УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Ј О В А Н Г А Л И Ћ

ПРЕПОЗНАВАЊЕ МУЛТИМОДАЛНОГ ГОВОРА ЗАСНОВАНО НА СТАТИСТИЧКОМ ПРИСТУПУ

ДОКТОРСКА ДИСЕРТАЦИЈА

Београд, 2019.

UNIVERSITY OF BELGRADE

SCHOOL OF ELECTRICAL ENGINEERING

J O V A N G A L I Ć

THE RECOGNITION OF MULTIMODAL SPEECH BASED ON STATISTICAL APPROACH

DOCTORAL DISSERTATION

Belgrade, 2019.

МЕНТОР:

_____

др Драгана Шумарац Павловић, редовни професор, Електротехнички факултет, Београд

ЧЛАНОВИ КОМИСИЈЕ:

_____

др Миомир Мијић, редовни професор, Електротехнички факултет, Београд

_____

др Јелена Ћертић, доцент, Електротехнички факултет, Београд

_____

др Милан Војновић, научни сарадник, Центар за унапређење животних активности, Београд

_____

др Жељко Ђуровић, редовни професор, Електротехнички факултет, Београд

_____

др Ана Гавровска, доцент, Електротехнички факултет, Београд

ЗАХВАЛНИЦА

Захваљујем се **проф. др Слободану Јовичићу**, ментору студијског истраживачког рада, на свесрдној помоћи у дугогодишњем истраживачком раду као и стрпљењу и истрајности при руковођењу почетним научно-истраживачким активностима.

Захваљујем се менторки, **проф. др Драгани Шумарац Павловић** на срдачној помоћи и корисним сугестијама при изради докторске дисертације.

Захваљујем се **проф. др Влади Делићу** на сарадњи током наставног процеса на матичном факултету и тежњи да се повећа број стручњака у области говорних технологија за ијекавско говорно подручје српског језика.

Захваљујем се **колегицама и колегама** са Катедре за телекомуникације на Електротехничком Факултету Универзитета у Бањој Луци на колегијалности и испомоћи у заједничким наставним активностима за вријеме израде дисертације. Такође, захваљујем се **проф. др Татјани Пешић-Брђанин**, редовном професору на Катедри за електронику, на помоћи приликом писања почетних научних радова.

Захваљујем се својим родитељима, **мајци Бориславки и оцу Нађи**, као и **брату Милану** што су ми помогли да истрајем у многим животним изазовима као и да се бавим послом којем сам тежио и који волим.

Посљедње, и ништа мање важно, захваљујем се **супрузи Драгани и ћеркама Дуњи и Вери** на пруженој љубави и разумијевању. Оне су ми помогле да се период студирања на докторском студију учини лакшим и дугорочно гледано кориснијим. Њима посвећујем ову докторску дисертацију.

Драгани, Дуњи и Вери

Draganu, Dunji i Veri

ПРЕПОЗНАВАЊЕ МУЛТИМОДАЛНОГ ГОВОРА ЗАСНОВАНО НА СТАТИСТИЧКОМ ПРИСТУПУ

Резиме

Шапат представља специфичан начин говорне комуникације карактерисан одсуством глоталних вибрација и шумном побудом вокалног тракта. И поред отежаног напора у перцепцији, разумљивост шапата у комуникацији између људи је врло висока. Огромна разлика на акустучком нивоу између нормалног говора (говора уобичајеног интензитета) и шапата је главни разлог што перформансе модерних препознавача говора имају снижене перформансе када се примене на шапат.

Статистички приступ у препознавању говора, који је заснован на скривеним Марковљевим моделима (енгл. *Hidden Markov Models* - HMM), има битну улогу у савременим препознавачима говора. У овој дисертацији је низом експеримената показана успешност HMM препознавача у препознавању мултимодалног говора, како зависно тако и независно од говорника. Експерименти су урађени у усаглашеним ("нормалан/нормалан" и "шапат/шапат") и неусаглашеним ("нормалан/шапат" и "шапат/нормалан") обука/тест сценаријима. Због своје нарочите важности у практичним применама, посебна пажња је посвећена препознавању шаптавог говора са обуком на нормалном говору ("нормалан/шапат" сценарио). Коришћена је говорна база изолованих речи у нормалном говору и шапату.

У иницијалном експерименту је анализирана успешност у препознавању за MFCC (енгл. Mel-Frequency Cepstral Coefficients) векторе обележја и 3 јединице за моделовање: фонеме независне од контекста (монофоне), фонеме зависне од контекста (трифоне) и целе речи. У препознавању шапата монофони су се показали као најуспешнији. У наредним експериментима је повећана успешност у

препознавању нормалног говора и шапата коришћењем динамичких обележја, лексикона изговора са 32 монофона и модела иницијализације који користи анотиран део говорне базе. Достигнута је успешност у препознавању шапата од 81.38% (зависно од говорника) и 87.42% (независно од говорника). Помоћу новог алгоритма за добијање кепстралних коефицијената базираних на модификованој фреквенцијској скали додатно је повећана успешност у препознавању шапата, а да при том нису деградиране перформансе у препознавању нормалног говора. Са предложеним векторима обележја је добијена успешност у препознавању од 87.50% (зависно од говорника) и 90.92% (независно од говорника). Урађена је и анализа успешности препознавања са препознавачем базираним на методи потпорних вектора (енгл. *Support Vector Machines*). Код препознавања зависно од говорника резултати SVM и HMM препознавача су упоредиви док је препознавање базирано на HMM алгоритму било знатно успешније код препознавања независно од говорника.

На крају је урађена анализа успешности HMM препознавача са мултимодном базом са обуку, која садржи изговоре и нормалног говора и шапата. Извршена је упоредна анализа успешности препознавача са измешаном базом за обуку (енгл. *Multi-Style Training*) и класификацијом говорног мода.

Кључне речи: Мултимодални говор, скривени Марковљеви модели, метода потпорних вектора, вектори обележја, класификација

Научна област: Електротехника

Ужа научна област: Техничка акустика

УДК: 621.3

THE RECOGNITION OF MULTIMODAL SPEECH BASED ON STATISTICAL APPROACH

SUMMARY

Whisper is a specific mode of speech communication characterized by an absence of glottal vibrations and noisy excitation of the vocal tract. Despite an increased effort in speech perception, the intelligibility of whisper in human communication is very high. An enormous acoustic mismatch between normal (normally phonated) and whispered speech is the main reason why modern speech recognizers have significant drop of performances when applied to whisper.

Statistical approach in automatic speech recognition, which is based on Hidden Markov Models (HMM), plays important role in state-of-the-art speech recognizers. The experiments conducted in this dissertation had shown the success in HMM recognition of multimodal speech, in both the speaker dependent (SD) and speaker independent (SI) cases. The experiments were done in matched ("normal/normal" and "whisper/whisper") and mismatched ("normal/whisper" and "whisper/normal") train/test scenarios. Because of the greatest importance in practical applications, special attention is paid in whispered speech recognition using models trained on normal speech ("normal/whisper" scenario). The speech database of isolated words in normal and whisper phonation was exploited.

Using MFCC (Mel-Frequency Cepstral Coefficients) feature vectors, three modeling units were examined in initial experiment: phonemes independent from context (monophones), phonemes dependent from context (triphones) and whole words. In whispered speech recognition using monophones was with the greatest success. In the following experiments, further improvement in recognition rate was achieved with dynamic feature vectors, pronunciation dictionary with 32 monophones and initialization using annotated part of speech database. The word recognition rate (WRR) of 81.38% (SD) and 87.42% (SI) was achieved. Using the new algorithm for cepstral coefficients based on modified frequency scale, the success in whisper

recognition was additionally improved without drop of performance in normal speech recognition. With proposed feature vectors, the WRR of 87.50% (SD) and 90.92% (SI) was obtained. The analysis in recognition was done using recognizer based on Support Vector Machines (SVM), as well. In the SD case, SVM and HMM performances were comparable, whereas HMM recognizer was with greater success in the SI case.

Finally, an analysis of HMM-based recognizer with multimodal training database (that contains utterances of both normal and whispered speech) was performed. A comparative analysis of speech recognizer performance based on multi-style training and classification of speech mode was conducted.

Keywords: Multimodal speech, Hidden Markov Models, Support Vector Machines, feature vectors, classification

Scientific area: Electrical Engineering

Scientific subarea: Technical acoustics

UDC number: 621.3

С А Д Р Ж А Ј

СПИСАК СЛИКА

СПИСАК ТАБЕЛА

1. УВОД

И поред значајног напретка информационо-комуникационих технологија у последње две деценије, говорна комуникација још увек недвосмислено представља најприроднији и најпогоднији начин комуникације између људи. Као таква, она ће заузимати значајно место у будућим интерфејсима између човека и машине. Према начину продуковања, говор се може класификовати у пет модова: шапат, тихи говор, говор уобичајеног интензитета (нормални говор), гласни говор и вика [1]. Шапат представља специфичан вид говорне комуникације који је у све чешћој употреби. Користи се у ситуацијама када је потребно остварити дискретну атмосферу у комуникацији (нпр. у позориштима, читаоницама, на пословним састанцима, итд.) или када се жели сакрити информација од других присутних особа. Такође, шапат може да буде артикулисан и као последица измењеног здравственог стања (ларингитис и ринитис), а није ретка употреба шапата у криминалним активностима.

Говорне технологије се могу поделити у две групе: аутоматско препознавање говора (енгл. *Automatic Speech Recognition - ASR*) и синтеза говора на основу текста (енгл. *Text to Speech - TTS*). Реч је о технологијама за чији развој је због комплексности проблема потребан мултидисциплинарни приступ у решавању.

Системи за аутоматско препознавање говора имају значајан низ предности у односу на остале интерфејсе човек-машина, које се превасходно огледају у следећем:

- руке и очи остају слободне за обављање других активности,
- микрофон има значајно мање габарите од нпр. тастатуре тако да је могуће коришћење у мањим уређајима (паметне наочари, паметни часовници, итд.),

- кориснике система није потребно дуго обучавати, и
- говор је најприроднији облик комуникације.

Ипак, и поред огромног напретка још увек имају низ недостатака који се највише огледају у значајно сниженим перформансама приликом различитих услова при обуци и тестирању (бучно окружење, различити дијалекти и начини говора говорника, различита емотивна стања говорника, итд.). Такође, на данашњем нивоу говорних технологија, није могућа примена са задовољавајућом тачношћу у ситуацијама где је потребно јако тихо саопштавање (често заступљено код заштите приватности корисника).

ASR системи су намењени препознавању типичног мода говора, а то је нормални говор. Из тог разлога перформансе система значајно опадају у препознавању било ког нетипичног мода говора. Од побројаних пет модова, највећа деградација перформанси система је управо при шапату, као говорном моду који се по својим акустичким карактеристикама највише дистанцира од преостала 4 мода.

Препознавање у ASR системима се може обављати зависно и независно од говорника. Уколико се параметри препознавача подешавају према говорнику који ће бити тестиран препознавање је зависно од говорника, док код препознавања независно од говорника изговори тестираног говорника не учествују у обуци система. Практичност употребе ASR система намеће природан услов препознавања независно од говорника. За исту или упоредиву величину базе за обуку, системи зависни од говорника имају знатно боље перформансе. Често се у системима независним од говорника ради постизања веће тачности врши адаптација на тестираног говорника, коришћењем одвојеног дела базе снимака дотичног говорника посебно намењених адаптацији. Показано је да системи са адаптацином имају боље перформансе како од препознавача зависних од говорника тако и од препознавача независних од говорника [2].

1.1.Мотивација рада, предмет и циљеви истраживања

Говорне технологије су много зависне од језика којем су намењене, те се као такве не могу једноставно увести из других земаља и применити на било ком језику. Српски језик спада у групу релативно малог броја светских језика за које

су постигнути задовољавајући резултати у препознавању говора који омогућују практичну примену [3].

Мотивација за истраживањима које обухвата ова дисертација је проистекла из нарастајуће потребе да се говорна комуникација човек-машина у ASR системима на српском језику подигне на виши ниво, који ће укључити и комуникацију са машином говором другачије артикулисаним од говора уобичајеног интензитета. За потребе истраживања формирана је прва верзија говорне базе *Whi-Spe* (енгл. ***Whi****spered* ***Spe****ech*) која је прва база бимодалне експресије говора (нормални говор - шапат) на српском језику [4]. Обим говорне базе је довољан за истраживачке активности на развоју препознавача како зависног тако и независног од говорника.

С обзиром да подаци у моду шапата нису доступни (бар у износу потребном за квалитетну обуку) посебан изазов представља препознавање шапата уколико је систем обучен на нормални говор (неусаглашени обука/тест сценарио). Циљ истраживања је оптимизација препознавача у препознавању изолованих речи у бимодалном говору, уколико за обуку нису на располагању снимци у говорном моду шапата. Узимајући у обзир чињеницу да је шапат у свакодневном говору значајно ређи од нормалног говора, циљ којем се тежи је уз услов да перформансе у препознавању нормалног говора не буду деградиране. Основни циљеви истраживања су:

- одређивање најпогодније јединице за моделовање у препознавању базираном на скривеним Марковљевим моделима (енгл. *Hidden Markov models - HMM*). Основне јединице за моделовање су фонем независан од контекста, фонем независан од контекста и цела реч.
- побољшање перформанси препознавача помоћу иницијализације параметара HMM модела лабелирањем дела базе за обуку;
- побољшање перформанси система у препознавању шапата новим алгоритмом за добијање вектора обележја базираним на модификованој фреквенцијској скали;
- упоредна анализа перформанси са препознавачем базираним на методи потпорних вектора (енгл. *Support Vector Machines - SVM*).

1.2.Кратак опис садржаја рада

У другој Глави је дата упоредна анализа акустичких карактеристика говорних модова са акцентом на сличности и разлике говора уобичајеног интензитета и шапата. Дат је преглед литературе у вези са истраживањима у препознавању бимодалног говора и опис говорних база коришћених у истим. Детаљно је описано снимање, обрада и контрола квалитета говорне базе *Whi-Spe*.

Трећа Глава описује методологију анализираних препознавача говора. Методологија заснована на HMM је у језгру свих савремених препознавача спонтаног говора управо из разлога што узима у обзир динамичку природу говорног сигнала за који су се статички класификатори показали неодговарајућим. Описани су основни проблеми у вези са евалуацијом, декодовањем и естимацијом параметара HMM модела, као и могућности њиховог решавања. Такође је описана и методологија занована на SVM препознавачу.

Опис експерименталне поставке и софтверских алата који су коришћени у обуци и тестирању су дати у четвртој Глави. У истраживањима при развоју HMM препознавача је коришћен конвенционални софтверски пакет у ASR системима HTK (енгл. *Hidden Markov model toolkit*). Имплементација софтвера самог препознавача је урађена у програмском пакету MATLAB. Глава садржи и опис новог алгоритма за генерисање кепстралних коефицијената који су показали већу робустност у препознавању шапата у односу на традиционална MFCC и PLP (енгл. *Perceptual Linear Prediction*) обележја.

У петој Глави су приказани резултати експеримената који анализирају избор оптималног броја мешавина као и јединице за моделовање у препознавању зависно и независно од говорника. Анализирана је могућност побољшања успешности у препознавању бимодалног говора при обуци само у моду нормалног говора помоћу лабелирања (мануелног и аутоматског) мањег дела говорне базе (10% говорне базе у нормалном изговору). На крају су дати резултати упоредне анализе HMM и SVM препознавача.

У шестој Глави је дата упоредна анализа препознавања бимодалног говора између технике са измешаном базом за обуку (енгл. *Multi-style training*) и препознавања са класификацијом говорног мода. Испитана је успешност класификације говорног мода као и раст перформанси у препознавању шапата у

зависности од процента додатих снимака у том моду. Дато је и поређење резултата са истраживањима у препознавању бимодалног говора за друге говорне базе.

Последње поглавље даје закључак, преглед и најзначајније доприносе дисертације са описом праваца будућих истраживања.

У прилогу су дати лексикон коришћене говорне базе и записи скрипт фајлова у *MATLAB*у који су коришћени у експериментима, као и резултати иницијалног експеримента.

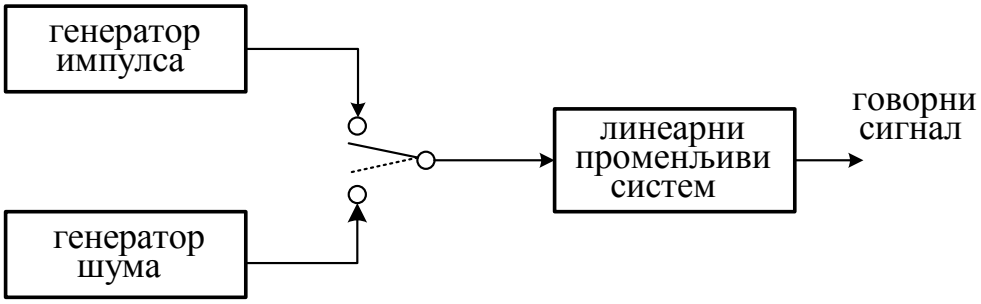
2. УВОД У ПРЕПОЗНАВАЊЕ МУЛТИМОДАЛНОГ ГОВОРА

2.1 Појам мултимодалног говора

Под уобичајеном говорном комуникацијом подразумева се комуникација две здраве особе нормалним интензитетом гласа и израженим емоцијама, без напора и стреса, и у амбијенту у којем бука не представља сметњу у комуникацији. У свим осталим случајевима, реч је о неуобичајеној или ређе коришћеној комуникацији [5].

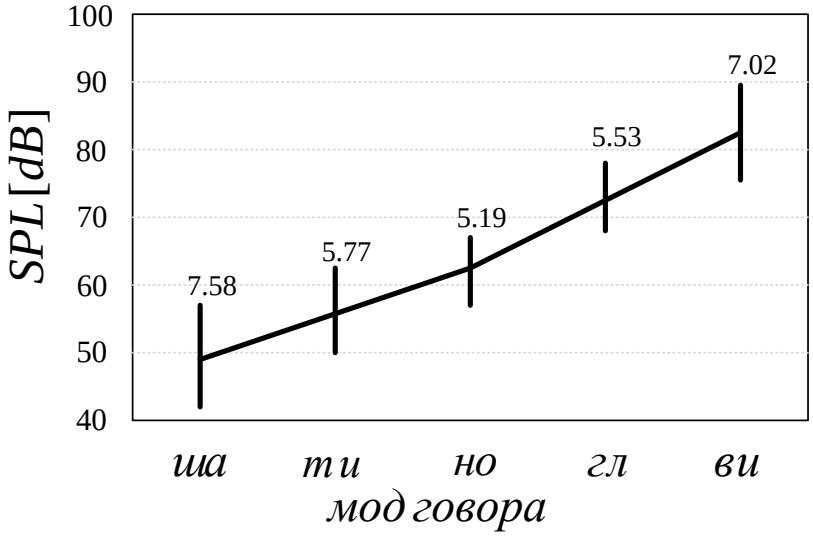
Говорна комуникација укључује два битна физикална процеса: генерисање говора и перцепцију говора. Процес генерисања говора започиње у централном нервном систему говорника, формулисањем поруке. Порука се преводи у неки језички кôд чиме се прелази из дискретног домена лингвистичких симбола у континуални домен покрета респираторних, фонаторских и артикулационих органа и акустичких појава у вокалном тракту говорника. У респираторне или дисајне органе спадају плућа са трахејом, мишићи грудног коша, дијафрагма и трбушни мишићи. У фонаторске органе спада ларинкс са гласним жицама док у артикулационе органе спадају ждријело, усна и носна дупља.

Процес перцепције се одвија у слушном механизму спектралном анализом акустичког таласа који изазива треперење базиларне мембране и стварање одговарајућих неуронских потенцијала. На слици 2.1 је приказан најчешће коришћен модел генерисања говорног сигнала [6]. Преклопник одређује тип екситације која је звучна (генератор глоталних импулса) или безвучна (генератор шума). Линеарни променљиви систем представља утицај вокалног тракта и зрачења на уснама.

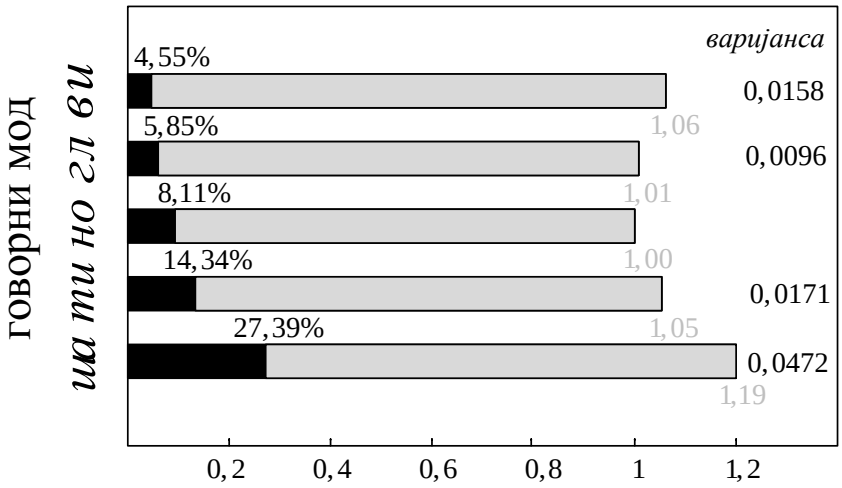


Слика 2.1 – Модел генерисања говорног сигнала

Према начину генерисања, говор се може поделити у пет модова: шапат, тихи говор, говор уобичајеног интензитета (нормални говор), гласни говор и вика. Основни параметри за дистинкцију који карактеришу поједине модове су ниво звука, трајање изговора и паузе и нагиб дуговременог спектра [1]. На слици 2.2 су приказани средња вредност и девијација нивоа звука (енгл. *Sound Pressure Level* - SPL) за 5 говорних модова, док су на слици 2.3 приказани средња вредност и варијанса трајања реченице (нормализована у односу на говор уобичајеног интензитета) и проценат заступљености тишине. У дотичном истраживању је учествовало 12 говорника енглеског језика (којима је то матерњи језик) снимањем са 3 микрофона: блиски, удаљени и микрофон на врату. Поред говорног сигнала, ради калибарције је снимљен и тест тон фреквенције 1 kHz и нивоа 75 dB.



Слика 2.2 – Ниво звука за пет говорних модова (ша - шапат; ти - тихи говор; но - нормални говор; гл - гласни говор; ви - вика). Вертикалне линије представљају стандардну девијацију нивоа звука. Подаци су преузети из [1].



Слика 2.3 – Средња вредност трајања реченице за пет говорних модова (ша - шапат; ти - тихи говор; но - нормални говор; гл - гласни говор; ви - вика). Трајање реченице је нормализовано у односу на нормални говор. Тамни део представља проценат застпљености тишине. Подаци су преузети из [1].

Као што се види на слици 2.2 средња вредност нивоа звука се мења од 50 dB (за шапат) до 83 dB (за вику). Веће дистанцирање од говора уобичајеног интензитета доприноси и већој девијацији нивоа звука, при чему је највећа за шапат (7,58 dB). Такође, трајање реченице је највише повећано за шапат (19%) као и заступљеност тишине у изговору (27,39% у односу на нормални говор са 8,11%). За шапат је добијена и највећа вредност варијансе трајања реченице. Исто истраживање је испитивало и просечни нагиб дуговременог спектра (енгл. *spectral tilt*). Резултати су приказани у табели 2.1 [1]. Најмањи нагиб (по апсолутној вредности), односно најравнији спектар, са -2,86 dB/октави је добијен за шапат.

ТАБЕЛА 2.1: ПРОСЕЧНИ НАГИБ ДУГОВРЕМЕНОГ СПЕКТРА (ИЗРАЖЕН У ДЕЦИБЕЛИМА ПО ОКТАВИ)

Говорни мод	Просечни нагиб
Шапат	-2,86
Тихи говор	-6,71
Нормални говор	-8,29
Гласни говор	-8,27
Вика	-7,51

Резултати приказани у истраживању показују да је по питању основних параметара за класификацију говорних модова највећа дистанца у односу на нормални говор управо за шапат. Због побројаних акустичких карактеристика

препознавање говора у моду шапата представља велики изазов за модерне ASR системе.

Поред побројаних параметара, велику тачност у класификацији тихог, нормалног и гласног говора показали су [7]:

- однос амплитуда консонаната и вокала,
- однос амплитуда консонаната и полувокала, и
- однос амплитуда вокала и полувокала.

2.2 Препознавање шапата

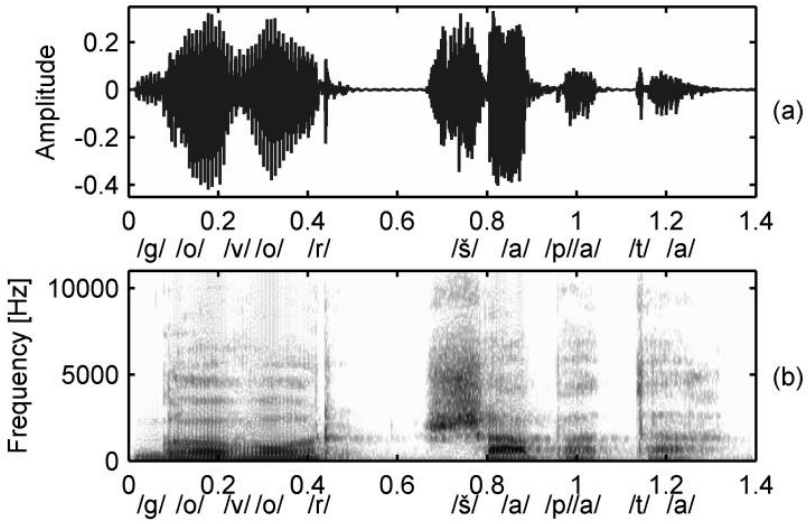
Основна карактеристика шаптавог говора је одсуство глоталних вибрација, односно шумна побуда вокалног тракта. Извор звука код генерисања шапата је апериодичан, турбулентан проток ваздушне струје између гласница које не вибрирају [8]. Мјерење тродимензионалног облика вокалног тракта помоћу магнетне резонанце је показало сужење дела вокалног тракта око вентрикуларних набора, што резултује знатно слабијој акустичкој спрези између субглоталног и супраглоталног система [9].

Неколико мултидисциплинарних истраживачких студија бавило се истраживањем шапата и то у физици за одређивање аеродинамике вокалног тракта при шапату [10] и у медицини за испитивање активности мозга код афоничних особа [11]. Такође, анализирана је могућност реконструкције нормалног говора из шапата [12]. Одређено је да су формантне учестаности у шапату померене навише [13,14]. У експериментима чији је био циљ анализа помераја формантних учестаности у шапату у односу на нормални говор, за изговоре 5 вокала у српском језику (са 10 говорника) добијени су помераји прва 2 форманта навише за вокале /и/, /а/, /е/ и /о/ и наниже за вокал /у/ [13]. Субјективни утисак основне фреквенције (енгл. *pitch*) код вокала у шапату врло је близак другом форманту [15].

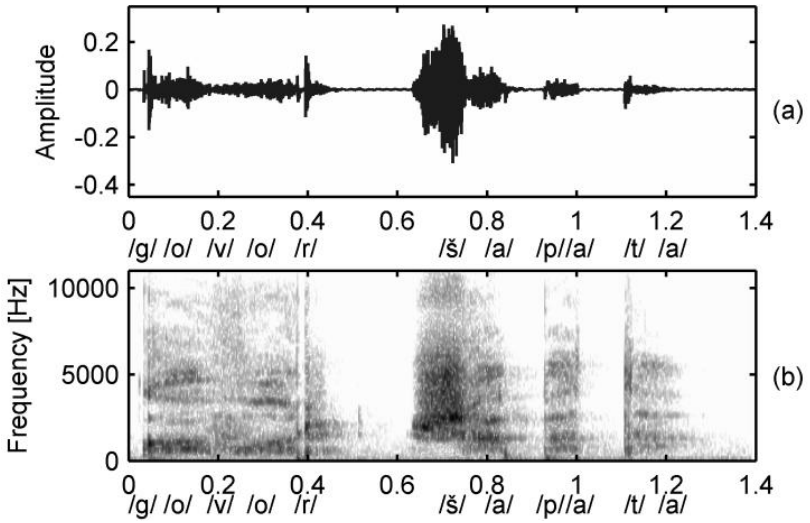
Ипак, и поред нешто отежане перцепције, разумљивост шапата у комуникацији између људи је веома висока. Препознавање фонема у логатомима типа консонант-вокал-консонант у српском језику је преко 80% код свих вокала и преко 2/3 консонаната [16], што је веома добра логатомска разумљивост, којој одговара реченичка разумљивост од преко 95%. Код вокала, минималну разумљивост око 88% има вокал /о/, а највећу (99%) имају вокали /а/ и /и/.

Разумљивост консонаната зависи од фонетске групе и највећа конфузија или потпуна замена је у оквиру истих фонетских група и то на нивоу парова звучни/безвучни (б/п, д/т, з/с, ж/ш). Разумљивост вокала у речима /hVd/ (V означава вокал) којих има 10 у енглеском језику је са процентом идентификације од 82% [17]. Такође, могуће је чак са значајним успехом идентификовати одређена емотивна стања (љутња, туга и страх) у шаптавом говору, што је било у фокусу истраживања за мандаринski језик [18]. С друге стране, у шапату је отежано препознавање нелингвистичких информација попут пôла говорника, старосне доби и идентитета.

На сликама 2.4 и 2.5 приказани су временски дијаграм и спектрограм кратке реченице "Говор шапата." изговорене говором уобичајеног интензитета и шапатом од стране истог говорника, респективно. Графици су подржани са фонетском транскрипцијом. Као што се види са слика, одређени делови фреквенцијског спектра су добро очувани у шапату. То је посебно изражено за безвучне консонанте, као што су фрикатив /ш/ и плозиви /п/ и /т/. Спектрограм вибранта /р/ такође има сличан фреквенцијски садржај у оба мода говора. Увидом у таласни облик и спектрограм се може приметити потпуни губитак хармонијске структуре код вокала у шапату.



Слика 2.4 – Таласни облик (а) и спектрограм (b) кратке реченице "Говор шапата." изговорене нормалним говором. На апсциси је време у секундама



Слика 2.5 – Таласни облик (а) и спектрограм (b) кратке реченице "Говор шапата." изговорене шапатом. На апсциси је време у секундама

Због побројаних значајних разлика између шапата и нормалног говора перформансе ASR система, који су превасходно намењени препознавању нормалног говора, значајно опадају у препознавању шапата. У распону говора од шапата до вике највећи негативни утицај на перформансе система за аутоматско препознавање говорника има управо шапат [1]. С обзиром да подаци у моду шапата најчешће нису доступни (бар у већој количини), највећи изазов представља препознавање шаптавог говора/говорника са обуком на говору уобичајеног интензитета. У вези са тим, неколико истраживачких студија се фокусирало на компензацију разлика између нормалног и шаптавог говора. У истраживању [19] постигнута је практична употреба шапата у ASR систему за препознавање говора независног од говорника коришћењем релативно мале количине шаптавог говора по говорнику за адаптацију (10 фонетски балансираних реченица). Адаптацијом која је заснована на трансформацији вектора средње вредности и коваријансне матрице (описани у Глави 3) је постигнуто побољшање препознавања шапата у апсолутном износу од чак 10% (са почетних 68% на 78%).

Методе за компензацију базиране на усклађивању спектра и мапирању обележја су представљени у [20] и [21], респективно. Значајно побољшање успешности препознавања је постигнуто редистрибуцијом банке филтара и смањењем димензије кепстралних обележја [22]. Перформансе система су побољшане и нормализацијом дужине вокалног тракта и транслацијом спектра за

генерисање тзв. псеудо-шапата који се користи за адаптацију [23], за ограничени речник од 160 речи.

С обзиром да је у континуалном говору уобичајеног интензитета понекад присутан и шапат (честа појава код саопштавања поверљивих информација) значајно је постићи прецизну детекцију сегмената шапата. У вези са тим, постигнута је детекција сегмената шапата у износу од 97,37% базирана на остатку линеарне предикције [24] и 96,85% базирана на ентропији спектра по фреквенцијским опсезима [25]. Оцена за детекцију је укључивала 2 типа грешке у класификацији говорног мода (промашена детекција и лажни аларм) и нормализовану грешку у временској детекцији.

Скорашња истраживања су демонстрирала могућност примене инверзног филтрирања у препознавању шапата са обуком на говору уобичајеног интензитета применом вештачких неуронских мрежа у комбинацији са HMM. Прије обуке модела врши се филтрирање говорног сигнала чиме се постиже уједначавање спектра нормалног говора и шапата. У препознавању зависно од говорника постигнуто је побољшање успешности у износу од 31% [26].

Технологије које су још коришћене за аутоматско препознавање бимодалног говора су:

- коришћење камере за добијање видео-обележја која се користе упоредо са аудио-обележјима [27],
- коришћење микрофонских низова [28],
- коришћење микрофона на врату (енгл. *throat microphone*) [29],
- примена вештачких неуронских мрежа [30] и
- примена DTW алгоритма [31].

Ипак, већина савремених ASR система је заснована на статистичком моделу говора, односно коришћењу скривених Марковљевих модела у комбинацији са мешавинама Гаусових расподјела [32] или дубоким неуронским мрежама [26].

2.3 Говорне базе коришћене за препознавање бимодалног говора

Главни разлог за изузетно ретку примену шапата у модерним ASR системима независним од говорника који имају задовољавајућу тачност је недостатак велике и систематски креиране говорне базе. Српски језик спада у јако малу групу светских језика за које су почела иницијална истраживања у препознавању

шапата, како са перцептивног аспекта тако и са аспекта коришћења у ASR системима. Преглед досад креираних дигитализованих говорних база у моду шапата је приказан у табели 2.2, сортираној према укупном броју говорника. Све говорне базе садрже изговоре нормалног говора и шапата и доступне су без накнаде за истраживање.

ТАБЕЛА 2.2: ПРЕГЛЕД ГОВОРНИХ БАЗА КОЈЕ САДРЖЕ СНИМКЕ ИЗГОВОРА У МОДУ ШАПАТА

Р.Б.	Назив	Број говорника	Величина [часова]	Фонетски балансирана	Језик
1.	CIAIR[33]	68М 55Ж	15	да	јапански
2.	iWhisper-Mandarin [34]	40М 40Ж	15	да	мандарински
3.	UTVE-II [23]	37М 35Ж	< 1	да	енглески
4.	wTIMIT [35]	25М 23Ж	15	да	енглески
5.	AV-Whisper [36]	20М 20Ж	< 10	да	енглески
6.	CHAINS [37]	18М 18Ж	< 3	да	енглески
7.	UTVE-I [25]	12М	< 1	да	енглески
8.	Whi-Spe [4]	5М 5Ж	2	да	српски

Као што се види из табеле 2.2, већина говорних база је на енглеском језику. За јапански језик је креирана говорна база *CIAIR* која садржи изговоре од укупно 123 говорника. Сваки говорник је изговарао 60 фонетски балансираних реченица и 50 реченица из новинских чланака. Снимање је вршено са 2 микрофона и 3 начина држања мобилног телефона (стандардни начин, са покривеним устима и са покривеним устима и доњим делом мобилног телефона).

За стандардни мандарински језик креирана је почетна варијанта говорне базе *iWhisper-Mandarin* која садржи изговоре 100 фонетски балансираних реченица од стране 80 говорника. Снимање је вршено у тихој соби *USB* микрофоном.

За енглески језик је креирано 5 говорних база под називима: *UTVE-II*, *wTIMIT*, *CHAINS*, *UTVE-I* и *AV-Whisper*.

Говорне базе *UTVE* (*I* и *II*) су досад најобимније и најсвеобухватније базе које садрже изговоре у моду шапата. База *UTVE-I* садржи изговоре 12 говорника у свих 5 говорних модова. Сваки говорник је шаптао 5 реченица из TIMIT говорне базе, као и једну минуту спонтаног говора. С друге стране, говорна база *UTVE-II* садржи далеко већи број говорника који су спонтано изговарали реченице нормалним говором са шаптавим сегментима у знатно краћем трајању.

Говорна база *CHAINS* (енгл. *Characterizing Individual Speakers*) је снимљена за потребе истраживања у вези са утицајем различитих начина изговора на систем за

аутоматско препознавање говорника. Има 36 говорника (по 18 оба пола) од којих је већина (28) из Ирске. Постоји 6 начина изговора:

1. соло говор - говорници читају текст умереним темпом
2. препричавање - говорници својим речима препричавају бајку
3. синхрони говор - пар говорника симултано и усаглашено читају текст
4. имитирани говор - говорници покушавају имитирати одабрани одељак из бајке
5. брзи говор - говорници читају цели текст брзим темпом
6. шаптави говор - говорници шапућу цели текст.

Аудио-визуелна говорна база *AV-Whisper* је поред микрофона снимана и видео-камером за добијање видео обележја. Њено креирање је обављано у лабораторијским условима при чему је звук сниман микрофоном и дигитализован фреквенцијом одмеравања 48 kHz. Говорници су у 3 одвојене сесије снимања изговарали читани текст, изоловане цифре и спонтани говор, у нормалном моду и шапату. За добијање видео обележја коришћене су камере високе дефиниције, по једна за фронтални и бочни поглед главе. Аудио снимци су сегментирани и аутоматски лабелирани.

Говорна база *wTIMIT* са 48 говорника (оба пôла) је у моду шапата и комплементна је *TIMIT* говорној бази у нормалном моду. Садржи изговоре 450 фонетски балансираних реченица из *TIMIT* говорне базе на 2 наречја енглеског језика: сингапурски и амерички. Снимања су вршена усмереним кондензаторским микрофоном.

2.4 Говорна база Whi-Spe

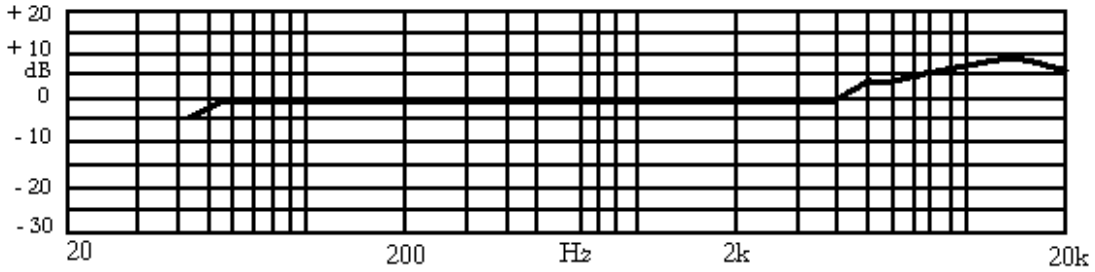
За потребе обуке и тестирања система за препознавање бимодалног говора на српском језику неопходно је снимити адекватну говорну базу. За потребе истраживања чији је предмет препознавање нормалног говора и шапата помоћу вештачких неуронских мрежа (ANN - *Artificial Neural Networks*), динамичког усклађивања времена (DTW - *Dinamic Time Warping*), скривених Марковљевих модела и методе потпорних вектора, креирана је почетна верзија говорне базе бимодалног говора, која је означена *Whi-Spe* (скраћеница од **Whi**spered **Spe**ech).

База садржи два дела: први део садржи изговоре шаптавог говора, док други део садржи изговоре нормалног говора. База садржи изговоре 10 говорника (5

женских и 5 мушких), а речник садржи 50 речи: 30 фонетски балансираних речи, 14 бројева и 6 боја. Балансиране речи су преузете из говорне базе емоција, експресија и ставова (ГЕЕС) [38] те задовољавају основне језичке критеријуме српског језика (расподела фонема, слоговна композиција, акценатска структура и консонантски скупови). Све речи говорници су изговарали у континуитету 10 пута, са паузом од неколико дана између суседних сесија. На крају, база садржи 10000 изговора (10 говорника x 50 речи x 10 изговора x 2 мода). Речи говорне базе су дате у Прилогу А1. Говорници су били студенти волонтери без говорних мана и са коректним слухом. База је снимана и обрађивана једну годину и доступна је за истраживања од стране других.

2.4.1 Снимање и обрада говорне базе

База је снимљена у лабораторијским условима у Високој школи техничких струковних студија у Чачку, квалитетним омнидирекционим микрофоном чија је фреквенцијска карактеристика приказана на слици 2.6.



Слика 2.6 – Фреквенцијска карактеристика микрофона коришћеног за снимање

Код нормалног говора микрофон је био удаљен око 25 cm од уста говорника (укосо постављен ради избегавања директног удувавања) док је код шапата био око 5 cm, како би снимци у шапату били што је могуће бољи. Снимци су у формату: моноканални PCM (енгл. *Pulse Code Modulation*), фреквенција одмеравања 22050 Hz и 16 бита по одмерку. Како би се обезбедио довољан број снимака са задовољавајућим квалитетом, организовано је више од 10 сесија снимања по говорнику. Током сесије, говорници су изговарали све речи прво нормалним говором, а потом и шапатом са паузом између речи од бар једне секунде. Прије снимања говорници су извежбали начин изговора речи у шапату, како исти не би био ни сувише тих ни пренаглашен. На крају је додата и једна реч

која није у речнику како не би дошло до пада интонационе контуре. Снимци су мануелно сегментирани у софтверском пакету *Adobe Audition* и успостављена је конзистентна ознака фајлова на следећи начин:

[група речи][редни број речи]_[редни број говорника]_[редни број изговора][мод].wav

При том вреди:

група речи $\in$ {boja, broj, rec}

редни број речи $\in \begin{cases} 1:30; \textit{ за балансиране речи} \\ 1:14; \textit{ за речи бројева} \\ 1:6; \textit{ за речи боја} \end{cases}$

редни број говорника $\in \mathbf{1}:\mathbf{10}$

редни број изговора $\in \mathbf{1}:\mathbf{10}$

мод $\in$ {n, s} n - нормални говор s - шапат

На пример, назив **boja3_2_7s.wav** означава снимак седмог изговора другог говорника у шапату, за трећу реч из скупа боја.

Након сегментације и означавања контролу квалитета снимака је радио фонетски експерт. За замену снимака лошег квалитета коришћени су резервни снимци.

2.4.2 Контрола квалитета снимака

И поред одређеног броја резервних снимака база је доснимавана у више наврата јер су контролама квалитета установљене различите грешке, субјективне и објективне природе [4]. Код нормалног говора најчешће је долазило до погрешног изговора дате речи или погрешне артикулације појединог гласа у тренутку изговора речи (што је честа појава и у свакодневној говорној комуникацији). Код снимања је понекад био изражен ефекат акустичког удара у микрофон.

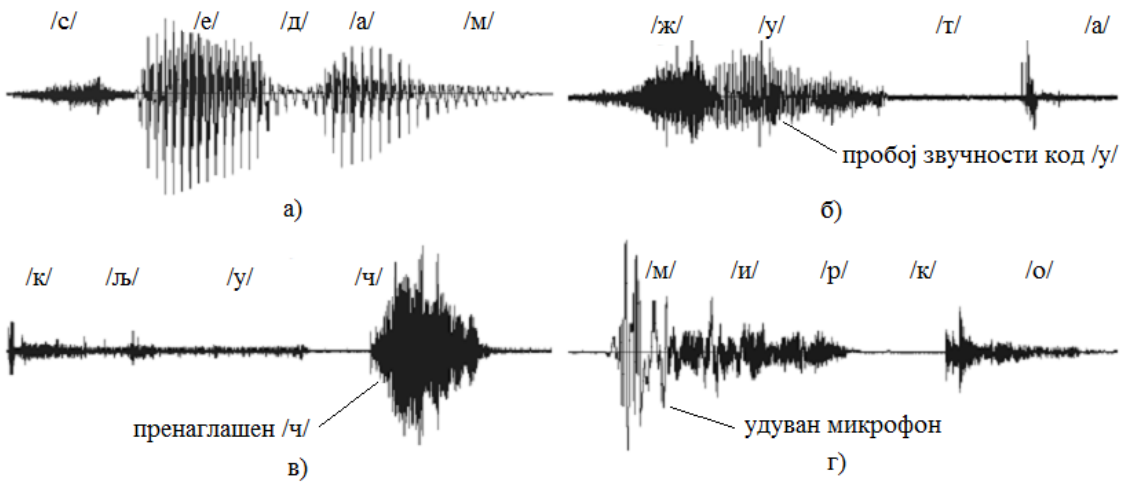
Ипак, највећи број лоших снимака се односио на речи изговорене шапатом. С обзиром да је циљ био добијање правилно изговорених речи шапатом најтипичније грешке код изговора и снимања шапата су:

1. сувише тих изговор шапата (значајно маскиран шумом),
2. сувише наглашен (исфорсиран до изобличења),
3. пробијање звучности код изговора шапата,
4. омисија гласова,
5. директно дување у микрофон (акустички удар),

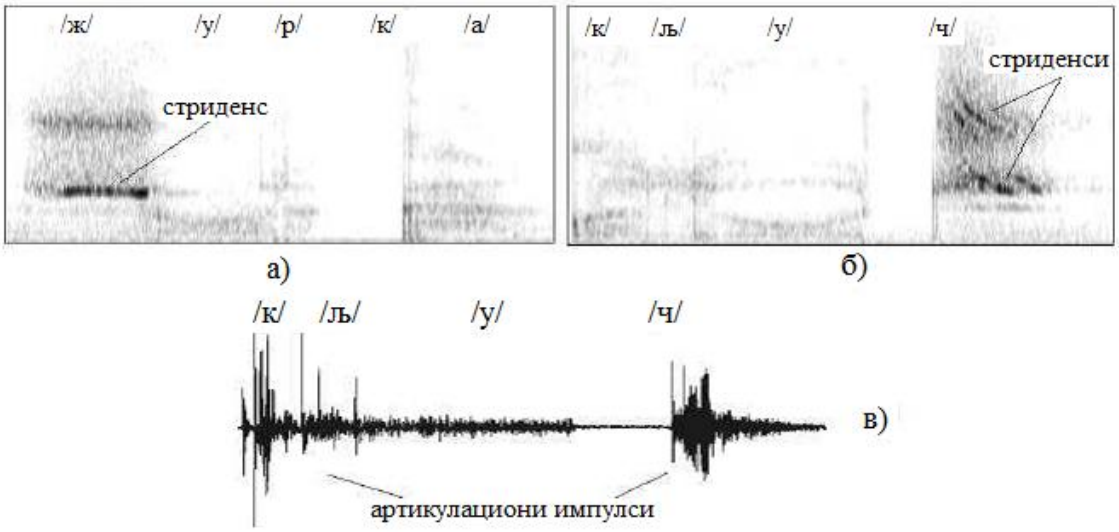
6. нерегуларност рада артикулатора (појава стриденса, тренутак одлепљивања језика од непца и сл.).

На слици 2.7 је приказан пример таласног облика нормално (звучно) изговорене речи и таласни облици снимака са типичним грешкама при изговору шапата.

Стриденс представља једну од атипчних релизација гласова српског језика која се првенствено јавља код изговора фрикатива и африката. У домену перцепције доживљава се као непријатан звук сличан звиждуку или као пискави или храпав звук који се јавља упоредо са нормалним изговором, односно са нормалном фрикцијом [39]. Спектрограми неправилних изговора са стриденсом при изговору речи које садрже фрикатив /ж/ и африкат /ч/ су приказани на слици 2.8(а) и 2.8(б), респективно. Са слика се могу јасно уочити хармонијски фреквенцијски концентрати енергије. На слици 2.8(в) је приказан таласни облик неправилног изговора ријечи /кључ/ у шапату са израженим артикулационим импулсима при одвајању језика од непца.



Слика 2.7 – Таласни облик: а) правилно изговорене речи (нормалним говором) и таласни облици лоших снимака речи изговорених шапатом



Слика 2.8 – Спектрограми изговора са стиденсом код а) фрикатива; б) африката и в) таласни облик неправилног изговора са импулсима при додиру језика и непца

2.5 Резиме

Интеракција између човека и машине путем говора има низ предности у односу на друге видове интеркација. Често се јавља потреба да се човек обрати машини говором другачије артикулисаним од говора уобичајеног интензитета (нормалног говора). Термин мултимодални говор обухвата следеће начине продуковања говора: шапат, тихи говор, нормални говор, гласни говор и вика. Шапат представља специфичан начин комуникације који је у све чешћој употреби, окарактерисан одсуством ларингеалних вибрација. И поред већег когнитивног напора при перцепцији, одликује га изузетно велика разумљивост.

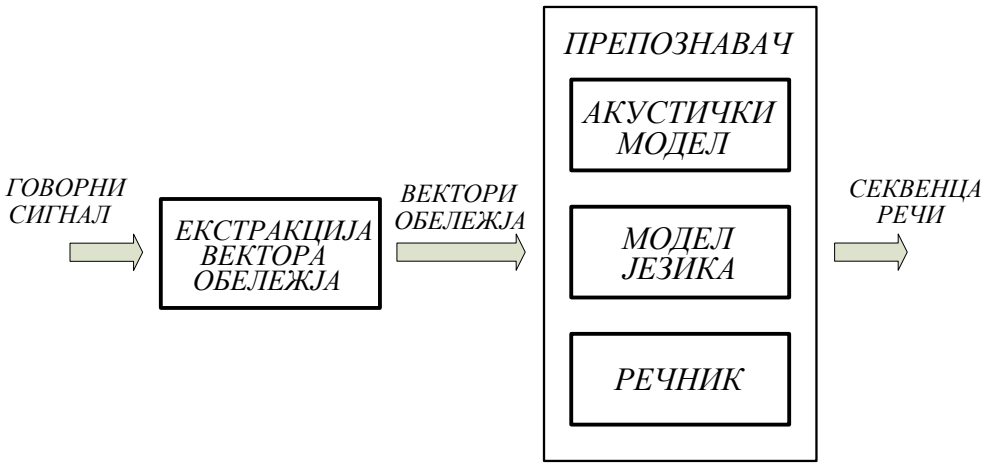
Компромис између тачности и робустности представља највећи изазов у савременим ASR системима. Због тога је проблем аутоматског препознавања мултимодалног говора изузетно сложен задатак. За сада не постоје софтверски алати који би са задовољавајућом тачношћу препознавали шаптави говор независно од говорника, са великим фондом речи. Главни разлог за то је непостојање велике, систематски обрађене говорне базе која садржи изговоре у различитим модовима говора. С обзиром да су говорне технологије изузетно зависне од језика за који су намењене, за потребе истраживања и вези са мултимодалним говором креирана је говорна база Whi-Spe. Садржи изговоре бимодалне експресије говора (нормалан говор - шапат) и прва је такве врсте на српском језику. У почетној форми је обима 10 говорника (5 женских и 5 мушких) са изговорима 50 изолованих речи (речи боја, бројева и фонетски балансиране речи). Укупно трајање почетне форме говорне базе је 2 сата. У овој глави су описани снимање, контрола квалитета и сегментација изговора, као и потешкоће током датих активности. Такође су описане најпознатије говорне базе мултимодалног говора које су доступне за истраживање, за друге светске језике.

Извесно је даље проширење говорне базе Whi-Spe и доступна је за коришћењење од стране других истраживача.

3. МЕТОДОЛОГИЈА ПРЕПОЗНАВАЧА ГОВОРА

3.1 Аутоматско препознавање говора

Приликом двосмерне комуникације са другом особом, човек има способност да поред садржаја речи са великим успехом препознаје и низ нелингвистичких информација као што су пол, емотивно стање, идентитет, итд. Задатак система за аутоматско препознавање говора (енгл. *Automatic Speech Recognition*) је да из говорног сигнала издвоји секвенцу речи која је изговорена. На слици је приказана општа блок шема ASR система базираног на HMM алгоритму.



Слика 3.1 – Блок шема ASR система

Пре свега, у блоку за издвајање обележја говорни сигнал се трансформише у низ вектора обележја. Задатак блока за екстракцију вектора обележја је да елиминише разне варијације узроковане варијацијама говорника, амбијента или

канала.

Задатак препознавача је да пронађе секвенцу речи која (по неком критеријуму) најбоље одговара оном што је изговорено. Са статистичког становишта, препознавач проналази низ од M речи $\hat{W}=w_1,w_2,...,w_M$ која максимизује апостериорну вероватноћу $P(W/X)$, при чему је $\hat{X}=x_1,x_2,...,x_T$ низ од T вектора обележја [40]. На основу Бајесовог правила је:

$$\hat{W}=\arg\max_W P(W/X)=\arg\max_W \frac{P(W)P(X/W)}{P(X)} \tag{3.1}$$

Пошто је $P(X)$ независно од W, следи:

$$\begin{aligned}\hat{W}&=\arg\max_W P(W/X)=\arg\max_W [P(W)P(X/W)]\\&=\arg\max_W [\ln P(X/W)]+\ln[P(W)]\end{aligned} \tag{3.2}$$

Релација 3.2 у формалном облику преставља задатак препознавача говора. Први сабирак представља акустички модел, док други сабирак представља модел језика.

- Акустички модел описује са статистичког аспекта понашање говора у простору вектора обележја. Акустички модел садржи сет скривених Марковљевих модела који репрезентује сваку јединицу за моделовање.
- Модел језика описује везу између речи узимајући у обзир граматику језика за који је препознавач намењен. На пример, у српском језику након речи "с обзиром" често следе "да" или "на" (као што у енглеском језику иза "in order" често следи "to"). За препознаваче скупа речи из великог речника (енгл. Large Vocabulary Continuous Speech Recognition - LVCSR) модел језика користи тзв. *n-gram* приступ, код којег се узимају у обзир условне вероватноће претходних *n-1* речи (често се користе триграми, односно *n=3*).

Акустичко моделовање и моделовање језика су раздвојене целине које се морају обучити пре коришћења. У овој дисертацији није анализиран модел језика, јер се ради о препознавању изолованих речи из ограниченог скупа.

Ако корисник има намеру да препознавач препознаје одређену реченицу,

потребно је дефинисати речи које су у истој садржане. За ASR систем речник (енг. *dictionary*) садржи лексикон и фонетску транскрипцију изговора за сваку поједину реч (енг. *pronunciation*). Пример речника базираног на монофонима дат је у табели 3.1. Са Y је означен фонем шва (тзв. мукло а).

ТАБЕЛА 3.1: ПРИМЕР РЕЧНИКА СА ТРИ РЕЧИ И ОДГОВАРАЈУЋОМ ТРАНСКРИПЦИЈОМ

Реч	Фонетска транскрипција
BELA	B E L A
MIRKO	M I R Y K O
BRATSKI	B R A C K I

У ASR системима основне јединице за моделовање су фонеми. Са фонетског становишта говорне комуникације, фонем представља најмању акустичку јединицу коју човек може да перцепира [5]. На пример, постоји фонетска разлика у речима "завет" и "савет" која их чини различитим (ако се изговарају нормалним говором). У препознавању независно од контекста, сваки фонем се независно моделује. Таква јединица за моделовање се назива монофон.

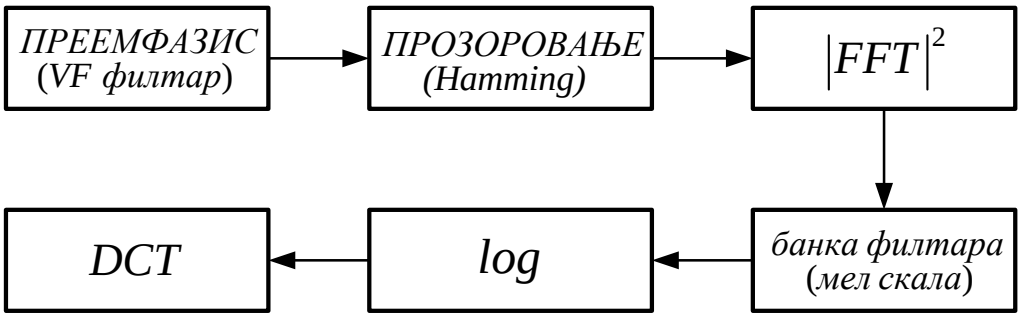
Са акустичког аспекта, услед коартикулације, изговор одређеног фонема у многоме зависи од суседних фонема. На пример, изговор фонема /j/ звучи различито у изговору имена "Јанко" и "Благоје". Стога коришћење фонема независних од контекста не може да обухвати све акустичке варијације фонема. Идеја моделовања фонема зависних од контекста је да се користи много већи број јединица за моделовање, који узимају у обзир и суседне фонеме. Обично се користе трифони[1] који су развијени из монофона на начин да се узима у обзир претходни и следећи фонем. На тај начин се добијају модели који су више специјализовани. Уколико је N број монофона, број трифона је N^3, што даје неколико десетака хиљада трифона (29791 за 31 монофон). Пошто трифони нису равномерно заступљени ређи трифони немају довољну количину података за обуку. Потреба за огромном базом за обуку је највећи недостатак коришћења трифона. Развијени су алгоритми који у одређеној мери превазилазе овај недостатак [41].

[1] Могу се користити и бифони (2 суседна фонема) али се користе много ређе

3.2 Екстракција вектора обележја

Препознавање говора се дели у две фазе: прва фаза представља тзв. обуку система, док друга фаза садржи тестирање (препознавање говора).

За обуку и тестирање заједничка фаза је издвајање вектора обележја. С обзиром да је информација о изговореном фонему садржана у обвојници амплитудског спектра намена вектора обележја је да описује исту. Пошто говор није стационаран случајни сигнал, он се дели на мање сегменте (фрејмове) у којима се може сматрати квазистационарним. Трајање фрејма је типично 20 – 30 ms. Да би се смањило "цурење спектра" користе се прозорске функције које значајно потискују тзв. бочне лобове (бар 40 dB), а код говорног сигнала користи се најчешће *Hamming*ов прозор. Фрејмови су у одређеном износу преклопљени како би се испратиле промене у говорном сигналу. У препознавању говора најчешће коришћена обележја су мел-фреквенцијски кепстрални коефицијенти (енгл. *Mel-frequency cepstral coefficients* - MFCC). Блок шема за екстракцију MFCC вектора обележја је приказана на слици 3.2.



Слика 3.2 – Блок шема за екстракцију MFCC вектора обележја

Први корак обично подразумева преемфазис, односно слабљење компоненти говорног сигнала на ниским и појачање на високим учестаностима (ради елиминисања ефекта зрачења на уснама). Реч је о FIR (енгл. *Finite impulse response*) филтру са преносном функцијом облика $F(z)=1-a\,z^{-1}$, при чему је параметар a између 0,95 и 0,99 (типично 0,97). Након прозоровања, односно поделе сигнала на фрејмове пондерисане *Hamming*овом прозорском функцијом амплитудски спектар (или спектар снаге) се добија брзом Фуријеовом трансформацијом (енгл. *Fast Fourier Transformation* - *FFT*). Потом се спектар

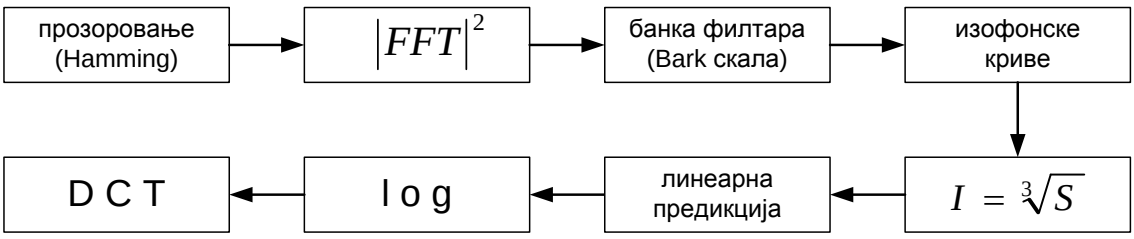
пропушта кроз банку филтара који описују рад базиларне мембране чији су филтри распоређени равномерно на мел-фреквенцијској скали. Веза између фреквенције у мелима и фреквенције у херцима је дата изразом 3.3.

$$f\left[mel\right] = 2595\log\left(1+\frac{f\left[Hz\right]}{700}\right) \tag{3.3}$$

Амплитудска карактеристика је скалирана како снага на излазу филтара не би зависила од броја тачака FFT. Уобичајена је троугаона карактеристика мада може бити и трапезна, у облику "подигнутог косинуса" или Гаусове криве [42]. Број филтара зависи од намењеног фреквенцијског опсега и обично је од 24 до 40.

Последња 2 блока представљају рачунање кепстра. Кепстар је трансформација која конволуцију пресликава у збир, а то је инверзна Фуријеова трансформација логаритма спектра снаге. Пошто је говорни сигнал реалан, амплитудски спектар је парна функција те се инверзна Фуријеова трансформација може заменити дискретном косинусном трансформацијом (енгл. *Discreat Cosine Transform* - DCT). Кепстром се у ствари настоји раздвојити побудни сигнал из глотиса и импулсни одзив вокалног тракта који садржи информацију шта је изговорено. Обично се користи 12 до 16 коефицијената, узимајући у обзир и нулти коефицијент који представља енергију.

Поред MFCC коефицијената у последње време се често користе и перцептивни линеарни предиктивни коефицијенти (PLP) јер су се у многим практичним применама показали робуснији од MFCC коефицијената. Блок шема за добијање PLP коефицијената је приказана на слици 3.3 [43].



Слика 3.3 – Блок шема за екстракцију PLP вектора обележја

Као што се види са слике, на почетку се рачуна спектар снаге по фрејмовима над којима је извршено *Hamming*ово прозоровање. Након тога се спектар

пропушта кроз *Bark* банку филтара код којих је пресликавање фреквенције на *Bark* скалу дато изразом 3.4.

$$f\left[Bark\right]=6\ln\left(\frac{f}{600}+\sqrt{\left(\frac{f}{600}\right)^{2}+1}\right) \tag{3.4}$$

Да би се симулирала осјетљивост човековог чула слуха *Hermanski* [44] је предложио модификацију спектра скалирањем излаза банке филтара помоћу кривих једнаке гласности (изофонске криве) и рачунање спектра помоћу *Stivens*овог закона (интензитет је једнак кубном корену гласности, $I=\sqrt[3]{S}$). На крају, PLP коефицијенти се добијају рачунањем кепстра предиктивних коефицијената модификованог спектра.

Моћна техника за повећање робустности система је нормализација средњом вредношћу кепстра (енгл. *Cepstral Mean Subtraction*). С обзиром на особине логаритамске функције (пресликавање производа у збир) одузимањем средње вредности кепстра постиже се у значајној мери раздвајање екситације и преносне функције вокалног тракта. Техника се показала изузетно плодоносна и у препознавању шапата [45], и дата је релацијом 3.5. Недостатак је што је потребно временско усредњавање па је јако ограничена примена у препознавању у реалном времену (енгл. *Real-Time*).

$$\vec{x}_{CMS}=\vec{x}-\frac{1}{N}\sum_{i=1}^{N}x_i\ ;\ \vec{x}=\left[x_1\,x_2\cdots x_N\right]' \tag{3.5}$$

У релацији 4.3, $\vec{x}$ је вектор статичких MFCC коефицијената димензије $N=13$.

Увођењем динамичких обележја боље се прате временске промене карактеристика говорног сигнала и постиже слабија корелација између суседних фрејмова. Тиме се постиже већа независност и оправдава коришћење дијагоналне коваријансне матрице код моделовања вишедимензионалне Гаусове расподеле. Најчешће се користе динамичка обележја која садрже статичка обележја и њихов први извод (делта обележја) и динамичка обележја која садрже статичка обележја заједно са њиховим првим и другим изводом (делта-делта обележја).

Делта обележја се добијају применом обрасца 3.6 [46]:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{3.6}$$

При том је d_t делта обележје у тренутку t, а параметар Θ означава померај фрејмова унапред и уназад и подразумевана вредност износи $\Theta = 3$. Делта-делта обележја се добијају применом формуле 3.6 на делта обележја.

3.3 Статистички приступ са Марковљевим моделима

У овом поглављу биће дат кратак преглед скривених Марковљевих модела. Марковљеви модели имају велике примене у препознавању узорака и облика (енгл. *Pattern Recognition*) код случајних процеса са дискретним временом. Названи су по руском математичару А.А. Маркову који је применио тада нови концепт на статистичку анализу Пушкинове поеме "Евгеније Оњегин" [47]. Неке од примена су у опису понашања системâ, препознавању говора, препознавању руком писаног текста, класификацији, итд.

Марковљеви ланци (енгл. *Markov chains*) спадају у групу случајних процеса са минималном меморијом. Нека је $X = (x_1, x_2, .., x_T)$ секвенца случајних променљивих чије вредности припадају коначном алфабету $\{1, 2, ..i, .. j, .., N\}$, при чему је N величина алфабета. Уколико је реч о Марковљевом ланцу првог реда, целокупна информација о прошлости случајног процеса која утиче на будућност је садржана у тренутној (садашњој) вредности, односно:

$$P(x_t = j / x_{t-1} = i, x_{t-2} = h, ...) = P(x_t = j / x_{t-1} = i) \tag{3.7}$$

Ако је случајни процес стационаран у ширем смислу прелазне вероватноће не зависе од времена, то јест:

$$P(x_t = j / x_{t-1} = i) = P(j / i) \tag{3.8}$$

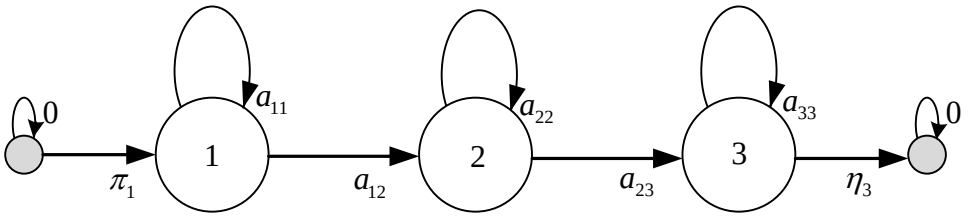
Ако је стање у тренутку t означено са x_t тада Марковљев ланац може бити описан са релацијама 3.9 и 3.10.

$$a_{ij} = P(x_t = j / x_{t-1} = i);\ 1 \le i, j \le N \tag{3.9}$$

$$a_{ij} \ge 0, \sum_{j=1}^{N} a_{ij} = 1;\ 1 \le i \le N \tag{3.10}$$

У релацији 3.10 a_{ij} представља прелазну вероватноћу из стања i у стање j.

Топологија мреже подразумева структуру дозвољених прелаза између појединих стања. Код ергодичног (потпуно повезаног) модела могући су прелази из сваког стања у било које друго стање. Ипак, због динамичке природе говорног сигнала у модерним ASR системима доминира серијска структура код којих су могуће транзиције у стања са индексима вишим од текућег. Такође, често су ASR системи значајно бржи, а без снижених перформанси уколико је серијска структура без прескока (енгл. *left-right topology without skips*). Неактивна стања (*null states*) су одређена почетком и крајем секвенце, као што је приказано у моделу чија је топологија приказана на слици 3.4 (неактивна стања су осенчена).



Слика 3.4 – Топологија серијске структуре без прескока са 3 активна стања

Улазне (иницијалне) вероватноће у тренутку t=1 за свако стање износе:

$$\pi_i = P(x_1 = i); \quad \sum_{i=1}^{N} \pi_i = 1 \tag{3.11}$$

Излазне (крајње) вероватноће у крајњем тренутку $t=T$ су слично дефинисане и износе:

$$\eta_i = P(x_T = i); 1 \le i \le N; \eta_i + \sum_{j=1}^{N} a_{ij} = 1; \forall i \tag{3.12}$$

Вероватноћа одређене секвенце стања $X = \{x_1, x_2, .., x_T\}$ за дати модел Λ износи:

$$P(X / \Lambda) = \pi_{x_1} (\prod_{t=2}^{T} a_{x_{t-1} x_t}) \eta_{x_T} \tag{3.13}$$

За разлику од Марковљевих ланаца код којих је видљива секвенца (опсервација) детерминистичка, код скривених Марковљевих модела та секвенца није видљива, већ је случајна променљива (дискретна или континуална). У ствари, скривени Марковљеви модели представљају двоструко-стохастички процес код којих је опсервација случајна променљива стања.

Вероватноћа генерисања дискретне опсервације $k \in \{1,2,..,K\}$ у стању i износи:

$$b_i(o_t) = P(o_t = k / x_t = i) \tag{3.14}$$

За случајни процес који генерише вредности из непребројивог (континуалног) скупа вероватноћа генерисања износи:

$$b_i(o_t) = P(o_t / x_t = i) \tag{3.15}$$

Прво ћемо размотрити дискретне HMM. Дискретни HMM модел Λ одређен са матрицом прелазних вероватноћа

$$A = \{\pi_j, a_{ij}, \eta_i\} = P(x_t = j / x_{t-1} = i); 1 \le i \le N;\ 1 \le j \le N \tag{3.16}$$

и матрицом вероватноћа емитовања

$$B = \{b_j(k)\} = \{P(o_t = k / x_t = j)\}; 1 \le j \le N, 1 \le k \le K \tag{3.17}$$

Сада је вероватноћа видљиве секвенце $o = \{o_1, o_2, .., o_T\}$ и секвенце стања (здружена вероватноћа), на основу дефиниције условне вероватноће:

$$P(o, X / \lambda) = P(X / \lambda) \cdot P(o / X, \lambda) \tag{3.18}$$

Тада је на основу формуле 3.13

$$P(o, X / \lambda) = \pi_{x_1} b_{x_1}(o_1) \cdot \left(\prod_{t=2}^{T} a_{x_{t-1} x_t} b_{x_t}(o_t) \right) \cdot \eta_{x_T} \tag{3.19}$$

Поред релације 3.7, још се претпоставља да је вероватноћа емитовања симбола o_t независна од претходних опсервација и зависна је само од стања у којем се систем тренутно налази x_t. Другим речима, претпоставља се некорелисаност видљивих стања у суседним тренуцима. И поред наведених претпоставки примена HMM се не сужава значајно, док се у значајној мери смањује број параметара који се требају естимирати у разумно кратком времену (на данашњем нивоу технологије). Имајући у виду опис HMM у практичним применама потребно је решити 3 главна проблема [41]:

1. **Проблем евалуације**

   За дати модел $\lambda = (A, B)$ колико износи вероватноћа генерисања добијене видљиве секвенце $o = \{o_1, o_2, .., o_T\}$, односно $P(o/\lambda)$?

2. **Проблем декодовања**

   За дати модел $\lambda = (A, B)$ и дату видљиву секвенцу $o = \{o_1, o_2, .., o_T\}$ која је највероватнија секвенца скривених стања која је генерисала добијену видљиву секвенцу?

3. **Проблем естимације (оцене) параметара**

   За дати модел $\lambda = (A, B)$ и скуп скривених и видљивих стања, како се може побољшати иницијални модел $\Lambda = \{\lambda\}$, тако да се максимизује здружена вероватноћа $\prod_o P(o/\lambda)$? Проблем се своди на естимацију (оцену) елемената матрица A и B, за дати скуп података (ансамбл) доступан при обуци.

И поред тога што су побројани проблеми међусобно уско повезани, размотрићемо могућност решавања за сваки понаособ.

3.3.1 Проблем евалуације

У претходном поглављу смо израчунали вероватноћу здружене секвенце опсервација и секвенце стања за дати модел Λ, једначине 3.18 и 3.19. За дати модел и видљиву секвенцу $o = \{o_1, o_2, .., o_T\}$ најједноставнији начин за добијање вероватноће $P(o/\lambda)$ је да се саберу вероватноће свих могућих секвенци скривених стања који генеришу дату секвенцу опсервација. На основу формуле тоталне вероватноће имамо:

$$P(o/\lambda) = \sum_{sve\,X} P(o, X/\lambda) = \sum_{sve\,\mathbf{X}_1^T} P(\mathbf{O}_1^T, \mathbf{X}_1^T/\lambda) \qquad (3.20)$$

Уколико је N број скривених стања, тада је број свих могућих "путања" N^T. Из израза 3.20 се види да је комплексност рачунања вероватноће видљиве секвенце $P(o/\lambda)$ експоненцијална и реда је $o(N^T \cdot T)$. Због тога је развијен рекурзивни алгоритам динамичког програмирања чија је комплексност рачунања квадратна.

Сада ћемо дефинисати вероватноћу "унапред". Нека за дату видљиву секвенцу $\alpha_t(j)$ представља вероватноћу да је модел у тренутку t у стању j, односно:

$$\alpha_t(j) = P(\mathbf{o}_1^t, x_t = j / \lambda) = \sum_{\{\mathbf{x}_1^{t-1}, x_t = j\}} P(\mathbf{o}_1^t, \mathbf{x}_1^t / \lambda) \tag{3.21}$$

Узимајући у обзир претпоставку да текуће стање зависи само од претходног и да вероватноћа емитовања зависи само од тренутног стања имамо:

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) \cdot P(x_t = j / x_{t-1} = i, \lambda) \cdot P(o_t / x_t = j, \lambda) \tag{3.22}$$

Из израза 3.22 видимо да се за одређивање вероватноће "унапред" може користити рекурзивни алгоритам којим се комплексност рачунања са експоненцијалне смањује на квадратну, односно $o(N^2T)$. Алгоритам спада у алгоритме динамичког програмирања. С обзиром на релацију 3.22 можемо написати псеудо-код алгоритма "напред" (енг. *forward algorithm*) [41].

Алгоритам 3.1: Алгоритам "напред"

Корак 1: Иницијализација за $t = 1$

$$\alpha_1(i) = \pi_i \cdot b_i(o_1); 1 \le i \le N$$

Корак 2: Рекурзија

$$\alpha_t(j) = \left(\sum_{i=1}^{N} \alpha_{t-1}(i) \cdot a_{ij} \right) \cdot b_j(o_t); 1 \le j \le N$$

Корак 3: Крај

$$P(o / \lambda) = \sum_{i=1}^{N} \alpha_T(i) \cdot \eta_i$$

На сличан начин, до вероватноће видљиве секвенце се може доћи и помоћу алгоритма "назад" (енгл. *backward algorithm*).

Дефинишемо вероватноћу "уназад":

$$\beta_t(i) = P\left(o_{t+1}{}^T / x_t = i, \lambda\right) \tag{3.23}$$

Можемо приметити да је за разлику од вероватноће "унапред" која је здружена, вероватноћа "уназад" је условна. На сличан начин можемо написати псеудо-код алгоритма "назад" [41].

Алгоритам 3.2: Алгоритам "уназад"

Корак 1: Иницијализација за $t = T$

$\beta_T(i) = \eta_i ; 1 \le i \le N$

Корак 2: Рекурзија

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j); 1 \le i \le N$$

Корак 3: Крај

$$P(o/\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$

Као што и сам назив каже, код овог алгоритма рачуна се од излазног чвора (*exit node*) ка улазном (*entry node*), односно уназад.

3.3.2 Проблем декодовања

За дати модел Λ и добијену секвенцу видљивих стања $o = \{o_1, o_2, .., o_T\}$, проблем декодовања поставља питање која секвенца скривених стања $X = (x_1, x_2, .., x_T)$ је највероватније генерисала добијену секвенцу? Одговор на то питање даје Витербијев алгоритам декодовања који секвенцу скривених стања одређује методом максималне веродостојности, односно ML (енгл. *Maximum Likelyhood*) критеријумом одлучивања датог изразом 3.24.

$$X^* = \arg\max_X P(o, X/\lambda) \tag{3.24}$$

При том је $P(o, X/\lambda)$ здружена вероватноћа дата изразом 3.19. Витербијев алгоритам је индуктивни метод који оптималну секвенцу скривених стања X^* проналази рачунањем максималне кумулативне вероватноће $\delta_t(j)$ за свако стање j, слично рачунању вероватноће "унапред".

Алгоритам 3.3: Витербијев алгоритам декодовања

Корак 1: Иницијализација за $t = 1$

$\delta_1(i) = \pi_i b_i(o_1)$

$\psi_1(i) = 0; 1 \le i \le N$

Корак 2: Рекурзија за $t = \{2,3,..,T\}$

$\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij} \right] b_j(o_t)$

$\psi_t(j) = \arg\max_i \left[\delta_{t-1}(i) a_{ij} \right]; 1 \le i \le N$

Корак 3: Завршетак трелиса

$x_T^* = \arg\max_i \left[\delta_T(i) \eta_i \right]$

Корак 4: Запис оптималне путање X^* уназад, за $t = \{T, T-1,..,2\}$

$x_{t-1}^* = \psi_t(x_t^*),\ X^* = \{x_1^*, x_2^*, .., x_T^*\}$

3.3.3 Проблем оцене (естимације) параметара

За прецизно описивање видљивих секвенци веома је важно оценити параметре модела $\lambda = (A, B)$. Од побројана 3 проблема проблем оцене параметара је најтежи јер до сад није познат аналитички метод у затвореној форми за максимизацију здружене вероватноће података у обуци [41].

Постоји неколико метода за естимацију параметара HMM модела [48]:

1. метода базирана на минимизацији средњеквадратне грешке (енг. *MMSE - Minimum Mean Square Error*)
2. метода базирана на максималној веродостојности (*ML* критеријум)
3. метода базирана на максималној апостериорној вероватноћи

Због тога што су у обуци код HMM приступа у обуци најчешће доступне функције густине вероватноће вектора обележја прибегава се најчешће другој методи.

У општем случају, желимо да одредимо параметре модела c који ће са највећом веродостојношћу генерисати секвенце података у обуци o_{train}. ML критеријум даје естимацију $\tilde{c}$ која се добија деривацијом $P(o_{train} / c) = 0$. Због монотоности логаритамске функције одређивање параметара се своди на:

$$\frac{\partial \ln P(o_{train} / c)}{\partial c} = 0 \tag{3.25}$$

Решавањем једначине дате изразом 3.25 добијамо оптималне вредности параметара модела за доступне податке у обуци. У препознавачима говора базираним на HMM параметри модела се могу добити помоћу Витербијевог алгоритма у обуци и методом максимизације очекивања (енг. *Expectation maximization - EM).*

За естимацију параметара модела можемо искористити оптималну путању коју смо добили Витербијевим алгоритмом (претходно поглавље):

$$P(o / \lambda) = \sum_{sve\,X} P(o, X / \lambda) \approx P(o, X^* / \lambda) \tag{3.26}$$

Коришћењем оптималне путање X^* можемо донијети грубу одлуку о заузећу стања (енг. *state occupation*) $q_t(i) \in \{0,1\}$ и према томе одредити реестимиране параметре модела. Овај начин реестимације се назива Витербијева обука са грубом проценом (енг. *hard state assignment*), а естимиране вредности су дате изразима 3.27 и 3.28. Овај начин се још назива форсирано поравнање (енг. *forced alignment*).

$$a_{ij} = \frac{\sum_{t=2}^{T} q_{t-1}(i) q_t(j)}{\sum_{t=1}^{T} q_t(i)};\ q_t(i) = \begin{cases} 1 & за\ o_t = k \\ 0 & иначе \end{cases} \tag{3.27}$$

$$\hat{b}_j(k) = \frac{\sum_{t=1}^{T} q_t(j) \omega_t(k)}{\sum_{t=1}^{T} q_t(j)};\ \omega_t(k) = \begin{cases} 1, & за\ i = x_t \\ 0 & иначе \end{cases}\ 1 \le j \le N; 1 \le k \le K \tag{3.28}$$

Слични су изрази и ако су у обуци на располагању изговори *R* фајлова. У прилогу Б је дат пример који за конкретан HMM и претходно одређену оптималну путању Витербијевом обуком даје оцену параметара модела.

За оцену параметара модела се далеко чешће користи обука ML критеријумом са максимизацијом очекивања (енг. *Expectation Maximization*) који оптимизује параметре модела меким одлучивањем (енг. *soft state assignment*) припадности опсервација стањима. Дефинишемо вероватноћу припадности стању (*occupation likelyhood*):

$$\gamma_t(i) = P(x_t = i / o, \lambda) \tag{3.29}$$

Коришћењем Бејсовог правила имамо:

$$\gamma_t(i) = \frac{P(o, x_t = i, \lambda)}{P(o/\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(o/\lambda)} \tag{3.30}$$

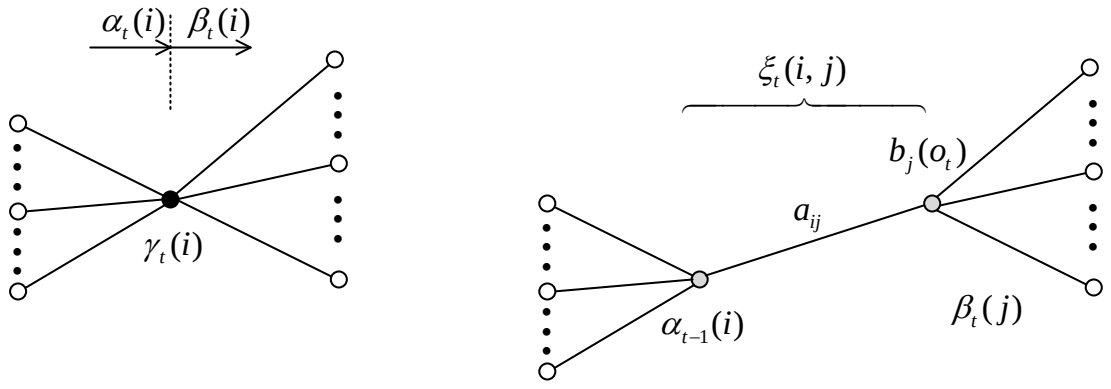
Из израза 3.30 видимо да је за добијање вероватноће $\gamma_t(i)$ довољно познавати вероватноће "унапред", "уназад" и вероватноћу видљиве секвенце (поглавље 3.3.1). На сличан начин дефинишемо вероватноћу прелаза (енг. *transition likelyhood*) за дату видљиву секвенцу:

$$\xi_t(i, j) = P(x_{t-1} = i, x_t = j / o, \lambda) \tag{3.31}$$

Применом дефиниције условне вероватноће имамо:

$$\begin{aligned} \xi_t(i, j) &= \frac{P(o, x_{t-1} = i, x_t = j, \lambda)}{P(o/\lambda)} = \\ &= \frac{P(\boldsymbol{o_1}^{t-1}, x_{t-1} = i / \lambda) \cdot P(\boldsymbol{o_t}, x_t = j / x_{t-1} = i, \lambda) \cdot P(\boldsymbol{o_{t+1}}^{T}, x_t = j, \lambda)}{P(o/\lambda)} \\ &= \frac{\alpha_{t-1}(i) a_{ij} b_j(o_t) \beta_t(j)}{P(o/\lambda)} \end{aligned} \tag{3.32}$$

За добијање вероватноће припадности и прелазне вероватноће може послужити илустрација дела трелис дијаграма, приказана на слици 3.5.



Слика 3.5 – Илустрација дела трелис дијаграма за добијање вероватноће припадности и прелазне вероватноће

Иницијално, усвоје се произвољне вредности за a_{ij} и b_{jk}. Побољшане вредности се добијају алгоритмом максимизације очекивања и методом Лагранжових

мултипликатора за одређивање локалног екстрема функције више променљивих [41]. Овај поступак се назива итеративни *Baum-Welch*ов алгоритам обуке.

Пажљивом анализом израза у BW алгоритму може се приметити да је ажурирана вероватноћа прелаза $\hat{a}_{ij}$ једнака количнику очекиваног броја прелаза из стања i у стање j и очекиваног броја прелаза из стања i. Такође, ажурирана вероватноћа емитовања $\hat{b}_j(k)$ једнака је количнику очекиваног броја опсервација симбола k у стању j и очекиваног броја опсервација у стању j.

Битно је нагласити да *Baum-Welch*ов алгоритам итеративним поступком не гарантује достизање глобалног максимума већ само локалног.

Алгоритам 3.4: *Baum-Welch*ов алгоритам реестимације параметара модела

Прелазне вероватноће

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)}; 1 \le i, j \le N$$

Вероватноће емитовања

$$\hat{b}_j(k) = \frac{\sum_{t=1}^{T} \gamma_t(j)\omega_t(k)}{\sum_{t=1}^{T} \gamma_t(j)}; 1 \le j \le N; 1 \le k \le K$$

Итеративни поступак се понавља док се повећава вероватноћа података у обуци $P(o_{train} / \hat{\lambda}) \ge P(o_{train} / \lambda)$

3.3.4 Скривени Марковљеви модели са континуалном расподелом и мешавинама

До сад анализирани HMM модели су претпостављали да видљиве секвенце долазе из дискретног (коначног или пребројивог) скупа. То значи да свака опсервација у дискретном тренутку долази има једну од *M* вриједности алфабета дискретних симбола. Да би се таква методологија могла применити на говорни сигнал неопходно је исти векторски квантизовати што доводи до грешке услед

квантизације. Због тога су уведени скривени Марковљеви модели са континуалном расподелом. Практично, сви препознавачи говора базирани на HMM (као и препознавач развијен у овој дисертацији) су овог типа. У овом случају, вероватноће емитовања дискретних симбола се мењају функцијама густине вероватноће опсервација.

$$B = \{b_i(o_t)\} = P\{o_t / x_t = i\} \tag{3.33}$$

Најчешће се користе Гаусове (нормалне) расподјеле вероватноћа. Постоје 2 типа HMM са континуалном расподелом:

1. једнодимензионални HMM (видљиве секвенце су скалари)
2. вишедимензионални HMM (видљиве секвенце су вектори димензије веће од 1)

Једнодимензионални HMM су одређени густином вероватноћа:

$$b_i(o_t) = \frac{1}{\sqrt{2\pi\Sigma_i}} \exp(\frac{-(o_t - \mu_i)^2}{2\Sigma_i}) \tag{3.34}$$

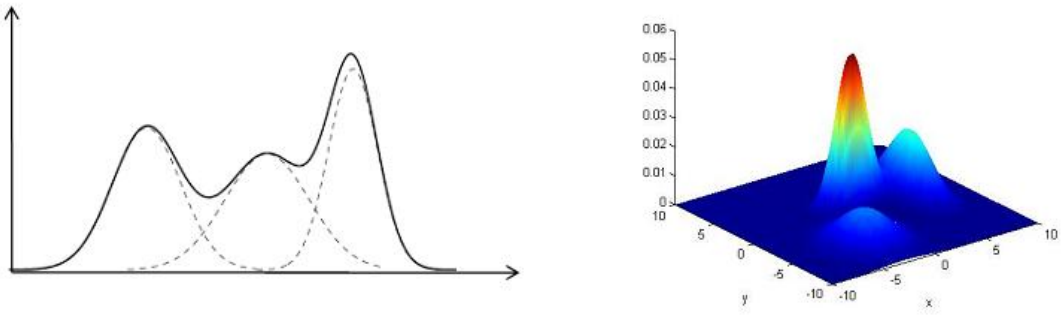
При том је μ_i средња вредност, а Σ_i варијанса.

Вишедимензионални HMM су одређени вектором средњих вреднсоти $\boldsymbol{\mu}_i$ (димензије $1 \times K$) и матрицом коваријансе $\boldsymbol{\Sigma}_i$ (димензије $K \times K$), при чему је K димензионалност простора опсервација, односно вектора обележја. Густина вероватноћа вишедимензионалних HMM је дата изразом 3.35.

$$b_i(\mathbf{o}_t) = \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}_i|}} \exp(-\frac{1}{2} (\mathbf{o}_t - \boldsymbol{\mu}_i) \boldsymbol{\Sigma}_i^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_i)^T) \tag{3.35}$$

С обзиром да се функција густине вероватноће било које случајне променњиве може апроксимирати сумом M Гаусових случајних променљивих, у HMM препознавачима претпоставља се вишедимензионална Гаусова расподела са мешавинама (енгл. *Multivariate Gaussian mixture*).

Пример функције густине вероватноће за расподелу са мешавинама (3 Гаусове компоненте) је приказана на слици 3.6.



а) б)

Слика 3.6 – Пример једнодимензиналне (а) и дводимензионалне (б) функције густине вероватноће за расподелу са мешавинама. Број мешавина је 3. Слика је преузета са [49].

Расподела са мешавинама је дата релацијом 3.36:

$$b_i(\mathbf{o_t}) = \sum_{m=1}^{M} c_{im} \mathcal{N}(\mathbf{o_t}; \boldsymbol{\mu}_{\mathbf{im}}, \boldsymbol{\Sigma}_{\mathbf{im}}) \tag{3.36}$$

При том је $\mathcal{N}(\bullet)$ вишедимензионална HMM са вектором средњих вредности $\boldsymbol{\mu}_{\mathbf{im}}$ и матрицом коваријансе $\boldsymbol{\Sigma}_{\mathbf{im}}$. Укупан број мешавина је M при чему је сума тежина једнака јединици.

$$\sum_{m=1}^{M} c_{im} = 1 \tag{3.37}$$

У случају HMM са континуалном расподелом може се показати (методом максималне веродостојности) да су оцене средње вредности и варијансе (за одређену средњу вредност) дате изразиме 3.38 и 3.39, респективно.

$$\hat{\mu}_{ML} = \frac{1}{T} \sum_{t=1}^{T} o_t \tag{3.38}$$

$$\hat{\Sigma}_{ML} = \frac{1}{T} \sum_{t=1}^{T} (o_t - \mu_{ML})^2 \tag{3.39}$$

У свим препознавачима говора базираним на HMM видљиве секвенце су вектори. За вишедимензионалне HMM са мешавинама оцене вектора средњих вредности, коваријансне матрице и тежине мешавина се могу добити методом Ларанжових мултипликатора и дате су изразима 3.41 – 3.43 [41].

$$\gamma_t(j,m) = \frac{\alpha_t(j,m)\beta_t(j)}{P(o/\lambda)} \tag{3.40}$$

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)\mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{3.41}$$

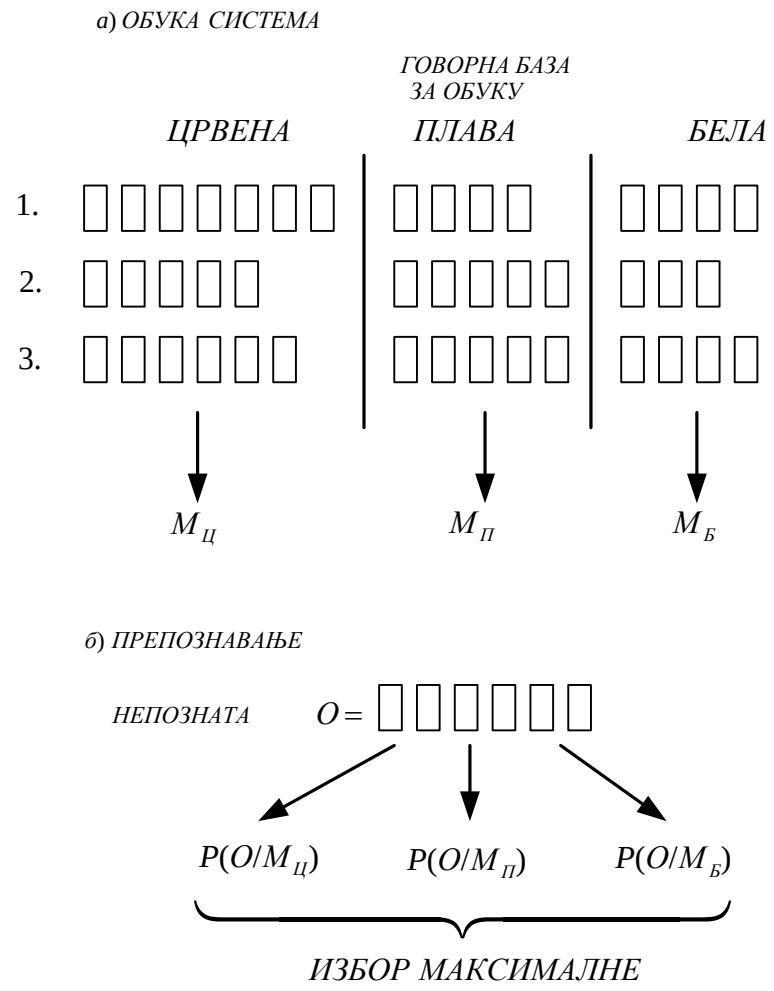
$$\hat{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)(\mathbf{o}_t - \boldsymbol{\mu}_{jm})(\mathbf{o}_t - \boldsymbol{\mu}_{jm})^T}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{3.42}$$

$$c_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{3.43}$$

3.4 Препознавање изолованих речи

На слици 3.7 је графички приказано коришћење HMM за препознавање изолованих речи. На пример, нека је у питању препознавач за речи три боје у српском језику: црвена, плава и бела. За сваку реч на располагању су по 3 изговора за обуку, који су у општем случају различите дужине [46].

У обуци се на основу изговора из говорне базе за обуку рачунају параметри HMM модела који репрезентују поједине речи ($M_Ц$,$M_П$,$M_Б$). Након обуке следи препознавање. За дати изговор O рачунају се вероватноће да је сваки појединачни HMM модел генерисао добијену секвенцу опсервација O. Излаз из препознавача представља реч за коју је добијена највећа вероватноћа.



Слика 3.7 – Пример коришћења HMM у препознавању изолованих речи

3.5 Препознавач базиран на методи потпорних вектора

Поред препознавања заснованог на статистичком притупу и вештачким неуронским мрежама, у многим практичним применама класификатор базиран на методи потпорних вектора је показао добру робустност у препознавању говора, нарочито уколико се користи у хибридној форми са HMM [50]. Због тога је у овој дисертацији анализирана могућност коришћења SVM класификатора у препознавању нормалног говора и шапата, зависно и независно од говорника.

Класификатор базиран на методи потпорних вектора је у почетним радовима Вапника [51] био осмишљен као бинарни класификатор за линеарно сепарабилне подскупове у n-димензионалном векторском простору. Идеја на којој почива је да се раздвајање класа постиже максимизацијом маргине између њих. На тај начин

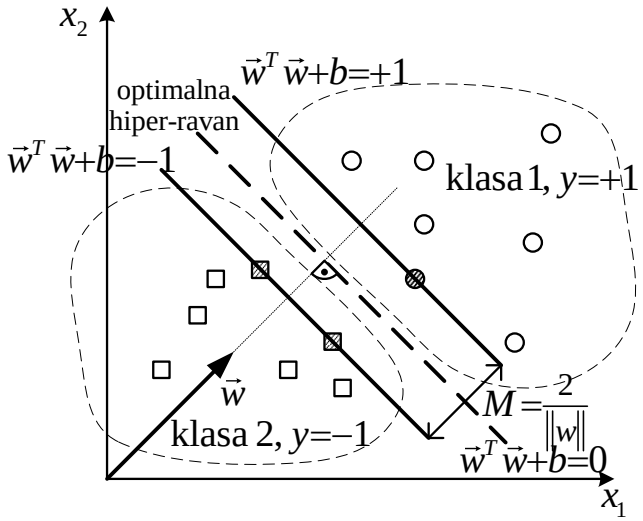
се очекује добра робустност, односно да узорак података доступан у обуци буде "репрезентативан" узорак дотичне класе. У n-димензионалном векторском простору обучавање SVM класификатора подразумева одређивање параметара тзв. хипер-равни (енг. *hyper-plane*). У 2-D простору реч је о правој а у 3-D простору реч је о равни. Хипер-раван се налази између две класе и потпуно је одређена појединим векторима из класа, тзв. потпорним векторима [52].

Нека је $X=\{\vec{x}_1,\vec{x}_2,..,\vec{x}_N\}$ скуп улазних вектора доступних у неком експерименту који имају само две могуће излазне вредности $y\in Y=\{-1,1\}$.

Једначина која описује хипер-раван дата је следећом једначином:

$$\vec{w}^T\vec{x}+b=0 \tag{3.44}$$

При том је $\vec{w}$ вектор који одређује нагиб хипер-равни а b удаљеност од исходишта координатног система. На слици 3.8 је приказана хипер-раван која раздваја двије линеарно сепарабилне класе. Оптимална хипер-раван је приказана задебљаном испрекиданом линијом, маргина је означена са M, а потпорни вектори су шрафирани.



Слика 3.8 – Пример одређивања хипер-равни у 2-D простору за линеарно сепарабилне класе [53]

Као што се види са слике, хипер-раван дели простор на два полупростора (у конкретном 2-D примеру на две полуравни) при чему је вредност израза $\vec{w}^T\vec{x}+b>0$ за све тачке са једне стране, а $\vec{w}^T\vec{x}+b<0$ за све тачке са друге стране хипер-равни. Из аналитичке геометрије у $\mathbb{R}^n$ и удаљености тачке од равни добија се маргина

[54], дата са:

$$M=\frac{2}{\|\vec{w}\|} \tag{3.45}$$

Максимизација маргине се своди на минимизацију норме вектора $\|\vec{w}\|$, односно израза $\vec{w}^T\vec{w}$. С обзиром да је $\vec{w}^T\vec{x}_i+b\geq 1$ (за $y_i=1$) и $\vec{w}^T\vec{x}_i+b\leq -1$ (за $y_i=-1$) за одређивање оптималне хипер-равни потребан услов је:

$$y_i(\vec{w}^T\vec{x}+b)\geq 1 \tag{3.46}$$

Решење оптималне хипер-равни се добија методом Лагранжових мултипликатора за одређивање условних екстрема [54].

$$\vec{w}=\sum\alpha_i y_i\vec{x}_i \tag{3.47}$$

Одговарајућа функција одлуке је дата изразом 3.48. Са *sgn* је означена сигнум функција.

$$d(x)=\text{sgn}(\vec{w}^T\vec{x}+b)=\text{sgn}(\sum\alpha_i y_i\vec{x}_i{}^T\vec{x}+b) \tag{3.48}$$

Претходно описана процедура је валидна за линеарно сепарабилне класе, односно класе без преклапања. У већини практичних примера то није задовољено. У том случају се користи класификатор са тзв. меком маргином, код којег се одређена грешка толерише.

У многим практичним применама ни увођење класификатора са меком маргином неће довести до повећања успешности класификатора. Због тога се скуп података у обуци нелинеарним пресликавањем пресликава у вишедимензионални простор у којем се очекује боља линеарна сепарабилност. Функције за пресликавање се називају кернели. Најчешће коришћени кернели, који су анализирани и у овој дисертацији, су:

- *RBF* (*Radial basis function*) кернел

  $K(x_1,x_2)=\exp(-\frac{\|x_1-x_2\|^2}{2\sigma^2});\ \gamma=\frac{1}{2\sigma^2}$

- полиномијални кернел

  $K(x_1,x_2)=(\alpha x_1{}^T x_2+c)^d;$

- линеарни кернел

  $K(x_1,x_2)=x_1{}^T x_2+c$

- сигмоидни кернел

$$K(x_1, x_2)=\tanh(\alpha x_1^T x_2+c)$$

Такође, било која функција која задовољава одређене услове (*Mercer*ова теорема) може да се користи као кернел [55].

3.6 Резиме

Статистички приступ у аутоматском препознавању континуалног говора из великог речника је заснован на познавању 3 блока: акустичког модела, модела језика и речника. У овој глави је укратко описана методологија таквог приступа. Препознавање говора садржи 2 одвојене фазе: обуку и тестирање (препознавање) говора. Пошто је екстракција вектора обележја заједничка за обе фазе, у овој глави је дата блок шема екстракције најчешће коришћених вектора обележја: MFCC и PLP обележја.

Скривени Марковљеви модели пружају основу за моћан и нимало једноставан математички апарат за препознавање облика код случајних процеса са дискретним временом. Савремени ASR системи су незамисливи без Марковљевих модела који су у комбинацији са моделом мешавина Гаусових расподела или са, у последње време веома популарним, дубоким неуронским мрежама. У овој дисертацији је коришћен први приступ.

У овој глави је дата основа скривених Марковљевих модела. Скуп параметара модела садржи:

- иницијалне вероватноће скривених стања,
- вероватноће прелаза између скривених стања,
- вероватноће емитовања видљивих стања (код HMM са дискретном расподелом) односно векторе средњих вредности, коваријансне матрице Гаусових расподела и тежине мешавина (код вишедимензионалних HMM са континуалном расподелом).

Описана су три основна проблема која је неопходно решити за практичну примену HMM: проблем евалуације, проблем декодовања и проблем естимације параметара модела. Рекурзивни алгоритми динамичког програмирања (*Baum-Welch*ов алгоритам "напред-назад" при естимацији и Витербијев алгоритам при декодовању) су омогућили да се нумеричка комплексност извршавања операција са експоненцијалне сведе на квадратну. Тиме је, уз стално повећање рачунарске моћи процесора, постигнуто решавање проблема у разумно кратком временском периоду; бржа обука система и брже декодовање непознатих изговора.

Обзиром да је дисертацијом обухваћена и упоредна анализа перформанси са препознавачем базираним на методи потпорних вектора, у овој Глави је дата

основа SVM методе која је у многим практичним применама показала добру робустност при класификацији. Међутим, за разлику од HMM препознавача захтева да улазни подаци буду фиксне дужине, тако да има ограничену примену у препознавању говора.

4. ОПИС ЕКСПЕРИМЕНТАЛНЕ ПОСТАВКЕ

У последње три деценије развијен је знатан број софтверских алата за аутоматско препознавање говора, који су намењени коришћењу у научно-истраживачке или комерцијалне сврхе. У прву групу спадају *HTK* (енгл. *Hidden Markov model toolkit*) [46], *Julius* [56] (оба писана у програмском језику *C*), *Sphinx-4* [57] (писан у програмском језику *Java*), *RWTH ASR* [58] (писан у програмском језику *C++*) и у последње време изузетно популарни *Kaldi* [59] и *Deepspeech* [60].

У другу групу спадају алати специфичне намене и ограничене интеракције са корисником и другим софтверским пакетима као што су *AT&T Watson* [61], *Microsoft Speech Server* [62], *Google Speech API* [63] и *Nuance Recognizer* [64].

Програмски пакет HTK представља референтни и конвенционални систем за препознавање говора, базиран на статистичком приступу и HMM. Развијен је у кооперацији Лабораторије за машинску интелигенцију Универзитета у *Cambridge*у и фирме *Entropic Ltd*. Има велику флексибилност и интеракцију са корисником путем конзоле (енгл. *CLI - command-line interface*). Могуће је алате позивати из програмског језика *MATLAB* (који је инжењерима често приступачнији) и извршити проширење истих с обзиром на доступност изворног кода. Иако је оптимизован за обраду говора, може се користити и у другим научним областима код препознавања облика. Бесплатан је за истраживање али је редистрибуција забрањена. Алати се могу користити под *Linux* и *Windows*® окружењем.

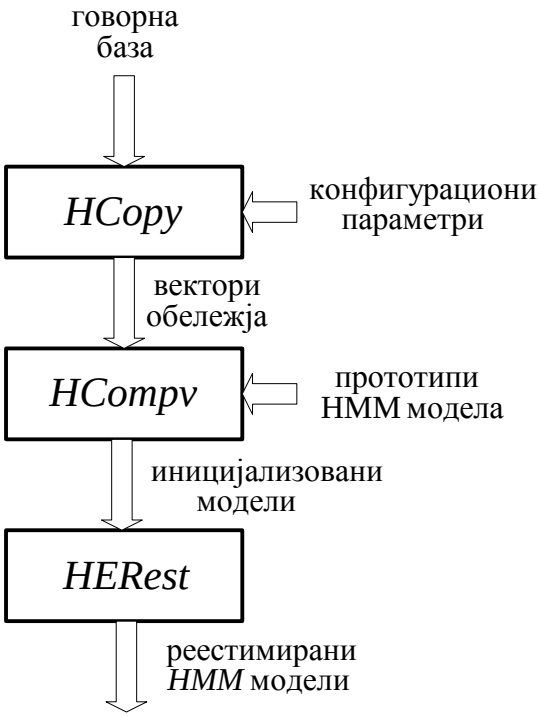
Препознавање говора се дели у две фазе: прва фаза представља тзв. обуку система, док друга фаза садржи тестирање (препознавање говора).

4.1 Обука система

У препознавању изолованих и повезано изговорених речи заснованом на статистичком приступу са HMM као јединице за моделовање се користе фонеми независни од контекста (монофони), фонеми зависни од контекста (трифони) и целе речи. Због тога ће посебна пажња бити посвећена моделовању и обуци сваке јединице понаособ у софтверском пакету HTK.

4.1.1 Обука фонема независних од контекста

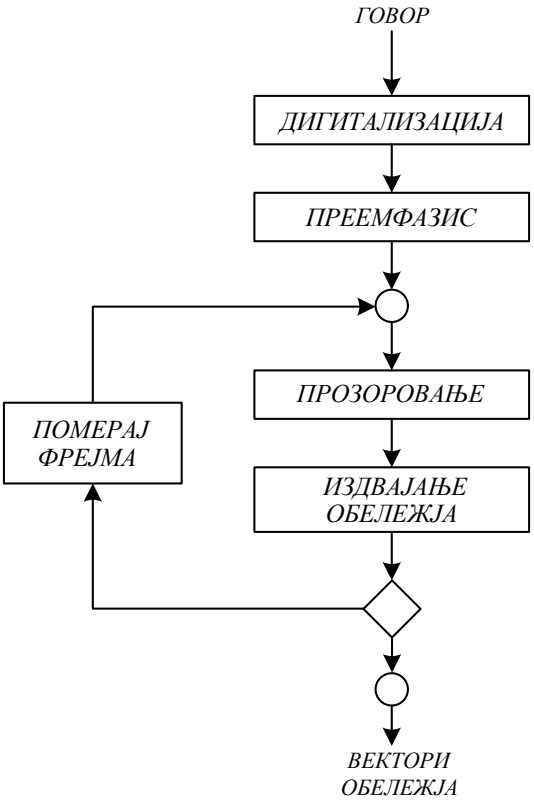
Уколико нису познате границе између фонема обука монофона садржи 3 HTK алата: *HCopy*, *HCompv* и *HERest*. Блок шема обуке монофона помоћу HTK алата је приказана на слици 4.1 [46].



Слика 4.1 – Блок шема коришћења алата у обуци фонема независних од контекста

- У процесу обуке пре свега је неопходно издвојити корисне информације из говорног сигнала које су значајне за препознавање говора, за што се користи алат *HCopy*. Улазни подаци за тај алат су говорна база у таласном (енг. *waveform*) облику и конфигурациони фајл

у којем се задају параметри конверзије (тип обележја, величина фрејма, померај, итд.). Запис конфигурационог фајла је дат у Прилогу А2 . Блок дијаграм којим се описује конверзија говорног сигнала у векторе обележја је приказана на слици 4.2.

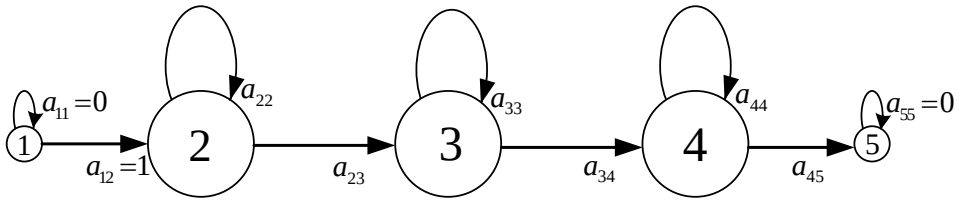


Слика 4.2 – Блок шема конверзије говорног сигнала у векторе обележја

Дигитализација говорног сигнала је извршена приликом креирања говорне (поглавље 2.4.1). Коришћен је преемфазис филтар са коефицијентом 0.97. У свим експериментима је коришћена величина прозора 24 ms, померај 8 ms (преклапање 2/3 прозора) и *Hamming*-ova прозорска функција.

- За иницијализацију модела монофона глобалном средњом вредношћу и глобалном варијансом користи се алат *HCompv*. Глобална варијанса говора је битна јер се као један од излазних параметара овог алата јавља вектор података који садржи 1% (или неки други износ процента) просечне вредности глобалне варијансе. Ти подаци се користе као праг да би се онемогућило да поједини модели услед мале количине

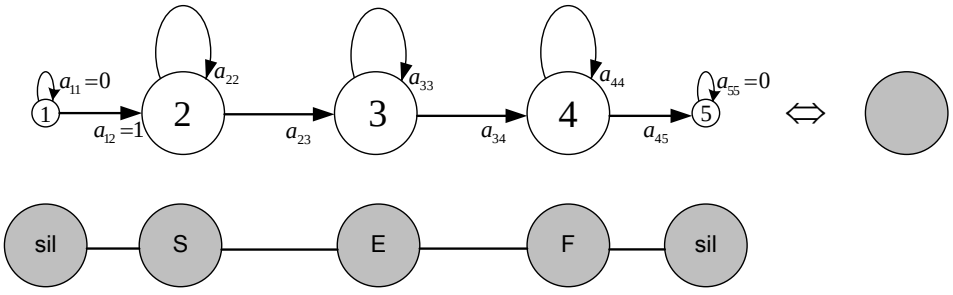
података за обуку имају мале вредности варијанси. Структура модела је приказана на слици 4.3 одакле се види да је реч о моделу са серијском структуром без прескока и 5 стања (од којих су 2 неемитујућа).



Слика 4.3 - Структура модела монофона

Листинг модела прототипа је дат у прилогу А3.

- Алат *HErest* се користи за естимацију параметара модела који представља имплементацију *Baum-Welch*овог алгоритма описаног у одељку 3.3.3. Алат симултано ажурира све HMM моделе користећи целу говорну базу немењену за обуку. Уз сваки говорни фајл постоји транскрипција на нивоу фонема за изговор који садржи дати фајл. По учитавању говорног фајла у меморију врши се креирање композитног модела једноставним спајањем модела појединачних фонема у складу са лабелом у транскрипцији. Након тога се врши описани *Baum-Welch*ов алгоритам. За реч /сеф/ (претпоследња реч у речнику говорне базе Whi-Spe) на слици 4.4 је приказан композитни модел монофона, при чему *sil* означава модел тишине (енгл. *silence*).



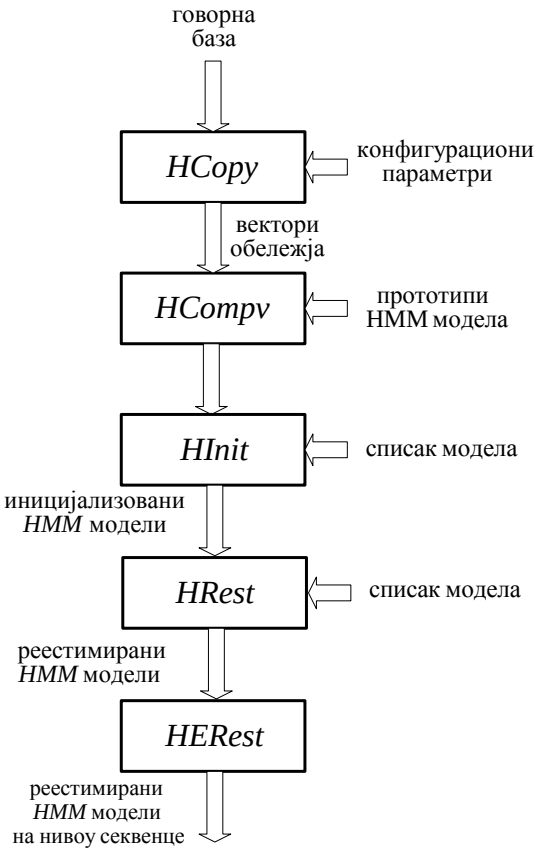
Слика 4.4 – Композитни модел фонема независног од контекста за реч /сеф/

Број итерација *Baum-Welch*овог алгоритма је могуће контролисати тако да се извршава све док се перформансе система побољшавају. У том случају може доћи до значајног повећања времена обуке али и до тзв. преобучавања (енгл. *overtraining*) када систем постаје мање робустан.

Због тога се у пракси примењује 2 до 5 итерација у коришћењу HERest алата [46].

За добре акустичке моделе потребна је што већа количина података за обуку. Потребно је неколико стотина изговора у препознавању зависно од говорника и неколико хиљада изговора у препознавању независно од говорника. За ограничење времена извршавања користи се тзв. поткресивање (енгл. *prunning*) постављањем минималне вредности опсега веродостојности при рачунању логаритма вероватноћа напред и назад. Транскрипција речи из говорне базе Whi-Spe за моделовање фонема независних од контекста је дата у прилогу А4.

Уколико су на располагању временске границе између фонема (добијене мануелно или аутоматски) у обуци фонема независних од контекста се још користе алати *HInit* и *HRest.* Блок шема коришћења алата је приказана на слици 4.5 [46].



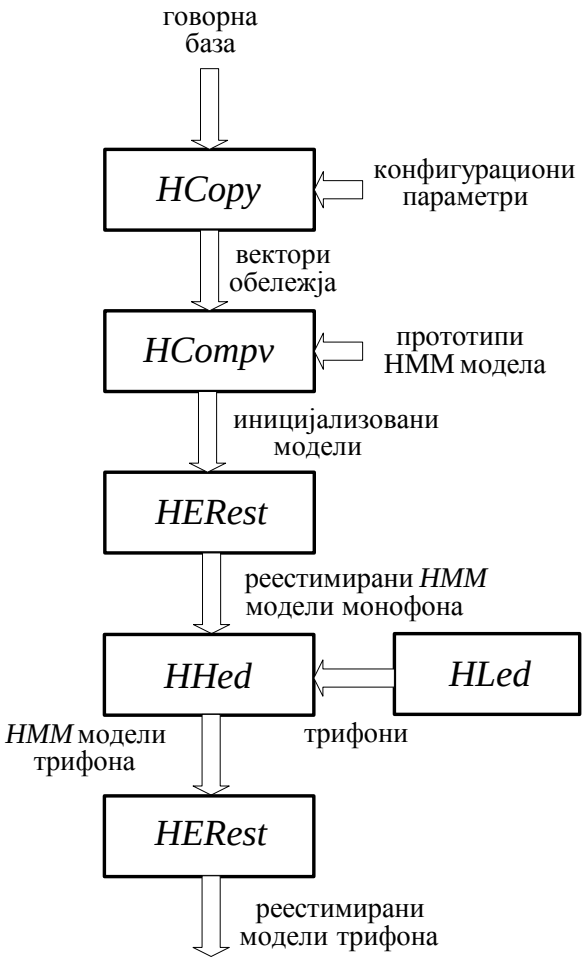
Слика 4.5 – Блок шема коришћења алата у обуци фонема независних од контекста ако су познате границе између фонема

Принцип функционисања алата HInit и HRest је детаљније описан у књизи за HTK [46]. Иницијализација помоћу HInit алата је заснована на концепту HMM као генератора вектора обележја [65]. Иницијализација параметара модела се врши Витербијевим алгоритмом (тврдо одлучивање). Алат *HRest* за иницијализоване моделе врши реестимацију *Baum-Welch*-овим поступком естимације параметара, описаним у одељку 3.3.3.

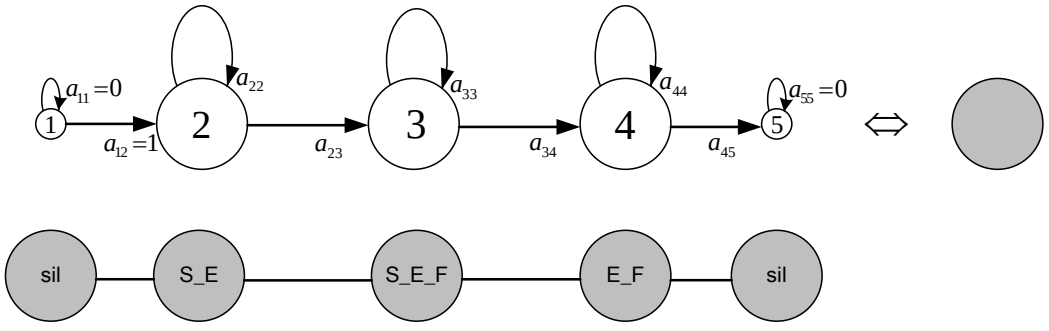
4.1.2 Обука фонема зависних од контекста

За добијање више специјализованих модела користе се фонеми зависни од контекста, од којих се најчешће користе трифони (триплети узастопних фонема). Тиме се добија много већи број модела чиме се смањују варијације услед коартикулације. Трифони могу бити некластеровани и кластеровани. Овде ће бити анализирани само некластеровани трифони. Блок шема добијања трифона помоћу HTK алата је приказана на слици 4.6.

- Алат *HLed* представља једноставан едитор за манипулацију са тзв. лабел фајловима [46]. Може да се користи за генерисање лабеле на нивоу трифона уколико су познате лабеле на нивоу монофона. Нпр. уколико је секвенца монофона ***sil S E F sil*** одговарајућа секвенца трифона је ***sil S_E S_E_F E_F sil***.
  Као што се види и са блок шеме, *HLed* представља помоћни алат за добијање HMM модела трифона.
- Алат *HHed* служи за манипулацију са HMM моделима. Може да се користи за повећање броја мешавина у корацима као и добијање иницијалних модела трифона. Иницијални параметри модела трифона се добијају на основу добијених иницијалних параметара модела монофона само променом назива модела.
- Иницијални модели трифона се побољшавају са још 2 додатна циклуса реестимације параметара трифона помоћу алата *HERest*. За реч /сеф/ на слици 4.7 је приказан композитни модел трифона.



Слика 4.6 – Блок шема коришћења алата у обуци фонема зависних од контекста

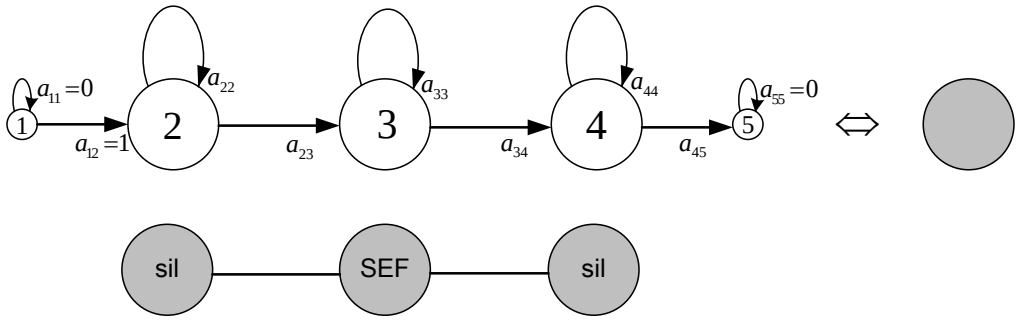


Слика 4.7 – Композитни модел фонема зависног од контекста за реч /сеф/

4.1.3 Обука модела целих речи

У препознавању изолованих и повезано изговорених речи као јединице за моделовање се поред фонема зависних и независних од контекста користе и целе

речи. За обуку модела целих речи користи се иста шема и скуп алата као код монофона, с тим да се цели изговор моделује самосталним HMM моделом (изузимајући моделе тишине на крајевима изговора). За реч /сеф/ на слици 4.8 је приказан композитни модел целе речи.



Слика 4.8 – Композитни модел целе речи за реч /сеф/

Транскрипција речи из говорне базе Whi-Spe за моделовање целих речи је дата у прилогу А5.

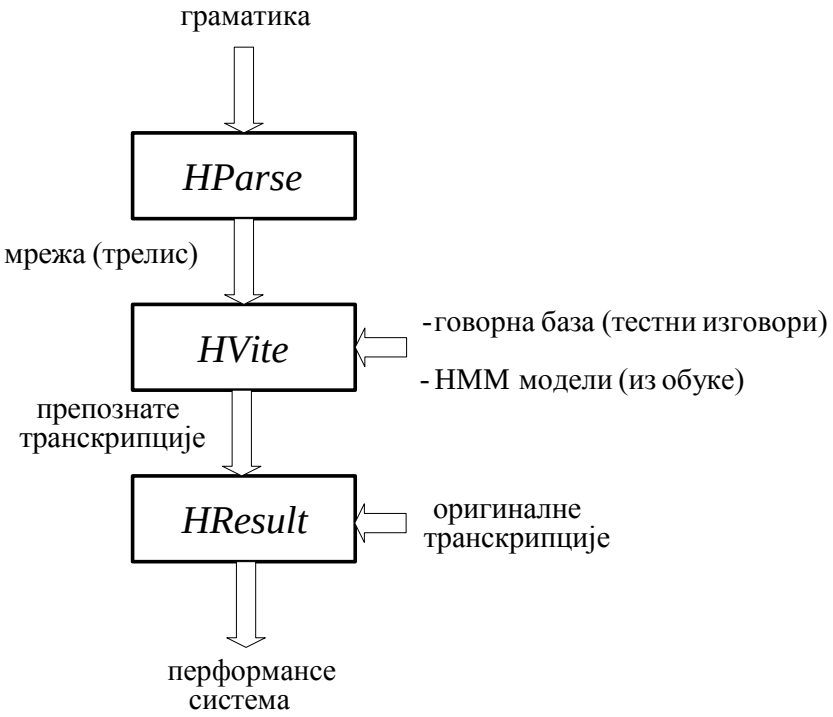
4.2 Тестирање система

Без обзира на који начин се врши обука ASR система тестирање се обавља помоћу 3 HTK алата. То су:

1. HParse
2. HVite
3. HResult

Употреба побројаних алата је приказана шематски на слици 4.9.

- На основу граматике која садржи секвенце речи које су дозвољене HTK алат *HParse* врши креирање мреже којом се добија трелис дијаграм потребан за препознавање. Мрежа представља граматику која се дефинише у тзв. *Backus-Naur* нотацији у којој се експлицитно дефинишу прелази између појединих речи. Фајл који садржи граматику се може креирати у било којем програму за обраду текста (нпр. *Notepad*). Изрази у *Backus-Naur* нотацији садрже алфанумеричке знакове заједно са матакарактерима чије је значење наведено у табели 4.1.



Слика 4.9 – Блок шема коришћења HTK алата у тестирању

ТАБЕЛА 4.1: МЕТАКАРАКТЕРИ КОЈИ СЕ МОГУ КОРИСТИТИ ЗА ЗАДАВАЊЕ ГРАМАТИКЕ

Метакарактер	Значење
|	Различите алтернативе
[ ]	опционо коришћење
{}	Нула или више понављања
< >	Једно или више понављања

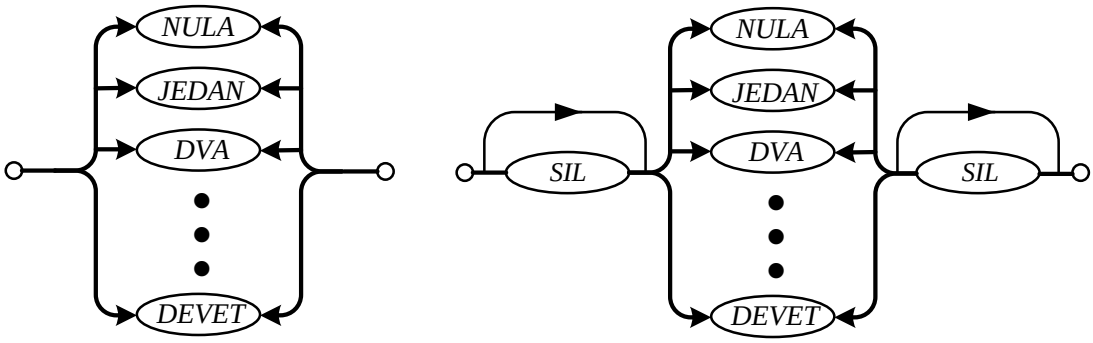
На слици 4.10а је приказана мрежа за препознавање изоловано изговорених речи цифара у српском језику, док је на слици 4.10б приказана мрежа за препознавање речи цифара са опционим моделом тишине (означен са *sil*) на почетку и крају.

Граматика за мрежу приказану на слици 4.10а би била:

(

NULA | JEDAN | DVA | TRI | CHETIRI | PET | SHEST | SEDAM | OSAM | DEVET

)



а) *б*)

Слика 4.10 – Примери мрежа за препознавање изолованих речи цифара (а) и изолованих речи цифара са опционим моделом тишине (б)

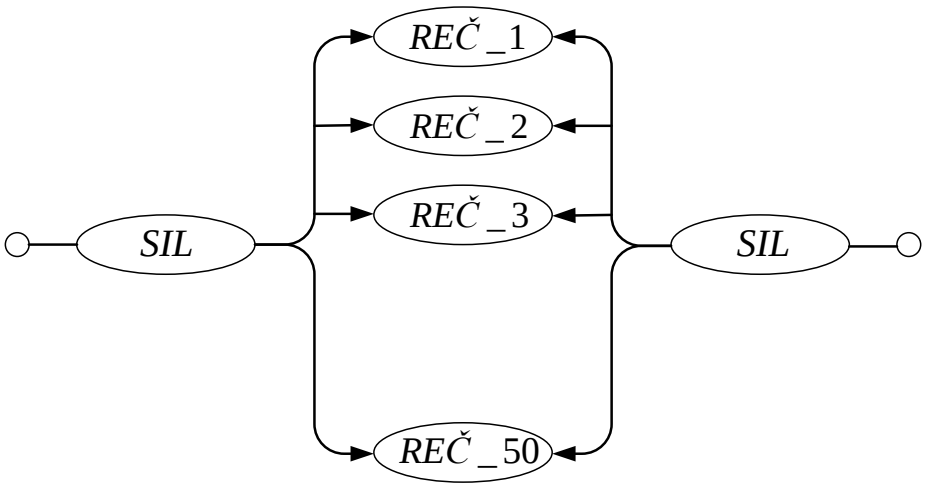
Аналогно, граматика за мрежу са опционим моделом тишине (4.10б) је:

(

[sil] (NULA | JEDAN | DVA | TRI | CHETIRI | PET | SHEST | SEDAM | OSAM | DEVET) [sil]

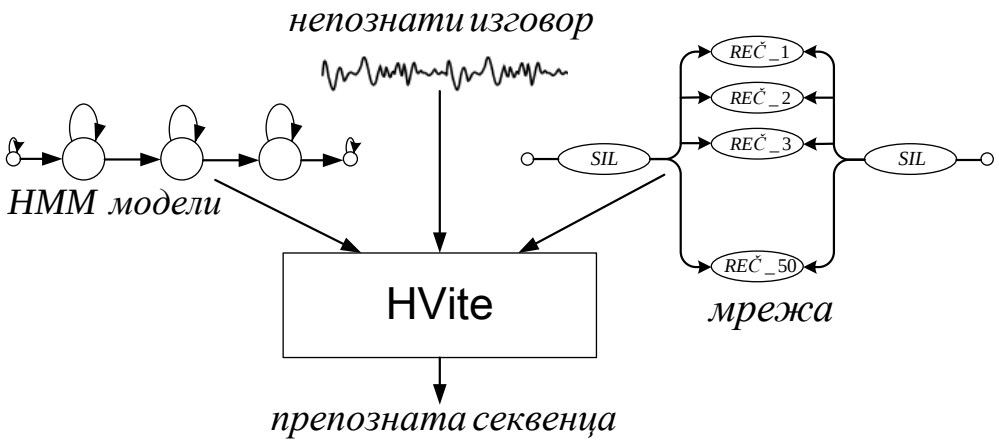
)

Запис фајла којим се задаје граматика која је коришћена за препознавање изолованих речи из говорне базе Whi-Spe (означених са REČ_1 до REČ_50) је дата у прилогу А6, а на слици 4.11 је приказана мрежа за препознавање.



Слика 4.11 – Мрежа која је коришћена за препознавање изолованих речи из базе *Whi - Spe*. Тишина је означена са *SIL*.

- Основна намена HTK алата HVite је одређивање непознатих тестних изговора. Као што је блоковски приказано на слици 4.12 за дату мрежу (претходно генерисану алатом HParse), параметре HMM модела који су одређени у обуци и непознати изговор може се одредити вероватноћа било које путање скривених стања. Задатак декодера је да одреди најизгледнију (најверодостојнију) путању, коришћењем алгоритма пропагације жетона (енгл. *Token Passing*) [46].



Слика 4.12 – Блок шема коришћења алата HVite

Жетон представља парцијалну путању кроз трелис дијаграм у распону од тренутка $t = 0$ до тренутка *t*. У почетном тренутку сваки почетни чвор поседује жетон. Сваки пут приликом досезања емитујућег стања зауставља се пропагација жетона и врши се увећање лог-вероватноће у складу са параметрима HMM модела. С обзиром да сваки чвор може да садржи највише *N* жетона (подразумевана вредност је 1) при сваком заустављању чворови задржавају *N* жетона са највећом акумулисаном лог-вероватноћом (аналогно Витербијевом алгоритму декодовања конволуционих кодова у Теорији информација).

Узимајући у обзир да велике мреже имају велики број чворова, ради контроле времена извршавања корисно је да се при пропагацији задржавају жетони који потенцијално могу да буду најбољи (садрже путању са максималном лог-вероватноћом). Овај поступак се назива поткресивање (енгл. *prunning*). Оптимални распон за поткресивање је

компромис између брзине препознавања и грешака услед губитка исправне путање. Ако је премален распон могуће је да путања која на крају може да буде најбоља, буде одбачена при одређеном заустављању. Подразумевана вредност распона лог-вероватноће у поткресивању износи 250.

Битно је нагласити да се алат *HVite* може користити и у фази обуке тако што се врши поравнање података у обуци (енгл. *forced alignment*). Као један од излаза могуће је добити анотиране фонеме са означеним тренуцима почетка и краја сваког фонема, који се могу користити у наредним фазама обуке.

- За евалуацију перформанси препознавача користи се HTK алат *HResult*. Алат пореди оригиналну транскрипцију и транскрипцију на нивоу речи добијену као излаз алата *HVite* и генерише различите оцене препознавача. За оптимално поравнање оригиналне и препознате транскрипције алат користи DTW алгоритам рачунањем мере поравнања (енгл. *a score for the match*). При том идентичне лабеле имају меру 0 (нула), убачене и обрисане лабеле имају меру 7 (седам), а замењене лабеле имају меру 100 (сто). Секвенца препознатих речи са најмањом мером је оптимално поравната.

Уколико је укупан број лабела у оригиналној транскрипцији N, број убачених лабела I, број обрисаних лабела D и број замењених лабела S алат *HResult* даје проценат коректности (*Cor*) и проценат тачности (*Acc*) који су дефинисани релацијама 4.1 и 4.2, респективно.

$$Cor = \frac{N - S - D}{N} \cdot 100[\%] \tag{4.1}$$

$$Acc = \frac{N - S - D - I}{N} \cdot 100[\%] \tag{4.2}$$

Као што се види из релације 4.2, проценат тачности узима у обзир и грешке услед убацивања лабела.

С обзиром да ова дисертација обухвата препознавање изолованих речи као перформанса препознавача ће бити рачунат проценат успешно препознатих речи (енгл. *Word Recognition Rate*) који узима у обзир само замењене лабеле. Проценат успешно препознатих речи је дефинисан релацијом 4.3.

$$WRR = \frac{N - S}{N} \cdot 100[\%] \tag{4.3}$$

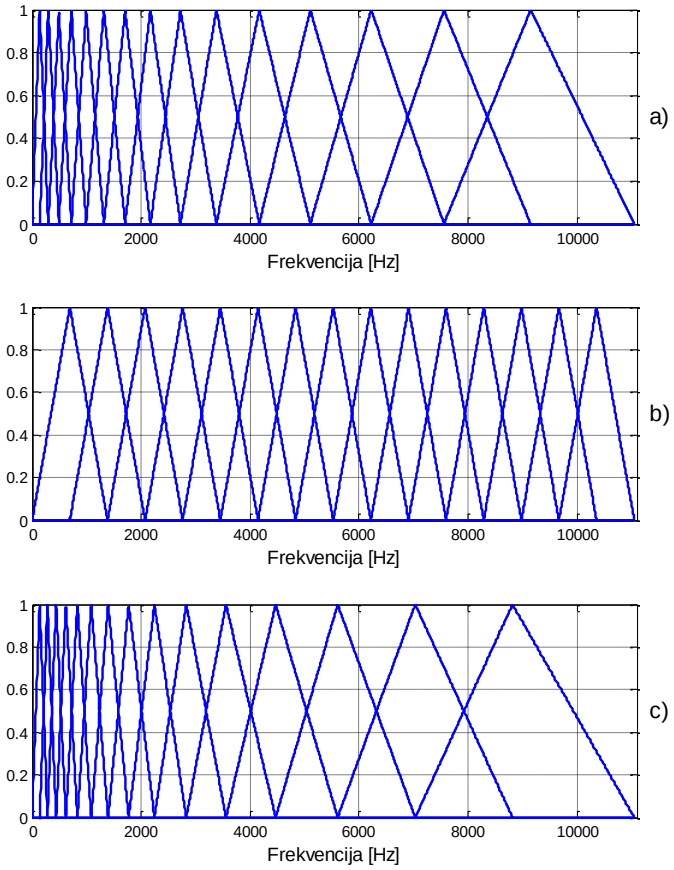
4.3 Предлог алгоритма за добијање кепстралних коефицијената са модификованом фреквенцијском скалом

Мел-фреквенцијски кепстрални коефицијенти су традиционални и најчешће коришћени вектори обележја у ASR системима. Алгоритам за екстракцију MFCC вектора је приказан на слици 3.2. Користи мелодијску (мел) фреквенцијску скалу која математички описује перцепцију човековог чула слуха и рад базиларне мембране. Међутим, у одређеним ASR применама MFCC вектори обележја показују слабу робустност, односно показују снижене перформансе уколико услови при тестирању одступају (мање или више) од услова при обуци.

Експерименти у препознавању говорника који шапуће су показали да се добија већа успешност са кепстралним коефицијентима који користе линеарну фреквенцијску скалу [66]. Ти вектори обележја су означени са LFCC (енгл. *Linear Frequency Cepstral Coefficients*). Разлика у односу на MFCC обележја је само у распореду банке филтара на фреквенцијској скали. Код LFCC карактеристика филтара је равномерно распоређена на фреквенцијској скали у херцима док је код MFCC у мелима.

Перцептивни линеарни предиктивни коефицијенти (PLP), чији је алгоритам екстракције приказан на слици 3.3 користе *bark* фреквенцијску скалу. У одређеним применама је добијена већа робустност са PLP векторима обележја у односу на MFCC. *Bark* скала описује перцепцију засновану на кривама једнаке гласности (тзв. изофонске криве).

Карактеристике банке филтара базиране на мел, линеарној и *bark* фреквенцијској скали су приказане на слици 4.13.



Слика 4.13 – Карактеристике банке филтара за мел (a), линеарну (b) и *bark* (c) фреквенцијску скалу и 15 троугаоних филтара.

Као што је напоменуто у поглављу 2.1, због одсуства глоталних вибрација, спектар шапата је значајно равнији у односу на спектар нормалног говора. Последица тога је да шаптави говор има значајан део информација у горњем опсегу фреквенција говорног сигнала, у којем мел и *bark* скала имају слабу резолуцију (као што се може видети на слици 4.13). На тај начин, препознавачи који користе векторе обележја са мел и *bark* скалом (које "испуштају" значајан део битних спектралних компонената) значајно деградирају перформансе када се примене на шапат. С друге стране, у односу на две поменуте скале, линеарна фреквенцијска скала има добру резолуцију у горњем делу спектра говорног сигнала, али истовремено нарушава добру резолуцију у доњем делу спектра (слика 4.13b).

У потрази за оптималном фреквенцијском скалом у препознавању шаптавог говора, размотрена је могућност комбиновања мел и линеарне фреквенцијске

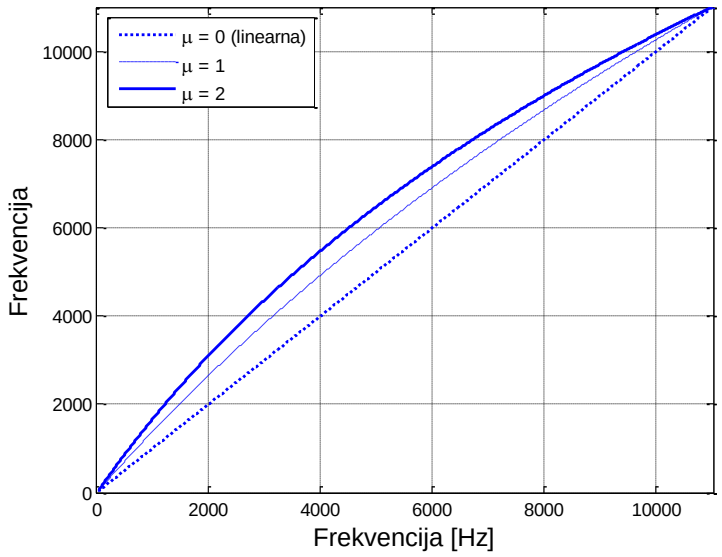
скале; добре резолуције у доњем опсегу (за мел скалу) и горњем опсегу (за линеарну скалу).

Због тога је предложена нова карактеристика фреквенцијског пресликавања (енгл. *frequency warping*) која је оригинално коришћена за компресију и експанзију говорног сигнала у дигиталним телекомуникацијама у северној Америци и Јапану (тзв. μ-закон компресије) [67].

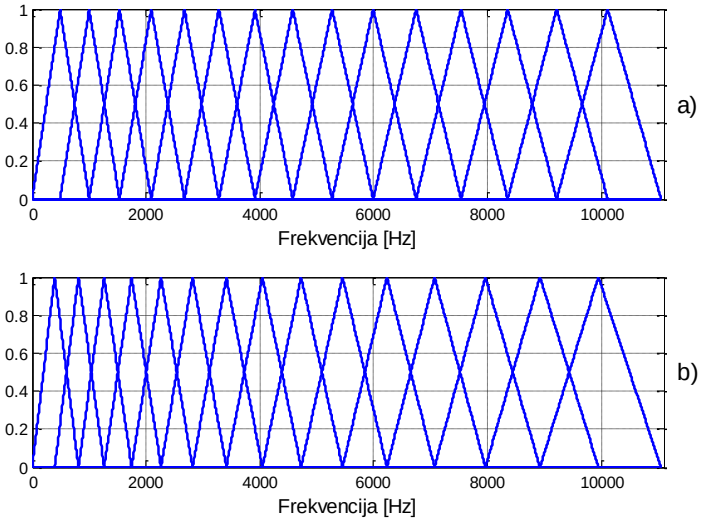
Пресликавање је дато једначином 4.4.

$$warp = f_N \frac{\ln\left(1+\mu\frac{f}{f_N}\right)}{\ln(1+\mu)} \tag{4.4}$$

При том је $f_N = f_s/2$ (f_s је фреквенција одмеравања) Никвистова фреквенција, μ позитивна реална константа, а ln означава природни логаритам. Као што је очигледно из израза 4.4, криве пресликавања секу јединичну праву за фреквенције $f = 0$ и $f = f_N$. Криве пресликавања су приказане на слици 4.14 за три вредности коефицијента $\mu \in \{0,1,2\}$, а одговарајућа карактеристика банке филтара је приказана на слици 4.15. Из облика кривих приказаних на слици 4.14 видљиво је да параметар μ одређује степен конкавности истих. Може се показати (применом *L'Hospital*овог правила) да за $\mu \to 0$ криве конвергирају ка јединичној правој $warp = f$, односно линеарној фреквенцијској скали.



Слика 4.14 – Криве мапирања према μ-фреквенцијској скали за 3 вредности коефицијента μ.



Слика 4.15 – Карактеристике банке филтара за μ-фреквенцијску скалу и 15 троугаоних филтара за вредности коефицијента μ=1 (а) и μ=2 (б).

Резолуција банке филтара према μ-закону је јасно видљива на слици 4.15. Уколико се пажљивије посматра распоред филтара на фреквенцијској скали и упореди са карактеристиком мел и линеарне скале (слика 4.13) може се приметити да скала по μ-закону има бољу резолуцију од мел скале (у горњем делу спектра) и линеарне скале (у доњем делу спектра). Постављена је хипотеза да коришћење кепстралних коефицијената са фреквенцијском скалом по μ-закону као вектора обележја може да побољша препознавање шапата. Због једноставности, ови вектори обележја су означени са μFCC.

Генерисање MFCC вектора обележја је урађено према процедури и блок шеми описаној у [46] коришћењем софтверског алата *MATLAB*. Параметри су дати у табели 4.2.

ТАБЕЛА 4.2: ВРЕДНОСТИ ПАРАМЕТАРА ЗА ЕКСТРАКЦИЈУ MFCC ВЕКТОРА ОБЕЛЕЖЈА

Параметар	Вредност
Прозоровање	Hamming
Трајање прозора	24 ms
Померај	8 ms
Број филтара	20
Преемфазис коефицијент	0.97

Параметри који су су коришћени за добијање LFCC и μFCC вектора обележја су исти. Разлика је само у томе што је коришћена друга фреквенцијска скала за мапирање.

Екстракција PLP вектора обележја је урађена према процедури описаној у [68] коришћењем доступног кода у *MATLAB*у [69]. Такође, анализирани су и PLP вектори са линеарном фреквенцијском скалом (означени са LPLP). Разлика у генерисању PLP и LPLP је само у фреквенцијској скали (остали параметри су исти).

Истраживачка група из центра за робусне системе у Даласу (САД) је у експериментима са препознавањем шаптавог говора (обученог на нормални говор) независно од говорника показала да се највећа тачност добије са модификованим PLP векторима обележја. Модификација PLP обележја се састоји у томе што су блокови за изофонске криве и рачунање интензитета (блокови 4 и 5 на слици 3.3) премоштени (изостављени). Такође, у банци филтара се користи линеарна скала, али у опсегу фреквенција од 0 до 5800 Hz [70]. Ови вектори обележја су означени са LPLP(mod) и генерисани су у *MATLAB*у модификацијом кода за PLP векторе обележја.

4.4 Развој препознавача базираног на методи потпорних вектора

Као што је описано у одељку 3.5, метода потпорних вектора је релативно једноставна техника за класификацију и препознавање узорака (енгл. *Pattern Recognition*). Метода је показала добре могућности класификације података уколико је ограничена количина расположивих података за обуку система. Пошто је то управо случај са изговорима шапата у овој дисертацији је анализирана могућност употребе SVM методе у аутоматском препознавању речи из говорне базе Whi-Spe, како зависно тако и независно од говорника.

С обзиром да SVM класификатор захтева као улазне податке векторе фиксне дужине (за разлику од HMM препознавача који је динамички) у раду су MFCC обележја из говорног сигнала издвојена на посебан начин. Могуће су две варијанте поделе говорног узорка: величина временског прозора да буде промењива и фиксна (са промењивим померајем фрејмова). Показана је приметно већа успешност у препознавању изолованих речи са прозорима промењивог трајања и одређен оптималан број временских прозора по изговору

који износи 13 (за другу говорну базу) [71]. У почетном експерименту сваки говорни сигнал из коришћеног дела говорне базе је сегментиран по својој дужини на 13 *Hamming*ових прозора са преклапањем 2/3.

Коришћена су динамичка обележја код којих се боље прате временске промене карактеристика говорног сигнала и постиже слабија корелација суседних фрејмова. Број статичких коефицијената износи 13 (са нултим коефицијентом), тако да је са делта и делта-делта обележјима укупан број коефицијената 39. То значи да је сваки узорак био представљен са вектором дужине 507 (39x13). Ти подаци су коришћени као улаз у SVM класификатор ради њихове обуке и тестирања. Због значајно веће робустности система коефицијенти су нормализовани средњом вредношћу кепстра и скалирани на опсег [-1,1].

Препознавач говора је развијен у програмском језику *Python* (верзија 3.6) [72] у интегрисаном развојном окружењу *Anaconda*.

4.5 Резиме

У аутоматском препознавању говора најчешће су заступљене 3 методе: први (и конвенционални) је статистички приступ заснован на Марковљевим моделима, други је заснован на неуронским мрежама, а трећи на методи потпорних вектора. Могућа су и хибридна решења побројаних метода. У овој дисертацији је доминантно коришћен први приступ при чему је препознавач за обуку и тестирање програмски реализован у софтверу MATLAB и користи скуп алата HTK. Софтверска реализација препознавача базираног на методи потпорних вектора је урађена у програмском језику *Python*.

У овој Глави детаљно су описани експериментална поставка и пратећи алати коришћени у три начина препознавања изолованих речи: фонема независних од контекста (монофона), фонема зависних од контекста (трифона) и целих речи. Укратко, коришћени су модели са 5 стања (од којих су 2 неемитујућа) и серијском структуром без прескока. При дигитализацији говорног сигнала коришћени су прозори дужине 24 ms са померајем 8 ms, са *Hamming*-овим прозоровањем и преемфазис коефицијентом 0,97. Број итерација у *Baum-Welch*овој реестимацији је фиксан и износи 5. При естимацији параметара модела коришћене су дијагоналне коваријансне матрице. У тестирању је коришћен Витербијев алгоритам за одређивање највероватније секвенце (путање) скривених стања, а самим тим и највероватнијег тестног изговора. Као перформанса квалитета препознавача је узет проценат успешно препознатих речи. С обзиром на слабу робустност MFCC и PLP вектора обележја у препознавању шапата, дат је предлог алгоритма за векторе обележја са кепстралним коефицијентима и модификованом фреквенцијском скалом (тзв. μ-фреквенцијска скала).

На крају Главе је дат опис експерименталне поставке препознавача базираног на методи потпорних вектора.

5. РЕЗУЛТАТИ ЕКСПЕРИМЕНАТА

Истраживање обухваћено овом дисертацијом садржи два истраживачка правца. Први правац, који обухвата ова Глава, садржи резултате експеримената у препознавању бимодалног говора (нормалног говора и шапата) уколико су на располагању у обуци изговори искључиво једног говорног мода (нормални говор *или* шапат). Дакле, могућа су 4 сценарија у препознавању:

1. нормалан/нормалан и шапат/шапат (Н/Н и Ш/Ш) - препознавач се обучава на изговорима нормалног говора или шапата и тестира се на изговорима истог говорног мода; ови сценарији се називају усаглашени.
2. нормалан/шапат и шапат/нормалан (Н/Ш и Ш/Н) - препознавач се обучава на изговорима нормалног говора или шапата и тестира на изговорима другог говорног мода; ови сценарији се називају неусаглашени.

С обзиром да подаци у говорном моду шапата често нису доступни, а до креирања великих говорних база у моду шапата може проћи и неколико година посебан изазов у савременим ASR системима представља препознавање бимодалног говора, али са обуком искључиво на нормалном говору. Због тога ће овим сценаријима (Н/Ш и Н/Н) бити посвећена посебна пажња, као и повећању робустности система обученог на нормални говор у препознавању шапата. Поред резултата успешности препознавања са HMM алгоритмом приказане су и перформансе препознавача базираног на методи потпорних вектора. Упоредна анализа успешности препознавача који користе векторе обележја са модификованом фреквенцијском скалом са MFCC и PLP обележјима је приказана на крају Главе.

Други истраживачки правац обухвата препознавање бимодалног говора уколико су у обуци на располагању изговори оба говорна мода. Резултати експеримената са мултимодном базом за обуку (зависно и независно од говорника) који обухватају упоредну анализу препознавања заснованог на измешаној бази за обуку и препознавања заснованог на класификацији говорног мода су приказани у Глави 6.

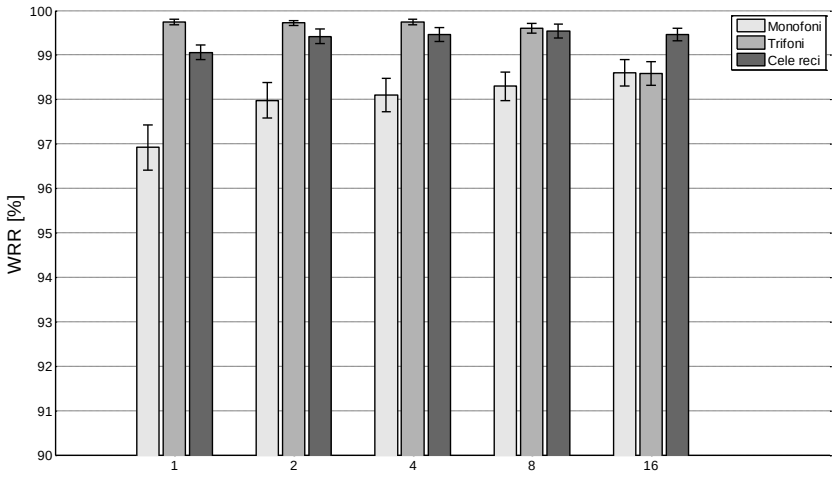
5.1 Иницијални експеримент

Циљ иницијалног експеримента је одређивање најпогодније јединице за моделовање (фонем независан од контекста, фонем зависан од контекста и цела реч) у препознавању бимодалног говора. У иницијалном експерименту су коришћени статички MFCC коефицијенти, којих је са нултим укупно 13.

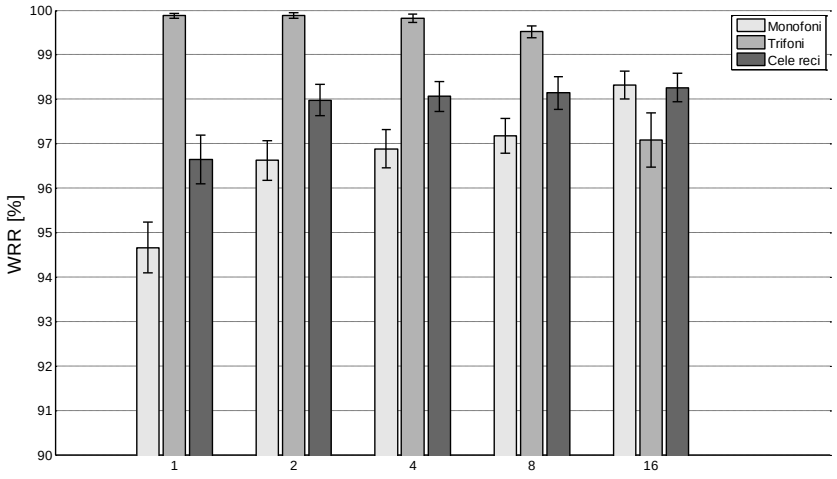
Препознавач је у потуности реализован коришћењем HTK алата. Генерисање текстуалних фајлова (конфигурациони, скрипт, фајлови за иницијализацију, клонирање и фонетску транскрипцију) је аутоматизовано коришћењем програмског пакета MATLAB. Исти је коришћен за евалуацију перформанси препознавача, коришћењем HTK алата. Број итерација у *Baum-Welch*овој реестимацији је фиксан и износи 5. Коришћена ја равномерна иницијализација глобалном средњом вредношћу и глобалном варијансом (енгл. *flat-start*). Број Гаусових компонената (мешавина) је постепено повећаван у корацима. У иницијалном експерименту је коришћена транскрипција која је коришћена у ASR системима на српском језику [73] и садржи 48 монофона. Посебно су моделовани наглашени и ненаглашени део код вокала, део за оклузију и експлозију код плозива и део за оклузију и фрикцију код африката. Такође, фонем шва је додат уколико се вибрант /р/ налази у окружењу сугласника. Модел тишине је додат на почетку и на крају сваког изговора. Експерименти су урађени у моду зависно од говорника за 5 вредности мешавина: један, два, четири, осам и шеснаест.

Ради добијања меродавније оцене препознавања зависно од говорника, у експериментима је коришћена петострука унакрсна провера (енгл. *crossvalidation*). На тај начин, у усаглашеним сценаријима део за обуку садржи 80% изговора одговарајућег говорника и мода, док се преосталих 20% користи за тестирање. Конзистентно поређење перформанси препознавача захтева да се и у

неусаглашеним сценаријима користи исти проценат расположиве базе за обуку. Стога је и у неусаглашеним сценаријима 80% изговора једног мода говора коришћено за обуку а сви изговори другог говорног мода су коришћени за тестирање. Проценат успешно препознатих речи тестираног говорника је добијен као аритметичка средина пет тестова добијених унакрсном провером. Средњи проценат успешно препознатих речи је добијен усредњавањем по говорницима. На слици 5.1 су графички приказани резултати у усаглашеним сценаријима, и то нормалног говора (а) и шапата (б). Тачне вредности су приказане табеларно у прилогу Б.



а)



б)

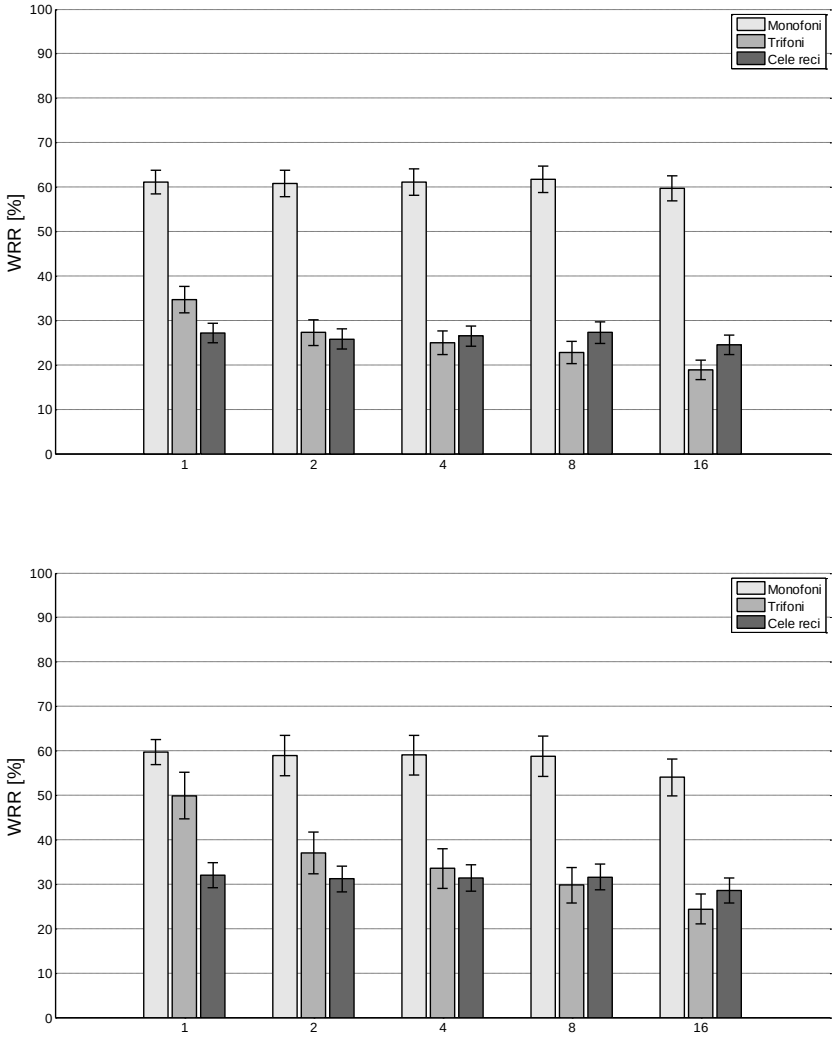
Слика 5.1 – Проценат успешно препознатих речи из говорне базе Whi-Spe у усаглашеним сценаријима за монофоне, трифоне и целе речи; (а) нормални говор и (б) шапат. На апсциси је приказан број мешавина.

Ради бољег прегледа и поређења резултата издвојен је део ординатне осе већи од 90%. Такође, приказана је и вертикална линија која представља стандардну грешку одступања од средње вредности (енгл. *standard error of the mean*).

Као што се могло и очекивати, најбољи резултати су постигнути са трифонима и то у износу 99.74% (за нормални говор, 4 мешавине) и 99.88% (за шапат, 1 мешавина). Повећање броја мешавина преко 8 резултује приметном смањењу перформанси због тога што модели трифона постају сувише специјализовани и долази до тзв. преобучавања (енгл. *overtraining*). Експерименти су показали и јако добре перформансе препознавача заснованог на моделима целих речи. Достигнута је успешност препознавања нормалног говора у износу 99.54% и шапата 98.26%. Модели монофона су допринели нешто лошијем препознавању у поређењу са моделима трифона и целих речи. И поред тога, постигнут је успех у препознавању од 98.60% и 98.32% за нормални говор и шапат, респективно. Позитиван утицај повећања броја мешавина код модела монофона се може уочити код препознавања оба мода говора. За разлику од модела трифона, препознавање шапата код модела монофона и целих речи је са приметно мањим успехом у односу на препознавање нормалног говора. То је и очекивано због специјализованости трифона и много мањег односа сигнал-шум код изговора шапата него код изговора нормалног говора.

На слици 5.2 су приказани резултати препознавања у неусаглашеним сценаријима. Успешност препознавања шапата са моделима обученим на нормални говор је приказано на слици 5.2(а) док је успех у препознавању нормалног говора са моделима обученим на шапат приказано на слици 5.2(б). У неусаглашеним сценаријима је добијен велики пад перформанси препознавача за моделе монофона и огроман пад перформанси за моделе трифона и целих речи. Успех у сценарију Н/Ш не прелази износ од 35% док је у сценарију Ш/Н максималан износ 50%, за моделе трифона. Препознавање засновано на моделима монофона је са значајно већим успехом и достиже успешност од 61.77% (у сценарију Н/Ш) и 59.68% (у сценарију Ш/Н). У неусаглашеним сценаријима је добијена и јако велика девијација успешности између појединих говорника. На пример, код модела монофона и 8 мешавина у сценарију Н/Ш успех у препознавању је у распону од 46.20% (за мушког говорника М1) до 70.36% (за

женског говорника F1). За утврђивање статистички значајних параметара који доприносе великој девијацији перформанси у неусаглашеним сценаријима потребан је већи број говорника.



Слика 5.2 – Успех у препознавању речи из говорне базе Whi-Spe у неусаглашеним сценаријима за монофоне, трифоне и целе речи; (а) сценарио Н/Ш (б) сценарио Ш/Н. На апсциси је приказан број мешавина.

Истраживања у вези са аутоматским препознавањем говорника који шапуће, са моделима обученим на нормални говор су такође показала огромну девијацију у перформансама између појединих говорника. Дотична говорна база је имала 28 говорника [66]. Разлог за такво понашање је приписан начину на који одређена особа шапуће. Утврђени су статистички значајни параметри појединачног

говорника који су у корелацији са перформансама препознавача: однос сигнал/шум, спектрални нагиб и однос енергије у фреквенцијском опсегу од 1 kHz до 2 kHz наспрам енергији у фреквенцијском опсегу од 1 kHz до 8 kHz.

И поред нешто мање успешности у усаглашеним сценаријима, због највеће робустности међу анализираним моделима, фонеми независни од контекста (монофони) су изабрани као најпогоднije јединице за моделовање. Због тога ће у експериментима који следе бити коришћени модели монофона.

5.2 Одређивање броја мешавина

У поглављу 3.3.4 је описан начин моделовања мултидимензионалних Гаусовових случајних променљивих са мешавинама. Циљ овог експеримента је одређивање броја мешавина.

Оптимални број мешавина је одређен имајући у виду да је тежиште истраживања фокусирано на препознавање бимодалног говора са моделима обученим на нормални говор. Због тога је у табели 5.1 приказана средња успешност у препознавању нормалног говора, шапата и њихова аритметичка средина, у зависности од броја мешавина (са моделима монофона).

ТАБЕЛА 5.1: УСПЕШНОСТ ПРЕПОЗНАВАЊА (У %) НОРМАЛНОГ ГОВОРА, ШАПАТА И АРИТМЕТИЧКА СРЕДИНА ЗА МОДЕЛЕ МОНОФОНА ОБУЧЕНЕ НА НОРМАЛНИ ГОВОР.

Број мешавина / Мод говора	1	2	4	8	16
Нормални говор	96,92	97,98	98,10	98,30	98,60
Шапат	61,16	60,78	61,09	61,77	59,68
Аритметичка средина	79,04	79,38	79,59	80,03	79,14

С обзиром да је аритметичка средина за осам мешавина највећа, у експериментима зависно од говорника су одабрани модели монофона са 8 Гаусових компонената (мешавина). У наредним поглављима ће бити анализирана могућност побољшања препознавања бимодалног говора коришћењем динамичких обележја, избором лексикона изговора са смањеним бројем монофона и одређивањем параметара иницијалних модела помоћу лабелираних података.

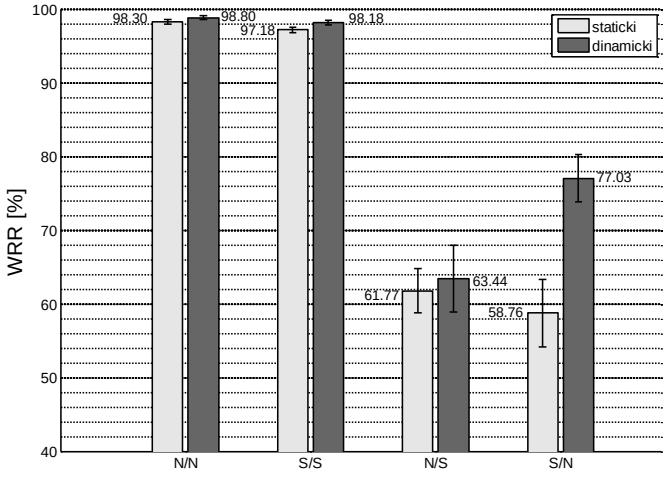
5.3 Анализа доприноса динамичких обележја

У поглављу 3.2 је описана процедура добијања динамичких вектора обележја. На слици 5.3 је графички приказана успешност препознавања у 4 сценарија за статичка и динамичка (делта-делта) обележја. Са слике се може уочити приметан допринос динамичких обележја у свим сценаријима, с тим да је у сценарију Ш/Н тај допринос јако велик. Допринос успешности препознавања динамичких обележја у односу на препознавање са статичким обележјима из иницијалног експеримента у апсолутном износу у процентима је дат у табели 5.2, за све 4 сценарија.

ТАБЕЛА 5.2: ДОПРИНОС ДИНАМИЧКИХ (ДЕЛТА-ДЕЛТА) ОБЕЛЕЖЈА У ПРОЦЕНТИМА (У АПСОЛУТНОМ ИЗНОСУ У ОДНОСУ НА ПРЕПОЗНАВАЊЕ СА СТАТИЧКИМ ОБЕЛЕЖЈИМА

Сценарио	Н/Н	Ш/Ш	Н/Ш	Ш/Н
Повећање [%]	0.50	1.00	1.67	18.27

Са слике 5.3 се може уочити да је увођење динамичких обележја резултовало асиметричношћу пероформанси између неусаглашених сценарија Н/Ш и Ш/Н (препознавање шапата са моделима обученим на нормални говор са делта-делта обележјима је скоро 14% мање него препознавање нормалног говора са моделима обученим на шаптави говор). Асиметричност перформанси у неусаглашеним сценаријима у врло блиском износу је показана и у експериментима са неуронским мрежама и DTW алгоритмом, са истом говорном базом [74], [75].



Слика 5.3 – Успех у препознавању речи из говорне базе Whi-Spe са статичким и динамичким (делта-делта) обележјима у различитим сценаријима

Због значајног побољшања у односу на статичка обележја у свим сценаријима, у наредним експериментима ће бити коришћена динамичка (делта - делта) обележја.

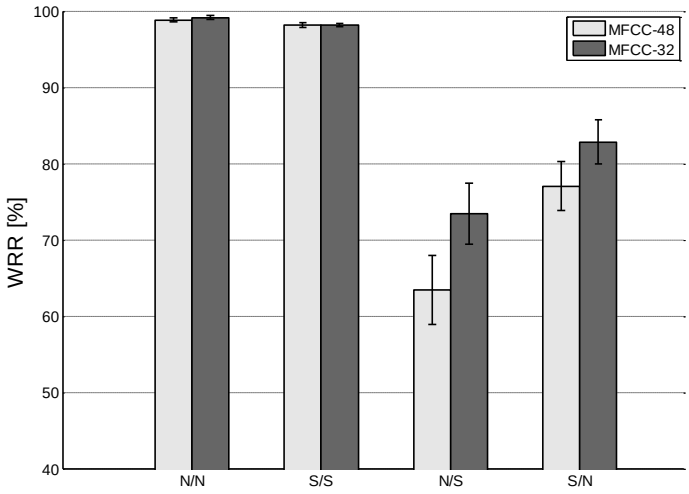
5.4 Анализа утицаја лексикона изговора

Одређен број научних радова је показао да се погодним избором лексикона изговора (енгл. *pronunciation dictionary*) могу побољшати перформансе ASR система за препознавање континуалног говора, независног од говорника [76]. Штавише, утицај је посебно изражен за морфолошки веома богате језике (у [76] је анализиран утицај на чешки и руски језик) у које свакако спада и српски језик.

С обзиром да су у истраживањима у вези са овом дисертацијом коришћени HTK алати, код моделовања плозива сви модели оклузија имају исти десни контекст, а модели експлозија имају исти леви контекст (исти је случај са моделима оклузија / фрикција код африката). Због тога што коришћена говорна база по обиму није велика (величине 2 сата) не постоји довољан број инстанци за обуку сваког посебног контекста. Промена лексикона изговора је извршена тако што су обједињени модели за оклузију и експлозију код плозива и оклузију и фрикцију код фрикатива. Такође, нису посебно коришћени модели за наглашене вокале. На тај начин, са првобитно коришћених 48 монофона, лексикон изговора је сведен на 32 монофона (30 фонема који одговарају 30 слова у азбуци српског језика, шва и модел тишине).

На слици 5.4 су приказани резултати препознавања са динамичким MFCC обележјем у све 4 сценарија, за лексикон изговора са 48 и 32 монофона.

Као што се може видети са слике 5.4, перформансе у усаглашеним сценаријима се нису значајно промениле. Препознавање нормалног говора са лексиконом изговора у редукованом обиму износи 99.14% и шапата 98.16%.



Слика 5.4 – Успех у препознавању речи из говорне базе Whi-Spe са MFCC обележјем и лексиконом изговора са 48 монофона (MFCC-48) и 32 монофона (MFCC-32). На апсциси је означен сценарио обука/тест.

С друге стране, перформансе препознавача у неусаглашеним сценаријима су се у значајној мери побољшале. У сценарију Н/Ш успешност препознавања износи 73.42% (повећана за 9.98% у апсолутном износу), док је у сценарију Ш/Н успех са износом 82.81% (повећан за 5.78% у апсолутном износу).

Следећи експерименти користе лексикон изговора са 32 монофона јер коришћењем истог перформансе препознавача се нису деградирале у усаглашеним сценаријима а значајно му је повећана робустност (успешност препознавања у неусаглашеним сценаријима).

5.5 Анализа утицаја модела за иницијализацију

За квалитетну обуку HMM модела и одређивање вероватноћа прелаза и емитовања веома је битно имати добре иницијалне параметре. На тај начин, ре-естимациони *Baum-Welch*ов поступак ће омогућити достизање глобалног максимума функције здружене вероватноће описане у одељку 3.3.3 [41].

У софтверском алату HTK постоје две могућности за иницијализацију полазних модела. Један начин је да се, уколико нису на располагању временске границе између фонема, за почетне вредности средње вредности и варијансе узму глобалне средње вредности и варијансе (равномерна иницијализација). Ако су

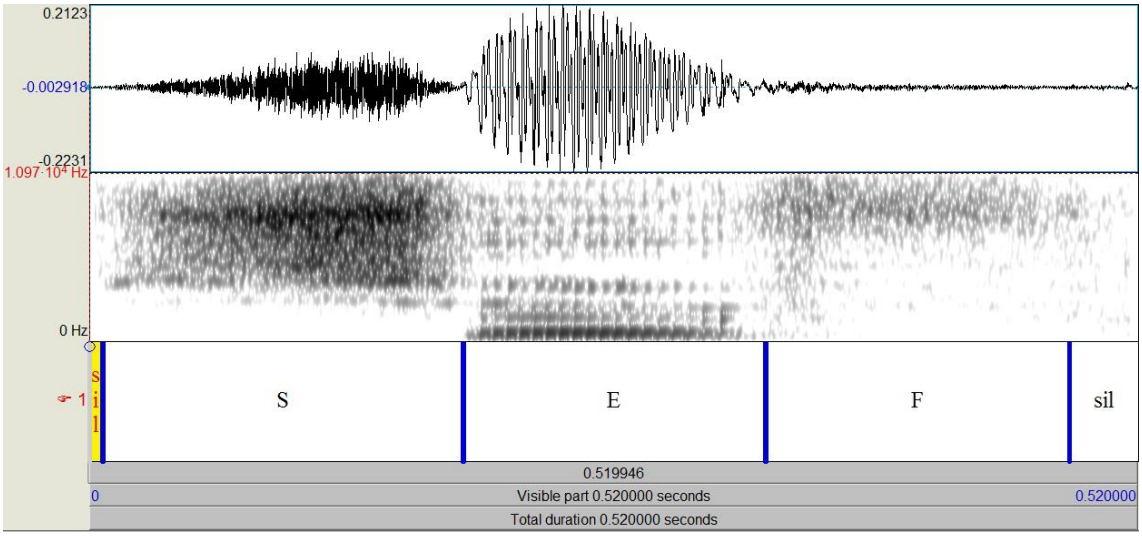
познате границе између фонема, други начин је да се коришћењем алгоритма k-средњих вредности (енгл. *k-means algorithm*) одреде почетне вредности елемената вектора средњих вредности и коваријансне матрице. На овај начин се постижу много бољи крајњи резултати [77]. С обзиром са је у свим претходним експериментима коришћена равномерна иницијализација, у овом поглављу је анализиран допринос у препознавању бимодалног говора (зависно од говорника) са обуком на нормалном говору, при чему су границе између фонема добијене на 3 начина [78]:

1. мануелном сегментацијом и лабелирањем;
2. аутоматском сегментацијом и лабелирањем који су добијени поравнањем података у обуци помоћу Витербијевог алгоритма уграђеног у HTK; и
3. аутоматском сегментацијом и лабелирањем који су добијени помоћу препознавача за препознавање континуалног говора (на српском језику) базираног на скупу алата KALDI [79].

Мануелна сегментација представља најзахтевнији део припреме за иницијализацију, бар што се тиче времена потребног за извршавање. Сегментација је вршена коришћењем софтверског пакета PRAAT [80], који је погодан за овакву врсту обраде јер постоји могућност постављања граница као и симултаног увида у таласни облик и спектрограм сегмената сигнала. У изговору су сегментирани и лабелирани и делови тишине на почетку и крају.

Комбиновањем аудитивног (слушањем), временског (увидом у таласни облик) и фреквенцијског метода (увидом у спектрограм) сегментирано је и лабелирано 10% говорне базе Whi-Spe у нормалном изговору. Прецизније, за сваког говорника сегментиран је први изговор (од укупно 10) за сваку реч у нормалној фонацији. Критеријуми постављања граница између фонема су преузети из искустава сегментације говорне базе на српском језику **S70W100S120** [81]. На крају, границе између фонема су записиване у формат текстуалног фајла потребног за HTK (у формату тзв. *mlf* фајла).

На слици 5.5 су приказани сегменти за реч /сеф/ заједно са таласним обликом и спектрограмом.



Слика 5.5 – Означени сегменти за реч /сеф/ након мануелне сегментације у прозору програмског пакета PRAAT. Са *sil* је означен део за тишину.

У овој дисертацији су поред мануелне сегментације анлизирани доприноси две технике аутоматске сегментације и лабелирања (анотације) који су добијени помоћу софтверских алата HTK и KALDI.

Сегментација и лабелирање у софтверском пакету HTK су добијени поравнавањем података за обуку, односно аудио фајлова и транскрипција на нивоу фонема. То се постиже помоћу модификованог Витербијевог алгоритма у обуци (енгл. *forced alignment*). Разлика у односу на Витервбијев алгоритам који се користи при декодовању је простор претраге [82]. Код модификованог алгоритма простор претраге се дефинише на основу транскрипција на нивоу речи, док је при декодовању простор претраге дефинисан на основу граматике или модела језика.

Сегментација помоћу софтверског пакета KALDI[2] је добијена коришћењем препознавача за српски језик публикованог у [79]. Није згорег напоменути да је препознавач искоришћен само за добијање анотираних података за део говорне базе у нормалном говору, док је за декодовање и оцену перформанси искоришћен HTK, као код мануелне анотације. Ради конзистентних услова при обуци, код аутоматске сегментације је искоришћен исти проценат говорне базе у нормалном говору (10%). Препознавач у софтверу KALDI користи модел језика базиран на

[2] аутоматско лабелирање базе Whi-Spe у KALDI сам урадио у сарадњи са др Бранославом Поповићем са Факултета техничких наука у Новом Саду на чему му се захваљујем

триграмима са 25000 гаусијана и говорну базу у трајању од 95 сати (90 сати за обуку и 5 за тест). Постиже тачност препознавања од преко 98%.

Такође, могуће је побољшање перформанси система уколико се уместо фиксног броја стања по моделу монофона користи променљиви број, који је сразмеран трајању фонема. У [82] је предложен број стања по моделу монофона из проширене листе, укључујући наглашене вокале и посебно моделоване делове за оклузију и експлозију код плозива, као и оклузије и фрикције код африката. У овој дисертацији је предложен укупан број стања (од којих су два неемитујућа) за сваки монофон, који је приказан у табели 5.3.

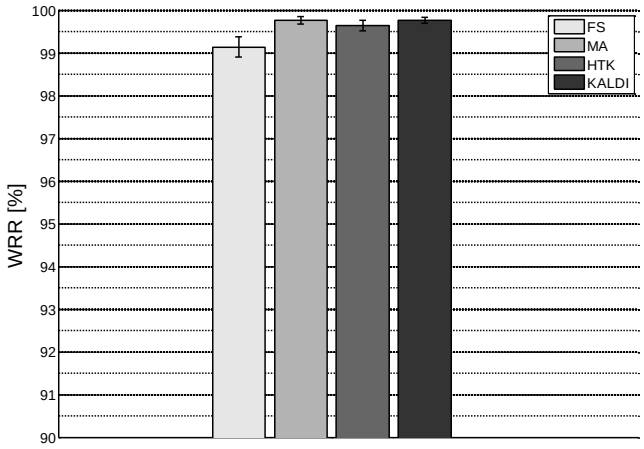
ТАБЕЛА 5.3: БРОЈ СТАЊА ПО МОДЕЛУ МОНОФОНА (ОД КОЈИХ СУ 2 НЕЕМИТУЈУЋА)

Број стања	монофон
6	/а/, /е/, /и/, /о/, /у/, /б/, /п/, /д/, /т/,/г/, /к/, /ц/, /ч/, /ћ/, /џ/, ђ/, /с/, /ш/, /з/, /ж/, /ф/, /х/, /м/, /н/, /њ/
5	/ј/, /л/, /љ/, /в/
4	/р/, /шва/
3	/sil/

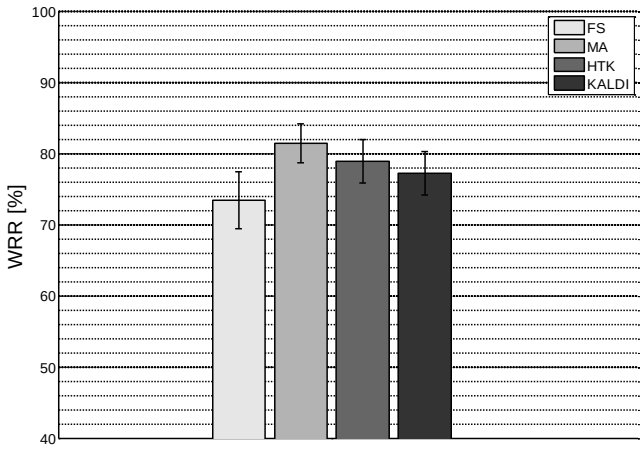
Број стања за плозиве и африкате је добијен сабирањем броја стања за делове за оклузију и експлозију (код плозива) односно делова за оклузију и фрикцију (код африката).

Резултати успешности препознавања нормалног говора и шапата (са обуком на нормалном говору) у зависности од типа одређивања параметара иницијалних модела су приказани на сликама 5.6 и 5.7, респективно. Коришћени су динамички MFCC вектори обележја и 4 типа иницијализације:

- равномерна иницијализација (означена са FS);
- иницијализација коришћењем мануелне анотације (означена са MA);
- иницијализација коришћењем аутоматске анотације добијене помоћу софтвера HTK (означена са HTK); и
- иницијализација коришћењем аутоматске анотације добијене помоћу софтвера KALDI (означена са KALDI).



Слика 5.6 – Успех у препознавању нормалног говора (сценарио Н/Н) у зависности од типа иницијализације



Слика 5.7 – Успех у препознавању шапата (сценарио Н/Ш) у зависности од типа иницијализације

Увидом у резултате успешности препознавања који су приказани на слици 5.6 могу се извести следећи закључци:

- Без обзира на тип иницијализације (са мануелном или аутоматском анотацијом) успешност у препознавању нормалног говора је побољшана за бар пола процента и износи преко 99.60%.
- Због ефекта засићења (енгл. *ceiling effect*) не може се веродостојно тврдити који тип иницијализације даје најбоље резултате.

Такође, резултати који су графички приказани на слици 5.7 као и услови при којима су добијени омогућују нам да констатујемо:

- Допринос успешности препознавања шапата је приметан за сваки тип иницијализације.
- Највећи допринос успешности препознавања шапата (у односу на препознавање са равномерном иницијализацијом) је остварен са мануелном анотацијом (7.96%) у износу од 81.38%.

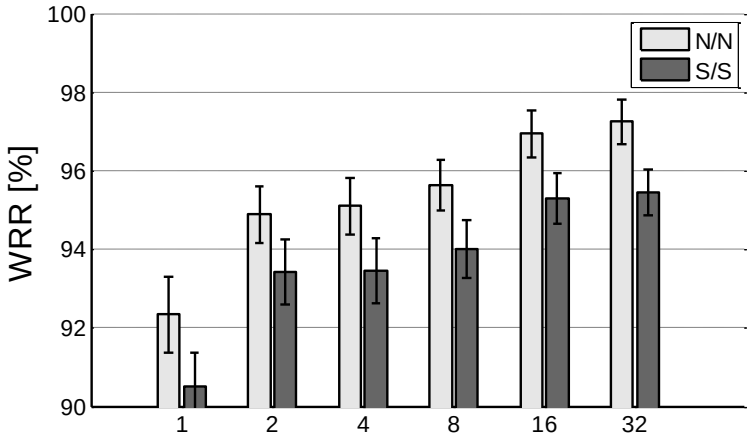
5.6 Препознавање бимодалног говора независно од говорника

Препознавање зависно од говорника захтева да се за тестираног говорника обезбеди скуп података за обуку у значајном обиму, што је ограничавајући фактор за практичне апликације. По дефиницији, препознавање независно од говорника подразумева потпуно одсуство изговора тестираног говорника из базе за обуку. За исту (или упоредиву) количину података у обуци, препознавање зависно од говорника има боље перформансе. Циљ овог поглавља је испитивање перформанси препознавања речи из базе Whi-Spe независно од говорника, како у усаглашеним тако и у неусаглашеним сценаријима. У свим експериментима су коришћени модели фонема независни од контекста (са 32 монофона) и MFCC делта-делта вектори обележја (укупно 39 коефицијената).

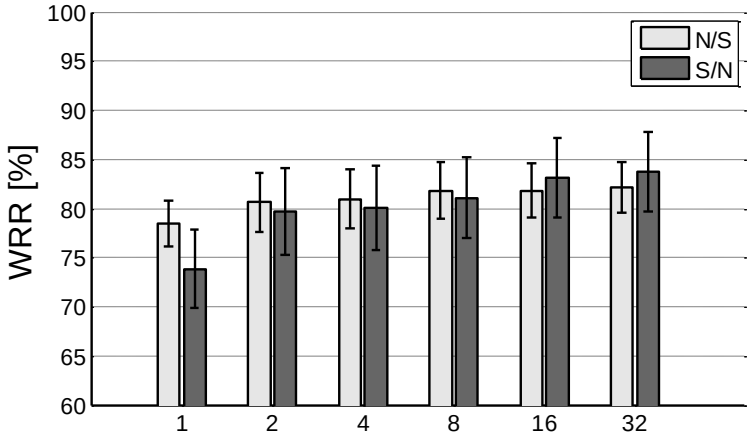
5.6.1 Одређивање броја мешавина

У почетном експерименту је одређен проценат успешно препознатих речи (WRR) у 4 сценарија. Експерименти су урађени за 6 вредности броја Гаусових компонената (1, 2, 4, 8, 16 и 32) при чему је база за обуку искоришћена у пуном капацитету. То значи да су сви изговори свих говорника(-ца) осим тестираног у обуци (енгл. *leave-one-out crossvalidation*). Као оцена перформанси је узета средња вредност успеха препознавања међу говорницима. На сликама 5.8 и 5.9 приказани су резултати у усаглашеним и неусаглашеним сценаријима, респективно.

Као што је било за очекивати, утицај повећања броја мешавина у усаглашеним сценаријима је као код препознавања зависно од говорника, односно доприноси побољшању перформанси. Код препознавања нормалног говора највећа успешност износи 97.24% док је најбољи успех у препознавању шапата 95.44% (за 32 мешавине у оба сценарија).



Слика 5.8 – Успех у препознавању говора у усаглашеним сценаријима независно од говорника. На апсциси је број мешавина.



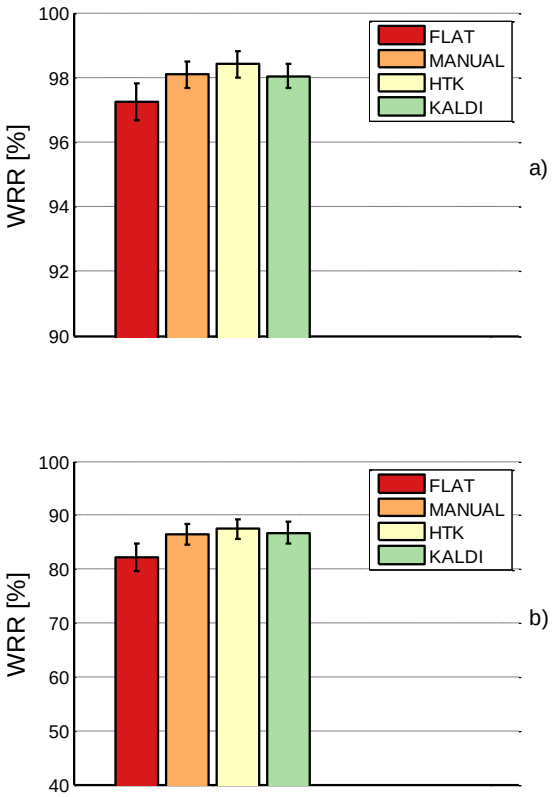
Слика 5.9 – Успех у препознавању говора у неусаглашеним сценаријима независно од говорника. На апсциси је број мешавина.

Као што се види са слике 5.9, препознавање у неусаглашеним сценаријима је значајно деградирало перформансе у односу на препознавање у усаглашеним сценаријима. У оба сценарија је највећи успех постигнут за 32 мешавине и то у износу 82.14% (сценарио Н/Ш), односно 83.70% (сценарио Ш/Н). Као што је већ речено, оптимизација препознавача се своди на оптимизацију препознавања бимодалног говора са обуком на нормалном говору. Пошто је у оба сценарија са обуком на нормалном говору (Н/Н и Н/Ш) највећа успешност постигнута за 32 мешавине, у наредним експериментима независно од говорника коришћен је тај број мешавина.

5.6.2 Анализа утицаја модела за иницијализацију

Као што је анализирано за случај препознавања зависно од говорника (поглавље 5.5), анализа утицаја модела за иницијализацију је извршена и за случај препознавања независно од говорника. Аналогно, одређена је успешност препознавања нормалног говора и шапата (са обуком на нормалном говору) за 4 типа иницијализације као и у случају зависно од говорника. На слици 5.10(a) је приказана успешност у препознавању нормалног говора, док је на слици 5.10(b) приказана успешност препознавања шапата.

Као што се може видети на слици 5.10, експерименти су показали допринос успешности препознавања бимодалног говора и у моду независно од говорника. Уколико се параметри иницијалних модела одређују помоћу анотираног дела базе у нормалном говору коришћењем променљивог броја стања по моделу монофона, повећање процента препознавања бимодалног говора у зависности од типа иниицијализације (у односу на равномерну иницијализацију) је приказано у табели 5.4.



Слика 5.10 – Успех у препознавању (WRR са стандардном грешком) нормалног говора (a) и шапата (b) у зависности од типа иницијализације код препознавања независно од говорника.

ТАБЕЛА 5.4: ПОВЕЋАЊЕ ПРОЦЕНТА (У АПСОЛУТНОМ ИЗНОСУ) ПРЕПОЗНАВАЊА НЕЗАВИСНО ОД ГОВОРНИКА У ЗАВИСНОСТИ ОД НАЧИНА АНОТАЦИЈЕ

Начин анотације/ Сценарио	Мануелна	Аутоматска (HTK)	Аутоматска (KALDI)
Н/Н	0.84%	1.16%	0.80%
Н/Ш	4.34%	5.28%	4.60%

Нумеричке вредности приказане у табели 5.6 показују знатан допринос побољшању препознавања оба мода говора са било којим начином иницијализације. Код препознавања нормалног говора највећи допринос успешности је постигнут са аутоматском анотацијом урађеном у HTK, са средњим процентом успешности од 98.40%. С обзиром да је успешност препознавања нормалног говора са равномерном иницијализацијом 97.24% (на слици 5.10(а) означено са FLAT) повећање успешности износи 1.16%. Разлика у успешности препознавања између мануелне анотације и аутоматске анотације добијене у софтверу KALDI није значајна.

Такође, у значајном износу је повећана и успешност препознавања шапата. За разлику од препознавања зависно од говорника код које је мануелна анотација резултовала највећем побољшању, код препознавања независно од говорника највећи допринос је са аутоматском анотацијом. Као у препознавању нормалног говора, опет је анотација урађена у HTK имала највећи допринос побољшању, са средњим WRR у износу од 87.42% (повећање у апсолутном износу 5.28%) [78].

5.7 Препознавање базирано на методи потпорних вектора

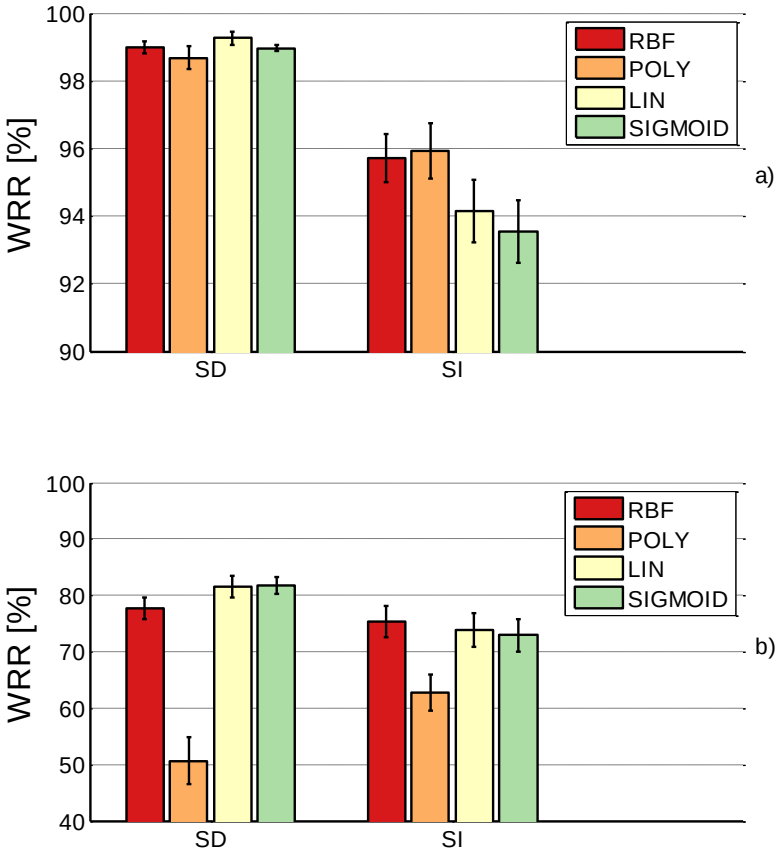
Експерименти су урађени у моду зависно (SD) и независно (SI) од говорника. Анализирана је успешност препознавања нормалног говора и шапата (са обуком на нормалном говору) за 4 типа кернела:

- *RBF* кернел
- полиномијални кернел (са степеном $d=3$)
- линеарни кернел
- сигмоидни кернел

Резултати успешности су приказани на слици 5.11.

Као што се може видети са слике 5.11(а), успешност препознавања нормалног говора је приметно мања у односу на HMM препознавач (слике 5.6 и 5.10). У SD случају највећи проценат успешности је добијен за линеарни кернел (99.26%). У

SI случају пад перформанси у односу на HMM (препознавач) је још већи са највећом успешношћу у износу 95.93% за полиномијални кернел.



Слика 5.11 – Успех у препознавању (WRR са стандардном грешком) нормалног говора (a) и шапата (b) зависно (SD) и независно од говорника (SI) у зависности од типа кернела; RBF (RBF), POLY (полиномијални), LIN (линеарни) и SIGMOID (сигмоидни).

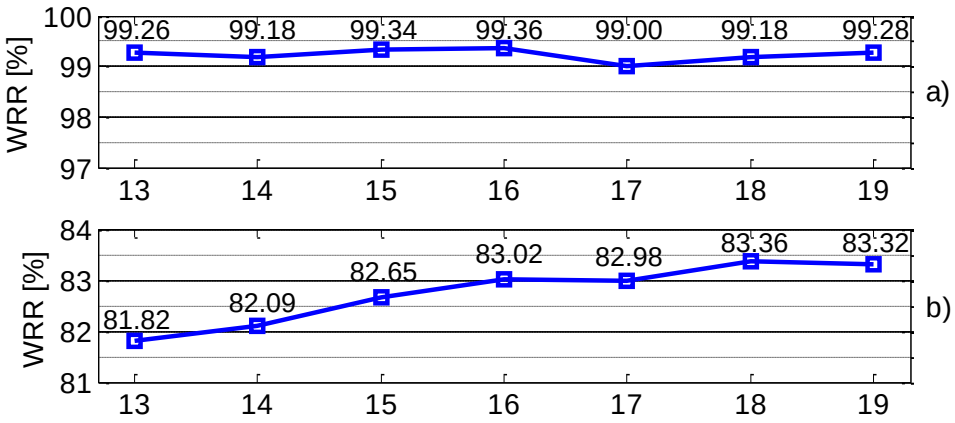
Експерименти у препознавању шапата чији су резултати приказани на слици 5.11(b) су показали да је у SD случају (зависно од говорника) највећа успешност (81.82%) добијена за сигмоидни кернел. Ипак, разлика у односу на линеарни кернел је занемарљива. Приметно мањи успех је постигнут са RBF кернелом док је полиномијални кернел практично неупотребљив.

У препознавању независно од говорника највећи WRR је добијен са RBF кернелом у износу од 75.29%. Узимајући у обзир да је највећи постигнути WRR са HMM препознавачем 87.42% (слика 5.10b) може се констатовати да је примена SVM препознавача у препознавању шапата (N/W сценарио) ограничена на случај зависно од говорника. У том случају SVM препознавач даје упоредиве резултате (у одређеним случајевима чак и боље) у односу на HMM препознавач.

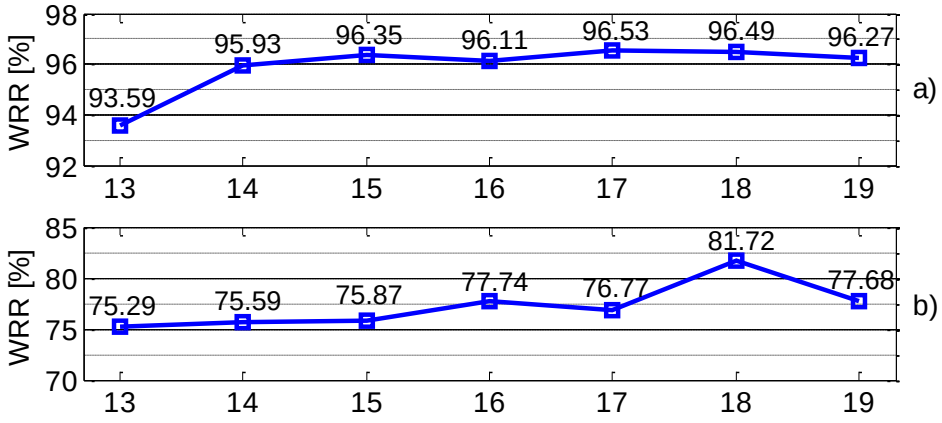
У поглављу 4.4 је напоменуто да је сегментација изговора извршена на 13 преклапајућих прозора (фрејмова). Опсег фонема по изговору речи из говорне базе Whi-Spe је у распону од 3 до 13 (просечна вреднсост је 5.58 јер су дуже речи веома ретке). Коришћење 13 фрејмова по изговору даје 1 фрејм по фонему за најдуже речи, односно 2 до 3 фрејма по фонему за краће речи.

Даље је извршена анализа утицаја броја фрејмова (у распону од 13 до 19) у препознавању речи из говорне базе Whi-Spe како зависно тако и независно од говорника [78].

Резултати су графички приказани на слици 5.12 (зависно од говорника) и 5.13 (независно од говорника).



Слика 5.12 – Успех у препознавању (WRR) зависно од говорника у препознавању нормалног говора (a) и шапата (b) за SVM препознавач у зависности од броја прозора.



Слика 5.13 – Успех у препознавању (WRR) независно од говорника у препознавању нормалног говора (a) и шапата (b) за SVM препознавач у зависности од броја прозора.

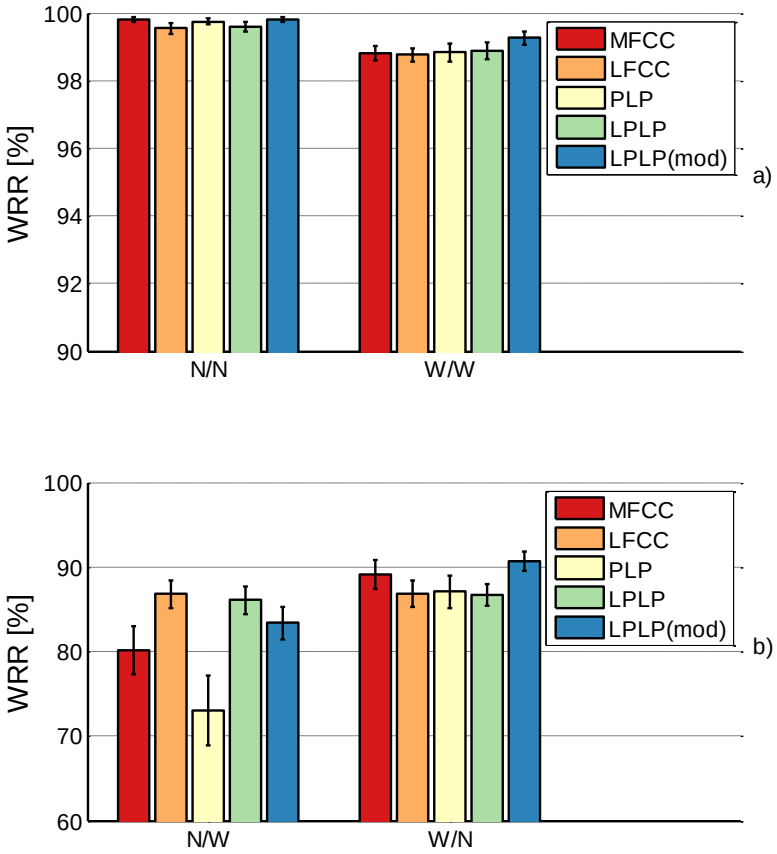
Експерименти су показали да допринос броја фрејмова у препознавању нормалног говора у SD случају није статистички значајан (слика 5.12a). С друге стране, препознавање шапата је повећано за додатних 1.54% у случају 18 преклапајућих фрејмова по изговору (слика 5.12b) са просечним WRR у износу од 83.36%.

Анализа у SI случају је показала допринос у препознавању нормалног говора за додатних 3% (приближно) када се користи сегментација изговора на 17 и 18 преклапајућих прозора. Истовремено, повећано је и препознавање шапата за 6.43%, при чему се као и у SD случају показао најуспешнијим број преклапајућих прозора 18.

5.8 Препознавање мултимодалног говора са кепстралним коефицијентима и модификованом фреквенцијском скалом

Коришћени су динамички вектори (статички + Δ + ΔΔ) којих је укупно 39. Извршена је нормализација средњом вредношћу. Систем за препознавање, који је развијен у *MATLAB*у коришћењем HTK алата је описан у поглављу 4.3, са аутоматском иницијализацијом модела урађеној у HTK. Урађена је десетострука кросвалидација. У иницијалном експерименту су анализиране перформансе препознавача за 5 побројаних вектора обележја (MFCC, LFCC, PLP, LPLP и LPLP(mod)). За препознавање зависно од говорника резултати су приказани на слици 5.14.

Ради боље прегледности, у усаглашеним сценаријима је издвојен део ординатне осе изнад 90%, а у неусаглашеним изнад 60%. Као што се могло и очекивати, у усаглашеним сценаријима препознавање шапата је са мањим успехом у односу на препознавање нормалног говора. За све анализиране векторе обележја препознавање нормалног говора је изнад 99.5%, с тим да разлика између појединих обележја није веома изражена. С друге стране, препознавање шапата је са приметним успехом веће за модификоване PLP векторе обележја са оствареним средњим процентом успешно препознатих речи (WRR) у износу од 99.26%.



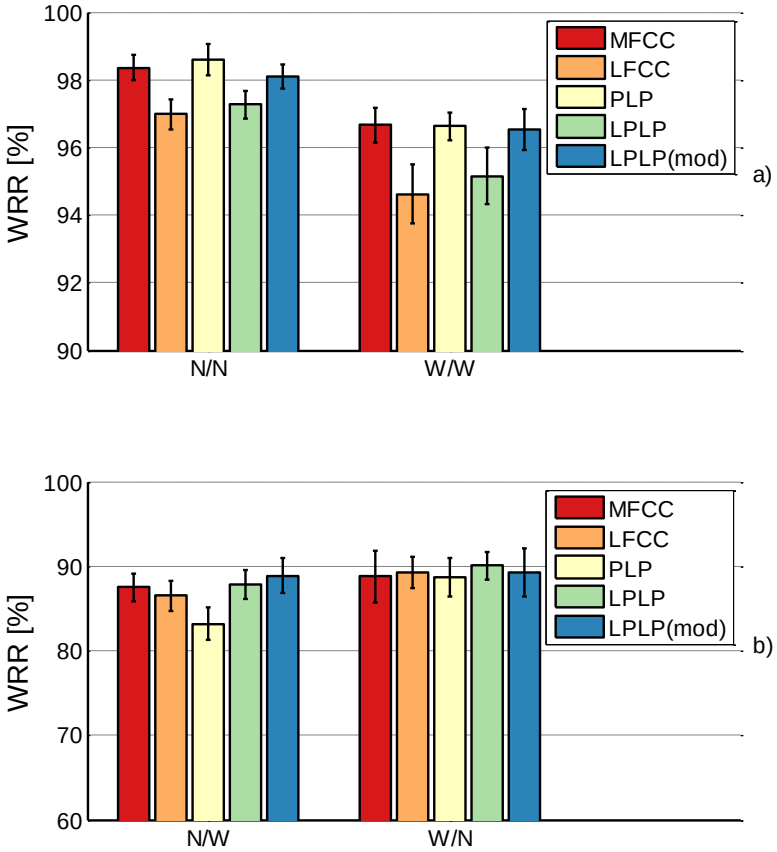
Слика 5.14 – Успех у препознавању зависно од говорника (WRR са стандардном грешком) у усаглашеним (нормалан/нормалан и шапат/шапат) (a) и неусаглашеним (нормалан/шапат и шапат/нормалан) (b) сценаријима у зависности од вектора обележја.

У неусаглашеним сценаријима је изражена асиметрија између N/W и W/N сценарија осим за обележја са линеарном скалом LFCC и LPLP. С обзиром на већ речену посебну значајност N/W сценарија, највећи успех је постигнут за LFCC обележје у износу од 86.80%.

Резултати препознавања независно од говорника су приказани на слици 5.15.

Као што се може видети на слици 5.15(a) перформансе система у усаглашеним сценаријима су значајно деградиране за LFCC и LPLP векторе обележја. Овај резултат иде у прилог тези да увођење линеарне скале може да у одређеним неусаглашеним сценаријима повећа робустност ASR система, али да у условима у којима тестни услови одговарају условима при обуци стандардни MFCC и PLP векторе дају боље резултате од LFCC и LPLP вектора обележја, респективно. У препознавању нормалног говора пад перформанси за LFCC у односу на MFCC је 1.38%, док је пад успешности LPLP у односу на PLP 1.34%. Највећа успешност у

препознавању нормалног говора је добијена за PLP векторе у износу од 98.60%. У препознавању шапата пад перформанси је 2.04% (LFCC наспрам MFCC) и 1.46% (LPLP наспрам PLP). Највећи WRR је добијен за MFCC обележје (96.66%), са неприметно мањим успехом за PLP обележје (96.62%).



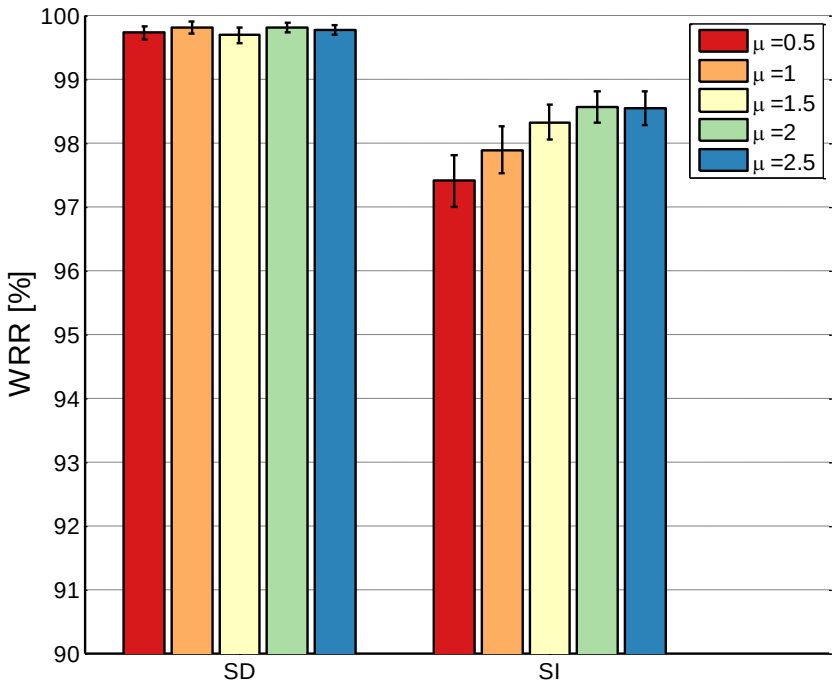
Слика 5.15 – Успех у препознавању независно од говорника (WRR са стандардном грешком) у усаглашеним (нормалан/нормалан и шапат/шапат) (a) и неусаглашеним (нормалан/шапат и шапат/нормалан) (b) сценаријима у зависности од вектора обележја.

У N/W сценарију је добијена највећа успешност са PLP(mod) обележјима (као и у истраживању [23]) са постигнутим WRR у износу од 88.86%.

5.8.1 Резултати експеримената са кепстралним коефицијентима и модификованом фреквенцијском скалом

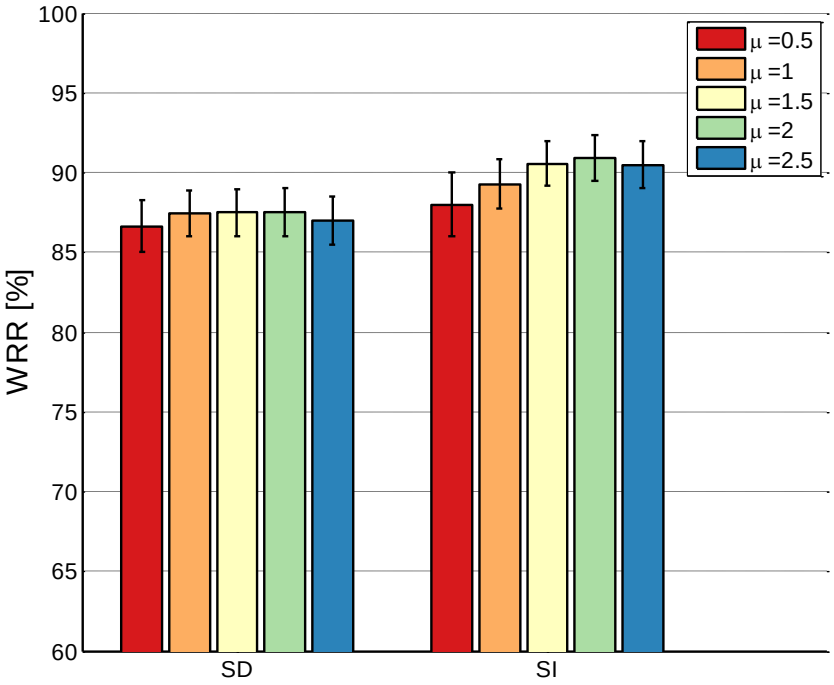
Експерименти су урађени за 5 вредности параметра μ у интервалу од 0.5 до 2.5 (са кораком 0.5), како зависно тако и независно од говорника [83]. За препознавач

обучен на нормални говор резултати су графички приказани за препознавање нормалног говора (слика 5.16) и шапата (слика 5.17).



Слика 5.16 – Успех у препознавању нормалног говора (WRR са стандардном грешком) зависно од говорника (SD) и независно од говорника (SI) са μFCC кепстралним коефицијентима у зависности од параметра μ.

Као што се може видети на слици 5.16, параметар μ нема значајан утицај на перформансе прпеознавача зависно од говорника (SD). За све анализиране вредности, успешност препознавања је преко 99.60%. Истовремено, највећа успешност у препознавању независно од говорника је добијена за вредност коефицијента μ=2 у износу од 98.56%.



Слика 5.17 – Успех у препознавању шапата (WRR са стандардном грешком) зависно од говорника (SD) и независно од говорника (SI) са μFCC кепстралним коефицијентима у зависности од параметра μ.

Резултати експеримената у препознавању шапата (слика 5.17) показују врло сличну тенденцију у погледу утицаја параметра μ на перформансе препознавача. У препознавању независно од говорника (SI) највећа успешност је опет добијена за вредност коефицијента μ=2 у износу од 90.92%. У препознавању зависно од говорника (SD) веома је мала разлика у перформансама препознавача за вредности коефицијента од μ=1 до μ=2, са максимално постигнутим WRR у износу од 87.50% за вредност μ=2.

Да бисмо потврдили ефикасност коришћења μFCC вектора обележја у препознавању шапата потребни су статистички тестови. Двострани *Wilcoxon* тестови (енгл. *two-tailed signrank test*) су потврдили да је побољшање са кепстралним коефицијентима који користе μ-фреквенцијску скалу статистички значајно. У табели 5.5 је приказан WRR за μFCC векторе обележја (за вредност μ=2) као и за 5 вектора обележја који су анализирани у иницијалном експерименту. Одговарајући опсег p-вредности је обележен звездицом.

ТАБЕЛА 5.5:ПРОСЕЧНИ WRR ЗА РАЗЛИЧИТЕ ВЕКТОРЕ ОБЕЛЕЖЈА У N/W СЦЕНАРИЈУ

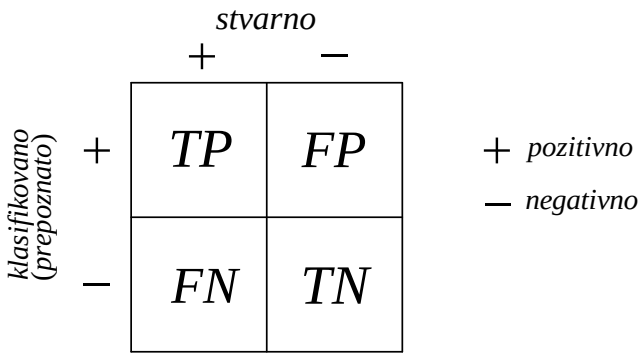
Обележје	WRR [%]	
	Зависно од говорника	**Независно од говорника**
μFCC	87.50	90.92
MFCC	80.14**	87.48**
LFCC	86.80**	86.50**
PLP	73.07**	83.16**
LPLP	86.05**	87.80**
LPLP(mod)	83.38**	88.86*

(p<0.05 *; p<0.005 **; Интервал поверења = 95%)

Као што се види из табеле 5.5, тестови су показали статистичку значајност побољшања μFCC вектора обележја у односу на свих 5 анализираних вектора обележја. И поред тога што је повећање успешности μFCC вектора у односу на LFCC векторе свега 0.70% *Wilcoxon* статистички тест је показао висок ниво значајности са p<0.005.

У свим досадашњим експериментима, метрика за опис квалитета ASR система је проценат успешно препознатих речи (WRR). Често се као метрике оцене ASR система приказују прецизност (енгл. *Precision*) и одзив (енгл. *Recall*) ASR система, односно њихова хармонијска средина, као мера компромиса између прецизности и одзива.

Нека је дат класификатор са матрицом конфузије на слици 5.18 (енгл. *Matrix confusion*).



Слика 5.18 – Матрица конфузије класификатора

При том је:

- TP – број тачно препознатих позитивних узорака
- TN – број тачно препознатих негативних узорака

- FP – број погрешно (нетачно) препознатих позитивних узорака
- FN – број погрешно (нетачно) препознатих негативних узорака

Прецизност и одзив се дефинишу изразима 5.4 и 5.5, респективно [84].

$$P=\frac{TP}{TP+FP} \tag{5.4}$$

$$R=\frac{TP}{TP+FN} \tag{5.5}$$

Из израза 5.4 се може видети да је прецизност у ствари удео тачно класификованих узорака у скупу позитивно класификованих узорака.

Такође, из израза 5.5 се може видети да је одзив (негде се зове и осетљивост) удео тачно класификованих узорака у скупу свих позитивних узорака.

Пошто су захтеви за великом презизношћу и одзивом опречни, као компромис се узима њихова хармонијска средина, дата изразом 5.6.

$$F=\frac{1}{\frac{\alpha}{P}+\frac{1-\alpha}{R}} \tag{5.6}$$

При том је α тежински фактор који је између 0 и 1 и зависи од тога да ли нам је битнија прецизност или одзив (за $\alpha=0$ имамо одзив а за $\alpha=1$ имамо прецизност). Типично се узима тзв. балансирана F-мера (која се често означава са F_1) за вредност $\alpha=0.5$ и дата је изразом 5.7 [85].

$$F_1=\frac{2PR}{P+R} \tag{5.7}$$

У N/W сценарију за сваку реч из говорне базе Whi-Spe су анализом излазног HTK фајла одређени прецизност и одзив, а потом израчуната балансирана F-мера која је усредњена за сваког говорника (у SD и SI случају). Резултати су приказани у табели 5.6. Говорници означени од 1 до 5 су мушки говорници а говорници од 6 до 10 су женски говорници.

ТАБЕЛА 5.6:БАЛАНСИРАНА F-МЕРА ЗА СВЕ ГОВОРНИКЕ У N/W СЦЕНАРИЈУ

	SD		SI	
Говорник	**MFCC**	**μFCC**	**MFCC**	**μFCC**
Говорник 1	0.9763	0.9695	0.9053	0.9453
Говорник 2	0.7629	0.8667	0.8630	0.8899
Говорник 3	0.8631	0.9093	0.8592	0.9186
Говорник 4	0.7324	0.8495	0.8449	0.9290
Говорник 5	0.7703	0.8227	0.7750	0.8199
Говорник 6	0.7274	0.8352	0.9444	0.9596
Говорник 7	0.8636	0.9119	0.9311	0.9438
Говорник 8	0.8273	0.8871	0.9130	0.9352
Говорник 9	0.8108	0.9152	0.9054	0.9497
Говорник 10	0.8560	0.9130	0.8895	0.8992
Просек	0.8190	0.8880	0.8831	0.9190

Добијени резултати показују да је за μFCC обележја добијена већа и балансирана F-мера, у односу на MFCC обележја. У случају зависно од говорника, F-мера је већа за приближно 7% (у апсолутном износу) и око 3.5% за случај независно од говорника.

5.9 Резиме

У овој глави су приказани резултати низа експеримената у препознавању нормалног говора и шапата у различитим обука/тест сценаријима. Иницијални експеримент за HMM-препознавач је показао најбоље перформансе препознавача за моделе фонема независних од контекста (монофона) у погледу компромиса између тачности у усаглашеним сценаријима и робустности. И поред тога што су модели фонема зависних од контекста (трифона) и целих речи били нешто успешнији у усаглашеним сценаријима, огроман пад перформанси у неусаглашеним сценаријима за трифоне и целе речи је био довољан разлог да се за препознавање бимодалног говора изаберу монофони, код којих је деградација перформанси значајно мања. Код препознавања бимодалног говора међу анализираним бројем мешавина, најбољи резултати су добијени за 8 мешавина (зависно од говорника) и 32 мешавине (независно од говорника).

Наредни екперименти су показали да се увођење динамичких вектора обележја, уз коришћење MFCC вектора обележја и модела језика са 32 монофона перформансе препознавача зависно од говорника могу значајно побољшати. У односу на резултате у иницијалном експерименту, кумулативни допринос (у апсолутном износу) побројаних техника је дат у табели 5.7, за све 4 сценарија.

ТАБЕЛА 5.7: ПОВЕЋАЊЕ ПРОЦЕНТА (У АПСОЛУТНОМ ИЗНОСУ) ПРЕПОЗНАВАЊА У ЗАВИСНОСТИ ОД СЦЕНАРИЈА

Сценарио	Н/Н	Ш/Ш	Н/Ш	Ш/Н
Повећање успешности	0,84%	0,98%	11,65%	24,05%

Даље је анализиран допринос иницијализације модела коришћењем мануелне и аутоматске анотације у односу на равномерну иницијализацију (за коју нису потребне границе између фонема). Највећи допринос код препознавања зависно од говорника је добијен за мануелну анотацију (око 0.5% за нормални говор и 8% за шапат), као и аутоматску анотацију код препознавања независно од говорника (око 1.2% за нормални говор и 5.3% за шапат).

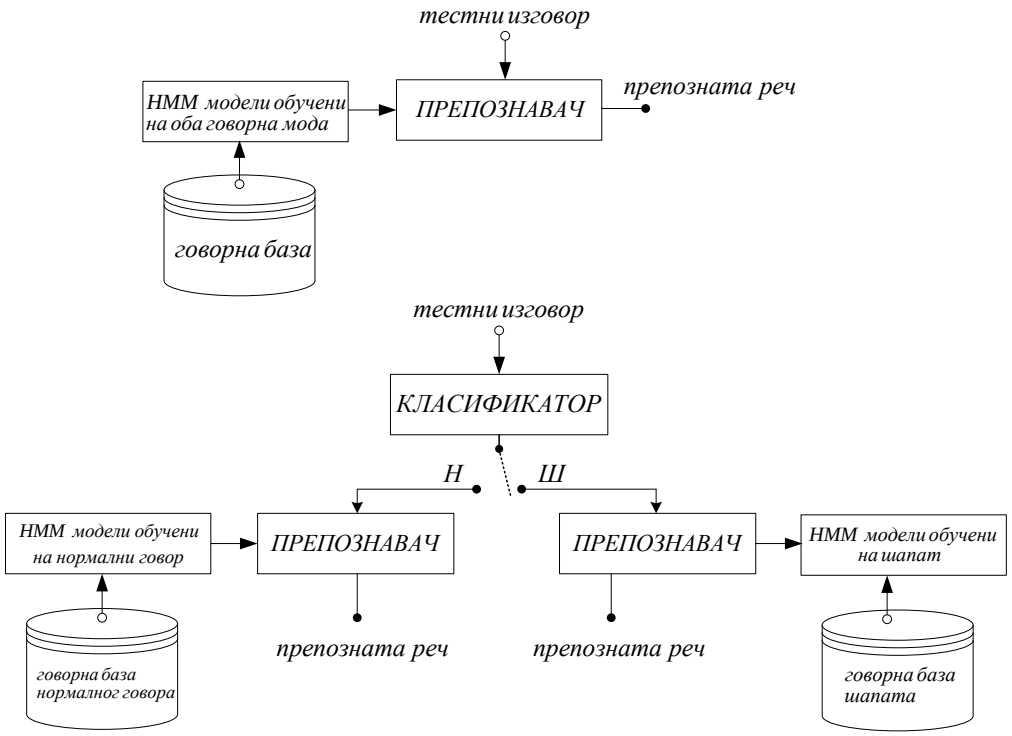
Урађени су и експерименти за препознавач нормалног говора и шапата базиран на методи потпорних вектора (SVM-препознавач) који је статички препознавач, за разлику од HMM-препознавача. У препознавању независно од

говорника, HMM-препознавач је по питању успешности значајно премашио SVM-препознавач. За случај зависно од говорника и моделе обучене на нормални говор, SVM-препознавач је показао боље перформансе у препознавању шапата, док је HMM-препознавач успешнији у препознавању нормалног говора.

На крају Главе су приказани резултати експеримената са предложеним векторима обележја са μ-фреквенцијском скалом. Експерименти су показали приметно побољшање перформанси у препознавању шапата док су истовремено задржане добре перформансе у препознавању нормалног вовора. Статистички тестови су потврдили побољшање успешности перформанси препознавача са новим векторима обележја, за препознавање зависно и независно од говорника.

6. РЕЗУЛТАТИ ЕКСПЕРИМЕНАТА СА МУЛТИМОДНОМ БАЗОМ ЗА ОБУКУ

Тежиште другог истраживачког правца који обухвата ова дисертација је упоредна анализа перформанси препознавача са измешаном базом за обуку (ИБ) и класификацијом говорног мода, зависно и независно од говорника. Оба приступа су заснована на претпоставци да су за обуку на располагању изговори оба говорна мода (нормални говор и шапат). Као што је шематски приказано на слици 6.1а препознавач са ИБ користи HMM моделе добијене обуком у којој симултано учествују изговори оба говорна мода [86].



Слика 6.1 – Блок шема препознавача са измешаном базом за обуку (а) и препознавача базираног на класификацији говорног мода (б)

За разлику од тог приступа, препознавање бимодалног говора базирано на класификацији (слика 6.1б) користи класификатор који одлучује о говорном моду тестног изговора. Уколико је реч о моду нормалног говора, препознавач користи HMM моделе обучене на нормални говор, у супротном користи моделе обучене на шапат. Наравно, перформансе препознавача су поред квалитета препознавача много зависне и од тачности класификације класификатора. Због тога ће посебна пажња бити посвећена избору класификатора говорног мода.

С обзиром да је у предобради извршена нормализација средњом вредношћу кепстра за цели изговор, сви класификатори су на нивоу изговора (а не фрејма). Значи, одлучивање се врши након обраде целог изговора (фајла који садржи звучни запис). Одлучивање класификатора се постиже минимизацијом грешке погрешне класификације. У програмском пакету MATLAB класификација је урађена коришћењем наредбе *classify*.

6.1 Избор параметра за класификацију говорног мода

У одељку 2.1 наведени су говорни модови: шапат, тихи говор, говор уобичајеног интензитета (нормални говор), гласни говор и вика. Основни параметри за дистинкцију говорних модова су ниво звука, трајање изговора и заступљеност тишине, расподела енергије по фрејмовима и спектрални нагиб. У [30] је анализирана расподела нултог кепстралног коефицијента c_0 (који репрезентује енергију) и првог кепстралног коефицијента c_1 (који репрезентује спектрални нагиб) за део говорне базе Whi-Spe. За c_0 показана је приближно иста вршна вредност нормализоване расподеле за нормални говор и шапат, што је последица различите удаљености микрофона од уста говорника при снимању. Због тога се енергија не може користити као параметар за класификацију говорног мода, али може однос сигнал/шум. Поред тога, дотично истраживање је показало значајну дислоцираност расподеле коефицијента c_1 за шаптави говор у односу на нормални говор. Узимајући у обзир чињеницу да је шапат у потупности девокализован још је анализиран класификатор базиран на основној фреквенцији говорног сигнала. У наставку ће бити анализирана успешност класификације сваког параметра понаособ.

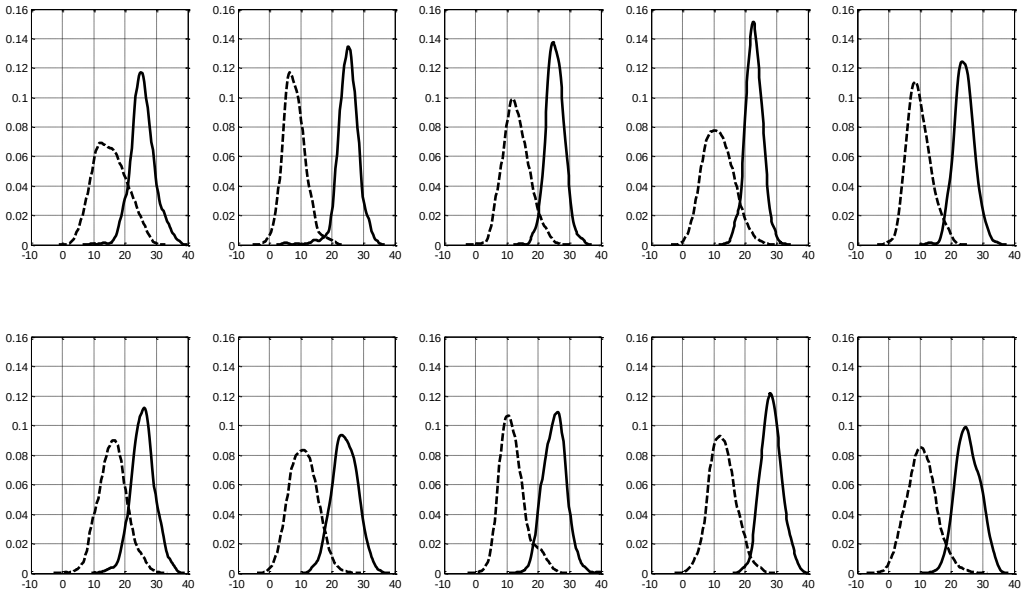
6.1.1 Класификатор базиран на односу сигнал/шум

Однос сигнал/шум (енгл. *Signal to Noise Ratio*) представља значајан параметар који битно утиче на перформансе ASR система. Дефинисан је релацијом 6.1, при чему је са P означена снага сигнала, а са N снага шума. Најчешће се изражава у децибелима.

$$SNR[dB] = 10\log\frac{P}{N} \tag{6.1}$$

Због чињенице да човек при континуалном говору прави паузе 30% времена [5], могуће је у тим временским интервалима одредити снагу шума. Пошто је говорна база коришћена у истраживањима база изолованих речи мануелна сегментација сесије изговорâ вршена је на начин да изговор сваке речи садржи и бар 20 ms тишине. Тај неактивни интервал је искоришћен за одређивање снаге шума амбијенталне буке.

На слици 6.2 приказана је нормализована учестаност појављивања односа *SNR* за женске говорнике (а) и мушке говорнике (б).



Слика 6.2 – Нормализована учестаност појављивања односа сигнал/шум за (а) женске говорнике и (б) мушке говорнике говорне базе Whi-Spe за нормални говор (пуна линија) и шапат (испрекидана линија)

За нормални говор, код женских говорника вршна вредност расподеле се креће од 22 до 25 dB, а код мушких говорника од 23 до 28 dB. За шапат је у распону од 6 до 12 dB за женске говорнике и од 10 до 16 dB за мушке говорнике. У табели 6.1 приказани су проценти класификације нормалног говора и шапата уколико је параметар однос *SNR*.

ТАБЕЛА 6.1: ПРОЦЕНАТ КЛАСИФИКАЦИЈЕ МОДА ГОВОРА ЗА ОДНОС SNR

Препознати мод / Мод тестног изговора	Нормалан говор	Шапат
Нормалан говор	**92,71%**	7,29%
Шапат	7,91%	**92,09%**

Средња грешка класификације је израчуната релацијом 6.2:

$$P_E = \frac{1}{2}\left[P(Н / Ш) + P(Ш / Н)\right] = 7,6\% \tag{6.2}$$

6.1.2 Класификатор базиран на првом кепстралном коефицијенту

Дуговремени усреднњени спектар шаптавог говора је равнији у односу на спектар нормалног говора. Како је први мел-фреквенцијски кепстрални коефицијент c_1 индикатор спектралног нагиба усреднњена вредност истог је мања за шапат него за нормални говор, што је експериментално потврђено за говорну базу *UT-Vocal Effort II* [1]. Такође, за говорну базу *Whi-Spe* је анализирана нормализована расподела првог кепстралног коефицијента при чему су добијене врло сличне расподеле [87].

У табели 6.2 приказан је проценат класификације мода говора уколико се као параметар користи средња вредност коефицијента c_1 по изговору.

ТАБЕЛА 6.2: ПРОЦЕНАТ КЛАСИФИКАЦИЈЕ МОДА ГОВОРА ЗА ПРВИ MFCC КОЕФИЦИЈЕНТ

Препознати мод / Мод тестног изговора	Нормалан говор	Шапат
Нормалан говор	**85,31%**	14,69%
Шапат	2,46%	**97,54%**

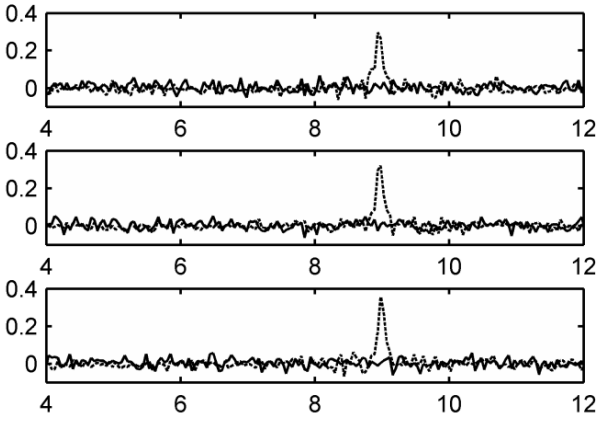
Средња грешка класификације је израчуната релацијом 6.3:

$$P_E = \frac{1}{2}\left[P(Н / Ш) + P(Ш / Н)\right] = 8,575\% \qquad (6.3)$$

6.1.3 Класификатор базиран на основној фреквенцији говорног сигнала

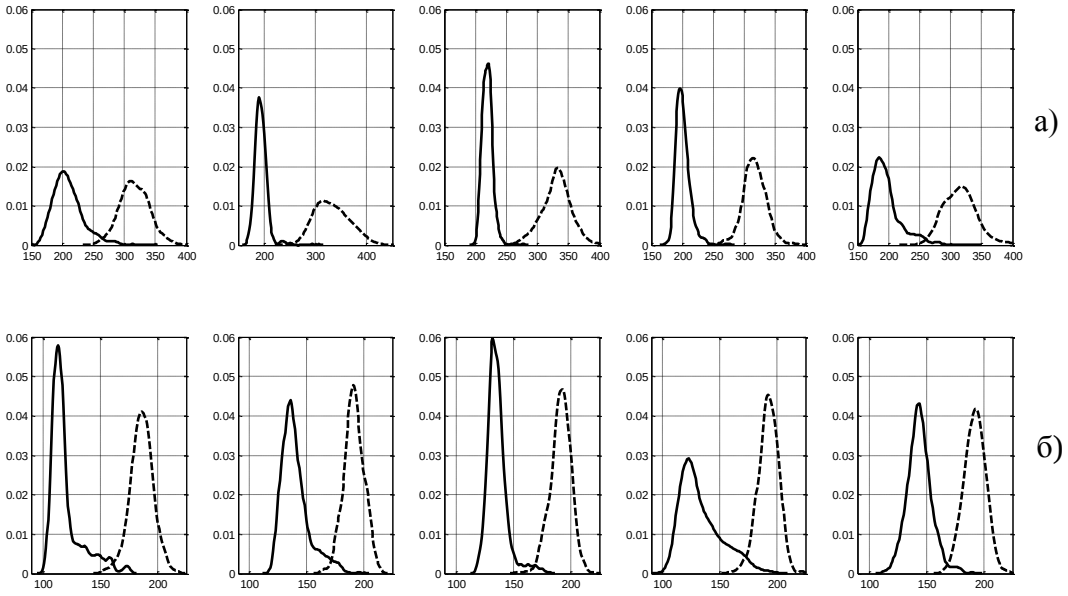
Једно од најзначајнијих индивидуалних акустичких обележја говорника јесте његова основна фреквенција гласа (f_0). Она је директно повезана са физичким карактеристикама ларинкса, односно гласних жица говорника. Основна фреквенција гласа није стабилан интраговорнички параметар тј. не може се рећи да свака особа има егзактну, фиксну вредност основне фреквенције гласа. Она се мења у току говора, што значи да се под овим термином подразумева нека средња вриједност, најчешће аритметичка средина. Од изузетног је значаја тачност одређивања основне фреквенције говорног сигнала с обзиром на велику осетљивост људског перцептивног механизма на њену вредност. За одређивање основне фреквенције говорног сигнала може се користити више различитих метода [88]-[91]. Међу њима су најпознатије аутокорелациона, кроскорелациона и кепстрална метода. Кепстар звучног фонема има изузетно изражен максимум који одговара периоду осциловања гласница [6] (у области које одговарају типичном опсегу основних фреквенција за одређени пол). Кепстралну методу карактерише велика тачност и већа нумеричка комплексност [92].

На слици 6.3 приказан је кепстар 3 узастопна фрејма (са *Hamming*овим прозоровањем) за нормални говор и шапат за реч /бела/ изговорену од једног мушког говорника из говорне базе Whi-Spe. Јако изражен кепстрални максимум је очигледан у кепстру нормалног говора.



Слика 6.3 – Кепстар 3 узастопна фрејма за вокал /e/ у речи /бела/ за нормални говор (испрекидана линија) и шапат (пуна линија). На апсциси је дата квефренција у ms.

За естимацију основне фреквенције коришћен је интервал квефренција од 2 до 6 ms за женске говорнике и 6 до 12 ms за мушке говорнике. На слици 6.4 приказана је нормализована расподела за женске говорнике (а) и мушке говорнике (б) уколико је параметар основна фреквенција говорног сигнала. Као класификатор је коришћена статистичка средња вредност (медијана) свих фрејмова у изговору због веће тачности у односу на аритметичку средину [93].



Слика 6.4 – Нормализована учестаност појављивања основне фреквенције за (а) женске говорнике и (б) мушке говорнике говорне базе Whi-Spe.

У табели 6.3 приказан је проценат класификације мода говора уколико се као параметар користи статистичка медијана основне фреквенције по изговору.

ТАБЕЛА 6.3: ПРОЦЕНАТ КЛАСИФИКАЦИЈЕ МОДА ГОВОРА ЗА ОСНОВНУ ФРЕКВЕНЦИЈУ

Препознати мод / Мод тестног изговора	Нормалан говор	Шапат
Нормалан говор	**96,76%**	3,24%
Шапат	0,42%	**99,58%**

Средња грешка класификације је израчуната релацијом 6.4:

$$P_E = \frac{1}{2}\left[P(Н/Ш) + P(Ш/Н)\right] = 1{,}83\% \qquad (6.4)$$

Прегледом бројних вредности средње грешке класификације за анализирана 3 параметра види се да је најмања вредност добијена за основну фреквенцију f_0.

Због најмање средње грешке класификације у препознавању говора са класификацијом говорног мода је коришћен класификатор базиран на основној фреквенцији.

6.2 Препознавање зависно од говорника

У поглављу 5.5 при испитивању утицаја модела за иницијализацију на препознавање бимодалног говора са обуком на нормалном говору највећи успех у препознавању зависно од говорника је добијен са мануелном анотацијом. Стога, услови експеримента у препознавању са ИБ и класификацијом говорног мода су следећи:

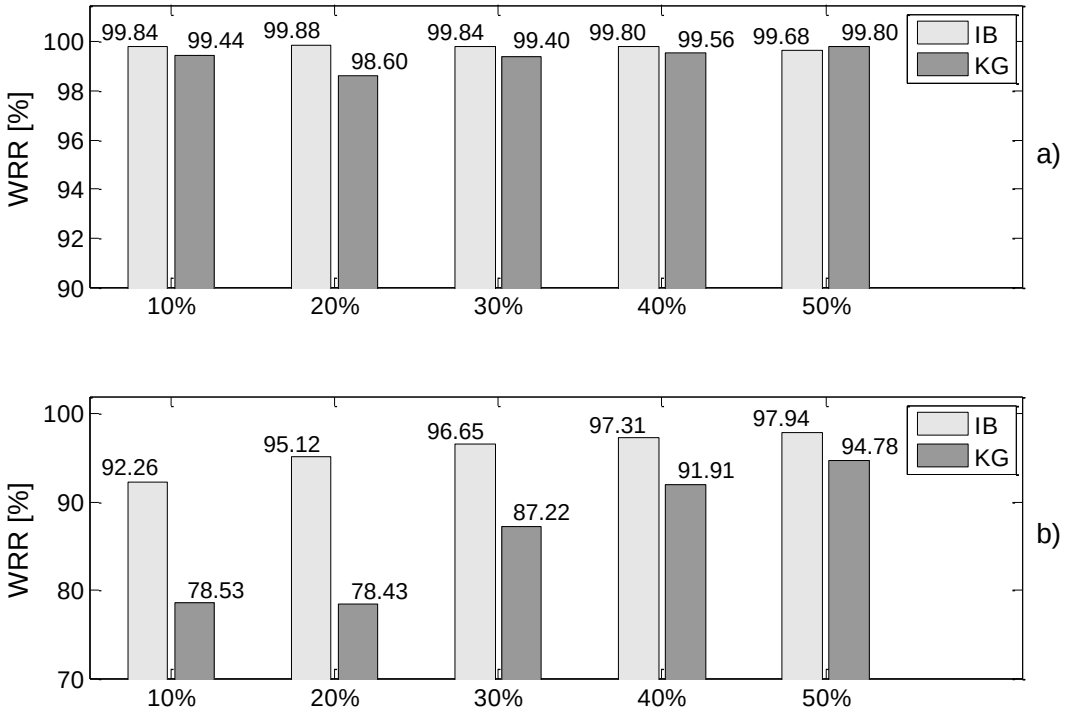
- Коришћени су модели фонема независних од контекста (монофони).
- Иницијални модели су добијени коришћењем мануелне анотације.
- Коришћен је променљив број стања по моделу монофона.
- Вектори обележја су PLP делта-делта нормализовани средњом вредношћу кепстра (генерисани коришћењем HTK алата).
- Број итерација у *Baum-Welch*овој реестимацији је 5 и број мешавина 8 (оптималан за препознавање зависно од говорника).

Циљ експеримента је анализа препознавања бимодалног говора уколико се у обуци расположивој бази у моду нормалног говора додаје одређени проценат базе у моду шапата. Због тога је у обуци коришћено 90% изговора нормалног говора а удео шапата се мењао од почетних 10% до крајњих 50%, са инкрементом 10%. Преостали изговори су коришћени за оцену перформанси при тестирању.

При препознавању са класификацијом говорног мода обука класификатора је вршена на првом изговору сваке речи (по 50 изговора за нормални говор и шапат). Тестни изговори су бирани на случајан начин 5 пута, а успешност препознавања за сваког говорника је рачуната као аритметичка средина. На крају, оцена препознавача је добијена усредњавањем по говорницима.

На слици 6.5 приказана је средња успешност препознавања нормалног говора (а) и шапата (б) у зависности од удела шапата у обуци.

У препознавању нормалног говора, експерименти су показали боље перформансе препознавача са ИБ уколико је удео изговора шапата мањи од 50%. За удео шапата од 40% препознавач са ИБ је незнатно успешнији (за 0.24% у апсолутном износу), док је за удео изговора шапата у износу 50% препознавач са класификацијом незнатно успешнији од препознавача са ИБ за 0.12% (99.80% наспрам 99.68%).



Слика 6.5 – Успешност препознавања нормалног говора (а) и шапата (б) са измешаном базом за обуку (IB) и класификацијом говорног мода (KG), зависно од говорника. На апсциси је приказан удео расположиве базе са изговорима у моду шапата, у процентима.

Такође, препознавање шапата је успешније са ИБ него са класификацијом говорног мода, поготово за мање проценте удела шапата у обуци. Познато је да је алгоритам максимизације очекивања којим се одређују параметри HMM модела много осетљив на тачност иницијалних модела [94]. С обзиром да су параметри иницијалних модела добијени коришћењем једино изговора нормалног говора,

као и јако мали удео шапата у обуци (за 10% и 20%), добијене су лоше перформансе препознавања са класификацијом говорног мода. Даљим повећавањем удела шапата у обуци (већим од 20%) успешност препознавања шапата се знатно повећава. Тренд повећања успешности препознавања шапата је добијен и са ИБ са највећим успехом од скоро 98% за удео шапата у обуци од 50%.

6.3 Препознавање независно од говорника

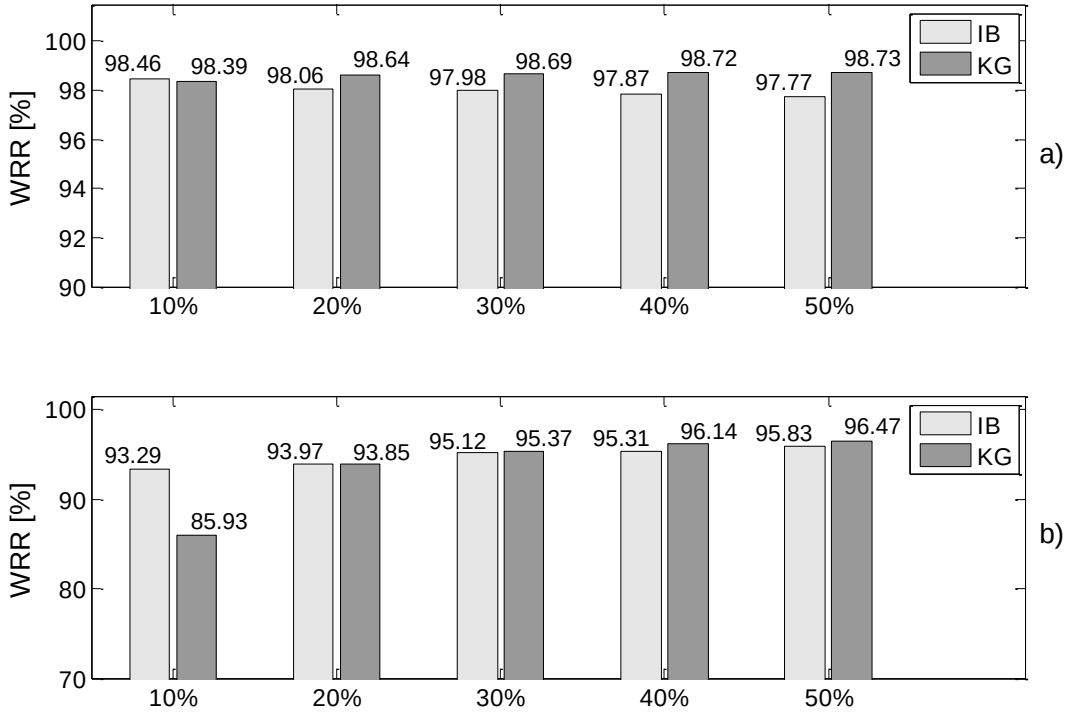
Из разлога боље успешности препознавања бимодалног говора (са обуком на изговорима нормалног говора) коришћењем иницијалних модела са аутоматском анотацијом (поглавље 5.6.2) услови експеримента у препознавању са ИБ и класификацијом говорног мода су следећи:

- Коришћени су модели фонема независни од контекста (монофони).
- Иницијални модели су добијени коришћењем аутоматске анотације.
- Коришћен је променљиви број стања по моделу монофона.
- Вектори обележја су PLP делта-делта нормализовани средњом вредношћу кепстра (генерисани коришћењем HTK алата).
- Број итерација у *Baum-Welch*овој реестимацији је 5 и број мешавина 32 (оптималан за препознавање независно од говорника).

Циљ експеримента је анализа препознавања бимодалног говора независно од говорника при чему се у обуци расположивој бази у моду нормалног говора додаје одређени проценат базе у моду шапата. Ради конзистентних услова са експериментима зависно од говорника (претходно поглавље, 6.2) део за обуку садржи 90% изговора нормалног говора (сви говорници изузев тестираног) док се удео шапата мењао од почетних 10% (1 говорник) до 50% (5 говорника). Наравно, изговори тестираног говорника не учествују у обуци јер је реч о препознавању независно од говорника. Говорници су бирани на случајан начин 5 пута, а успех препознавања је рачунат као аритметичка средина. Коначно, оцена препознавача је добијена усредњавањем по говорницима. Као и код препознавања зависно од говорника, за обуку класификатора је коришћен први изговор сваке речи тестираног говорника.

На слици 6.6 приказана је средња успешност препознавања нормалног говора (а) и шапата (б) у зависности од удела изговора у моду шапата у обуци.

Као што се може видети на слици 6.6а препознавање нормалног говора је успешније са класификатором него са ИБ, за све уделе изгоговора шапата веће од 10%. Због контаминације базе за обуку, препознавач са ИБ има тенденцију благог опадања успешности. Успех у препознавању се креће од 98.46% (за удео шапата у износу 10%) до 97.77% (за удео шапата од 50%). Препознавач са класификацијом има већу успешност у препознавању нормалног говора која се креће од 0.58% до 0.96%, зависно од удела изговора шапата у обуци.



Слика 6.6 – Успешност препознавања нормалног говора (а) и шапата (б) са измешаном базом за обуку (IB) и класификацијом говорног мода (KG), независно од говорника. На апсциси је приказан удео расположиве базе са изговорима у моду шапата, у процентима.

Због повећања удела изговора шапата у обуци препознавање шапата има тендецију повећања успешности, при чему се може уочити "ефекат засићења" код обе технике прпеознавања. Препознавање шапата је успешније са ИБ за уделе шапата у бази за обуку у износу 10% (93.29% наспрам 85.93%) и 20% (незнатно,

93.97% наспрам 93.85%). Када удео шапата у обуци постане довољно обиман (у овом случају 30%) препознавање са класификацијом говорног мода почиње да премашује препознавање са ИБО. Премашај износи неколико десетих делова процента, у зависности од удела изговора шапата у обуци.

6.4 Поређење са истраживањима у свету

Приликом поређења перформанси различитих ASR система потребна је доследност у навођењу свих битних параметара на које је систем веома осетљив: коришћена технологија, језик на којем се врши обука и тестирање, величина речника, обим базе за обуку, начин препознавања (зависно или независно од говорника), начин на који су извршени обука и декодовање, као и коришћени вектор обележја.

У истраживањима у вези са аутоматским препознавањем говорника који шапуће [21], у којем је учествовало 28 женских говорника добијена је тачност препознавања говорника од 99.10% (у сценарију Н/Н) и 94.41% (у сценарију Ш/Ш). Тачност препознавања говорника у сценарију Н/Ш је била 79.29% при чему је добијена велика зависност успешности између појединих говорника. Неки говорници су имали успешност у препознавању упоредиву са препознавањем у усаглашеним сценаријима (преко 90%) док су неки имали изузетно лоше препознавање (испод 50%). У раду су анализирани и објективни параметри за постојање таквих "добрих" и "лоших" шаптача. Тачност препознавања говорника са спонтаним говором је око 6% мања (апсолутно) од препознавања говорника који чита. Коришћени су динамички MFCC динамички вектори обележја (19 димензија) и препознавање базирано на моделима Гаусових мешавина.

Веома обимно истраживање препознавања бимодалног континуалног говора (независно од говорника) урађено је за јапански језик [19]. У обуци је учествовало 80 говорника (по 40 мушких и женских) са укупно 8000 фонетски балансираних реченица (по 4000 у оба говорна мода). Препознавач је базиран на Марковљевим моделима и користи трифоне који деле 500 стања од којих сваки има Гаусову расподелу са 16 мешавина. За моделовање језика се користе триграми, а речник има 20000 речи. Коректност и тачност су приказане као оцене препознавача.

Урађено је и препознавање са ИБО и класификацијом говорног мода (са грешком класификације око 5% за нормални говор и занемарљивом грешком за шапат).

Тачност препознавања нормалног говора је око 82% (са обуком на нормалном говору) и око 53% (са обуком на шапату). Препознавање нормалног говора са ИБ и класификацијом говорног мода је са приближно истим успехом у износу 80%.

С друге стране, тачност препознавања шапата је са успехом од око 68% (са обуком на шапату) и 20% (са обуком на нормалном говору). Препознавање са ИБ и класификацијом говорног мода је дало сличне резултате са успехом од 66%. Коректност препознавања је од 3-10% веће. У раду је показано да се коришћењем адаптације са свега 10 реченица по говорнику у моду шапата тачност може повећати за 10% (апсолутно).

И у истраживањима из којих је проистекла ова дисертација је добијена значајна асиметрија успешности препознавања у неусаглашеним сценаријима, у корист Ш/Н сценарија. Такође, уколико је на располагању довољан обим базе за обуку у моду шапата добијају се приближне перформансе препознавања бимодалног говора са ИБ и класификацијом. Овде је добијена већа успешност у препознавању јер је реч о препознавању изолованих речи из малог речника.

У истраживању [34] анализирано је препознавање нормалног говора и шапата за мандарински и енглески језик, независно од говорника, коришћењем вектора обележја који обухвата динамичка MFCC обележја и нека артикулаторна обележја (укупно 61 димензија). Препознавање је урађено коришћењем:

- модела Гаусових мешавина и естимације трифона са препознавањем помоћу максималне веродостојности (енгл. *Maximum Likelyhood Estimation* - MLE), и
- дубоких неуронских мрежа (енгл. *Deep Neural Network* - DNN)

Препознавање помоћу дубоких неуронских мрежа је било са значајно већим успехом, са значајно лошијим успехом у сценарију Н/Ш у односу на обрнут случај.

У експериментима са говорном базом *UT-Vocal Effort II* [23] извршена је анализа препознавања бимодалног говора независно од говорника, али са речником ограниченим на 160 речи, за MFCC и PLP векторе обележја. При том, у

банци филтара нису коришћене *mel* и *bark* фреквенцијска скала него линеарна, јер се показала као боља у неусаглашеним сценаријима.

Најмања учестаност грешака на нивоу речи (енгл. *Word Error Rate* - WER) у сценарију је добијена за PLP вектор и износи 16,9% (што даје средњи успех препознавања речи 83,1%).

6.5 Резиме

Уколико су за обуку препознавача бимодалног говора на располагању изговори оба говорна мода најједноставнији начин је обједињавање појединачних база за обуку, односно измешана база за обуку. Међутим, неизбежна последица коришћења те технике (код анализе са HMM) је приметна деградација перформанси у односу на препознавање бимодалног говора са одговарајућом базом за обуку. Због тога алтернативни приступ подразумева коришћење класификатора говорног мода и коришћење модела обученог на мод говора који је класификатор препознао. У овој Глави је извршена упоредна анализа побројане две технике, зависно и независно од говорника. Међу анализираним класификаторима (заснованим на односу сигнал/шум, првом кепстралном коефицијенту и основној фреквенцији говорног сигнала) најмања грешка класификације је добијена за класификатор базиран на основној фреквенцији. Грешка класификације износи 3.24% за нормални говор и 0.42% за шапат.

Код препознавања зависно од говорника препознавање нормалног говора је веома добро код обе технике, док је препознавање шапата боље са ИБ. Са порастом удела изговора шапата у обуци постигнут је значајан раст перформанси код препознавања шапата, са обе технике.

Код препознавања независно од говорника препознавање нормалног говора је са већим успехом коришћењем класификације, поготову за веће проценте заступљености шапата у обуци. С друге стране, препознавање шапата је успешније са ИБ за мање уделе изговора шапата, док је боље са класификацијом ако је заступљеност изговора шапата у обуци већа од 20% расположиве базе за обуку у моду шапата. Свакако, коришћење класификатора са мањом грешком класификације би резултовало још бољим перформансама препознавача који користи класификатор говорног мода.

На крају Главе приказано је поређење успешности препознавања бимодалног говора који су добијени у експериментима са актуелним истраживањима за друге говорне базе и језике.

7. ЗАКЉУЧАК

Огромна улагања последње две деценије у развој говорних технологија за највеће светске језике (енглески и кинески) је допринео препознавању континуалног говора независно од говорника са великом тачношћу, у реалном времену. За енглески језик и читани текст недавно је постигнута успешност боља од човекове (WER=5,2%). Међутим, поред зависности од језика којем је намењено, препознавање говора је изузетно осетљиво уколико услови при тестирању (препознавању) и обуци одступају једно од другог. То неизбежно доводи до значајне деградације перформанси система, тако да се јавила потреба да се комуникација човек - машина подигне на виши ниво. Поред препознавања говора при различитим емотивним стањима, посебан изазов у савременим ASR системима представља аутоматско препознавање говора другачије артикулисаног од говора уобичајеног интензитета. То се пре свега односи на шапат који по својим акустичким карактеристикама има најјача дискриминативна својства и највише се дистанцира од преостала 4 мода говора (тихи говор, нормални говор, гласни говор и вика).

У овој дисертацији је низом експеримената анализирана проблематика препознавања мултимодалног говора (нормални говор и шапат) за изоловане речи из ограниченог речника, на српском језику. Коришћена је говорна база Whi-Spe, која је наменски креирана за истраживања у вези са бимодалном експресијом говора. Садржи изговоре 50 речи у нормалном говору и шапату, сваку по 10 пута. У почетној форми је обима 10 говорника (5 женских и 5 мушких). Методологија препознавања је заснована на конвенционалном статистичком приступу са скривеним Марковљевим моделима (HMM) у комбинацији са мешавинама Гаусових расподела. Експерименти су урађени у моду зависно и независно од

говорника, са фокусом на оптимизацију препознавача бимодалног говора са моделима обученим на нормални говор. Оптимизација подразумева компромис између тачности и робустности, тј. повећање препознавања шапата али да се препознавање нормалног говора не деградира значајно. Два су разлога за то: прво, за задовољавајуће перформансе препознавања шапата независно од говорника потребне су веома обимне говорне базе у том говорном моду (а креирање истих је веома дуготрајан процес); и друго, са практичног аспекта, често се намеће потреба да се говорник испред ASR система намењеног препознавању говора уобичајеног интензитета машини обрати шапатом.

Поред HMM препознавача, први пут је анализирана примена методе потпорних вектора (SVM) у препознавању шапата.

7.1 Преглед резултата

Главни задатак који је постављен пред истраживања из којих је проистекла ова дисертација је развој ASR система за препознавање мултимодалног говора базираног на HMM алгоритму који ће пре свега имати задовољавајућу тачност у препознавању шапата. Како је већ речено, постављени задатак је уз услов да перформансе препознавача у препознавању нормалног говора не буду снижене. Тежња том компромису је довела до избора фонема независних од контекста (монофона) као јединица за моделовање у препознавању мултимодалног говора. Упркос већој тачности у препознавању нормалног говора за моделе трифона и целих речи (преко 99.5% зависно од говорника) у односу на монофоне (98.6%) значајно већа робустност је добијена за моделе монофона. Препознавање шапата у почетним експериментима је са монофонима (укупно 48) досегло успешност од скоро 62%, док је за моделе трифона и целих речи испод 35%. У почетним експериментима су коришћени статички MFCC вектори обележја (13 коефицијената). Анализирана су 4 обука/тест сценарија: 2 усаглашена (нормалан/нормалан и шапат/шапат) и 2 неусаглашена (нормалан/шапат и шапат/нормалан).

У експериментима који су потом следили повећање робустности у апсолутном износу од 11,65% је остварено са динамичким векторима обележја (39 коефицијената) и редукованим бројем монофона (укупно 32). Додатно повећање

препознавања шапата за 7,96% је остварено рачунањем параметара иницијалних модела са мануелном анотацијом. На тај начин је остварено препознавање шапата зависно од говорника од 81,38%. Истовремено, постигнуто је и повећање препознавања нормалног говора на износ 99,6%. У експериментима независно од говорника, најбоља успешност је остварена иницијализацијом уз аутоматску анотацију и то у износу од 98,40% (за нормални говор) и 87,42% (за шапат).

Урађена је и анализа препознавања заснована на методи потпорних вектора. Анализирана је успешност препознавања нормалног говора и шапата за 4 типа кернела (RBF, линеарни, полиномијални и сигмоидни). Код препознавања зависно од говорника, највећи WRR је постигнут у износу од 99,36% (за нормални говор) и 83,36% (за шапат). Код препознавања независно од говорника успешност препознавања је 96,53% (за нормални говор) и 81,72% (за шапат). Практично сви савремени ASR системи су намењени препознавању независно од говорника. Узимајући то у обзир, проведена анализа указује да је препознавање бимодалног говора успешније са HMM него са SVM препознавачем.

Накнадно повећање успешности HMM препознавача је остварено коришћењем нових вектора обележја; кепстралних коефицијената са μ-фреквенцисјком скалом. Између анализираних вредности параметра μ највеће побољшање је постигнуто за μ=2. За ту вредност параметра препознавање шапата је 87,50 (зависно од говорника) и 90,92% (независно од говорника). Истовремено су задржане добре перформансе у препеознавању нормалног говора.

Дисертација завршава упоредном анализом успешности препознавања HMM препознавача базираног на измешаној бази за обуку и препознавача са класификацијом говорног мода. Анализиран је раст перформанси препознавања шапата у зависности од раста удела изговора шапата у обуци (зависно и независно од говорника).

7.2 Доприноси дисертације

Развој ASR система независног од говорника који ће имати добре перформансе у препознавању говора другачије артикулисаног од уобичајеног је област којом се баве многи стручњаци савремених говорних технологија. Због

тога је препознавање шапата актуелна истраживачка тема. Основни доприноси ове дисертације се огледају у следећем:

- Говорне технологије су јако зависне од језика којем су намењене те се не могу увести и применити на било ком језику. Због тога је систематски креирана говорна база Whi-Spe на српском језику која се може користити и у другим врстама истраживања бимодалне експресије говора.
- Развијен је препознавач базиран на статистичком приступу и скривеним Марковљевим моделима. Показано је да између анализираних јединица за моделовање фонеми независни од контекста имају највећу робустност. Одређен је оптималан број мешавина у препознавању зависно и независно од говорника.
- Повећање робустности препознавача у значајној мери је остварено коришћењем динамичких обележја уз редукован број монофона и иницијализацију модела са мануелном (зависно од говорника) и аутоматском анотацијом (независно од говорника). При том су повећане перформансе препознавача и за нормални говор.
- Развијен је нови препознавач (први пут у препознавању шапата) базиран на методи потпорних вектора и извршена анализа перформанси у зависности од типа кернела и броја преклапајућих прозора.
- Дат је алгоритам за екстракцију нових вектора обележја заснованих на кепстралним коефицијентима и модификованом скалом мапирања. Пресликавање је извршено према квазилогаритамској функцији (тзв. μ-закон пресликавања) која је прилагођена шапату на тај начин што комбинује резолуцију мел и линеарне фреквенцијске скале. Експерименти су потврдили повећање робустности HMM-препознавача са новим векторима обележја у оба начина препознавања (зависно и независно од говорника).
- Развијен је бинарни класификатор мода говора (нормалан говор / шапат) базиран на односу сигнал/шум, првом кепстралном коефицијенту и основној фреквенцији говорног сигнала. Између

анализираних, најмања грешка класификације је добијена за класификатор баиран на основној фреквенцији.

- Извршена је упоредна анализа перформанси HMM-препознавача базираног на измешаној бази за обуку и класификацији мода говора у зависности од удела изговора шапата расположивих у обуци.

7.3 Правци даљих истраживања

Ова дисертација представља наставак истраживања препознавања шапата као специфичног начина говорне комуникације који је у све чешћој употреби. И поред тога што је заједно са анализом препознавања базираној на DTW алгоритму [95] и вештачким неуронским мрежама [87] заокружена целина машинског препознавања изолованих речи у шапату зависно од говорника, постоји простор за даља истраживања. То се односи пре свега на препознавање независно од говорника. Могући правци се огледају у следећем:

- За препознавање већег корпуса речи независно од говорника неопходно је извршити проширење говорне базе Whi-Spe како по броју речи тако и по броју говорника.
- Испитати могућности адаптације на говор шапата и упоредити резултате са препознавањем са измешаном базом за обуку и класификацијом мода говора. У прелиминарним истраживањима су добијени бољи резултати са адаптацијом базираној на максималној веродостојности (енгл. *Maximum Likelyhood Linear Regression - MLLR*) [96].
- Испитати могућности развоја хибридног HMM/SVM система у препознавању шапата; иако SVM метода има боља дискриминативна својства од HMM недостатак исте је што као улазе захтева податке фиксне дужине. Због тога се у препознавању говора приступа хибридном решењу где се SVM класификатор користи у делу рачунања параметара модела при обуци, а HMM препознавач у алгоритму динамичког програмирања и препознавању.
- Анализирати могућност употребе скривених Марковљевих модела у препознавању говорника који шапуће.

- Испитати перформансе коришћењем савремених софтвера за препознавање говора (нпр. Deepspeech, KALDI, Sphinx, итд).

Решавањем побројаних задатака наставио би се даљи рад на машинском препознавању говора/говорника који би унапредио квалитет постојећих ASR система на српском језику.

ЛИТЕРАТУРА

[1] C. Zhang and J. H.L. Hansen, “Analysis and Classification of Speech Mode: Whisper through Shouted,” in *Proc. of Interspeech 2007*, Antwerp, Belgium, 2007, pp. 2289–2292.

[2] X. Huang and K.F. Lee, “On speaker-independent, speaker-dependent, and speaker adaptive speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 150–157, 1993.

[3] V. Delić, M. Sečujski, N. Jakovljević, M. Janev, R. Obradović, and D. Pekar, “Speech Technologies for Serbian and Kindred South Slavic Languages,” in *Advances in Speech Recognition*, 2010.

[4] B. Marković, S. Jovičić, J. Galić, and Đ. Grozdić, “Whispered Speech Database: Design, Processing and Application,” in *Proc. of 16th International Conference TSD*, Pilsen, Czech Republic, 2013, pp. 591–598.

[5] S. Jovičić, *Govorna komunikacija: fiziologija, psihoakustika i percepcija*. Beograd: Nauka, 1999.

[6] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice-Hall Signal Processing Series, 1978.

[7] J.H.L. Hansen, “Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Speech Recognition,” PhD. thesis, Georgia Institute Technology, Atlanta, 1988.

[8] J.C. Catford, *Fundamental problems in phonetics*. Edinburgh: Edinburgh University Press, 1977.

[9] M. Matsuda and H. Kasuya, “Acoustic nature of the whisper,” presented at the Eurospeech, 1999, pp. 137–140.

[10] W. Meyer-Eppler, “Realization of prosodic features in whispered speech,” *Journal of Acoustical Society of America*, pp. 104–106, 1957.

[11] K. Tsunoda, S. Sekimoto, and T. Baer, “Brain activity in aphonia after a coughing episode: different brain activity in healthy whispering and pathological aphonic conditions,” *Journal of Voice*, vol. 26, no. 5, p. 668.e11–668.e13, 2012.

[12] R.W. Morris and M.A. Clements, “Reconstruction of speech from whispers,” *Medical Engineering & Physics*, vol. 24, pp. 515–520, 2002.

[13] S.T. Jovičić, “Formant feature differences between whispered and voiced sustained

vowels," *Acustica-acta*, vol. 84, no. 4, pp. 739–743, 1998.

[14] Y. Swerdlin, J. Smith, and J. Wolfe, "The effect of whisper and creak vocal mechanisms on vocal tract resonances," *Journal of Acoustical Society of America*, no. 127, pp. 2590–2598, 2010.

[15] I.B. Thomas, "Perceived Pitch of Whispered Vowels," *Journal of Acoustical Society of America*, vol. 46, no. 2, pp. 468–470, 1969.

[16] S.T. Jovičić and M.M. Đorđević, "Percepcija fonema u šapatu: identifikacija i konfuzija.," in *Govor i jezik: interdisciplinarna istraživanja, II*, Beograd: Centar za unapređenje životnih aktivnosti i Institut za eksperimentalnu fonetiku i patologiju govora, 2008.

[17] V. Tartter, "Identifiability of vowels and speakers from whispered syllables," *Perception & Psychophysics*, vol. 49, no. 4, pp. 365–372, 1991.

[18] G. Chenghui, Z. Heming, Z. Wei, and W. Min, "A Preliminary Study on Emotions of Chinese Whispered Speech," presented at the International Forum on Computer Science-Technology and Applications (IFCSTA), Chongqing, China, 2009, vol. 2, pp. 429–433.

[19] T. Ito, K. Takeda, and F. Itakura, "Analysis and Recognition of Whispered Speech.," *Speech Communication*, vol. 45, pp. 129–152, 2005.

[20] X. Fan and J.H.L. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," presented at the Conference of the International Speech Communication Association, 2008, vol. 1, pp. 1313–1316.

[21] X. Fan and J. H.L. Hansen, "Speaker Identification with whispered speech based on modified LFCC parameters and feature mapping," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 4553–4556.

[22] S. Ghaffarzadegan, H. Boril, and J. H. L. Hansen, "Model and Feature Based Compensation for Whispered Speech Recognition," in *Proc. of Interspeech 2014*, Singapore, 2014, pp. 2420-2424.

[23] S. Ghaffarzadegan, H. Boril, and J. H.L. Hansen, "UT-Vocal Efort II: Analysis and constrained-lexicon recognition of whispered speech," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 2544–2548.

[24] C. Zhang and J. H.L. Hansen, "Advancements in whisper-island detection using the linear predictive residual," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 5170–5173.

[25] C. Zhang and J. H.L. Hansen, “Whisper-Island Detection Based on Unsupervised Segmentation With Entropy-Based Speech Feature Processing,” *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 4, pp. 883–894, 2011.

[26] Đ. Grozdić, S.T. Jovičić, and M. Subotić, “Whispered speech recognition using deep denoising autoencoder,” *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 15–22, 2017.

[27] X. Fan, C. Busso, and J. H.L. Hansen, “Audio-visual isolated digit recognition for whispered speech,” presented at the European Signal Processing Conference (EUSIPCO), 2011, pp. 1500–1503.

[28] C. Zhang, T. Yu, and J. H.L. Hansen, “Microphone array processing for distance speech capture: A probe study on whisper speech detection,” in *Proc. of the Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, 2010, pp. 1707–1710.

[29] S. Jou, T. Schultz, and E. Waibel, “Adaptation for Soft Whisper Recognition Using a Throat Microphone,” in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, 2004, pp. 1493–1496.

[30] Đ. Grozdić, S. Jovičić, J. Galić, and B. Marković, “Application of inverse filtering in enhancement of whispered speech,” in *Proc. of Neural Network Applications in Electrical Engineering (NEUREL)*, Belgrade, Serbia, 2014, pp. 157–162.

[31] B. Marković, J. Galić, Đ. Grozdić, S. Jovičić, and M. Mijić, “Whispered speech recognition based on gammatone filterbank cepstral coefficients,” *JOURNAL OF COMMUNICATIONS TECHNOLOGY AND ELECTRONICS*, vol. 62, no. 11, pp. 1255–1261, 2017.

[32] M. Gales and S. Young, “The Application of Hidden Markov Models in Speech Recognition,” *Foundation and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[33] N. Kawaguchi *et al.*, “Ciair speech corpus for real world applications,” presented at the The International conference Committee for the Co-ordination and Standardization of Speech Databases and Assesment Techniques, 2002.

[34] P.X. Lee, D. Wee, H. Si ZinToh, B.P. Lim, N. Chen, and B.A. Ma, “Whispered Mandarin Corpus for Speech Technology Applications,” in *Proc. of Interspeech 2014*, Singapore, 2014, pp. 1598–1602.

[35] B.P. Lim, “Computational differences between whispered and non-whispered speech,” PhD. thesis, University of Illinois at Urbana Champaign, 2010.

[36] T. Tran, S. Marioryad, and C. Busso, “Audiovisual corpus to analyze whisper speech,” presented at the International Conference on Acoustics, Speech and Signal Processing ICASSP, 2013.

[37] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, “The chains corpus: Characterizing individual speakers,” in *Proc. of International Conference on Speech and Computer SPECOM*, St. Petersburg, Russia, 2006, pp. 421–435.

[38] S. Jovičić, Z. Kašić, M. Đorđević, and M. Rajković, “Serbian emotional speech database: design, processing and evaluation,” presented at the SPECOM, 2004, pp. 77–81.

[39] R. Bilibajkić, “Prepoznavanje artikulaciono-akustičkih odstupanja glasova u patološkom govoru,” PhD. thesis, Beograd, 2016.

[40] K. Yu-Ting, “DISTINCT ACOUSTIC MODELING FOR AUTOMATIC SPEECH RECOGNITION,” PhD. thesis, The Hong Kong University of Science and Technology, Hong-Kong, 2014.

[41] X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*. Prentice-Hall. Inc, 2001.

[42] N. Jakovljević, “Upoznavanje sa HTK alatima.”, Laboratorijske vježbe, 2013.

[43] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, “Revising Perceptual Linear Prediction (PLP),” presented at the Interspeech, 2005, pp. 2997–3000.

[44] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *Journal of Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[45] Đ. Grozdić, S. Jovičić, D. Šumarac Pavlović, J. Galić, and B. Marković, “Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition,” *Advances in Electrical and Computer Engineering (AECE) Journal*, vol. 17, no. 1, pp. 21–26, 2017.

[46] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland. *The HTK Book (for version 3.4). Cambridge University Engineering Department*, 2009.

[47] M. Merkle, *Verovatnoća i statistika*. Beograd: Akademska misao, 2006.

[48] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. Wiley, 2000.

[49] http://computerrobotvision.org/2010/tutorial_day/GMM_said_crv10_tutorial.pdf, pristupljeno u novembru 2018.

[50] S.A. Hejazi, R. Kazemi, and S. Ghaemmaghami, “ISOLATED PERSIAN DIGIT RECOGNITION USING A HYBRID HMM-SVM,” *Intelligent Signal Processing and Communications Systems*, 2009. DOI: 10.1109/ISPACS.2009.4806757

[51] C. Cortes and V. Vapnik, *Support-Vector Networks*, vol. 20. 1995.

[52] Z. Ćirković and Z. Banjac, "Jedna primena SVM klasifikatora u verifikaciji govornika nezavisno od teksta," in *Zbornik radova konferencije Infoteh Jahorina*, 2013, vol. 11, pp. 833–837.

[53] J. Galić, D. Šumarac Pavlović, S. Jovičić, B. Marković, and Đ. Grozdić, "Prepoznavanje bimodalnog govora bazirano na metodi potpornih vektora," in *Zbornik radova konferencije TELFOR*, 2017.

[54] V. Kecman, *Learning And Soft Computing - Support Vector Machines, Neural Networks, And Fuzzy Logic Models*. The MIT Press, 2001.

[55] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer US, 1995.

[56] A. Lee,T. Kawahara, K. Shikano, *Julius – an open source real-time large vocabulary recognition engine*. *EUROSPEECH*, pp. 1691–1694, 2001.

[57] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, J. Woelfel, *Sphinx-4: A flexible open source framework for speech recognition. Sun Microsystems Inc., Technical Report* SML1 TR2004-0811, 2004.

[58] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Loof, R. Schluter, H. Ney, The *RWTH Aachen University Open Source Speech Recognition System*. *INTERSPEECH*, strane 2111–2114, 2009.

[59] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, *The Kaldi Speech Recognition Toolkit*. *Workhop on Automatic Speech Recognition and Undertanding*, 2011.

[60] A. Hannun, C. Case, and J. Casper et al., "*Deep Speech: Scaling up end-to-end speech recognition*," in *eprint arXiv:1412.5567*, 2014.

[61] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tur, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, M. Saraclar, *The AT&T WATSON speech recognizer*. *ICASSP*, 2005.

[62] M. Dunn, *Pro Microsoft Speech Server 2007: Developing Speech Enabled Applications with .NET (Pro). Apress*, 2007.

[63] J. Adorf, *Web Speech API. Technical report, KTH Royal Institute of Technology*, 2013.

[64] Nuance Communication: Speech recognition solutions., 2014. URL https://www.nuance.com/mobile/speech-recognition-solutions.html, pristupljeno u septembru 2018.

[65] L. Berbakov, "Zavisnost performansi sistema za prepoznavanje govora na srpskom jeziku od izbora obeležja,", Master rad, Fakultet tehničkih nauka, Novi Sad, 2007.

[66] X. Fan and J. H.L. Hansen, "Speaker Identification within Whispered Speech Audio Streams," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1408–1421, 2011.

[67] B. Sklar, *Digital Communications: Fundamentals and Applications*, Second Edition, Los Angeles: Prentice-Hall, 1988.

[68] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994. doi: 10.1109/89.326616

[69] D. Ellis, *PLP and RASTA (and MFCC, and inversion) in Matlab*. URL https://labrosa.ee.columbia.edu/matlab/rastamat/, pristupljeno u novembru 2017.

[70] S. Ghaffarzadegan, H. Boril, and J. H.L. Hansen, "Generative Modeling of Pseudo-Whisper for Robust Whispered Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1705–1720, 2016.

[71] J. Miguel Garcia-Cabellos, C. Peleaz Moreno, A. Gallardo Antolin, F. Perez-Cruz, and F. Diaz-de-Maria, "SVM classifiers for ASR: A discussion about parameterization," presented at the European Signal Processing Conference, 2004, pp. 2067–2070.

[72] *Python Software Foundation, Python Language Reference*.

[73] S. Sovilj-Nikić, V. Delić, I. Sovilj-Nikić, and M. Marković, "Tree-based phone duration modeling of the serbian language," *Electronics and Electrical Engineering (Elektronika ir Elektrotechnika)*, vol. 20, no. 3, pp. 1192–1215, 2014.

[74] Đ. Grozdić, B. Marković, J. Galić, and S. Jovičić, "Application of neural networks in whispered speech recognition," *Telfor Journal*, vol. 5, no. 2, pp. 103–106, 2013.

[75] B. Marković, S. Jovičić, J. Galić, and Đ. Grozdić, "Recognition of the Multimodal Speech Based on the GFCC features," presented at the IcETRAN, Srebrno Jezero, Srbija, 2015, p. AK1 1.3 1-5.

[76] I. Oparin, "Language models for Automatic Speech Recognition of inflectional languages," PhD. Thesis, University of West Bohemia, Pilsen, Czech Republic, 2008.

[77] LJ. Jovanov, D. Pekar, and R. Obradović, "Prepoznavač kontinualnog govora baziran na programskom paketu HTK," in *Zbornik radova konferencije DOGS*, Novi Sad, Serbia, 2002.

[78] J. Galić, B. Popović, and D. Šumarac Pavlović, "Whispered Speech Recognition using Hidden Markov Models and Support Vector Machines," *Acta Politechnica Hungarica*, Vol.15(5), 2018. *DOI: 10.12700/APH.15.5.2018.5.2*

[79] B. Popović, E. Pakoci, S. Ostrogonac, and D. Pekar, "Large Vocabulary Continuous Speech Recognition for Serbian Using the Kaldi Toolkit," in *Zbornik radova konferencije DOGS*, Novi Sad, Serbia, 2014, pp. 31–34.

[80] P. Boersma and D. Wenink, *Doing phonetics by computer*. 2018.

[81] Ž. Bojović, D. Pekar, and V. Delić, "Iskustva iz segmentacije govorne baze S70W100S120," in *Zbornik radova konferencije TELFOR*, Beograd, Srbija, 2001.

[82] N. Jakovljević, "Primena retke reprezentacije na modelima Gausovih mešavina koji se koriste za automatsko prepoznavanje govora," PhD. Thesis, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Srbija, 2013.

[83] J. Galić, S. Jovičić, V. Delić, B. Marković, D. Šumarac Pavlović, and Đ. Grozdić, "HMM-based Whisper Recognition Using μ-law Frequency Warping," *SPIIRAS Proceedings*, vol. 3, no. 58, pp. 27–52, 2018.

[84] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2007.

[85] E. Zhang and Y. Zhang, "F-measure," in *Encyclopedia of Database Systems*, Boston, MA: Springer US, 2009, p. 1147.

[86] R. Lippmann, E. Martin, and D.B. Paul, "Multi-style training for robust isolated-word speech recognition," presented at the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1987, vol. 12, pp. 705–708.

[87] Đ. Grozdić, "Primena neuralnih mreža u prepoznavanju šapata," doktorska diseracija, Beograd, 2017.

[88] H. Boril and P. Pollak, "Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions," in *Proceedings of European Signal Processing Conference*, 2004, pp. 1003–1006.

[89] W.J. Hess, *Pitch Determination of Speech Signals*, Springer. New York, 1993.

[90] A. M. Noll, "Cepstrum Pitch Determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[91] J. Galić and T. Pešić-Brđanin, "Uporedna analiza metoda za estimaciju osnovne frekvencije govornog signala u prisustvu bijelog šuma," in *Zbornik radova 55. Konferencije za ETRAN*, Banja Vrućica, Republika Srpska, Bosna i Hercegovina, 2011, p. AK2.1-1-4.

[92] J. Galić, T. Pešić-Brđanin, and I. Janković, "Statistička analiza osnovne frekvencije kod vokala srpskog jezika," in *Zbornik radova konferencije INDEL*, Banja Luka, Republika Srpska, Bosna i Hercegovina, 2010, pp. 236–239.

[93] J. Galić and T. Pešić-Brđanin, "The Voice Fundamental Frequency Statistical Parameters under Noisy Conditions with the Cepstrum Method," presented at the International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services TELSIKS, Niš, Srbija, 2011, vol. 2, pp. 769–772.

[94] A. Sankar, "Experiments with a Gaussian Merging-Splitting Algorithm for HMM Training for Speech Recognition," in *Proceedings of the Broadcast News Transcription and Understanding Workshop*, 1998, pp. 99–104.

[95] B. Marković, "Analiza obeležja u govornom signalu za potrebe prepoznavanja multimodalnog govora," doktorska disertacija, Univerzitet u Beogradu, Elektrotehnički fakultet, Beograd, Srbija, 2018.

[96] J. Galić, S. Jovičić, B. Marković, D. Šumarac Pavlović, and Đ. Grozdić, "Speaker dependent recognition of whispered speech based on MLLR adaptation," in *Zbornik radova konferencije DOGS*, Novi Sad, Srbija, 2017, pp. 29–32.

ПРИЛОЗИ

Прилог А1: Лексикон говорне базе Whi-Spe

Ознака	Реч	Ознака	Реч
boja1	бела	rec6	пијаца
boja2	жута	rec7	падавине
boja3	црна	rec8	понедељак
boja4	црвена	rec9	година
boja5	плава	rec10	представа
boja6	зелена	rec11	компјутери
broj1	нула	rec12	иностранство
broj2	један	rec13	дрво
broj3	два	rec14	Мирјана
broj4	три	rec15	море
broj5	четири	rec16	киша
broj6	пет	rec17	зграде
broj7	шест	rec18	клинци
broj8	седам	rec19	Милан
broj9	осам	rec20	резултати
broj10	девет	rec21	телефон
broj11	десет	rec22	светло
broj12	сто	rec23	прозор
broj13	хиљаду	rec24	руке
broj14	милион	rec25	локал
rec1	Мирко	rec26	кључ
rec2	журка	rec27	сунце
rec3	Петар	rec28	паре
rec4	демонстрације	rec29	сеф
rec5	стандард	rec30	блок

Прилог А2: Садржај конфигурационог фајла

```
SOURCEKIND = WAVEFORM
SOURCEFORMAT = WAV
SOURCERATE = 454
TARGETKIND =MFCC_0_D_A_Z % TARGETKIND =PLP_0_D_A_Z
TARGETRATE = 80000.0
SAVECOMPRESSED = F
SAVEWITHCRC = F
WINDOWSIZE = 240000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
USEPOWER = T
```

Прилог А3: Листинг модела прототипа

```
~o <VecSize> 36 <MFCC_0_D_A_Z> <StreamInfo> 1 36
~h "proto"
<BeginHMM>
    <NumStates> 5
    <State> 2
        <Mean> 36
            0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0
        <Variance> 36
            1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1
    <State> 3
        <Mean> 36
            0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0
        <Variance> 36
            1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1
    <State> 4
        <Mean> 36
            0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0
        <Variance> 36
            1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1
    <TransP> 5
        0 1 0 0 0
        0 0.6 0.4 0 0
        0 0 0.6 0.4 0
        0 0 0 0.7 0.3
        0 0 0 0 0
<ENDHMM>
```

Прилог А4: Транскрипција речи за моделовање монофона

```
BELA Bo Be Es L A
ZHUTA ZH Us To Te A
CRNA  Co Ce Y R Y N A
CRVENA Co Ce Y R Y V Es N A
PLAVA Po Pe L As V A
ZELENA Z Es L E N A
NULA N Us L A
JEDAN J Es Do De A N
DVA Do De V As
TRI To Te Y R Is
CHETIRI CHo CHe Es To Te I R I
PET Po Pe Es To Te
SHEST SH Es S To Te
SEDAM S Es Do De A M
OSAM Os S A M
DEVET Do De Es V E To Te
DESET Do De Es S E To Te
STO S To Te O
HILJADU H Is LJ A Do De U
MILION M Is L I O N
MIRKO M Is R Y Ko Ke O
ZHURKA ZH Us R Y Ko Ke A
PETAR Po Pe Es To Te A R Y
DEMONSTRACIJE Do De E M O N S To Te R As Co Ce I J E
STANDARD S To Te As N Do De A R Y Do De
PIJACA Po Pe Is J A Co Ce A
PADAVINE Po Pe As Do De A V I N E
PONEDELJAK Po Pe O N Es Do De E LJ A Ko Ke
GODINA Go Ge Os Do De I N A
PREDSTAVA Po Pe Y R Es Co Ce To Te A V A
KOMPJUTERI Ko Ke O M Po Pe J Us To Te E R I
INOSTRANSTVO I N O S To Te Y R As N S To Te V O
DRVO Do De Y R Y V O
MIRJANA M Is R Y J A N A
MORE M Os R E
KISHA Ko Ke Is SH A
ZGRADE Z Go Ge Y R As Do De E
KLINCI Ko Ke L Is N Co Ce I
MILAN M Is L A N
REZULTATI R E Z U L To Te As To Te I
TELEFON To Te E L Es F O N
SVETLO S V Es To Te L O
PROZOR Po Pe Y R O Z O R Y
RUKE R Us Ko Ke E
LOKAL L Os Ko Ke A L
KLJUCH Ko Ke LJ Us CHo CHe
SUNCE S Us N Co Ce E
PARE Po Pe As R E
SEF S Es F
BLOK Bo Be L Os Ko Ke
silence sil
```

Прилог А5: Транскрипција речи за моделовање целих речи

```
BELA BELA
ZHUTA ZHUTA
CRNA CRNA
CRVENA CRVENA
PLAVA PLAVA
ZELENA ZELENA
NULA NULA
JEDAN JEDAN
DVA DVA
TRI TRI
CHETIRI CHETIRI
PET PET
SHEST SHEST
SEDAM SEDAM
OSAM OSAM
DEVET DEVET
DESET DESET
STO STO
HILJADU HILJADU
MILION MILION
MIRKO MIRKO
ZHURKA ZHURKA
PETAR PETAR
DEMONSTRACIJE DEMONSTRACIJE
STANDARD STANDARD
PIJACA PIJACA
PADAVINE PADAVINE
PONEDELJAK PONEDELJAK
GODINA GODINA
PREDSTAVA PREDSTAVA
KOMPJUTERI KOMPJUTERI
INOSTRANSTVO INOSTRANSTVO
DRVO DRVO
MIRJANA MIRJANA
MORE MORE
KISHA KISHA
ZGRADE ZGRADE
KLINCI KLINCI
MILAN MILAN
REZULTATI REZULTATI
TELEFON TELEFON
SVETLO SVETLO
PROZOR PROZOR
RUKE RUKE
LOKAL LOKAL
KLJUCH KLJUCH
SUNCE SUNCE
PARE PARE
SEF SEF
BLOK BLOK
sil sil
```

Прилог А6: Запис фајла којим се задаје граматика

```
/*
* Task grammar
*/
$WORD = BELA | CRNA | CRVENA | PLAVA | ZELENA | ZHUTA | NULA |
JEDAN | DVA | TRI | CHETIRI | PET | SHEST | SEDAM | OSAM | DEVET
| DESET | STO | HILJADU | MILION | MIRKO | ZHURKA | PETAR |
DEMONSTRACIJE | STANDARD | PIJACA | PADAVINE | PONEDELJAK |
GODINA | PREDSTAVA | KOMPJUTERI | INOSTRANSTVO | DRVO | MIRJANA
| MORE | KISHA | ZGRADE | KLINCI | MILAN | REZULTATI | TELEFON |
SVETLO | PROZOR | RUKE | LOKAL | KLJUCH | SUNCE | PARE | SEF |
BLOK;
( silence $WORD silence )
```

Прилог Б: Резултати иницијалног експеримента

а) сценарио Н/Н

	Број мешавина				
Јединице за моделовање	1	2	4	8	16
Монофони	96.92	97.98	98.10	98.30	98.60
Трифони	99.74	99.72	99.74	99.60	98.58
Целе речи	99.06	99.42	99.46	99.54	99.46

б) сценарио Ш/Ш

	Број мешавина				
Јединице за моделовање	1	2	4	8	16
Монофони	94.66	96.62	96.88	97.18	98.32
Трифони	99.88	99.88	99.82	99.52	97.08
Целе речи	96.64	97.98	98.06	98.14	98.26

в) сценарио Н/Ш

	Број мешавина				
Јединице за моделовање	1	2	4	8	16
Монофони	61.16	60.78	61.09	61.77	59.68
Трифони	34.68	27.27	24.93	22.76	18.78
Целе речи	27.17	25.78	26.45	27.25	24.46

г) сценарио Ш/Н

	Број мешавина				
Јединице за моделовање	1	2	4	8	16
Монофони	59.68	58.9	59.04	58.76	54.04
Трифони	49.92	37	33.49	29.78	24.41
Целе речи	32.02	31.16	31.32	31.60	28.57

Прилог B: Биографија

Јован Галић је рођен 7. новембра 1981. године у Травнику, Република Босна и Хецеговина. Основну школу је завршио у Бањој Луци са Вуковом дипломом. Средњу електротехничку школу је завршио 2000. године у Бањој Луци. Током школовања, неколико пута је учествовао на Републичким такмичењима из Математике.

У марту 2007. године је завршио студије на Одсеку за електронику и телекомуникације Електротехничког факултета у Бањој Луци, са просечном оценом положених испита 8,80 (на дипломском испиту оцена 10). Непосредно по завршетку студија запослио се на Катедри за телекомуникације, где и сада ради. У звање вишег асистента на Катедри је изабран у фебруару 2017. године. Екипа из Телекомуникација, чији је био вођа, остваривала је запажен резултат на међународним сусретима студената електротехнике. Учесник је неколико пројеката Министарства Науке и технологије Републике Српске и једног међународног пројекта. Коаутор је уџбеника из Дигиталних телекомуникација.

Докторске студије на Електротехничком факултету Универзитета у Београду је уписао под менторством студијског истраживачког рада проф. др Слободана Т. Јовичића. Просечна оцена положених испита на докторским студијама је 9,90.

Члан је професионалног удружења IEEE. Ожењен је и отац две ћерке.

Прилог Г: Изјаве аутора

Изјава о ауторству

Име и презиме аутора Јован Галић
Број индекса 5036/16

Изјављујем

да је докторска дисертација под насловом

„Препознавање мултимодалног говора засновано на статистичком приступу”

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 18.12.2018.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Јован Галић
Број индекса 5036/16
Студијски програм Телекомуникације
Наслов рада "Препознавање мултимодалног говора засновано на статистичком приступу"
Ментор др Драгана Шумарац Павловић, редовни професор Електротехничког факултета, Универзитет у Београду

Потписани Јован Галић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**
Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.
Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 18.12.2018.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

“Препознавање мултимодалног говора засновано на статистичком приступу“ која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 18.12.2018.

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.
2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.
3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране

аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.