и 6 – горњи део експоненцијалне функције).
1, 3, 5 – немодификовани критични опсези,
2, 4, 6 – примењена корекција енергије.

Посматрањем хистограма на сликама 5.11 и 5.12 уочава се прилично уједначена тачност у оквиру истог теста, када није примењена корекција енергије критичних опсега, за

једну стрмину претпостављених експоненцијалних чујних критичних опсега. Такође евидентно је да се тестови над групом говорних снимака "Речи", тестови 4 – 14, одликују тачношћу преко 95% (табеле: 5.11, 5.12 и 5.13). У тесту 2 евидентирана је тачност између 90 и 95%, у тесту 1 тачност је била око 20% мања, док је тачност у тесту 3 била најнижа и приближно је око 18 – 20% мања у односу на тачност препознавања у тесту 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$\Sigma/14$
$a=1$	74.4	93.4	52.9	95	99.2	97.5	98.3	99.2	99.2	100	96.7	98.3	99.2	96.7	92.9
$a=1$ са корекцијом енергије	76.9	89.3	58.7	97.5	99.2	96.7	97.5	99.2	98.3	99.2	96.7	96.7	99.2	98.3	93.1
$a=2$	76.9	94.2	52.9	94.2	99.2	97.5	97.5	99.2	100	100	98.3	98.3	100	99.2	93.4
$a=2$ са корекцијом енергије	79.3	92.6	53.7	98.3	99.2	97.5	98.3	99.2	99.2	100	98.3	96.7	99.2	99.2	93.6

Табела 5.11. Упоредни приказ постигнуте тачности за тестове 1 – 14 над Говорном базом 1 при апроксимацији чујних критичних опсега доњим деловима експоненцијалних кривих у зависности од стрмине коришћене експоненцијалне функције.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$\Sigma/14$
$a=1$	72.7	93.4	52.9	95	99.2	97.5	97.5	99.2	99.2	100	96.7	99.2	99.2	96.7	92.7
$a=1$ са корекцијом енергије	51.2	86.8	37.2	96.7	98.3	94.2	98.3	99.2	97.5	100	90.1	95.9	100	97.5	88.8
$a=2$	74.4	95	48.8	95	99.2	97.5	96.7	99.2	100	100	97.5	98.3	100	99.2	92.9
$a=2$ са корекцијом енергије	55.4	85.9	35.5	96.7	99.2	95	95.9	99.2	98.3	98.3	92.6	95.9	100	98.3	89

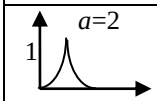
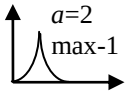
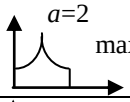
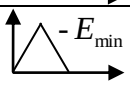
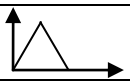
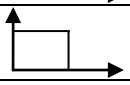
Табела 5.12. Упоредни приказ постигнуте тачности за тестове 1 – 14 над Говорном базом 1 при апроксимацији чујних критичних опсега горњим деловима експоненцијалних кривих спуштеним за 1 на доле у зависности од стрмине коришћене експоненцијалне функције.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$\Sigma/14$
$a=1$	74.4	93.4	52.9	95	99.2	97.5	98.3	99.2	99.2	100	96.7	98.3	99.2	96.7	92.9
$a=1$ са корекцијом енергије	46.3	76.9	34.7	95.9	98.3	94.2	99.2	98.3	97.5	100	87.6	94.2	99.2	97.5	87.1
$a=2$	76.9	94.2	52.9	94.2	99.2	97.5	97.5	99.2	100	100	98.3	98.3	100	99.2	93.4
$a=2$ са корекцијом енергије	53.7	82.6	34.7	95.9	99.2	95	97.5	99.2	98.3	99.2	89.3	95.9	100	97.5	88.4

Табела 5.13. Упоредни приказ постигнуте тачности за тестове 1 – 14 над Говорном базом 1 при апроксимацији чујних критичних опсега горњим деловима експоненцијалних кривих у зависности од стрмине коришћене експоненцијалне функције.

Примена корекције логаритма енергије у чујним критичним опсезима у одређеним тестовима резултовала је значајним смањењем тачности препознавања. При примени чујних критичних опсега заснованих на горњим деловима одговарајућих експоненцијалних

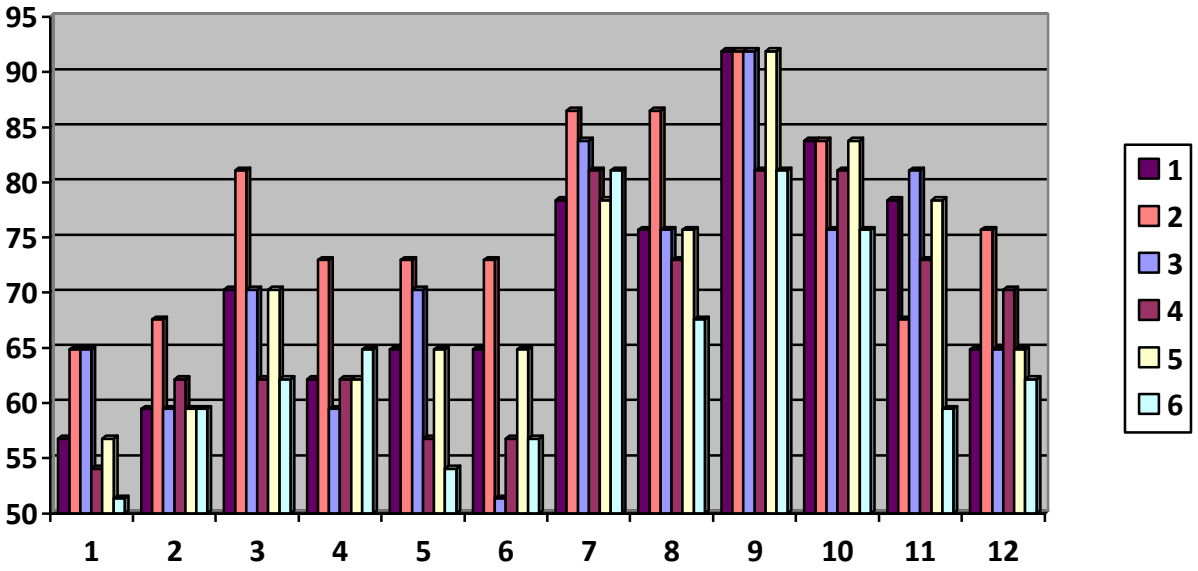
функција мала повећања тачности препознавања при примени енергијске корекције су примећена у тестовима 4 и 7 док су најизраженија смањења тачности препознавања евидентна у тестовима 1, 2 и 3, тестовима у којима је и пре примене енергијске корекције постигнута најмања тачност препознавња. Када су чујни критични опсези апроксимирани доњим деловима експоненцијалних функција евидентно је побољшање тачности применом енергијске корекције, тестови 1, 3, 4 и тест 14 при стрмини критичног опсега $a=1$. Највеће повећање тачности од око 6% примећује се у тесту 3 при коришћењу експоненцијалних критичних опсега стрмине $a=1$. Просечне тачности за сваки скуп од 14 тестова над Говорном базом 1 без примењене енергијске поправке на експоненцијалне чујне критичне опсеге (табеле: 5.11, 5.12 и 5.13, колона $\Sigma/14$) су прилично уједначене док је примена енергијске поправке врло мала побољшања донела углавном у случају чујних критичних опсега апроксимираних доњим делом експоненцијалне функције.

Тип критичног опсега	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	76.9	94.2	52.9	94.2	99.2	97.5	97.5	99.2	100	100	98.3	98.3	100	99.2
	74.4	95	48.8	95	99.2	97.5	96.7	99.2	100	100	97.5	98.3	100	99.2
	76.9	94.2	52.9	94.2	99.2	97.5	97.5	99.2	100	100	98.3	98.3	100	99.2
	65.3	90.9	44.6	93.4	100	98.3	97.5	100	94.2	100	96.7	95.9	98.3	95.9
	65.3	92.6	43.8	90.9	100	97.5	97.5	100	96.7	99.2	96.7	96.7	99.2	94.2
	59.5	81	39.7	86.8	100	98.3	93.4	99.2	97.5	98.3	95	94.2	100	95.9

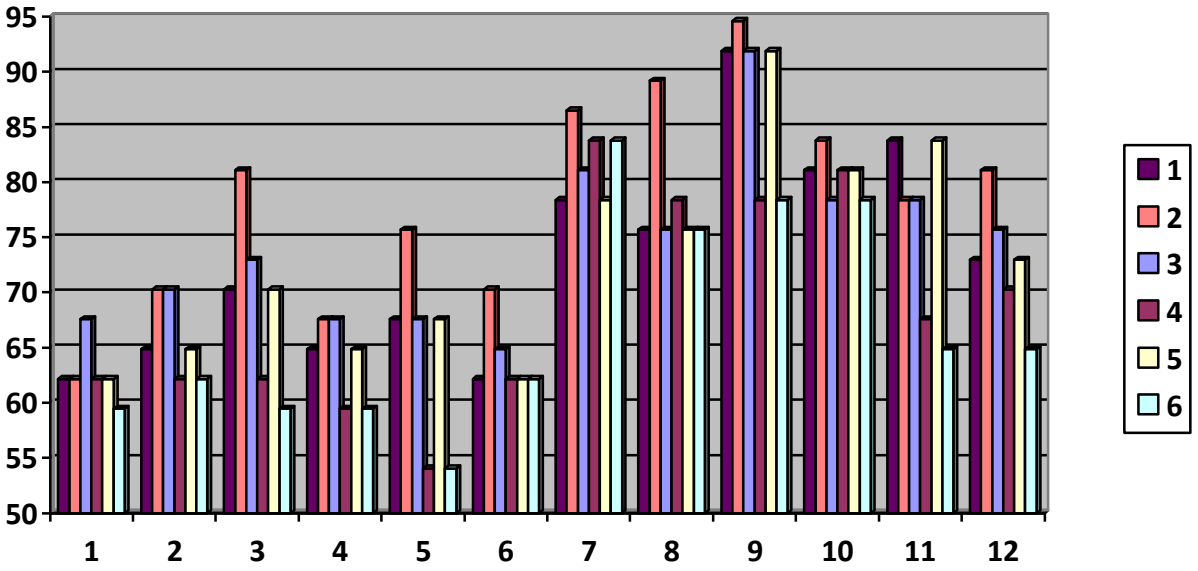
Табела 5.14. Упоредни приказ постигнуте тачности аутоматског препознавања говорника над Говорном базом 1 у зависности од теста и типа коришћеног критичног опсега (фактор стрмине $a=2$), вектор обележја садржи првих 19 MFCCs.

Примена експоненцијалних чујних критичних опсега повећала је тачност аутоматског препознавања говорника над Говорном базом 1. Тачност препознавања у оквиру теста 2 се приближила нивоу тачности који се постиже над скупом снимака "Речи". Примена експоненцијалних критичних опсега у односу на почетну примену правоугаоних побољшала је тачност препознавања за око 14%, чиме је и тачност у тесту 2 достигла ниво од 95%. Тачност у тесту 1 је и даље за око 20% мања у односу на тест 2. У поређењу са постигнутом тачношћу када су примењени правоугаони критични опсези и у овом тесту је остварено побољшање препознавања од приближно 15% (табела 5.14). Постигнута тачност у тесту 3 се задржала на око 50% у зависности какав се тип експоненцијалног критичног опсега примени. Евидентно је релативно повећање тачности и у овом тесту при промени облика чујних критичних опсега (табела 5.14) као и достизање највеће вредности при примени експоненцијалних критичних опсега. Ипак остварено повећање није толиких размера да би тачност у овиру овог теста значајније приближило оствареној тачности над говорним снимцима из групе "Речи". У тестовима {5, 6, ..., 14} који су се и при примени правоугаоних критичних опсега одликовали прилично високом тачношћу препознавања, изнад 90%, остварено је мање процентуално побољшање. Само у тестовима 5 и 6 препознат је по један

говорник мање у односу на исте тестове када је примењен правоугаони облик критичних опсега.



Слика 5.13. Тачност препознавања над Говорном базом 2 у зависности од посматраног тест фајла, 19 MFCCs, 20 експоненцијалних критичних опсега, фактор стрмине $a = 1$ (1 и 2 – доњи део експонецијалне функције, 3 и 4 – горњи део експоненцијалне функције спуштен за један; 5 и 6 – горњи део експоненцијалне функције).
1, 3, 5 – немодификовани критични опсези,
2, 4, 6 – примењена корекција енергије.



Слика 5.14. Тачност препознавања над Говорном базом 2 у зависности од посматраног тест фајла, 19 MFCCs, 20 експоненцијалних критичних опсега, фактор стрмине $a = 2$ (1 и 2 – доњи део експонецијалне функције, 3 и 4 – горњи део експоненцијалне функције спуштен за један; 5 и 6 – горњи део експоненцијалне функције).
1, 3, 5 – немодификовани критични опсези,
2, 4, 6 – примењена корекција енергије.

	1	2	3	4	5	6	7	8	9	10	11	12	$\Sigma/12$
$a=1$	56.76	59.46	70.27	62.16	64.86	64.86	78.38	75.68	91.89	83.78	78.38	64.86	70.94
$a=1$ са корекцијом енергије	64.86	67.57	81.08	72.97	72.97	72.97	86.49	86.49	91.89	83.78	67.57	75.68	77.03
$a=2$	62.16	64.86	70.27	64.86	67.57	62.16	78.38	75.68	91.89	81.08	83.78	72.97	72.97
$a=2$ са корекцијом енергије	62.16	70.27	81.08	67.57	75.68	70.27	86.49	89.19	94.59	83.78	78.38	81.08	78.38

Табела 5.15. Упоредни приказ постигнуте тачности за тестове 1 – 12 над Говорном базом 2 при апроксимацији чујних критичних опсега доњим деловима експоненцијалних кривих у зависности од стрмине коришћене експоненцијалне функције.

	1	2	3	4	5	6	7	8	9	10	11	12	$\Sigma/12$
$a=1$	64.86	59.46	70.27	59.46	70.27	51.35	83.78	75.68	91.89	75.68	81.08	64.86	70.72
$a=1$ са корекцијом енергије	54.05	62.16	62.16	62.16	56.76	56.76	81.08	72.97	81.08	81.08	72.97	70.27	73.44
$a=2$	67.57	70.27	72.97	67.57	67.57	64.86	81.08	75.68	91.89	78.38	78.38	75.68	74.32
$a=2$ са корекцијом енергије	62.16	62.16	62.16	59.46	54.05	62.16	83.78	78.38	78.38	81.08	67.57	70.27	68,47

Табела 5.16. Упоредни приказ постигнуте тачности за тестове 1 – 12 над Говорном базом 2 при апроксимацији чујних критичних опсега горњим деловима експоненцијалних кривих спуштеним за 1 на доле у зависности од стрмине коришћене експоненцијалне функције.

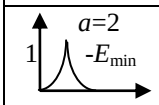
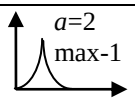
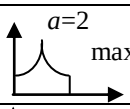
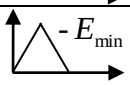
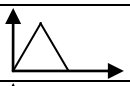
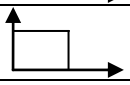
	1	2	3	4	5	6	7	8	9	10	11	12	$\Sigma/12$
$a=1$	56.76	59.46	70.27	62.16	64.86	64.86	78.38	75.68	91.89	83.78	78.38	64.86	70.94
$a=1$ са корекцијом енергије	51.35	59.46	62.16	64.86	54.05	56.76	81.08	67.57	81.08	75.68	59.46	62.16	64.64
$a=2$	62.16	64.86	70.27	64.86	67.57	62.16	78.38	75.68	91.89	81.08	83.78	72.97	72.97
$a=2$ са корекцијом енергије	59.46	62.16	59.46	59.46	54.05	62.16	83.78	75.68	78.38	78.38	64.86	64.86	66.89

Табела 5.17. Упоредни приказ постигнуте тачности за тестове 1 – 12 над Говорном базом 2 при апроксимацији чујних критичних опсега горњим деловима експоненцијалних кривих у зависности од стрмине коришћене експоненцијалне функције.

Највећа средња тачност над Говорном базом 2 остварена је при примени експоненцијалних критичних опсега заснованих на доњем делу одговарајуће експоненцијалне функције (табеле: 5.15, 5.16 и 5.17). Као што се примећује на сликама 5.13 и 5.14 примена енергијске корекције је при примени оваквих чујних критичних опсега у највећем броју тестова допринела повећању тачности препознавања изнад тачности постигнуте за остала два тестирана типа експоненцијалних критичних опсега. Упоређујући тачност остварену применом спуштених експоненцијалних чујних критичних опсега у односу на тачност остварену када су претпостављени подигнути експоненцијални критични опсези уочава се да је у већини тестова боља тачност постигнута применом спуштених експоненцијалних чујних критичних опсега.

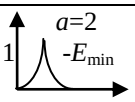
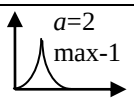
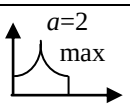
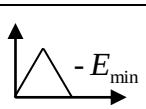
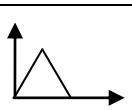
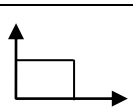
Примена корекције на логаритме енергија унутар посматраних чујних критичних опсега заснованих на горњем делу експоненцијалне функције у већини тестова је и над Говорном базом 2 као и над Говорном базом 1 резултовала смањењем постигнуте тачности у односу на случај без њене примене.

У већини тестова над Говорном базом 2 при примени експоненцијалних чујних критичних опсега постигнута је највећа тачност у односу на примену правоугаоних односно троугаоних (табела 5.18). У односу на примену правоугаоних чујних критичних опсега постигнута су значајна повећања тачности која су у највећем броју случајева у интервалу 15 – 20%. Примена експоненцијалних критичних опсега резултовала је мањим повећањима тачности у односу на случај када су примењени троугаони чујни критични опсези.

Тип критичног опсега	1	2	3	4	5	6	7	8	9	10	11	12
	62.16	70.27	81.08	67.57	75.68	70.27	86.49	89.19	94.59	83.78	78.38	81.08
	67.57	70.27	72.97	67.57	67.57	64.86	81.08	75.68	91.89	78.38	78.38	75.68
	62.16	64.86	70.27	64.86	67.57	62.16	78.38	75.68	91.89	81.08	83.78	72.97
	51.35	56.76	70.27	64.86	70.27	62.16	83.78	83.78	86.49	81.08	67.57	67.57
	45.95	51.35	51.35	54.05	62.16	43.24	86.49	64.86	78.38	75.68	70.27	64.86
	40.54	40.54	54.05	48.65	54.05	43.24	64.86	62.16	70.27	64.86	62.16	56.76

Табела 5.18. Упоредни приказ постигнуте тачности аутоматског препознавања говорника над Говорном базом 2 у зависности од теста и типа коришћеног критичног опсега (фактор стрмине a=2), вектор обележја садржи првих 19 MFCCs.

Као што је евидентно из табеле 5.18 тестови 7 и 8 резултовали су смањењем тачности препознавања при преласку са троугаоних на експоненцијалне чујне критичне опсеге засноване на горњем делу експоненцијалне функције. Посматрајући промене облика чујних критичних опсега, правоугаони – троугаони – експоненцијални, при тестовима над Говорном базом 2 највећа тачност препознавања остварена је у тесту 9 при примени експоненцијалних чујних критичних када је од 37 говорника аутоматски препознавач тачно препознао 34 односно 35 говорника при примени енергијске корекције на експоненцијалне критичне опсеге засноване на доњем делу експоненцијалне функције.

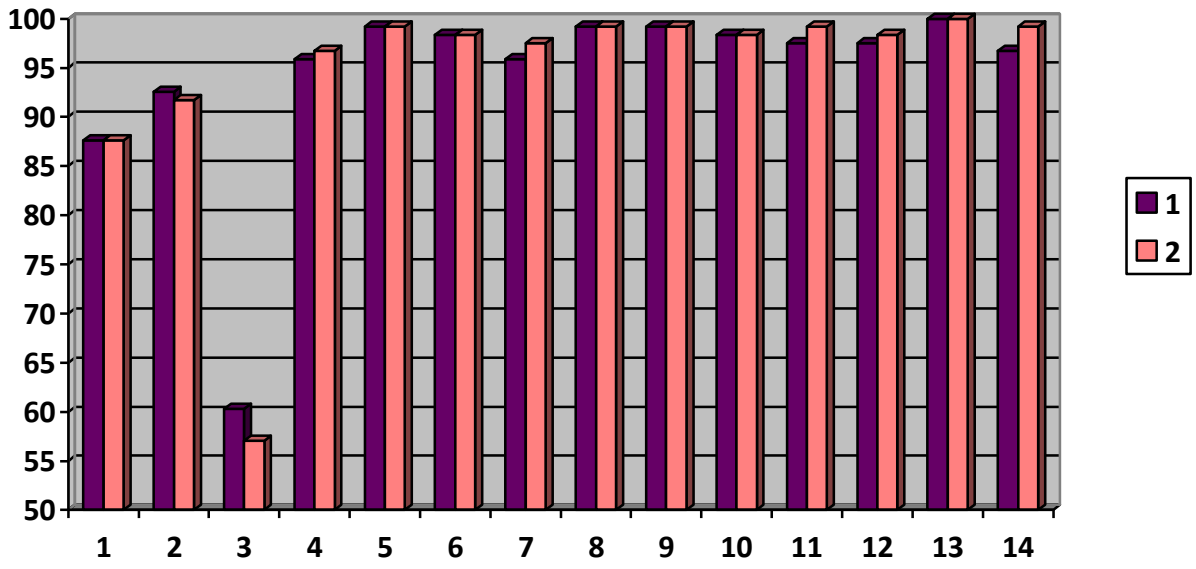
Тип чујног критичног опсега						
Просечна тачност	78.38	74.32	72.97	70,49	62.39	55.18

Табела 5.19. Упоредни приказ просечне тачности препознавања остварене над Говорном базом 2 у зависности од примењеног чујног критичног опсега.

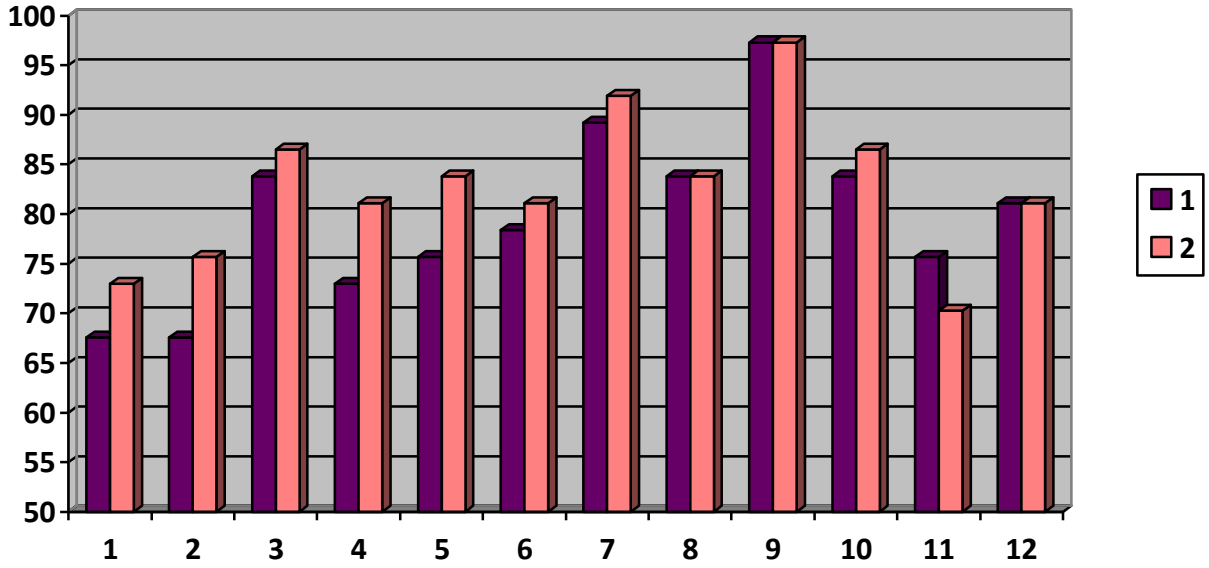
Примена енергијске корекције у случају коришћења троугаоних чујних критичних опсега значајније је поправила тачност препознавања у многим тестовима чиме их је по

тачности приближила случају када је претпостављен експоненцијални облик чујних критичних опсега (табела 5.19). У односу на почетну примену правоугаоних чујних критичних опсега, примена енергијске корекције на експоненцијалне чујне критичне опсеге засноване на доњем делу експоненцијалне функције поправила је тачност препознавања у просеку од 20 до 25%.

Резултати препознавања показују да је примена експоненцијалних чујних критичних опсега резултовала повећањем просечне тачности аутоматског препознавања говорника над обе говорне базе. У досадашњим тестовима у оквиру вектора обележја посматрано је 19 MFCCs. Посматрајући бројне вредности у оквиру израчунатих коваријансних матрица које представљају моделе говора посматраних говорника као и моделе њиховог тест говора уочено је истакнуто разликовање између коваријансних матрица за истог говорника на местима која одговарају варијанси 19-ог MFCC. Међусобна апсолутна разлика елемента $\Sigma_{19,19}$ у матрици добијеној на основу снимака намењених прављењу модела посматраног говорника и елемента на истој позицији у коваријансној матрици која одговара тест говорном снимку истог говорника износила је често више од 10 док је апсолутна разлика елемената који нису последица 19-ог MFCC често имала вредност мању од 1. У приметном броју случајева апсолутно разликовање елемената коваријансних матрица за истог говорника који нису последица 19-ог MFCC била је мања од 0.1 што је указивало да је 19. MFCC давао значајан допринос разликовању модела говорника добијеног при обуци и модела његовог тест говора. У циљу елиминације овог утицаја испитана је тачност аутоматског препознавања говорника при коришћењу чујних критичних опсега онога облика који је у претходним тестовима резултовао највећом тачношћу препознавања, тј. задржан је експоненцијални облик чујних критичних опсега заснованих на доњем делу експоненцијалне функције при чему је вектор обележја садржао 18 MFCCs (слике 5.15 и 5.16).



Слика 5.15. Тачност препознавања над Говорном базом 1 у зависности од посматраног тест фајла, 18 MFCCs, 20 експоненцијалних критичних опсега, фактор стрмине $a = 2$, доњи део експоненцијалне функције,
1 – немодификовани критични опсези,
2 – примењена корекција енергије.



Слика 5.16. Тачност препознавања над Говорном базом 2 у зависности од посматраног тест фајла, 18 MFCCs, 20 експоненцијалних критичних опсега, фактор стрмине $a = 2$, доњи део експонецијалне функције,
1 – немодификовани критични опсези,
2 – примењена корекција енергије.

Смањење броја обележја на 18 MFFCs као и задржавање примене експонецијалних чујних критичних опсега поправило је тачност препознавања у обе говорне базе. У два теста који се истичу по најмањој односно највећој тачности у оквиру тестова над Говорном базом 1 односно Говорном базом 2 тачност је повећана. Наиме тачност у тесту 3 над Говорном базом 1 је повећана на 60% док је тачност у тесту 9 над Говорном базом 2 повећана на око 97% што значи да је у том тесту препознавач од 37 говорника 36 говорника тачно препознао.

5.3. ПРАЋЕЊЕ ПРОМЕНЉИВОСТИ ЕЛЕМЕНАТА МОДЕЛА И ЊИХОВ УТИЦАЈ НА ТАЧНОСТ АУТОМАТСКОГ ПРЕПОЗНАВАЊА ГОВОРНИКА

У претходним тестовима над говорним базама 1 и 2 након промене облика чујних критичних опсега од правоугаоних ка експоненцијалним и избацивању нултог и 19-тог MFCC из вектора обележја најмања тачност препознавања остварена је у тесту 3 над Говорном базом 1. Такође евидентно је да је у тестовима над Говорном базом 2 након претходно наведених побољшања тачност препознавања од преко 90% остварена у тестовима 9 и 7.

У претходним тестовима праћења утицаја промене облика чујних критичних опсега на тачност аутоматског препознавања говорника такође праћено је и понашање модела истих говорника у случају скупа говорних снимака намењених обуци и тесту. На основу тога у претходним експериментима су изузети нулти и деветнаести MFCC што је у већини тестова резултовало повећањем тачности препознавања. Ради детаљније анализе мере разликовања елемената модела истих говорника дефинисана је матрица разликовања, D_t, на нивоу одговарајућег теста, $t \in \{1,2,...,14\}$ за Говорну базу 1, односно $t \in \{1,2,...,12\}$ у оквиру тестова над Говорном базом 2. Елемент одговарајуће матрице разликовања, $D_t(i,j)$, представља укупно апсолутно разликовање у посматраном тесту узимајући у обзир тренинг односно тест коваријансне матрице, тј. моделе истих говорника, на месту (i,j) односно:

$$D_t(i,j)=\sum_{n=1}^{N}\left|\Sigma_{(i,j)(n)}^{тренинг}-\Sigma_{(i,j)(n)}^{тест}\right|, \tag{5.6}$$

при чему су *i* и *j* ознаке посматраног елемента у тренинг и тест коваријансним матрицама и *N* је број говорника у говорној бази у оквиру које се врши тестирање.

У табели 5.20 приказане су вредности укупног апсолутног разликовања за $\Sigma_{0,0}$ односно $\Sigma_{19,19}$ елементе у тренинг и тест коваријансним матрицама истих говорника у оквиру групе тестова 1-14 над Говорном базом 1 рачунате на основу једнакости 5.6, релативно у односу на најмању вредност у одговарајућим матрицама разликовања. Евидентне су изразите величине ових релативних разликовања одакле и оправданост за изузимање ових елемената модела из коначног модела говорника, као елемената који нису зависни само од говорника над посматраном говорном базом.

Тест (*t*)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
dmin	3.48	2.23	2.54	2.19	1.81	2.23	2.01	2.06	1.68	1.97	1.67	2.12	2.09	1.62
$D_t(0,0)/dmin$	3213	2238	2060	2065	3113	2367	2809	1860	3218	2413	3136	2195	1576	2555
$D_t(19,19)/dmin$	722	1008	593	263	221	260	215	160	274	302	602	242	183	451

Табела 5.20. Упоредни приказ укупног апсолутног разликовања за $\Sigma_{0,0}$ и $\Sigma_{19,19}$ елементе у разматраним тестовима при примени 20 правоугаоних чујних критичних опсега (слика 5.1). d_{min} представља минимално апсолутно разликовање у матрицама разликовања извршених тестова.

Табела 5.20 приказује релативне вредности $D_t(0,0)$ и $D_t(19,19)$ одговарајућих матрица разликовања у случају примењених правоугаоних чујних критичних опсега. Претходни експериментални резултати су показали да промена облика чујних критичних опсега од троугаоних ка експоненцијалним доприноси повећању тачности препознавања али проблем разликовања елемената тест и тренинг модела једног истог говорника, посматраних на истим местима у коваријансним матрицама ће и даље бити присутан, тј. модели једног истог говорника показиваће разликовања при моделовању различитих снимака. Имајући у виду да су као обележја говорника коришћени MFCCs рачунати на основу једнакости 4.19 следи да ово није изненађујућа чињеница. Наиме MFCCs по својој дефиницији израчунавања зависе од амплитудског спектра разматраног говорног сигнала тј. од његове енергије. Енергија говорног сигнала поред тога што зависи од гласности изговарања која је последица како самог говорника и његовог тренутног начина изражавања, што може бити и последица његових тренутних осећања, такође последица је и текстуалне садржине говора. На основу ове чињенице може се закључити да MFCCs као обележја говорника не испуњавају у потпуности први услов постављен пред њих на основу ког се обележја говорника требају одликовати испољавањем великих међуговорничких разлика и малих промена за једног посматраног говорника. Ове неидеалности MFCCs се пресликавају у примењене моделе говорника тј. у коваријансне матрице којима је моделована одређена група снимака или појединачни снимак посматраног говорника.

Грубо решење проблема разликовања модела једног истог говорника може се решити утврђивањем елемената који се значајно разликују у моделима и занемаривањем тих елемената у коначном моделу говорника. Ипак, иако се ови елементи значајно разликују те као такви може се рећи да нису директна последица посматраног говорника, они носе и извесну количину информација о говорнику ком модел припада. Пошто ови елементи показују значајну променљивост од модела до модела једног истог говорника то значи да они носе релативно малу количину информацију о припадајућем говорнику, односно следи да је значајност ових елемената модела при одлучивању релативно мала у односу на елементе модела који показују мању променљивост. Стога је боље решење задржати све елементе

модела и при томе сваки елемент модела посматрати кроз призму његове значајности у моделовању говорника, тј. увести одговарајуће тежинске коефицијенте елемената модела говорника.

Мера важности појединих елемената модела тј. тежински коефицијенти $W(i,j)$ који им одговарају последица су вредности одговарајућих елемената $D(i,j)$ у матрици растојања ***D***. Што је нагомилано растојање на одређеном месту у матрици растојања веће то је вредност тежинског коефицијента који му одговара мања, тј.

$$D(i,j) \geq D(k,l) \;\Rightarrow\; W(i,j) \leq W(k,l). \tag{5.7}$$

На основу овог правила постављене су границе важности појединих елемената тежинске матрице:

$$D(i,j) > \frac{\max}{5} \;\Rightarrow\; W(i,j) = 0.05, \tag{5.8а}$$

$$\frac{\max}{10} < D(i,j) \leq \frac{\max}{5} \;\Rightarrow\; W(i,j) = 0.3, \tag{5.8б}$$

$$\frac{\max}{20} < D(i,j) \leq \frac{\max}{10} \;\Rightarrow\; W(i,j) = 0.6, \tag{5.8в}$$

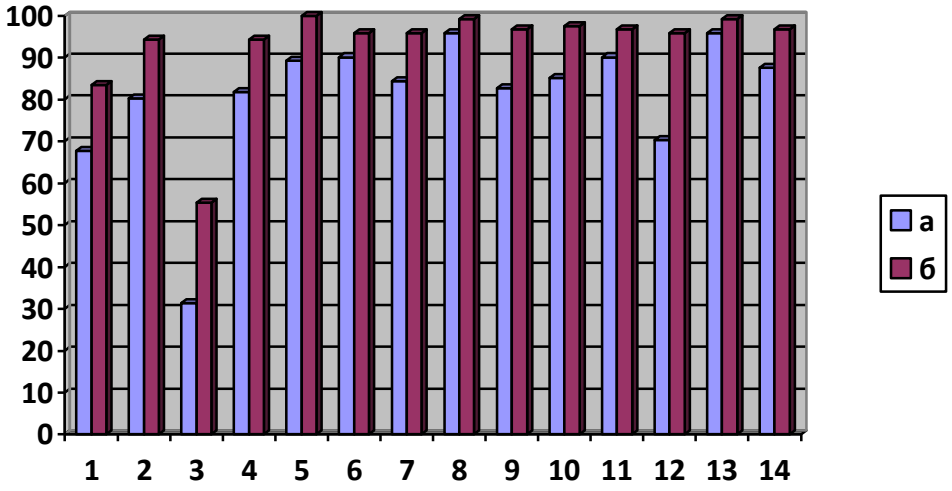
$$\frac{\max}{40} < D(i,j) \leq \frac{\max}{20} \;\Rightarrow\; W(i,j) = 1.0, \tag{5.8г}$$

$$D(i,j) \leq \frac{\max}{40} \;\Rightarrow\; W(i,j) = 1.9. \tag{5.8д}$$

Границе су дефинисане у односу на највећу вредност, *max*, у одговарајућој матрици разликовања *D*. Резултантни елемент модела, $\Sigma^{рез}(i,j)$, рачунат је као производ његове првобитно процењене вредности, $\Sigma(i,j)$, и одговарајуће вредности у тежинској матрици, $W(i,j)$:

$$\Sigma^{рез.}(i,j) = \Sigma(i,j) \cdot W(i,j), \tag{5.9}$$

Примена овог поступка прво је извршена за 20 правоугаоних чујних критичних опсега при чему је вектор обележја садржао нулти и првих 19 MFCCs (слика 5.17).

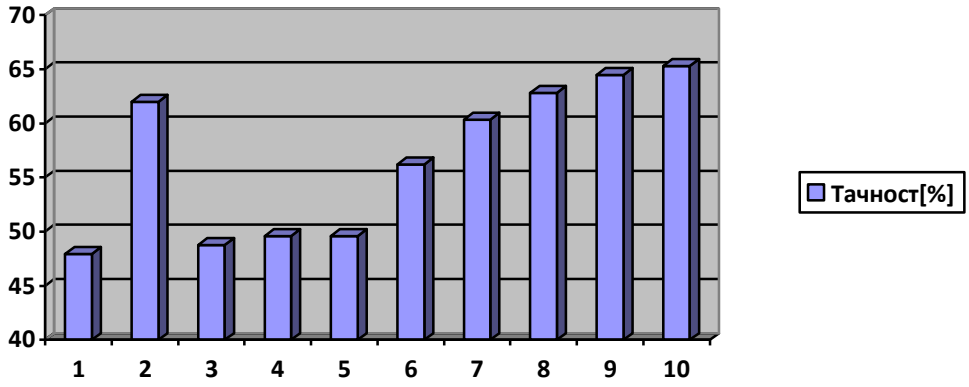


Слика 5.17. Упоредни приказ тачности препознавања при примени 20 правоугаоних чујних критичних опсега и вектора обележја који садржи нулти и првих 19 MFCCs у случајевима: **а** – основног начина моделовања помоћу коваријансних матрица, **б** – примене трансформације елемената модела.

Евидентно је значајно побољшање тачности препознавања применом трансформације 5.8а-д и када су приликом израчунавања MFCCs примењени правоугаони критични опсези.

Тест 1 је поновљен за случај примене 20 експоненцијалних критичних опсега заснованих на доњем делу експоненцијалне функције при чему је извршена и енергијска корекција над енергијама унутар посматраних чујних критичних опсега. Првобитно, што би одговарало случају **а** на слици 5.17, постигнута је тачност препознавања 68.595% (83/121). Након примене трансформације на моделе говорника, једнакости 5.8а-д, тачност препознавања је повећана на 90.0826% (109/121), што такође представља и повећање тачности у односу на случај примене вектора обележја од првих 18 MFCCs, слика 5.15.

У свим претходним тестовима, тачност препознавања у тесту 3 је значајно мања од тачности у осталим тестовима, што издваја овај тест као подесан и репрезентативан за испитивање успешности могућих трансформација на моделе засноване на примени MFCCs као обележја говорника.



Слика 5.18. Тачност препознавања у тесту 3 над говорном базом 1 у зависности од начина моделовања говорника: **1** – моделовање извршено коваријансним матрицама за расположиве векторе обележја који садрже нулти и првих 19 MFCCs; **2** – примењене трансформације у једнакостима 5.8а-д на моделе добијене у **1**; **3** – примењене трансформације као у једнакостима 5.10а-д на моделе из **1**; **4**- примењен логаритам на чланове модела веће од max/1.1; **5** – примењена сигмоид функција на чланове модела веће од max/1.7; **6** – примењена сигмоид функција на све елемента модела из **1**; **7** – 18MFCCs; **8** – 18MFCCs и сигмоид на све чланове модела; **9** – 18MFCCs и сигмоид на чланове модела веће од max/2; **10** – 18MFCCs и сигмоид на чланове модела веће од max/1.7. Примењено је 20 експоненцијалних критичних опсега заснованих на доњем делу експоненцијалне функције фактора стрмине 2.

Израчунавање претходно дефинисане матрице разликовања подразумева познавање тест скупа. Ово је изводљиво у случају тестирања над говорним базама, када се унапред дефинишу тренинг и тест скуп, али не одговара случају стварне примене при препознавању говорника из снимка говора за који се унапред не зна коме припада. Евидентно је да ће се тренинг и тест модел истог говорника разликовати, при чему је циљ да се на неки начин то разликовање што више умањи.

Уколико елемент модела има изгледе да има велику вредност тада такође је вероватније да на том месту тренинг и тест модели показују велико разликовање. Ради умањења могућег разликовања између тест и тренинг модела истих говорника на поменутим местима потребно је на одређени начин умањити изразито велике елементе модела. Попут трансформације која је исказана низом једнакости 5.8а-д, то би се могло постићи применом истих односних једнакости али сада на сваки модел понаособ и у односу на највећи његов елемент:

$$\Sigma(i,j) > \frac{\max}{5} \quad \Rightarrow \quad W(i,j) = 0.05, \tag{5.10а}$$

$$\frac{\max}{10} < \Sigma(i,j) \le \frac{\max}{5} \quad \Rightarrow \quad W(i,j) = 0.3, \tag{5.10б}$$

$$\frac{\max}{20} < \Sigma(i,j) \le \frac{\max}{10} \quad \Rightarrow \quad W(i,j) = 0.6, \tag{5.10в}$$

$$\frac{\max}{40} < \Sigma(i,j) \le \frac{\max}{20} \quad \Rightarrow \quad W(i,j) = 1.0, \tag{5.10г}$$

$$\Sigma(i,j) \le \frac{\max}{40} \quad \Rightarrow \quad W(i,j) = 1.9. \tag{5.10д}$$

Сада су границе дефинисане у односу на највећу вредност, *max*, у одговарајућој матрици Σ која представља модел говора говорника. Тачност препознавања је за око проценат већа у односу на тачност препознавања при примени основног модела, слика 5.18 случај 3. Учинак трансформација уведених кроз једнакости 5.8а-д односно 5.10а-д огледа се и у сразмерном умањењу чланова модела односно коваријансне матрице. Сразмерно умањење подразумева веће умањење већих елемената модела. Поменуто се може такође постићи и применом неких нелинеарних функција на елементе модела, као што су логаритамска или сигмоид функција.

Сходно томе да примена логаритамске функције на велике вредности резултује знатним умањењем полазних вредности примена логаритамске функције на елементе модела је спроведена над елементима модела чија је величина близу највеће вредности посматраног модела на следећи начин:

$$\textit{ако} \quad \Sigma(i,j)^{\textit{старо}} > \frac{\max}{1.1} \quad \Rightarrow \quad \Sigma(i,j)^{\textit{ново}} = \log_{10}\left(\Sigma(i,j)^{\textit{старо}}\right). \tag{5.11}$$

На овај начин резултат препознавања је врло мало поправљен у односу на тачност остварену након примене трансформација 5.10а-д. По угледу на претходну примену логаритамске функције, на сличан начин примењена је и сигмоид функција:

$$\textit{ако} \quad \Sigma(i,j)^{\textit{старо}} > \frac{\max}{1.7} \quad \Rightarrow \quad \Sigma(i,j)^{\textit{ново}} = \frac{1}{1+e^{-\left(\Sigma(i,j)^{\textit{старо}}\right)}}. \tag{5.12}$$

Ово је повећало тачност препознавња, слика 5.18 случај 5, али је значајније повећање препознавања остварено применом сигмоид функције на све елементе модела, слика 5.18 случај 6.

Такође испитан је потпун као и делимичан утицај примене сигмоид функције на елементе модела говорника када је вектор обележја садржао првих 18MFCCs, слика 5.18 случајеви: 8, 9 и 10. Евидентно да је у том случају највеће препознавање остварено у случају 10, делимичне примене сигмоид функције на елементе модела у складу са условима постављеним у исказу 5.12. Дакле у случају коришћења модела добијених за векторе обележја који садрже нулти и првих 19 MFCCs највећа тачност препознавања је остварена применом сигмоид функције на све елементе модела, док је при моделовању говорника на основу вектора обележја који садрже првих 18 MFCCs највећа тачност осварена применом сигмоид трансформације на елементе модела који испуњавају услов 5.12. Обзиром да модели говорника који одговарају векторима обележја од првих 18 MFCCs не садрже изразито велике елементе евидентно је да сигмоид функцију није потребно применити на све елементе модела него само на оне елементе модела који су ближи највећој вредности у посматраном моделу, како је исказано у 5.12 добијеној на основу извршених експеримената.

6. ЗАКЉУЧАК

MFCCs се могу успешно користити као обележја при аутоматском препознавању говорника. При томе потребно је имати у виду природу њиховог израчунавања, тј. да се они израчунавају на основу краткотрајних временских сегмената говорних сигнала. Дакле они се свакако мењају од сегмента до сегмента говорног сигнала али коришћењем стохастичких модела, као што је у овом раду коришћена коваријансна матрица као репрезент одговарајуће претпостављене Гаусове вишедимензионалне расподеле, та њихова временска променљивост се на одређени начин упросечује те модел на одређени начин представља слику промене ових обележја. Стога резултати препознавања зависе од фонетског састава или текстуалног састава како тренинг тако и тест говора. У суштини зависе од мере њихове уједначености.

Ова чињеница не условљава текстуалну идентичност тренинг и тест говорних секвенци да би аутоматски препознавач показивао највеће или прихватљиве резултате препознавања. Обзиром да на количину информације о говорнику највише утиче количина звучности тј. заступљеност звучних елемената у говору који се анализира следи да се резултати примене аутоматског препознавања говорника такође могу посматрати и у светлу примењеног окружења. На овај начин лошији резултати препознавања могу се правдати недостатком потребних информација да би аутоматски препознавач говорника донео исправну одлуку препознавања.

У стварним применама аутоматског препознавача говорника потребно је на одређени начин обезбедити довољно потребних информација о гласу говорника да би се донела исправна одлука препознавања. Добар пример оваквог приступа су тестови 4-14 над Говорном базом 1, тестови над снимцима који садрже изговоре одређених низова речи. Оваквим скуповима речи обезбеђена је довољна количина информација о звучности говора говорника како у фази обуке тако и у фази тестирања. Наиме у оваквим случајевима не само да је реч о више расположивог фонетског материјала него се ради и о квалитету фонетског материјала. Као што је раније напоменуто снимци у оквиру Говорне базе 1 из групе Имена одликују се на известан начин спонтанијим изговором при чему иако сваки снимак са текстуалног становишта има најмање три речи: идентификациони број, име и презиме, са тачке становишта звука који се чује често се може рећи да се ради о две изговорене целине. Често се име и презиме изговарају спојено и онда неизбежно долази до скраћивања вокала те деградације тест модела, што се и одразило на тачност препознавања. С друге стране, у Говорној бази 2 присутност шума у споју са чињеницом да је текстуални садржај снимака сведен на изговорени низ 4 од могућих 10 цифара (0-9), такође често је утицала на деградацију тест модела што се такође одразило на резултате препознавања у многим тестовима над овом говорном базом.

Да би се предупредили ови недостаци како у препознавању над коришћеним говорним базама тако и у стварним применама реализованог аутоматског препознавача говорника експерименти су показали да се одређеним изменама у уобичајеним поступцима за израчунавање MFCCs може постићи тачније препознавање. Показало се да примена чујних критичних опсега заснованих на доњем делу експоненцијалне функције у просеку највише побољшава тачност препознавања у односу на случај када су примењени правоугаони или троугаони чујни критични опсези. Такође уведена поправка енергије унутар чујних критичних опсега дала је побољшања у постигнутој тачности препознавња, највише у тестовима над Говорном базом 2, обзиром да је у њеним снимцима изражено присуство шума.

Вредности међудимензионалних производа вектора израчунатих MFCCs као обележја говорника се пресликавају у вредности унутар коваријансних матрица које су коришћене ради моделовања говорника. Елементи унутар модела одсликавају меру корелисаности димензија унутар расположивих вектора обележја. Обзиром на чињеницу да MFCCs зависе од енергијске расподеле у говорном сигналу следи неизбежно разликовање тренинг и тест модела једног истог говорника. Примена анализе елемената који се значајно разликују у тест

и тренинг говору истих говорника на основу које је израчуната одговарајућа матрица разликовања у посматраном тесту и одређени тежински елементи појединих елемената модела допринела је тачнијем препознавању говорника. Имајући у виду начин добијања елемената матрице разликовања, који се може исказати правилом – елементима у моделима тест односно тренинг говора истих говорника који су нагомилали велику количину разликовања придаје се мала важност при одлучивању односно елементима који се одликују малим разликовањем придаје се велика важност при одлучивању, примена матрице тежинских коефицијената на матрицу која представља модел говора посматраног говорника може се схватити као нискофреквентна представа модела посматраног говорника односно представа модела на тај начин да резултујући модел показује релативно мале промене у времену. Стога примена сигмоид функције такође доприноси побољшању тачности препознавања и на тај начин дефинише могући начин побољшања тачности препознавања у применама аутоматског препознавача говорника када нису унапред познати и тест снимци говорника те није могуће проценити матрицу разликовања. Примена матрице тежинских коефицијената као и сигмоид функције на чланове коваријансних матрица модела смањује распон вредности у моделу. Имајући у виду да су различити модели, модели различитих снимака, истих говорника, међусобно више корелисани него модели различитих говорника, следи да примена поменутих трансформација на моделе поред тога што смањује распон вредности у резултујућем моделу такође и боље групише односно међусобно ближе групише елементе на истим местима у моделима истих говорника, што резултује већом тачношћу аутоматског препознавања говорника.

ЛИТЕРАТУРА

Bimbot Frédéric, Bonastre Jean-François, Fredouille Corinne, Gravier Guillaume, Magrin-Chagnolleau Ivan, Meignier Sylvain, Merlin Teva, Ortega-Garcia Javier, Petrovska-Delacrétaz Dijana, and Reynolds A. Douglas, (2004). "A Tutorial on Text-Independent Speaker Verification", *EURASIP Journal on Applied Signal Processing 2004:4*, pp. 430-451.

Campbell P. Joseph, Jr., (1997). "Speaker recognition: a tutorial", *Proceedings of IEEE,* Vol. 85, No. 9, pp. 1437-1462.

Che Wei Chi, Lin Qiguang, Yuk Dong-Suk, (1996). "An HMM approach to text-prompted speaker verification", *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference* – Volume 2, pp. 673-676.

Chou Wu, Juang Biing-Hwang, (2003). "Pattern Recognition in Speech and Language Processing", © 2003 by *CRC Press LLC*.

Делић Владо, Јованов Љубомир, (2002). "Препознавање говорника коришћењем програмског пакета HTK", *IV ДОГС,* pp. 69-72, Бечеј, мај 2002.

Делић Д. Владо, Сечујски С. Милан, Јаковљевић М. Никша, (2008). "Акциони модел говорне комуникације човек-машина", 16. Телекомуникациони форум *ТЕЛФОР 2008*, pp. 680-683, Србија, Београд, новембар 25.-27., 2008.

Gold Ben, Morgan Nelson, (2000). "Speech and Audio Signal Processing – Processing and Perception of Speech and Music", *Copyright © 2000 by John Wiley & Sons, Inc*.

Imperl Bojan, "Speaker recognition techniques", University of Maribor, http://wwwbox.uni-mb.si/Dsplab/Clanki/bojan/mmmmc_94.pdf

Јаковљевић Никша, (2002). "Препознавање говорника - експерименти", *Дипломски рад, Факултет техничких наука – Нови Сад, октобар 2002*.

Јовичић Т. Слободан, (1999). "ГОВОРНА КОМУНИКАЦИЈА *физиологија, психоакустика и перцепција*", *Издавачко презузеће Наука, Београд, 1999*.

Јокић Иван, Јокић Стеван, (2008). "Аутоматско препознавање говорника – идентификација и верификација гласа говорника, актуелни домети и правци истраживања", *ДОГС, Келебија, 2-3.10.2008*, pp. 46-49.

Јокић Иван, Јокић Стеван, (2009). "Проблематика аутоматског препознавања говорника", *INFOTEH – JAHORINA*, Vol. 8, Ref. B-III-11, p.237-241, March 2009.

Јокић Д. Иван, Добријевић Н. Томислав, Јаковљевић М. Никша, Делић Д. Владо, (2009). "Опис говорне базе за препознавање говорника на српском језику," 17. Телекомуникациони форум *ТЕЛФОР 2009*, Србија, Београд, новембар 24.-26., 2009., Зборник радова, ISBN 978-86-7466-375-2, стр. 1109-1112.

Јокић Иван, (2010). "Утицај телефонских канала на аутоматско препознавање говорника", Магистарски рад, Факултет техничких наука – Нови Сад, 02.11.2010.

Jokić Ivan, Jokić Stevan, Perić Zoran, Gnjatović Milan, Delić Vlado, "Influence of the Number of Principal Components used to the Automatic Speaker Recognition Accuracy", Electronics and Electrical Engineering, ISSN 1392-1215, Kaunas: Technologija, 2012., No. 7(123), pp. 83-86.

Јокић Д. Иван, Јокић Д. Стеван, Делић Д. Владо, "Један начин реализације аутоматског препознавања говорника", Телекомуникациони форум ТЕЛФОР (20, Београд; 2012), Друштво за телекомуникације, ISBN: 978-1-4673-2984-2, стр. 744-747.

Jokić Ivan, Delić Vlado, Jokić Stevan, Perić Zoran, "Influence of the discarding non-speaker specific model parameters and features to accuracy of automatic speaker recognition", Proceedings of the Second International Conference TAKTONS – Novi Sad, Serbia, November 13th-16th, 2013., pp. 96-99.

Kinnunen Tomi, Li Haizhou, (2010). "An overview of text-independent speaker recognition: From features to supervectors", Speech Communication 52, 2010, pp. 12-40.

Lee Chulhee, Hyun Donghoon, Choi Euisun, Go Jinwook, and Lee Chungyong, (2003). "Optimizing Feature Extraction for Speech Recognition", IEEE transaction on Speech and Audio Processing, Vol. 11, No. 1, January 2003., pp. 80-87.

Lyon Richard F., Katsiamis Andreas G., Drakakis Emmanuel M., (2010). "History and Future of Audio Filter Models", Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS 2010), May 30 – June 2, Paris, France, pp. 3809-3812.

Molau Sirko, Pitz Michael, Schlüter Ralf, and Ney Hermann, (2001). "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum", Proceedings International Conference on Acoustic, Speech and Signal Processing, Salt Lake City, UT, June 2001, Volume:1, pp. 73-76.

Mølgaard L Lasse, Jørgensen W Kasper, (2005). "Speaker Recognition – Special Course", IMM-DTU, December 14, 2005, http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4414/pdf/imm4414.pdf

Moreno J. Pedro, Ho P. Purdy, (2003). "A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels", *Published in Eurospeech 2003, 1-4 September 2003, Geneva, Switzerland*.

Quan Le, Bengio Samy, (2002). "Hybrid generative-discriminative models for speech and speaker recognition", Technical Report IDIAP-RR 02-06, *IDIAP*, March 2002.

Quatieri F. Thomas, (2002). "Discrete-Time Speech Signal Processing – Principles and Practice", ©2002 Prentice Hall PTR.

Staroniewicz Piotr, (2005). "Speaker Recognition for VoIP Transmision Using Gaussian Mixture Models", *Advances in Soft Computing*, *Springer Berlin/Heidelberg*, Vol. 30/2005, *Computer Recognition Systems*, pp.739-745.

Siafarikas Mihalis, Ganchev Todor, Fakotakis Nikos, Kokkinakis George, (2007). "Wavelet Packet Approximation of Critical Bands for Speaker Verification", International Journal of Speech Technology, ISSN 1381 – 2416, vol.10, no.4, Springer, pp.197-218.

Tobias Herbig, Franz Gerl, and Wolfgang Minker, (2011). "Self-Learning Speaker Identification – A System for Enhanced Speech Recognition", ISBN 978-3-642-19898-4, © 2011 Springer-Verlag Berlin Heidelberg.

Šalna B., Kamarauskas J., (2010.), "Evaluation of Effectiveness of Different Methods in Speaker Recognition", Electronics and Electrical Engineering, ISSN 1392-1215, Kaunas: Technologija, 2010., No. 2(98), pp. 67-70.

Wildermoth Richard Brett, (2001). "Text-Independent Speaker Recognition Using Source Based Features", *M. Phil. Thesis, Griffith University, Brisbane, Australia*, 2001.

Wu Dalei, Li Baojie and Jiang Hui, (2008). "Normalization and Transformation Techniques for Robust Speaker Recognition", Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert, ISBN 987-953-7619-29-9, pp. 550, 311-330, November 2008, I-Tech, Vienna, Austria.

Young Steve, Evermann Gunnar, Gales Mark, Hain Thomas, Kershav Dan, Liu Xunying (Andrew), Moore Gareth, Odell Julian, Ollason Dave, Povey Dan, Valtchev Valtcho, Woodland Phil, (2009). "The HTK Book (for HTK Version 3.4)", ©COPYRIGHT 1995-1999 *Microsoft Corporation*, ©COPYRIGHT 2001-2009 *Cambridge University Engineering Department*.

https://ccrma.stanford.edu/courses/422/projects/WaveFormat/

http://en.wikipedia.org/wiki/Resource_Interchange_File_Format

http://en.wikipedia.org/wiki/Two's_complement

http://www.sonicspot.com/guide/wavefiles.html