Универзитет у Новом Саду

Факултет техничких наука

Драгиша Мишковић

КОНТЕКСТНО ЗАВИСНО ПРЕПОЗНАВАЊЕ ГОВОРА У ИНТЕРАКЦИЈИ ИЗМЕЂУ ЧОВЕКА И МАШИНЕ

Докторска дисертација

Нови Сад, 2017

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

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, **РБР**:	
Идентификациони број, **ИБР**:	
Тип документације, **ТД**:	Монографска документација
Тип записа, **ТЗ**:	Текстуални штампани материјал
Врста рада, **ВР**:	Докторска дисертација
Аутор, **АУ**:	Драгиша Мишковић, дипл. инг. - мастер
Ментор, **МН**:	проф. др Владо Делић и доц. др Милан Гњатовић
Наслов рада, **НР**:	Контекстно зависно препознавање говора у интеракцији између човека и машине
Језик публикације, **ЈП**:	Српски
Језик извода, **ЈИ**:	Српски, енглески
Земља публиковања, **ЗП**:	Република Србија
Уже географско подручје, **УГП**:	Војводина
Година, **ГО**:	2017.
Издавач, **ИЗ**:	Ауторски репринт
Место и адреса, **МА**:	Трг Доситеја Обрадовића 6, Нови Сад
Физички опис рада, **ФО**: (поглавља/страна/цитата/табела/слика/графика/прилога)	7 поглавља/ 138 страна/ 60 референци / 11 табела/ 30 слика/ 3 прилога
Научна област, **НО**:	Електротехничко и рачунарско инжењерство
Научна дисциплина, **НД**:	Телекомуникације и обрада сигнала
Предметна одредница/Кључне речи, **ПО**:	Препознавање говора, контекст, фокусно стабло, разумевање говора, интеракција између човека и машине
УДК	
Чува се, **ЧУ**:	Библиотека Факултета техничких наука
Важна напомена, **ВН**:	
Извод, **ИЗ**:	Поред великог значаја контекстуалних информација при разумевању говора, њихова обрада и употреба у савременим системима за аутоматско препознавање говора је веома ограничена, што знатно нарушава перформансе препознавања у реалним условима употребе. Стога, уколико желимо да се карактеристике ових система приближе људским, неопходно је укључити контекст у адекватном обиму. У овој тези је представљен нови методолошки приступ контекстно зависном препознавању говора у интеракцији између човека и машине. На методолошком нивоу, овај приступ је хибридан, јер интегрише статистичке и симболичке методе, и когнитивно инспирисан, јер узима у обзир увиде у резулатате истраживања из области неурокогнитивних наука. Основни принцип је да се оцењивање хипотеза система за препознавање врши на основу њихове контекстуалне усклађености, информационог садржаја и семантичке исправности. Приступ је илустрован прототипским имплементацијама за конкретне домене интеракције.
Датум прихватања теме, **ДП**:	24.09.2015.

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

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Датум одбране, **ДО**:			
Чланови комисије, **КО**:	Председник:	др Бранислав Боровац редовни професор Факултет техничких наука, Нови Сад	
	Члан:	др Зоран Перић редовни професор Електронски факултет, Ниш	
	Члан:	др Марко Јанев научни сарадник Математички институт САНУ, Београд	
	Члан:	др Никша Јаковљевић доцент Факултет техничких наука, Нови Сад	Потпис ментора
	Члан, ментор:	др Милан Гњатовић доцент Факултет техничких наука, Нови Сад	
	Члан, ментор:	др Владо Делић редовни професор Факултет техничких наука, Нови Сад	

Образац **Q2.НА.06-05**- Издање 1

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

KEY WORDS DOCUMENTATION

Accession number, **ANO**:	
Identification number, **INO**:	
Document type, **DT**:	Monograph documentation
Type of record, **TR**:	Textual printed material
Contents code, **CC**:	PhD thesis
Author, **AU**:	Dragiša Mišković, M.Sc.E.E
Mentor, **MN**:	prof. Vlado Delić, PhD prof. Milan Gnjatović, PhD
Title, **TI**:	Context-Dependent Speech Recognition in Human-Machine Interaction
Language of text, **LT**:	Serbian
Language of abstract, **LA**:	Serbian, English
Country of publication, **CP**:	Serbia
Locality of publication, **LP**:	Vojvodina
Publication year, **PY**:	2017
Publisher, **PB**:	Faculty of Technical Sciences
Publication place, **PP**:	Trg Dositeja Obradovića 6, Novi Sad
Physical description, **PD**: (chapters/pages/ref./tables/pictures/graphs/appendixes)	7 chapters/ 138 pages/ 60 references/ 11 tables/ 30 figures/ 3 appendices
Scientific field, **SF**:	Electrical Engineering
Scientific discipline, **SD**:	Telecommunications
Subject/Key words, **S/KW**:	Speech recognition, context, focus tree, language comprehension, human-robot interaction
UC	
Holding data, **HD**:	The library of the Faculty of Technical Sciences, Novi Sad, Serbia
Note, **N**:	
Abstract, **AB**:	Although the importance of contextual information in speech recognition has been acknowledged for a long time now, it remained clearly underutilized even in state-of-the-art speech recognition systems. This thesis introduces a novel, methodologically hybrid approach to the research question of context-dependent speech recognition in human-machine interaction. To the extent that it is hybrid, the approach integrates aspects of both statistical and representational paradigms. The aim of this thesis is to extend the standard statistical pattern matching approach with a cognitively-inspired and analytically tractable model with explanatory power. This methodological extension allows for accounting for contextual information which is otherwise unavailable in speech recognition systems, and using it to improve post-processing of recognition hypotheses. The thesis introduces an algorithm for evaluation of recognition hypotheses, illustrates it for concrete interaction domains, and discusses its implementation within two prototype conversational agents.
Accepted by the Scientific Board on, **ASB**:	24.09.2015
Defended on, **DE**:	

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

KEY WORDS DOCUMENTATION

Defended Board, **DB**:	President:	Branislav Borovac, PhD full professor Faculty of Technical Sciences, Novi Sad	
	Member:	Zoran Perić, PhD full professor Faculty of Electronic Engineering, Niš	
	Member:	Marko Janev, PhD research associate Mathematical Institute of Sebian Academy of Sciences and Arts, Belgrade	
	Member:	Nikša Jakovljević, PhD assistant professor Faculty of Technical Sciences, Novi Sad	Menthor's sign
	Member, Mentor:	Milan Gnjatović, PhD assistant professor Faculty of Technical Sciences, Novi Sad	
	Member, Mentor:	Vlado Delić, PhD full professor Faculty of Technical Sciences, Novi Sad	

Obrazac **Q2.HA.06-05**- Izdanje 1

Садржај

Листа слика

Листа табела

Листа скраћеница

ДКТ Дискретна косинусна трансформација (енгл. *discrete cosine transform, DCT*)

ДФТ Дискретна Фуријеова трансф ормација (енгл. *discrete Fourier transform, DFT*)

ЕЕГ Електроенцефалографија (енгл. *electroencephalography, EEG*)

ЕП Евоцирани потенцијали (енгл. *event related potentials, ERP*)

МФКК Мел-фреквенцијски кепстрални коефицијенти (енгл. *mel-frequency cepstral coefficients, MFCC*)

ПЛП Перцептивна линеарна предикција (енгл. *perceptual pinear prediction, PLP*)

СММ Скривени Марковљеви модели (енгл. *hidden Markov models, HMM*)

Захвалница

Ова дисертација је резултат истраживања на пројектима Министарства просвете, науке и технолошког развоја бр. ТР32035 и ИИИ44008.

Велику захвалност дугујем својим менторима, доц. др Милану Гњатовићу и проф. др Влади Делићу. Дали су ми мотивацију и велику подршку у научно-истраживачком раду, сваки на свој начин.

Такође, захваљујем проф. др Браниславу Боровцу на несебичној помоћи у бројним ситуацијама. Захваљујући њему и колегама са Катедре за мехатронику, роботику и аутоматизацију, мој рад је нашао праву примену у свету роботике.

Желим да се захвалим и свим члановима комисије који су својим сугестијама унапредили ову дисертацију.

Ипак, највећу захвалност дугујем својој породици.

ФТН, Нови Сад, 2017. *Драгиша Мишковић*

Сажетак

Процеси људске перцепције, укључујући препознавање и разумевања природног говора, се никад не одвијају ван контекста. Приликом интерпретирања говорног сигнала, људи узимају у обзир различите аспекте интеракције – очекивање шта говорник може рећи у датом тренутку, опште знање о свету, претходно искуство слушаоца итд. На тај начин се у свакодневној интерпретацији говора превазилазе различите сметње у преносу говорног сигнала. Насупрот великом значају контекстних информација, њихова обрада и употреба у савременим системима за аутоматско препознавање говора је веома ограничена, што знатно нарушава перформансе препознавања у реалним условима употребе. Стога, уколико желимо да се карактеристике ових система приближе људским, неопходно је укључити контекст у адекватном обиму.

У овој тези је представљен нови методолошки приступ контекстно зависном препознавању говора у интеракцији између човека и машине. На методолошком нивоу, овај приступ је хибридан, јер интегрише статистичке и симболичке методе, и когнитивно инспирисан, јер узима у обзир увиде у резулатате истраживања из области неурокогнитивних наука. Основни принцип је да се оцењивање хипотеза система за препознавање врши на основу њихове контекстуалне усклађености, информационог садржаја и семантичке исправности. Важна особина овог

приступа је да је независан од дијалошког домена, што је у тези илустровано кроз изабране дијалошке домене при имплементацији прототипског система. Коначно, у тези се демонстрира и анализира повећање тачности аутоматског препознавања говора применом овог приступа.

Abstract

Although the importance of contextual information in speech recognition has been acknowledged for a long time now, it remained clearly underutilized even in state-of-the-art speech recognition systems. This thesis introduces a novel, methodologically hybrid approach to the research question of context-dependent speech recognition in human-machine interaction. To the extent that it is hybrid, the approach integrates aspects of both statistical and representational paradigms. The aim of this thesis is to extend the standard statistical pattern matching approach with a cognitively-inspired and analytically tractable model with explanatory power. This methodological extension allows for accounting for contextual information which is otherwise unavailable in speech recognition systems, and using it to improve post-processing of recognition hypotheses. The thesis introduces an algorithm for evaluation of recognition hypotheses, illustrates it for concrete interaction domains, and discusses its implementation within two prototype conversational agents.

Глава 1

Увод

"...when we listen to a person speaking or read a page of print, much of what we think we see or hear is supplied from our memory. We overlook misprints, imagining the right letters, though we see the wrong ones; and how little we actually hear, when we listen to speech, we realize when we go to a foreign theatre; for there what troubles us is not so much that we cannot understand what the actors say as that we cannot hear their words. The fact is that we hear quite as little under similar conditions at home, only our mind, being fuller of English verbal associations, supplies the requisite material for comprehension upon a much slighter auditory hint."

— W. James, Talks to Teachers on Psychology and to Students on Some of Life's Ideals, 1899

У последњих неколико деценија је дошло до значајних напредака у развоју система за аутоматско препознавање говора (енгл. *automatic speech recognition - ASR*), што је резултовало бројним применама ове технологије, нпр., у позивним центрима, системима за пружање информација или резервације карата, системима за гласовно бирање и диктирање, итд.

Међутим, перформансе ових система су још увек ограничене. Спектар појава које отежавају ефикасно препознавање је веома широк, и укључује позадинску буку, реверберацију, сметње на преносном каналу,

особености говорника, непланиране промене теме, итд. Да би се адекватно обрадили овакви језички феномени, који су инхерентно присутни у говорном дискурсу, неопходно је укључити шири опсег информација, попут акустичких, фонетских, семантичких, итд.

Важно је нагласити да још увек не постоји задовољавајуће решење за овај проблем. Кључна предност људског разумевања говора у односу на рачунарске системе је што људи користе контекстуалне информације које омогућавају тачнију интерпретацију реченог. Овом чињеницим је инспирисан приступ аутоматском препознавању говора представљен у тези.

1.1 Предмет истраживања

Сам процес препознавања говора од стране рачунара је веома сложен. За разлику од људског мозга, који паралелно обрађује различите стимулансе и, потпомогнут искуством, ефикасно претпоставља или предвиђа делове комуникације код којих постоје акустичке сметње, у дигиталном домену проблеми се решавају парцијално. Постојећи системи за препознавање говора решавају поједностављени проблем, тј. функционишу над ограниченим речником, препознају секвенце речи у оквиру одређених синтаксних правила, прилагођавају се циљаном говорнику итд.

У последње четири деценије, доминантна методологија у препознавању говора се заснива на статистичким приступима. У основи ових приступа, статистички алгоритми машинског учења примењују се над одабраним корпусима у циљу што бољег акустичког и језичког моделовања. Иако су овакви приступи постигли значајне резултате, они су ограничени, јер се не може очекивати да језички корпуси за обуку, без

обзира на то колико су велики и пажљиво креирани, садрже све манифестације релевантних феномена у говору.

Са друге стране, процес људске перцепције и разумевања говора се заснива на интерпретацији акустичког сигнала као носиоца различитих типова информација. Поред примарног, лингвистичког садржаја, различите информације о говорнику, окружењу и осталим аспектима интеракције саставни су део говорног сигнала. У складу са тим, људско интерпретирање говора се у великој мери ослања на очекивања шта говорник може рећи у датом тренутку. Другим речима, људска перцепција и разумевање говора узимају у обзир контекст интеракције, намеру говорника, опште знање о свету и претходно искуство.

У овом раду је представљен нови методолошки приступ моделовању контекста интеракције. Основна идеја предложеног приступа усмерена је у правцу креирања система који интегрише статистички препознавач говора и когнитивно инспирисано моделовање контекста и осталих аспеката интеракције. У оквиру обједињеног система, препознавање говора се врши у две фазе. У првој фази, статистички препознавач говора задужен је за генерисање резултата у виду скупа највероватнијих хипотеза (секвенце речи). Друга фаза је заснована на примени фокусног стабла које моделује контекст интеракције између човека и машине. Овај модел, заједно са алгоритмима предложеним у овој тези, омогућава контекстно зависно оцењивање појединачних хипотеза. Инспирисан људском способношћу паралелне обраде различитих информација у току разговора, предложени приступ процењује релевантност и валидност лингвистичког садржаја хипотезе на основу различитих критеријума.

1.2 Структура тезе

Свеобухватни циљ тезе је да представи развој софтверског модула за контекстно зависно препознавање говора, који ће омогућити стварање унапређеног система за аутоматско препознавање говора. Поглавља су организована тако да сукцесивно прикажу кораке у развоју оваквог система.

У другом поглављу је, кроз осврт на историју развоја изабраних алгоритама и модела у области обраде природних језика, приказан след догађаја који су обликовали развој данашњих система за аутоматско препознавање говора. Посебно су наглашене прекретнице узроковане применом скривених Марковљевих модела (СММ, енгл. *hidden Markov models, HMM*), и у скорије време, вештачких неуронских мрежа (енгл. *artificial neural networks*). У анализи савремених приступа, посебна пажња посвећена је алгоритмима намењеним приближавању перформанси аутоматских препознавача говора људским могућностима. Анализом основних методолошких поставки истакнута је њихова заједничка карактеристика — веома ограничене могућности моделовања контекста и тока интеракције. На основу идентификације проблема изложени су доприноси тезе — нови методолошки приступ и адекватно прилагођена архитектура система.

Треће поглавље детаљније приказује поједине делове актуелних система за препознавање говора. Поред основних теоријских поставки и математичких формализама на којима се базира статистички приступ, поглавље даје и увид у реалне проблеме који прате реализацију аутоматских препознавача говора. Изложена је њихова модуларна архитектура, након чега следи детаљни приказ сваког појединачног подсистема. Разматрање обухвата поступак обраде акустичког сигнала и издвајање обележја, акустичко моделовање уз осврт на статистички механизам

скривених Марковљевих модела, језичко моделовање засновано на н–грамима и процес декодовања као завршни корак обраде акустичког сигнала при препознавању говора.

У четвртом поглављу изложен је основни концепт фокусног стабла као приступа који омогућава моделовање контекста у интеракцији између човека и машине. Наведене су релације овог модела са ранијим истраживањима у областима неурокогнитивних наука и рачунарске лингвиситке. Посебна пажња посвећена је механизму фокуса пажње током процеса обраде лингвистичких стимуланса. У циљу илустрације овог процеса, дат је прикладни пример из домена интеракције између човека и робота.

Пето поглавље започиње освртом на резултате истраживања изабраних електроенцефалографских сигнала као индикатора когнитивних активности током људског разумевања говора. Актуелна тумачења тзв. евоцираних потенцијала послужила су као инспирација за развој алгоритма који, применом фокусног стабла, симулира когнитивно оптерећење слушаоца при разумевању природног говора. На основу тога, у поглављу је прво изложен поступак за одређивање параметара намењених процени комплексности дијалошких чинова. Након тога, дефинисани су информациони, семантички и лексички критеријуми за оцењивање хипотеза препознавача. Уз детаљно образложење алгоритма, дата је и његова прототипска илустрација на изабраном домену интеракције. Основна намена прототипског модела је да демонстрира свеобухватност приступа, те је у поглављу показана функционалност алгоритма за више карактеристичних момената интеракције.

Шесто поглавље разматра реализовани интегрисани систем. Да би се демонстрирала ефективност у реалним условима примене, извршено је оцењивање над знатно већим језичким корпусом, који покрива домен интеракције између корисника и мобилног телефона. Применом на

овом корпусу који садржи приближно 1500 изговора различитих говорника, демонстрирано је унапређење препознавања кроз упоредне резултате базног и предложеног система.

Закључак и општа анализа предложеног приступа дати су у седмом поглављу.

У првом прилогу је дата листа говорних чинова које су субјекти изговарали приликом продукције језичког корпуса коришћеног за тестирање интегрисаног система, представљеног у шестом поглављу. У другом прилогу је приказано фокусно стабло које моделује домен интеракције између корисника и интегрисаног система. У трећем прилогу су дати изабрани примери препознавања говорних чинова од стране интегрисаног система.

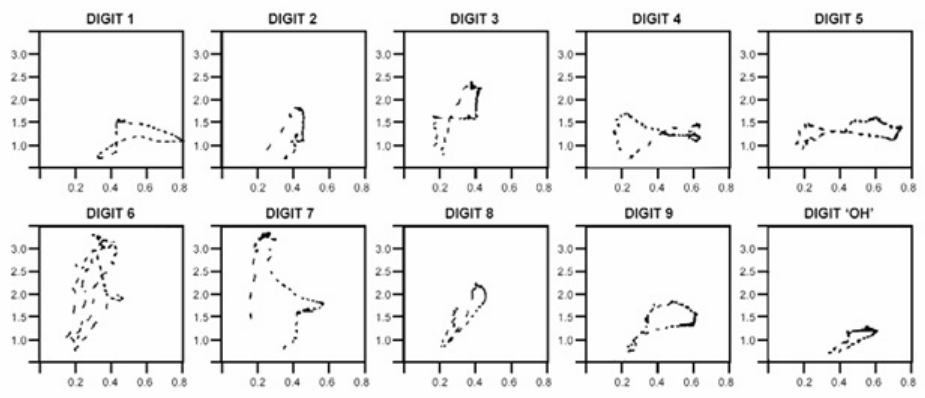
Глава 2

Преглед стања у области и доприноса тезе

У овом поглављу пружен је приказ историје досадашњег развоја система за аутоматско препознавање говора. Поред кључних момената у току развоја, анализирана је и методологија, са посебним освртом на ограничења присутна у савременим системима. Као један од начина за превазилажење ових ограничења, предложене су методолошке и архитектуралне промене у оквиру тезе.

2.1 Историја развоја аутоматског препознавања говора

Истраживања у области препознавања и перцепције људског говора бележе дугу историју [7]. Већ у првој половини двадесетог века, у Беловим лабораторијама се откривају везе између карактеристика појединих гласова и њиховог спектралног облика [8]. Значај овог открића огледа се у чињеници да се и данашњи системи за препознавање говора базирају на различитим варијантама процене амплитудског спектра сигнала.



Слика 2.1: Међусобне зависности првог и другог форманта при изговору цифара (преузето из [1])

Даљи развој је био заснован на покушајима да се на дискриминативни начин опише акустичка реализација основних јединица говора — фонема. Препознати су форматни (тј. формантне фреквенције) као региони у амплитудском спектру сигнала у којима је концентрисана енергија. Као последица, педесетих година се јавља први систем за препознавање изоловано изговорених цифара [1]. Могао је да препознаје искључиво једног говорника (оног на основу чијих исказа је обучен), и заснивао се на процени форманата из припадајућих самогласника садржаних у изговору цифара. На основу карактеристичних зависности између првог и другог форманта у изговору сваке цифре, реализован је уређај који је упоређивао облике (енгл. *pattern matching*) са референтним вредностима. Слика 2.1 приказује шаблоне чије је детектовање у основи овог препознавача.

Шездесетих година двадесетог века развијено је неколико система за аутоматско препознавање говора веома скромних могућности. Њихова заједничка карактеристика је да су препознавали издвојене сегменте

говора (нпр. самогласнике), без могућности да моделују њихову временску динамику. Увођењем метода динамичког програмирања и временског поравнања, створени су предуслови да препознавачи говора у процесу декодовања узму у обзир варијабилност трајања појединачних гласова. У складу са тим, у овом периоду се развијају основни алгоритми динамичког програмирања који ће довести до система за препознавање говора заснованих на Витербијевом алгоритму [9] и методама динамичког усклађивања у времену (енгл. *dynamic time warping*) [10].

Укључивање америчког министарства одбране као спонзора пројеката из области препознавања говора обележило је истраживања седамдесетих и осамдесетих година прошлог века. У оквиру Агенције за напредна истраживања (енгл. *Advanced Research Projects Agency, ARPA*), као циљ се поставља континуално препознавање говора са речником од 1000 речи. У ово истраживање се укључују бројне институције, што резултује новим алгоритмима који ће значајно унапредити перформансе будућих система (претрага по графу, увођење различитих нивоа знања кроз вишеструки процес декодовања, структура решетке са алтернативним хипотезама, итд.). Интересантно је поменути систем ХАРПИ (енгл. *HARPY*) развијен на Универзитету Карнеги Мелон [11], који је користио мрежу коначних аутомата за лексичку репрезентацију речи. У оквиру ове мреже, систем је претрагу ограничавао на одређени број „најбољих" путања по синтаксним и акустичким критеријумима (енгл. *beam search*). Приступ је омогућио да претрага буде линеарна функција времена, и тиме отворио врата за креирање употребљивих апликација у области препознавања говора. Заједно са бројним хеуристичким оптимизацијама, овај систем је већ тада демонстрирао оправданост укључивања виших нивоа информација у процес претраге. Њихова примена је у стању да компензује инхерентну некомплетност акустичког моделовања (јер акустички модели никад не садрже све акустичке варијације).

Ова унапређења резултовала су појавом система који препознају више стотина речи, што је омогућило њихову комерцијалну експлоатацију. Следећа велика прекретница настаје осамдесетих година двадесетог века. Алгоритми за препознавање еволуирају ка приступима заснованим на статистичким моделима, уместо на базичном препознавању шаблона. То, пре свега, укључује примену скривених Марковљевих модела (енгл. *hidden Markov models*), раздвајање акустичких и језичких модела, итд. Ови кораци су омогућили убрзани и континуирани развој препознавача говора у следећих пар деценија. Уместо препознавања изолованих речи, нови системи су препознавали континуални говор. Да би се редуковао простор претраге, у оквиру језичких модела се дефинишу механизми који одређеним секвенцама лексичких симбола (фонеме или речи) придружују вероватноћу појаве у датом ограниченом контексту. Намера је да се у процес уведу граматичка и синтаксна правила језика — у виду граматика (енгл. *grammar*), или у форми вероватноће појаве у тренутном лексичком контексту (применом модела н-грама — енгл. *n-gram model.*).

Укључивање CMM у процес препознавања говора мотивисано је потребом да се на адекванији и компактнији начин испрати природа говорног сигнала — нестационарни сигнал у ком је информација смештена у временским варијацијама амплитудског спектра [12]. Уместо метода за временско усклађивање, архитектура CMM са различитим бројем стања и припадајућим функцијама расподеле је омогућавала да се ове варијације детектују и временски моделују. Увођење концепта смеше расподела, уместо јединствене функције расподеле, придружене сваком стању CMM [13], значајно је поспешило примену CMM у препознавању говора, нарочито у системима независним од говорника. Такође, битно је нагласити и способност CMM за ефикасну интеграцију различитих

нивоа знања при развоју система. Ове особине су омогућиле да скривени Марковљеви модели заузму доминантну улогу у реализацији система за препознавање континуалног говора над великим речницима (енгл. *large-vocabulary continuous speech recognition*). Више детаља о њиховој примени дато је у следећем поглављу.

Упоредо са развојем препознавача базираних на примени скривених Марковљевих модела у комбинацији са смешама Гаусовим расподелама, деведесетих година двадесетог века се јавља још једна технологија у овој области — вештачке неуронске мреже. Њихова примена у системима за препознавање говора била је очекивани корак у настојањима да се перформансе поменутих система приближе људским. Захваљујући структури која „имитира“ мрежу неурона у људском мозгу, вештачка неуронска мрежа мења основну парадигму рачунарске обраде информација. Уместо ослањања на сложен процесор и брзу, локалну меморију, обрада података се врши паралелно кроз мрежу релативно простих елемената распоређених у слојеве, чије се везе моделују кроз процес обуке (инспирисано синапсама у људском мозгу).

У основи, примена неуронских мрежа у процесу препознавања говора започиње у области акустичког моделовања као замена за смеше Гаусових расподела. У оквиру ове намене, вештачке неуронске мреже превазилазе недостке модела са Гаусовим смешама, и омогућавају моделовање нелинеарности у простору улазних података (енгл. *non-linear manifolds*)[1]. У почетку, због изостанка рачунарских ресурса, неадекватних алгоритама и малих база за обуку, мреже су садржале само

[1]Наиме, обрада говора, као и осталих појава из реалног света, резултује издвајањем обележја високе димензионалности (нпр. најчешћа обележја у препознавању говора представљају податке у простору чија је димензионалност приближна 40). Ако, са друге, стране анализирамо процес артикулације, можемо запазити да он зависи од знатно мање параметара, који се током говора модулишу. Другим речима, кохерентна структура података узрокује високу корелисаност улазних вредности (нпр. корелисаност говорног сигнала из суседних сегмената). Као последица, често је присутно размештање података у облику закривљености у простору улазних

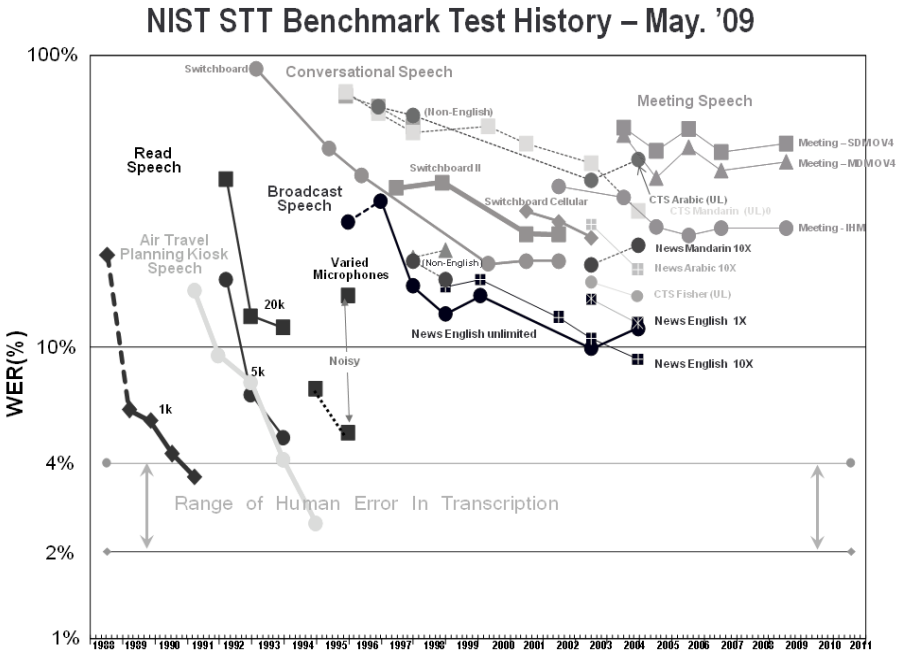
по један нелинеарни (скривени) слој. Услед тога, нису биле у стању да значајније побољшају моделовање у односу на Гаусове смеше. У скорије време, уз огромне базе за обуку и моћне рачунарске ресурсе, долази до значајније примене дубоких неуралних мрежа (енгл. *deep neural networks*) које садрже већи број скривених слојева [14] и које су у стању да изврше адекватно предвиђање стања скривених Марковљевих модела у процесу акустичког моделовања.

2.2 Анализа актуелних приступа

Алгоритми и поступци развијени осамдесетих и деведесетих година двадесетог века поставили су теоријске основе већине савремених система за препознавање говора. У већини система акустичка варијабилност говорног сигнала је моделована применом скривених Марковљевих модела у комбинацији са смешама Гаусових расподела или вештачким неуронским мрежама, док се језик моделује помоћу граматичких правила или н-грама (у случају препознавача са великим речницима). Заједничко за ове системе је представљање говора као секвенце акустичких обележја (опсервација) добијених анализом улазног аудио-сигнала. Процес обуке подразумева велике језичке корпусе, који садрже акустичке секвенце (за тренинг акустичких модела) односно текстуалне секвенце речи (за тренинг језичког модела).

На слици 2.2 је илустровано како су се током времена мењале перформансе и сложеност система за препознавање говора. Тестови вршени од стране Националног института за стандарде и технологију САД (енгл. *National Institute of Standards and Technology – NIST*) су оцењивали перформансе система у различитим областима примене и амбијенталним условима. Резултати у облику грешке на нивоу речи

обележја. Гаусове расподеле нису адекватне за моделовање оваквих просторних расподела података, што је повећало заступљеност неуронских мрежа.



Слика 2.2: Историјски преглед перформанси система за препознавање говора у различитим областима примене. Преузето са веб-сајта Лабораторије за информационе технологије Националног института за стандарде и технологију САД `https://www.nist.gov/itl` (мај 2016.)

(енгл. *word error rate*), показују да се у контролисаним условима (попут читања текста, дијалошких система са ограниченим доменом и предвидљивим током интеракције) карактеристике препознавача говора приближавају људским. Међутим, у реалним условима, перформансе система су знатно лошије, а истраживања у последњих пар деценија нису донела значајнија побољшања. Ипак, тренутно стање је такво да су статистички приступи и даље доминантни, а повећање робустности и превенција грешака су предмет истраживања у оквиру овог концептуалног приступа. Оваква истраживања су фокусирана на појединачне аспекте препознавања говора (проблеми препознавања говора изазвани

различитим амбијенталним условима, различитим говорницима и дијалектима, неговорним сегментима, речима изван речника, итд.) и не посматрају проблем препознавања интегративно.

У наставку ће бити размотрене основне методолошке поставке савремених приступа препознавању говора, као и њихова улога у оквиру система за дијалошку интеракцију између човека и машине.

2.2.1 Методолошки аспекти

Општа карактеристика свих статистичких приступа у оквиру машинског учења је генерализација знања на основу великих база података. Предности оваквог приступа су сасвим очигледне. Ако говоримо о препознавању говора, поменута методологија омогућава креирање употребљивих система за препознавање говора, без дубљег разумевања језика или принципа људског разумевања говора. Као резултат, добијамо препознаваче говора чије карактеристике у великој мери зависе од језичког корпуса (тј. база података које садрже снимке говора, текстуалне документе, итд.) на ком су обучавани. Уобичајени приступ је да се језички модели обучавају на великим текстуалним корпусима (књиге, новински чланци, итд.), а акустички модели на говорним корпусима. Овако развијени системи имају задовољавајуће карактеристике при условима коришћења за које су намењени, али знатно лошије перформансе у реалним условима, нпр. спонтани говор.

Као илустрација, у раду [2] дат је осврт на креирање корпуса спонтаног говора на јапанском језику. Мотивација за креирање овог корпуса је очигледна — савремени системи у тестовима над одговарајућим корпусима постижу тачност на нивоу речи од преко 95%, док је у случају спонтаног говора она практично преполовљена. Такође, приказани су и експерименти са постепеним повећањем обучавајућих скупова у циљу праћења њиховог утицаја на тачност крајњег система. Корпус за обуку

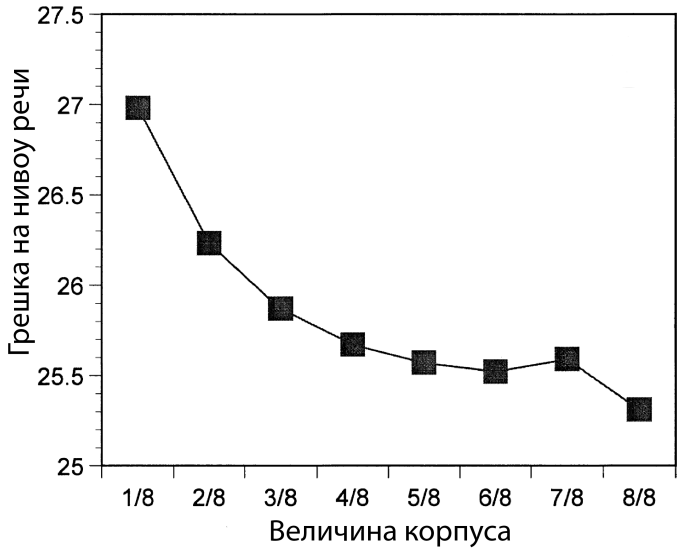
језичких модела (величине 6,84 милиона речи) и корпус аудио-снимака (510 сати говора) подељени су на мање целине, које су постепено укључиване у обуку. Слика 2.3 приказује резултате тестова који испитују зависност грешке на нивоу речи од величине обучавајућих скупова за акустичке моделе. Слични резултати добијени су и у случају обуке на целокупној аудио-бази, док се величина текстуалног корпуса постепено повећава. Поред приметног смањења грешке препознавања на нивоу речи при повећању скупова за обуку, експерименти су показали још једну карактеристику статистичког приступа. Након почетних позитивних ефеката, систем у одређеном моменту улази у засићење, након чега више није могуће остварити побољшање перформанси повећањем база за обуку[2].

Још један од битних недостатака актуелних система за препознавање говора је и немогућност моделовања контекста. Разлог је крајње очигледан — без обзира на величину, корпуси за обуку нису у стању да покрију све контекстно зависне језичке феномене [15]. Притом, концептуализација контекста у статистичким проступима је веома редукована, што ће бити размотрено у наставку.

(i) Акустички модели узимају у обзир контекст на нивоу једне речи.

Један од чинилаца који утичу на избор најбоље хипотезе при препознавању говора је степен усаглашености између улазног скупа акустичких опсервација и одговарајућих акустичких модела (ово ће бити детаљније размотрено у поглављу 3.2). У општем случају, акустичка репрезентација речи се своди на секвенцу скривених

[2]Приказ резултата на слици 2.3 може навести на закључак да употреба целог корпуса за обуку акустичких модела значајно смањује грешку на нивоу речи. Међутим, апсолутна вредност грешке од 25,3% у последњем мерењу није резултат повећања обучавајућег скупа, већ измењених услова експеримента. Више детаља може се наћи у [2].



Слика 2.3: Грешка на нивоу речи у зависности од величине корпуса за статистичку обуку акустичких модела. Систем је тестиран са спонтаним говором. Преузето из [2]. Корпус садржи 510 сати говора.

Марковљевих модела који представљају појединачне фонеме. При томе, узимају се у обзир ефекти коартикулације, те се креирају различити модели фонема у складу са положајем фонеме у речи и њеном најближом околином (тј., моделује се ефекат околних гласова на изговор појединаче фонеме). Типична реализација ових модела је у виду тзв. трифона, тј. модела фонема са припадајућим левим и десним контекстом. Другим речима, акустичко моделовање узима у обзир ограничени контекст који осликава само динамику вокалног тракта.

(ii) Језички модели узимају у обзир контекст на нивоу реченице.

Други критеријум при одређивању оптималних хипотеза препознавача је вероватноћа припадајуће секвенце речи. Овај параметар репрезентује природу самог језика и процењује се на основу н-грама (ово ће бити детаљније размотрено у поглављу 3.3). Основна претпоставка код овог моделовања је да вероватноћа појаве одређене речи зависи само од $n-1$ претходних при чему се у практичним реализацијама n своди на 2 или 3 (биграми или триграми). Последично, н-грамима редукујемо историју речи на две или три претходне речи, и тиме обухватамо само синтаксна правила која дефинишу уобичајени положај речи у реченици. Лексичке корелације између дијалошких чинова или семантичке улоге речи у реченици нису укључени оваквим видом моделовања.

(iii) Ниједан од ових модела на укључује контекст на нивоу дијалога.

Недостатак контекста на нивоу дијалога је директна последица описаног приступа за моделовање језика, и не може се решити повећањем корпуса за обуку. Иако савремени системи за тренинг користе огромне текстуалне корпусе са више милијарди речи [16], изостанак информације о ширем дијалошком контексту онемогућава корективне механизме у процесу препознавања. Као резултат, имамо смањену робустност система у условима примене који акустички или језички одступају од услова обучавајућих скупова. Акустичке сметње (нпр. бука, позадински шум), неуобичајен говор (измењен акценат, брз или испрекидан изговор итд.), су неки од узрока значајног повећања грешке препознавања [17] [18] [19].

У скоријим приступима препознат је значај контекстуалних информација, али се оне углавном уводе у складу са статистичком парадигмом или као парцијално решење одређеног проблема за

унапред дефинисани сценарио [20]. Карактеристично решење је и да се на основу корпуса који садржи историју бројних дијалога креира још један статистички алат, који предвиђа шта корисник може рећи у наставку интеракције. У раду [21], оваква методологија се користи да би се креирали додатни н-грами који садрже делове дијалога репрезентоване на високом семантичком нивоу. На основу тога, проверава се усклађеност садржаја хипотеза препознавача са током дијалога предвиђеним оваквим статистичким мехнизмом, и врши се њихово додатно оцењивање.

Такође, у раду [22] приказан је поступак креирања контекстно „осетљивог" језичког модела који узима у обзир историју дијалога. Пошто, осим у случају тривијалних дијалошких система, задатак праћења историје није једноставан, врши се кластеризација корисничких дијалошких чинова тј. груписање семантички сличних дијалошких чинова. Добијени скупови се користе за обуку специфичних језичких модела, који се користе у даљем препознавању.

У раду [23] описан је начин како да се на статистичким основама омогући семантичка, синтаксна, лексичка и контекстуална обрада резултата препознавања. Аутори предлажу формирање семантичко-синтаксних шаблона на основу статистичке анализе дијалошког корпуса, а као карактеристичани моменат истичу примену у ситуацијама када је тачност препознавања веома мала (нпр. почетак интеракције између човека и машине када систем није у могућности да примени прилагођене језичке моделе). Полазна основа овог приступа је груписање различитих инстанци дијалога, како би се креирали подскупови везани за одређене корисничке дијалошке чинове. Након тога, креирају се класе речи (свака представљена низом кључних речи), шаблони за граматичка правила, синтакстно-семантички модели (који репрезентују

генералну структуру реченице у одређеном домену) и лексички модели (на основу матрице конфузије речи).

Са становишта корективног механизма, и слично приступу изложеном у овој тези, аутори на основу скупљених информација врше проверу резултата класичног препознавача говора. Међутим, предложени систем практично обједињује разне статистичке алгоритме за унапређење препознавања говора, и његов је недостатак то што могућности за превазилажење грешака препознавања директно зависе од величине корпуса. Такође, зависност од дијалошког домена и потреба да се одређене лингвистичке информације уносе ручно (нпр. граматичка правила се креирају за сваки домен посебно), додатно ограничавају домете оваквог приступа.

Као одступање од ове доминантно статистичке линије истраживања, рад [24] представља систем који уводи опште знање о свету у процес препознавања говора. Уз осврт на недостатак класичног приступа моделовању језика (ограничена историја и условљеност језичким корпусом), аутори у свом приступу користе детаљно разрађену семантичку мрежу ентитета Концептнет (енгл. *ConceptNet*), креирану на Технолошком институту Масачусетс (енгл. Massachusetts Institute of Technology), као рачунарску репрезентације знања о свету. Након препознавања говора и генерисања скупа хипотеза, систем процењује валидност хипотеза и одбацује оне хипотезе које са становишта доступног знања немају смисла. На пример, прворангирана хипотеза статистичког препознавача:

my bike has a squeaky ***break***

може да буде одбачена као семантички некоректна, и замењена реченицом

my bike has a squeaky ***brake***

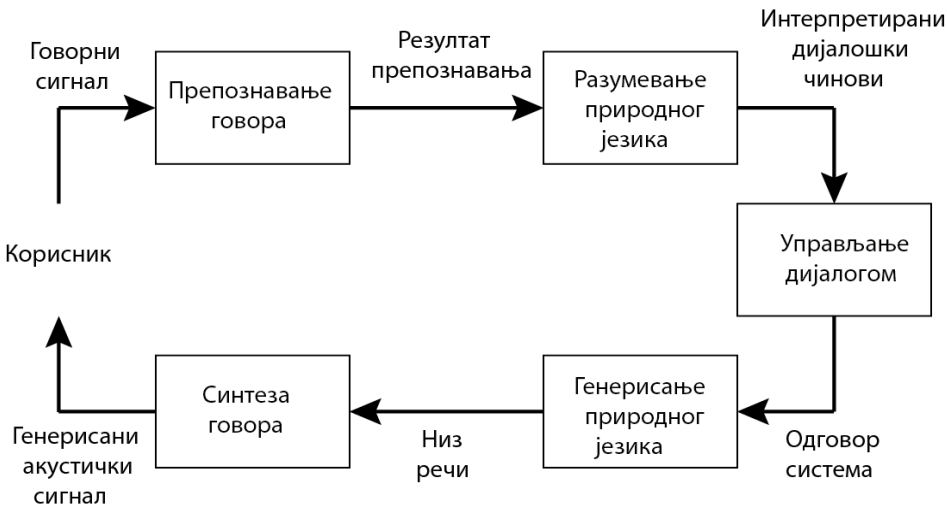
која представља лексички сличну и семантички исправну верзију почетне хипотезе. Другим речима, иако крајње речи у горњим секвенцама имају исти изговор, систем фаворизује другу секвенцу на основу знања о бициклима, које укључује семантичке ентитете *точак, седиште, гуме, кочница* (енгл. brake), итд.

Са становишта семантичке генерализације и проширеног контекста, овај приступ је близак алгоритму изложеном у тези. Разлика је у репрезентацији знања и механизму који, у случају Концептнет платформе, нема могућности за праћење тока дијалога, већ сваки дијалошки чин третира као засебни ентитет. На пример, приступ описан у [24] не може правилно да обради случај када корисник формулише сложени дијалошки чин као секвенцу непотпуних дијалошких чинова, због изостанка контекстуалних информација о историји интеракције.

2.2.2 Осврт на архитектуру савремених приступа

Основна архитектура актуелних дијалошких система приказана је на слици 2.4. Основни задатак дијалошког ситема — тј. интерпретација говорног чина и генерисање прикладног одговора — остварује се кроз пет засебних модула, при чему сваки од њих опонаша појединачни когнитивни аспект људског механизма за обраду говора:

- Модул за аутоматски препознавање говора конвертује аудио сигнал у низ речи.
- Модул за разумевање природног говора (енгл. *natural language understanding*) интерпретира препознати низ речи и креира семантичку репрезентацију изговореног.



Слика 2.4: Типична архитектура дијалошког система. [3, 4].

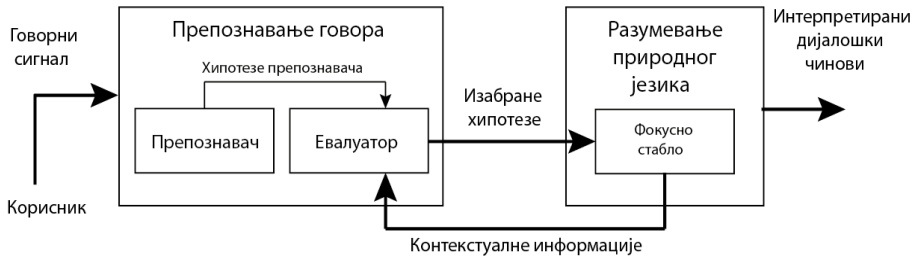
- Модул за управљање дијалогом (енгл. *dialogue manager*) моделује контекст интеракције и дијалог. На основу интерпретираног садржаја координира даље акције.
- Модул за генерисање природног језика (енгл. *natural language generation*) генерише секвенцу речи за синтезу говора.
- Модул за претварање текста у говор (енгл. *text-to-speech*) генерише изговор задате секвенце речи.

Оваква општа архитектура омогућава различите адаптације у складу са потребама различитих сценарија говорне интеракције. Најчешће варијације основне архитектуре се рефлектују кроз повећан број модула у циљу имплементације додатних функционалности. У скорије време, заступљена је надоградња овакве архитектуре модулима за препознавање емоција, мултимодалну интеакцију итд. Као резултат, добијају се системи оптимизовани за одређени сценарио, али тешко употребљиви ван њега.

Такође, важно је нагласити да модуларни концепт дијалошких система условљава да развој препознавача говора нема додирних тачака са архитектуром система у ком се препознавач користи. Јасне су предности оваквог приступа, али он истовремено сужава могућности система за препознавање говора да коригују грешке и повећају робустност препознавања. Док се у људској перцепцији говора при интерпретацији реченог интензивно користе разни когнитивни механизми који процес обогаћују информацијама о теми, контексту, намерама итд., аутоматски препознавачи говора функционишу издвојено од остатка система. Везе са осталим модулима своде се на прихватање аудио-сигнала на улазу и генерисање излаза у виду низа препознатих речи. Услед тога, унапређење перформанси овог модула узима у обзир једино скуп карактеристика везаних за сам изговор: величина речника, карактеристике канала, тип говорника итд. Оцена перформанси система је такође независна од примењене архитектуре. Своди се на тестирање над одређеним корпусом и рачунање квантитативних показатеља попут грешке на нивоу речи.

2.3 Предложени приступ и допринос тезе

Доприноси тезе обухватају нови методолошки приступ препознавању говора и адекватно прилагођену архитектуру система. У методолошком смислу, предложен је модел који, у циљу превазилажења ограничења статистичког препознавача говора, одступа од стандардних алгоритама и уводи нови, симболички приступ заснован на фокусном стаблу. Варијације основне намене овог модела (тј. управљање дијалогом у интеракцији између човека и машине) описане су у [25, 26, 27]. У оквиру ове тезе, модел фокусног стабла представља основу алгоритма



Слика 2.5: Предложена архитектура дијалошког система. Уводи се повратна спрега између модула за разумевање природног говора и препознавача.

који на основу контекста, процене информационог садржаја и усклађености са доменом интеракције оцењује валидност хипотеза препознавача.

Поред тога, у тези је уведена модификација стандардне архитектуре, тако да модули за препознавање говора и разумевање језика више нису независни. Ово је последица предложеног методолошког приступа, по којем оба модула деле јединствену репрезентацију домена интеракције и дијалошког контекста (тј. фокусно стабло). Ова модификација, реализована у облику повратне спреге између модула за препознавање говора и разумевање језика, приказана је на слици 2.5. Поред тога, у модификованој архитектури препознавач говора проширује своју функционалност, која сада укључује двостепену обраду говорног сигнала у оквиру следећих компоненти:

- стандардни препознавач говора заснован на статистичком приступу који на основу говорног сигнала генерише скуп хипотеза,
- контекстно зависни евалуатор хипотеза.

У оквиру тезе размотрен је развој оба модула, а примена је демонстрирана кроз два прототипска система. Први случај разматра концептуалну оправданост приступа а предложени алгоритам је примењен за

говорну интеракцију са хуманоидним роботом. Други случај квантитативно демонстрира ефикасност приступа, а алгоритам је примењен за говорну интеракцију између корисника и мобилног телефона.

Глава 3

Статистички приступ препознавању говора

Резултат препознавања говора базираног на скривеним Марковљевим моделима је највероватнија секвенца речи за дати улазни сигнал (тј. низ вектора акустичких обележја). Користећи Бајесово правило одлучивања (енгл. *Bayes' rule*)

$$P(W|X) = \frac{P(W)P(X|W)}{P(X)} \tag{3.1}$$

ово се може представити на следећи начин:

$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} P(W|X) = \underset{W \in \mathcal{L}}{\operatorname{argmax}} \frac{P(W)P(X|W)}{P(X)} \tag{3.2}$$

где су:

- $\hat{W}$ — препознати низ речи,
- X — низ акустичких опсервација,
- $P(X|W)$ — условна вероватноћа (изгледност, енгл. *likelihood*) да се за низ речи $W = \{w_1, w_2, ..., w_n\}$, моделован помоћу низа СММ, јавља низ акустичких опсервација $X = \{x_1, x_2, ..., x_T\}$ — тј. акустички модел,

- $P(W)$ – априорна вероватноћа јављања датог низа речи W — тј. језички модел,
- $P(X)$ – вероватноћа јављања низа акустичких опсервација X.

Приликом декодовања, вредност $P(X)$ се занемарује, јер је иста за све могуће секвенце речи W, па самим тим не утиче на то који ће низ речи бити изабран као највероватнији. Као резултат, препознавач на основу претраге по могућим секвенцама W генерише хипотезе које за задати низ акустичких опсервација максимизују вероватноћу $P(W|X)$:

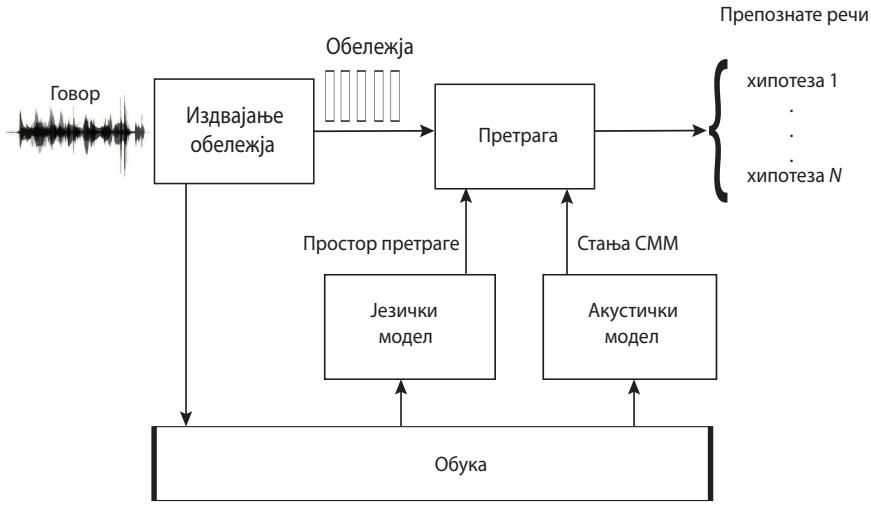
$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} P(X|W)P(W) \tag{3.3}$$

Хипотезе које максимизују вероватноћу $P(W|X)$ називамо оптималним. Из израза се види да овај приступ омогућава да одвојено моделујемо језичку ($P(W)$) и акустичку компоненту ($P(X|W)$). Слика 3.1 приказује стандардну архитектуру статистичког препознавача говора, а у наредним секцијама ће бити детаљније објашњене поједине компоненте.

У зависности од тежине задатка препознавања, перформансе система, изражене преко грешке на нивоу речи, варирају од једног до приближно четрдесет процената (грешка на нивоу речи расте са порастом броја речи, спонтаности и зашумљености говора, присутношћу других звукова итд.)

3.1 Издвајање акустичких обележја

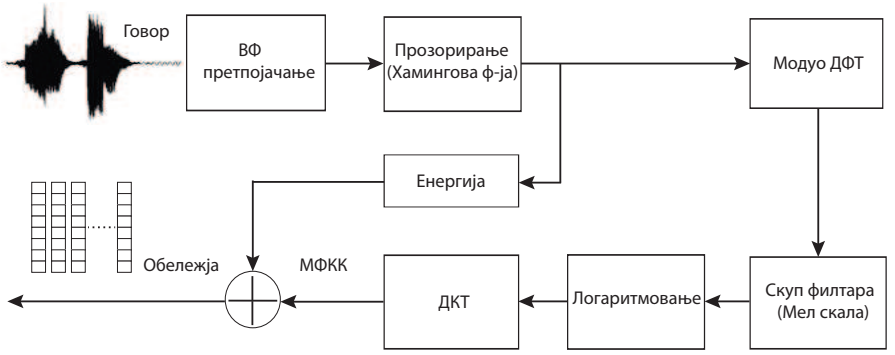
Први корак у реализацији аутоматског препознавања говора је издвајање акустичких обележја. Поред компактне репрезентације континуалног сигнала, овај процес укључује и раздвајање лингвистичког



Слика 3.1: Основна архитектура статистичког препознавача говора.

садржаја од осталих информација садржаних у аудио-сигналу. Актуелни системи користе акустичка обележја креирана у складу са перцептивним карактеристикама људског слушног апарата. Ту се пре свега мисли на мел-фреквенцијске кепстралне коефицијенте (МФКК, енгл. *mel-frequency cepstral coefficients, MFCC*) и линеарне предиктивне коефицијенте (енгл. *perceptual linear prediction*).

Детаљније ћемо приказати поступак издвајања доминантног обележја, МФКК, у модерним системима. Ови коефицијенти описују обвојницу амплитудског спектра, која је носилац информације о разликама између појединих гласова које треба препознати. Издвајање коефицијената започиње поделом говорног сигнала на мање сегменте (фрејмове). Уобичајено трајање сегмента је 20–40 милисекунди, што представља компромис између остваривања задовољавајуће фреквенцијске резолуције и претпоставке о стационарности сигнала унутар сегмента. Над сегментом се примењују одговарајуће прозорске функције, да би се смањили дисконтинуитети на ивицама. Најчешће се користи Хемингова



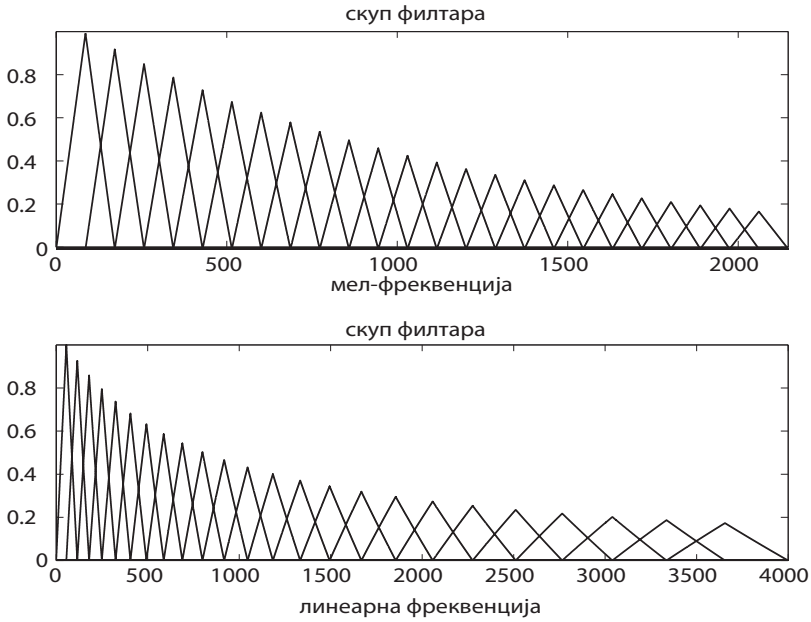
Слика 3.2: Блок шема екстракције акустичких обележја заснованих на мел-фреквенцијским коефицијентима. Значење скраћеница је следеће: ВФ — високофреквенцијски, ДФТ — дискретна Фуријеова трансформација, ДКТ — дискретна косинусна трансформација, МФКК — мел-фреквенцијски кепстрални коефицијенти

(енгл. *Hamming*) функција која смањује изобличења настала услед примене прозора у анализи аудио сигнала (тзв. цурење спектра). Сегменти се преклапају, и за сваки се издваја скуп коефицијената који образују тзв. вектор обележја. Слика 3.2 даје шематски приказ издвајања акустичких обележја заснованих на мел-фреквенцијским коефицијентима.

Први корак (претпојачање, енгл. *pre-emphasis*) представља филтрирање улазног сигнала са циљем да се оствари слабљење компоненти на ниским учестаностима, а појачање на високим. Ово се реализује филтером са импулсним одзивом коначног трајања (енгл. *Finite impulse response, FIR*). Овај поступак је оправдан, јер више спектралне компоненте у говору имају мању енергију, али и већи значај са аспекта разумљивости. Анализе показују да нпр. компоненте испод 1500Hz носе 65% разумљивости и приближно 95% енергије. Слични проценат разумљивости, са само 5% енергије, носе компоненте изнад 1500Hz [28].

Након прозорирања и израчунавања амплитудског спектра дискретне Фуријеове трансформације (ДФТ, енгл. *discrete Fourier transform, DFT*), сигнал се пропушта кроз скуп филтара, илустрован на слици 3.3.

Улога ових филтара је да процене снагу говорног сигнала у сваком од подопсега. Као што се може приметити на слици, амплитудска карактеристика је скалирана. У противном, процена снаге на излазу филтра би зависила од броја тачака спектра ДФТ које тај филтар обухвата (на линеарној фреквенцијској скали, филтрима одговарају различите ширине опсега). Број филтара зависи од ширине фреквенцијског опсега који желимо да покријемо у аудио-сигналу, и креће се у опсегу од 24 до 40.



Слика 3.3: Скуп филтара при екстракцији мел-фреквенцијских кепстралних коефицијената.

Вредности које представљају снагу говорног сигнала на појединим подопсезима се логаритмују, а затим се примењује дискретна косинусна трансформација (ДКТ, енгл. *discrete cosine transform, DCT*). Тиме се елиминише корелисаност између процењених вредности над суседним

подопсезима скупа филтара. У пракси се користи првих 12–16 мел-фреквенцијских коефицијената након ДКТ. Ови коефицијенти представљају споре промене у спектру тј. обвојницу спектра, и заједно са нормализованом енергијом чине тзв. статичка обележја. Поред њих, вектор обележја који описује сегмент аудио-сигнала садржи први и други извод статичких обележја (тзв. динамичка обележја). На овај начин, описана је трајекторија промене кепстралних коефицијената током времена.

Процес издвајања мел-фреквенцијских коефицијената се у великој мери заснива на имитирању људског чула слуха (слично важи и за линеарне предиктивне коефицијенте). Филтри имитирају начин функционисања базиларне мембране, а логаритмовање процењене снаге говорног сигнала је аналогна људској перцепцији јачине звука, која је текође логаритамска. Наиме, са повећањем фреквенције, потребна је све већа фреквенцијска дистанца између звукова да би их човек опазио као различите висине тона. На нижим фреквенцијама (нпр. испод 500Hz), човек може да перципира одвојено звукове који се разликују за само неколико херца.

Технике које се примењују за побољшање перформанси савремених система за препознавање говора се у значајној мери односе управо на издвајање акустичких обележја (тзв. *front end*). Ове технике полазе од чињенице да спектар аудио сигнала на улазу пепознавача говора представља производ траженог спектра говора и преносне функције комуникационог канала. Услед тога, у фреквенцијском домену (након логаритмовања) добијамо суперпонирање различитих спектралних компоненти на спектар говорног сигнала. Нормализација средњом вредношћу кепстрала (енгл. *cepstral mean normalization*) је стандардни поступак за потискивање шума и линерних изобличења, тј. елиминација утицаја комуникационог канала [29].

Са друге стране, приступ PACTA (енгл. *relative spectra*, RASTA) у издвајању обележја [30], се заснива на још једној особености људске перцепције говора — реаговање на релативне промене у спектру улазног сигнала. Услед тога, исправна детекција одређеног говорног сигнала је више условљена спектралним разликама између посматраног и претходног сегмента, него њиховим апсолутним вредностима. Због тога, људи не реагују на споре промене фреквенцијских карактеристика, што им омогућава да успешно комуницирају у присуству котинуираног шума. Осим тога, фреквенцијске промене у говору су дефинисане особинама вокалног тракта. За уклањање нелингвистичких компоненти говорног сигнала, приступ PACTA детектује и потискује компоненте које имају брже или спорије промене у односу на референтну.

Такође, методе прилагођавања говорнику или амбијенту се често односе на процес издвајања обележја [31]. За разлику од техника које модификују параметре модела [32], овде се врши трансформација добијених вектора обележја у складу са карактеристикама циљног модела.

3.2 Акустичко моделовање

У статистичком приступу препознавању говора, деловима звучног сигнала се придружују статистичке репрезентације. У претходној секцији разматран је процес конвертовања аудио-сигнала у секвенцу вектора обележја. Имајући у виду да ови вектори представљају тачке у високодимензионалном простору улазних обележја, успешна репрезентација аудио-сигнала захтева модел који ће интегрисати временску и просторну динамику секвенци података. Формализам скривених Марковљевих модела у виду двоструке функције вероватноће обезбеђује добру основу за такво моделовање.

Полазна претпоставка за примену скривених Марковљевих модела

у области акустичког моделовања је да се говор може представити као параметарски случајни процес. При томе, подразумевају се следеће карактеристике:

- Говорни сигнал је део-по-део стационаран.
- Суседни сегменти су међусобно независни.

Иако ова препоставка не одговара у потпуности природи говорног сигнала, ефикасне методе обуке и декодовања, обезбедиле су вишегодишњу доминацију скривених Марковљевих модела у реализовању акустичке компоненте препознавача говора.

3.2.1 Основе скривених Марковљевих модела

У општем случају, скривени Марковљеви модели су случајни процеси код којих није позната секвенца стања, већ само одговарајућа функција вероватноће. Основне теоријске поставке и детаљи о математичком формализму везаном за Марковљеве моделе могу се наћи у [33], [34] и [35]. Овде ће бити приказан статистички модел са становишта репрезентације акустичких догађаја.

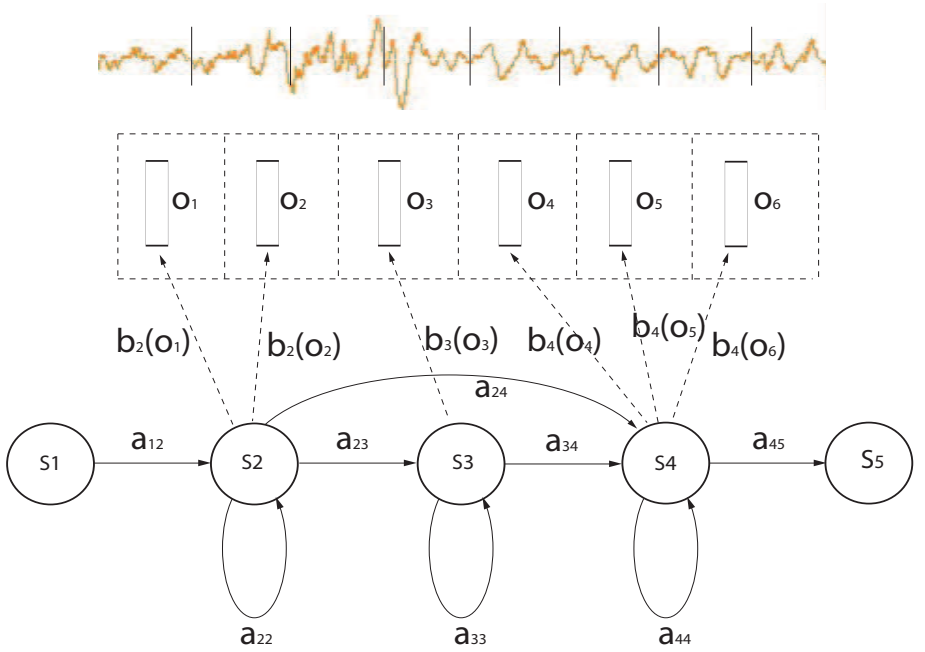
Слика 3.4 графички илуструје СММ уз пратећи скуп акустичких опсервација. У скупу стања модела постоје емитујућа (S_2, S_3 и S_4) и крајња, неемитујућа стања (S_1 и S_5), намењена повезивању основних модела у циљу формирања сложенијих модела речи.

Модел се описује уређеном четворком параметара

$$\lambda = (A, B, \pi, M) \tag{3.4}$$

при чему је значење параметара следеће:

- M — број стања у моделу,



Слика 3.4: Графички приказ скривеног Марковљевог модела са 5 стања. За свако стање је илустрован скуп опсервација које могу бити емитоване из посматраног стања, при чему се, у сваком тренутку, из активног стања емитује тачно једна опсервација.

- $A = [a_{ij}]_{M \times M}$ — матрица вероватноћа прелаза између стања (a_{ij} је вероватноћа прелаза из стања i у стање j),
- $B = [b_1 \ ... \ b_M]$ — вектор густина расподела емисионих вероватноћа ($b_j(o)$ је изгледност (енгл. *likelihood*) јављања опсервације o у стању j),
- π — вектор почетних вероватноћа.

У моделовању говора, СММ се користи као генеративни модел. То значи да низ опсервација $O = o_1 o_2 ... o_T$ посматрамо као резултат случајног процеса, генерисан секвенцом стања Марковљевог ланца. Ова секвенца стања је, за разлику од опсервација, невидљива. Веза између

стања модела и опсервација је дефинисана функцијама густине расподеле емисионих вероватноћа $b_j(o_t)$, које описују могућност да је стање j генерисало опсервацију o_t, и придружене су сваком емитујућем стању. Генерално говорећи, ове расподеле моделују спектралну варијабилност звучног догађаја, и заслужне су за дискриминативност (функционалност разликовања различитих говорних сегмената) и робустност модела (функционалност обраде варијација у суштини истих говорних сегмената, које су карактеристичне за природни говор).

Да би се постигле ове карактеристике, савремени системи за препознавање говора користе CMM са емитујућим вероватноћама у виду континуалне случајне променљиве. Најчешће се користи пондерисана сума Гаусових расподела (енгл. *continuous-density hidden Markov model*), па се емитовање опсервације o_t у стању j може представити на следећи начин:

$$(\forall\, j \in \{1, ..., M\})\;\; b_j(o_t) = \sum_{k=1}^{G_j} c_{jk}\; p(o_t \,|\, \theta_k^j) \;\wedge\; \sum_{k=1}^{G_j} c_{jk} = 1 \tag{3.5}$$

У горњем изразу, G_j је укупни број Гаусових расподела придружених стању j, θ_k^j је k-та компонента смеше, а c_{jk} је тежински коефицијент компоненте k за стање j. Задовољавањем услова:

$$\begin{gathered}(\forall\, j \in \{1, ..., M\})\; \sum_{k=1}^{G_j} c_{jk} = 1, \\ (\forall\, j \in \{1, ..., M\})(\forall\, k \in \{1, ..., G_j\})\; c_{jk} \geq 0\end{gathered} \tag{3.6}$$

постиже се да $b_j(o)$ задовољава услове за функцију густине расподеле вероватноће, тј. да важи:

$$(\forall\, j \in \{1, ..., M\})\; \int_{-\infty}^{\infty} b_j(x)dx = 1 \tag{3.7}$$

Израз 3.5 представља смешу Гаусових расподела (енгл. *Gaussian mixture model – GMM*). Уколико у овом изразу на место условне вероватноће $p(o_t \mid \theta_k^j)$ уврстимо израз за Гаусову расподелу са средњом вредношћу μ и коваријансном матрицом Σ, тј. $\mathcal{N}(\mu, \Sigma)$, коначни израз за густину расподеле вероватноће емитовања опсервације o_t у стању j је[1]:

$$b_j(o_t) = \sum_{k=1}^{G_j} \frac{c_{jk}}{\sqrt{(2\pi)^D |\Sigma_{jk}|}} e^{-\frac{1}{2}(o_t - \mu_{jk})^T \Sigma_{jk}^{-1} (o_t - \mu_{jk})} \tag{3.8}$$

У посматраном моделу, транзиционе вероватноће a_{ij} моделују временске варијације у говору. Дефинисане су као дискретне случајне промењиве, које одређују вероватноћу преласка из актуелног стања модела, у тренутку t, у неко од могућих стања, у тренутку $t+1$. Ове вероватноће задовољавају стандардне стохастичке критеријуме:

$$\begin{gathered} (\forall\ i \in \{1, ..., M\}) \sum_{j=1}^{M} a_{ij} = 1, \\ (\forall\ i, j \in \{1, ..., M\})\ a_{ij} \geq 0 \end{gathered} \tag{3.9}$$

Током процеса обуке, одређују се параметри модела (в. 3.5) и вредности транзиционих вероватноћа. На основу тога, акустички модели засновани на Марковљевим моделима могу да процене вероватноћу генерисања опсервационе секвенце O за дати модел λ. Процена ове вероватноће се, у општем случају, своди на сумирање вероватноћа јављања секвенце O по свим могућим секвенцама стања, пошто практично свака од њих може генерисати дату секвенцу опсервација. За дати опсервациони низ O и познату секвенцу стања S модела λ, вероватноћа генерисања опсервационе секвенце се може представити следећим изразом

[1] Коваријансна матрица се традиционално обележава знаком Σ, и та конвенција је усвојена и у овој тези. Читалац може да стекне утисак да се симбол Σ у овом изразу користи двосмислено: његово прво појављивљање се односи на поступак сумирања, а у остала два појављивљања представља коваријансну матрицу. Међутим, начини навођења опсега суме, односно индекса и степена коваријансне матрице, једнозначно одређују значење ове ознаке у формули.

[36]:

$$P(O, S \mid \lambda) = \pi_1 a_{12} b_2(o_1) a_{22} b_2(o_2) a_{23} b_3(o_3) ... \tag{3.10}$$

Пошто је секвенца стања непозната, вредност вероватноће генерисања секвенце O за модела λ се добија сумирањем вероватноћа генерисања опсервације O по свим могућим секвенцама стања, тј.:

$$P(O| \lambda) = \sum_{S=s_1 s_2 ... s_t} \pi_{s_1} b_{s_1}(o_1) \prod_{t=2}^{T} a_{s_{t-1} s_t} b_{s_t}(o_t) \tag{3.11}$$

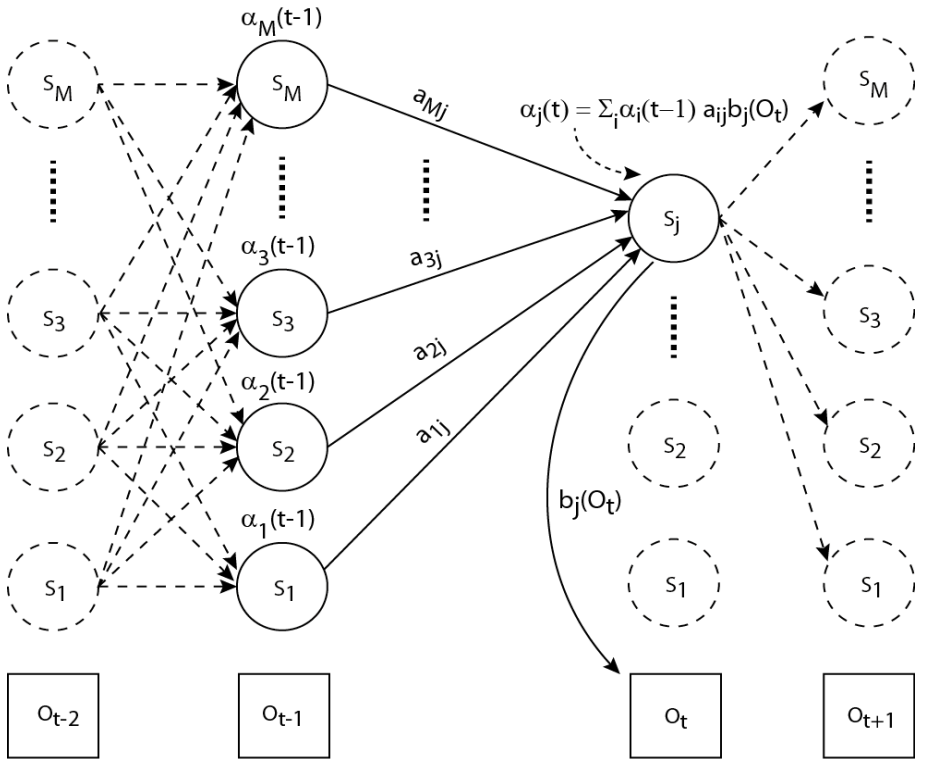
Комплексност овог израчунавања је $O(M^T T)$ (где су: M — број скривених стања, T — број опсервација), и оптимизује се процедуром тзв. израчунавања унапред (енгл. *forward procedure*). Смањење комплексности израчунавања се заснива на парцијалним израчунавањима вредности промењиве $\alpha_j(t)$, која представља вероватноћу да су регистровани опсервациони симболи $o_1 o_2 ... o_t$ и да ће се модел наћи у стању s_j у тренутку t:

$$\alpha_j(t) = P(o_1 o_2 ... o_t, S_t = j) \tag{3.12}$$

Илустрације ради, на слици 3.5 је приказан граф мрежасте структуре (енгл. *trellis*), чијим вертикалним чворовима су представљена стања СММ за различите временске тренутке. За сваки чвор j израчунава се вредност $\alpha_j(t)$ сумирањем вероватноћа по свим могућим путањама које воде до посматраног чвора у тренутку t:

$$\alpha_j(t) = \sum_i \alpha_i(t-1) a_{ij} b_j(o_t) \tag{3.13}$$

Ове вредности се даље пропагирају за тренутак $t+1$, итд. Вероватноћа целе опсервационе секвенце $P(O|\lambda)$ се израчунава итеративно, кроз следеће кораке:



Слика 3.5: Графичка илустрација поступка израчунавања унапред (енгл. *forward procedure*).

1. Иницијализација:

$$\alpha_i(1) = \pi_i b_i(o_1), 1 \leq i < M \tag{3.14}$$

2. Индуктивни корак:

$$\alpha_j(t) = \sum_{i=1}^{M} \alpha_i(t-1) a_{ij} b_j(o_t),\ 2 \leq t \leq T,\ 1 \leq j \leq M \tag{3.15}$$

3. Завршни корак:

$$P(O|\lambda) = \sum_{i=1}^{M} \alpha_i(T) \tag{3.16}$$

У првом кораку, вредност променљиве $\alpha_1(t)$ представља вероватноћу да се у почетном тренутку СММ налази у стања i и да је при томе емитовао опсервацију o_1. Индуктивни корак подразумева да је се у тренутку t модел налази у стању j, у које се прешао из стања i, у ком се налазио у тренутку $t-1$. Сумирањем вероватноћа по M стања, добија се вероватноће стања j уз опсервације $o_1, o_2 ... o_{t-1}$. Да би се добила вредност $\alpha_j(t)$, потребно је ову суму помножити вероватноћом емитовања o_t у стању j, тј. $b_j(o_t)$.[2]

На сличан начин се дефинише и процедура израчунавања уназад (енгл. *backward procedure*), а овакав поступак редукује комплексност израчунавања вероватноће $P(O|\lambda)$ на $O(M^2T)$.

Имајући у виду да је, у оквиру система за препознавање говора, свака реч замењена одређеним низом скривених Марковљевих модела, процес препознавања концептуално представљен једначином 3.3 (тј. одређивање највероватније речи која генерише посматрани низ опсервација) се своди на[3]:

$$P(O \mid w_i) = P(O \mid \lambda_i) \tag{3.17}$$

Ово омогућава успостављање везе између посматраног скупа опсервација и речи w_i, за дати СММ. Карактеристике модела λ_i (тј. транзиционе и емисионе вероватноће) се одређују у процесу обуке. Ово подразумева да се на довољном броју инстанци процене вредности параметара, тако да се робустно покрију све варијације у изговорима сегмента који се моделују.

[2]Велики број операција множења вероватноћа може у практичним реализацијама резултовати веома малим вредностима, које превазилазе могућност рачунарске репрезентације (енгл. *underflow*). Да би се ово избегло, углавном се уместо $P(O|\lambda)$ посматра $\log P(O|\lambda)$, чиме се множење вредности вероватноћа замењује сабирањем логаритмованих вредности вероватноћа.

[3]У циљу лакше илустрације, ово разматрање се односи на препознавање изоловано изговорених речи, не континуалног говора.

На процедури израчунавања унапред се заснива и поступак одређивања највероватније секвенце стања за дати низ акустичких опсервације $o_1 o_2 ... o_T$ (тј. декодовање). Овај поступак се детаљније разматра у секцији 3.4.

3.3 Моделовање језика

Основна формула статистичког препознавања говора 3.3 садржи две компоненте: акустичку и језичку, представљене вероватноћама $P(X|W)$ и $P(W)$, респективно. У претходној секцији је разматрано израчунавање вероватноће акустичке компоненте, а овде се разматра језичка компонента. $P(W)$ представља априорну вероватноћу да се одређена секвенца речи може појавити у говору. Основни задатак језичког моделовања је процена ове вероватноће.

Наиме, примена само акустичког модула у системима за препознавање говора није могућа без додатних информација, нпр:

- скуп речи које систем препознаје (речник) — овај податак одређује димензију простора претраге, укључујући и скуп CMM.
- типични редослед речи на нивоу реченице у посматраном језику — услед спектралних и временских варијација присутних у говору, може се очекивати низак ниво усклађености између акустичких опсервација и појединачних CMM, па је потребан додатни меаханизам који ће фаворизовати поједине хипотезе на основу знања о језику.

Због познатих ограничења детерминистичког представљања синтаксних правила регуларним граматикама, карактеристике језика се у системима за препознавање говора уобичајено описују статистичким моделима — н-грамима. Ови модели процењују условну вероватноћу појаве

речи након одређеног низа речи које јој непосредно претходе. Ако посматрамо низ речи $W = w_1 w_2 ... w_n$, израз за вероватноћу секвенце ове секвенце, P(W), може се представити на следећи начин:

$$P(W) = P(w_1 w_2 ... w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_n|w_1 w_2 ... w_{n-1}) \tag{3.18}$$

тј.,

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1 w_2 ... w_{i-1}) \tag{3.19}$$

где је $P(w_1)$ вероватноћа јављања речи w_1 на првом месту у секвенци речи.

Међутим, практично није могуће одредити вероватноћу појаве произвољно дуге секвенце речи, због ограничења језичког корпуса и временске захтевности израчунавања. Због тога се најчешће врши редуковање предисторије речи на неколико речи које јој непосредно претходе. Моделовањем помоћу н-грама реда N, вероватноћа секвенце речи се своди на:

$$P(w_1 w_2 ... w_n) \approx \prod_{i=1}^{n} P(w_i|w_{i-N+1} ... w_{i-2} w_{i-1}) \tag{3.20}$$

Конкретна примена оваквог моделовања језика у системима за препознавање говора се најчешће заснива на употреби триграма ($N = 3$), чиме се историја сваке речи ограничава на две претходне речи. Практични разлог за овако редуковано посматрање историје речи лежи у немогућности квалитетне обуке модела вишег реда.

Наиме, као и код осталих статистичких приступа, обука модела се врши обрадом великих корпуса (у овом случају текстуалних). Процена вероватноћа појединачних н-грама се врши на основу критеријума максималне изгледности (енгл. *maximum likelihood estimation*). На пример, за случај биграма вероватноћа $P(w_2|w_1)$ се израчунава на основу следеће формуле

$$P(w_2|w_1) = \frac{C(w_1w_2)}{\sum_w C(w_1w)} \tag{3.21}$$

У горњем изразу, $C(w_1w_2)$ представља број појављивања биграма w_1w_2 у корпусу, а $C(w_1w)$ број свих биграма у корпусу у којима је прва реч w_1 док је друга реч произвољна[4].

Може се приметити да је сума у имениоцу горњег израза једнака укупном броју појављивања речи w_1 у корпусу, па се претходна формула своди на

$$P(w_2|w_1) = \frac{C(w_1w_2)}{C(w_1)} \tag{3.22}$$

У општем случају н-грама, формула за процену максималне изгледности гласи:

$$P(w_n|w_{n-N+1}...w_{n-1}) = \frac{C(w_{n-N+1}...w_{n-1}w_n)}{C(w_{n-N+1}...w_{n-1})} \tag{3.23}$$

Обука језичких модела подразумева одговарајућу базу текстова који треба да буду репрезентативни у односу на језик за који се модел обучава. Као што је већ речено, употреба н-грама вишег реда је праћена немогућношћу креирања довољно великог корпуса, који би обезбедио адекватну обуку. Ово је нарочито изражено код инфлективних језика, као што је српски. Велики број изведених облика речи узрокује њихову ретку заступљеност у корпусу а тиме и лошу процену појединих н-грама. Услед тога, у пракси је заступљено креирање језичких модела за одређене стилове изражавања и намену.

[4]Да би се избегле нулте вероватноће н-грама у случајевима јављања речи ван речника, на ову формулу се примењује Лапласова корекција:

$$P(w_2|w_1) = \frac{C(w_1w_2) + 1}{\sum_w C(w_1w) + V}$$

где је V број различитих речи у корпусу.

Поред тога, мора се нагласити и основни недостатак модела, а то је изостанак лингвистичких информација (модел не узима у обзир врсту речи, синтаксна правила итд). Нпр., узимајући у обзир само појављивање одређених облика речи у корпусу, језички модели опште намене ће добро моделовати следећу реченицу:

Сутра нас очекује пад температуре уз могуће падавине

Међутим, у примени оваквог модела за препознавање говора у интеракцији између човека и машине, јавиће се следећи проблеми:

- слаба заступљеност одређених облика речи у корпусима (нпр. изостанак императива у текстуалним формама),
- изостанак информација о домену интеракције (нпр. речник и језичке форме карактеристичне за област права или медицине могу бити веома нетипичне ван датих области),
- недостатак информације о контексту и намери говорника.

Због тога ће секвенци која представља команду карактеристичну за говорни интерфејс између човека и мобилног телефона:

Хајде отвори јучерашњу гласовну поруку од Милана Петровића

бити додељена неадекватна вероватноћа у језичком моделу опште намене.

Наведени проблеми потенцирају потребу за концептуалним решењима која ће унапредити статистиче приступе у моделовању говора. Увођење знања о језику у виду синтаксних генерализација је једно од решења. Као резултат, добијамо језичке моделе класа речи који могу да врше предикцију појава речи којих нема у корпусу за обуку [37]. Такозвани класни н-грами моделују вероватноћу контекста припадајуће класе уместо појединачне речи. У зависности од карактеристика језика, категоризација речи може се вршити на више начина, узимајући

у обзир значење речи, синтаксне функције, морфолошке категорије итд. Као резултат, примена класног н-грам модела у контексту претходног примера би резултовала придруживањем једнаких вероватноћа следећим реченицама

Хајде отвори јучерашњу текстулану поруку од Милана Петровића
Хајде отвори јучерашњу гласовну поруку од Милана Петровића

иако обучавајући корпус не садржи реч *гласовну* у датом контексту.

Међутим, овај приступ истовремено елиминише дискриминативну улогу језичког модела тако да реченица:

Хајде одмори јучерашњу текстулану поуку код Милана Петровића

такође постаје прихватљива хипотеза препознавача говора.

У наредним поглављима ће бити предложен приступ који превазилази разматране проблеме.

3.4 Декодовање

У статистичким системима за препознавање говора, задатак декодера је да пронађе секвенцу речи чији акустички и језички модели највише одговарају улазном скупу опсервација. Ово се своди на проналажење секвенце $W = w_1 w_2 \dots w_n$ за коју је вероватноћа $P(W|O)$ максимална, за дати низ акустичких опсервација $O = o_1 o_2 \dots o_T$ (в. израз 3.3). Као што је већ напоменуто, једна од предности статистичког приступа у моделовању говора је могућност рашчлањивања сложеног модела, применом Бајесовог правила, на једноставније подмоделе — акустички и језички. Њихова улога је објашњена у секцијама 3.2 и 3.3, а задатак

декодера је да обједини ова два математичка формализма, чиме се проблем препознавања своди на претрагу простора дефинисаног речником, акустичким моделом и језичким моделом.

У општем случају, процес декодовања при препознавању говорног чина укључује:

- Претраживање простора могућих хипотеза — при томе, треба имати у виду да на димензионалност овог проблема не утиче само број речи у речнику, већ и бројни други параметри (нпр. варијације у изговору, временско поравнање унутар секвенце речи, сложеност модела итд.).
- Оцењивање (енгл. *score*) хипотеза на основу акустичких и језичких показатеља.
- Одабирање скупа највероватнијих хипотеза.

Ипак, практични декодери нису у могућности да у потпуности реализују наведене задатке. Хипотезе укључују све могуће комбинације речи из речника, изговора и различитих временских расподела, те је немогуће разматрати све варијације у реалном времену. Са друге стране, статистички приступ у препознавању говора укључује веома ограничено знање о језику, што ограничава могућност система да редукује комплексност претраге.

3.4.1 Дефинисање простора претраге

Полазна тачка за реализацију декодера је дефинисање топологије простора претраге. Наиме, код система за препознавање говора, јединица акустичког моделовања углавном није реч, већ мање целине, најчешће фонем у одређеном контексту. За пресликавање између речи

и ових мањих целина задужен је модул за генерисање изговора. Њиме се дефинишу правила на основу којих се, здруживањем основних јединица акустичког моделовања, долази до модела речи. Табела 3.1 приказује формиране изговоре за неколико речи. Као што се може видети, поступак претвара реч у низ фонема, при чему се узимају у обзир особености појединих изговора и карактеристике самог језика. Иако су код српског језика ортографска и фонетска транскрипција речи практично идентичне, генерисање изговора уводи и додатне информације о појединим гласовима у речи. Примера ради, акценат и начин изговора речи дефинишу њен наглашени део, што се манифестује продуженим трајањем вокала и већом енергијом у односу на ненаглашени остатак речи. Овакви делови се моделују посебним моделима за наглашене вокале (означени додатним суфиксом „s“ у изговору речи). Такође, плозиви и африкати садрже два акустички веома различита дела. Услед тога, њихово моделовање је рашчлањено на оклузије (означене суфиксом „o“) и експлозије (означене суфиксом „e“). Додатни задатак генератора изговора је да обезбеди подршку за различите изговоре истих речи (нпр. „четири“ и „чет’ри“).

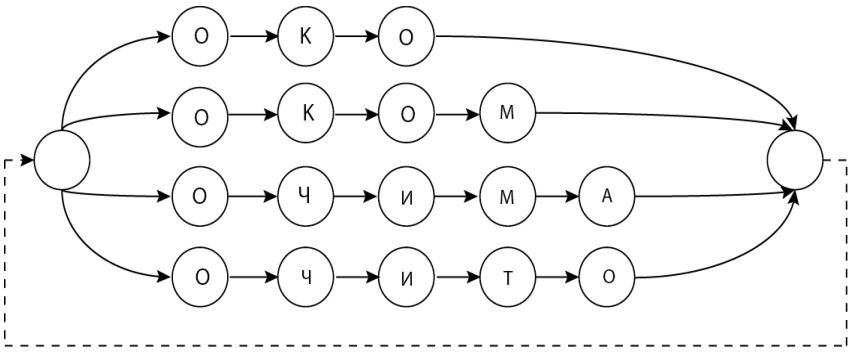
Табела 3.1: Пример фонетизације речи из речника препознавача.

Реч	Изговор
Лево	L Es V O
Левим	L Es V I M
Око	Os Ko Ke O
Оком	Os Ko Ke O M
Отвори	O To Te V Os R I
Четири	CHo CHe Es To Te I R I
Четири	CHo CHe Es To Te R I

Формирањем изговора се свака реч из речника система за препознавање говора трансформише у секвенцу нижег реда. Даљом експанзијом долазимо до секвенце стања СММ која репрезентује реч у простору претраге. Постоје две структуре података за представљање простора претраге [38]:

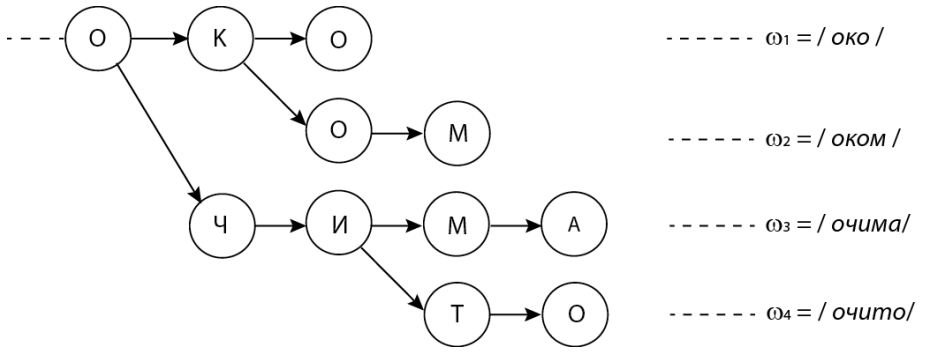
- Мрежа речи — простор претраге је формиран тако да је свака реч представљена као линеарна секвенца фонема, независно од других речи (слика 3.6). Оваква организација, у виду линеарног лексикона, користи се углавном за препознавање говора над малим речницима.

- Лексичко стабло — простор претраге се формира тако да речи деле заједничке подсеквенце фонема, тј. заједничке сегменте који се налазе на почетку речи (слика 3.7). Оваква организација обезбеђује компактну репрезентацију простора претраге, јер су, у случајевима великих речника, многе подсеквенце фонема заједничке. Ова структура се назива и фонетским префиксним стаблом.



Слика 3.6: Организација простора претраге у виду мреже речи. Свака путања од почетног до крајњег чвора представља модел једне речи.

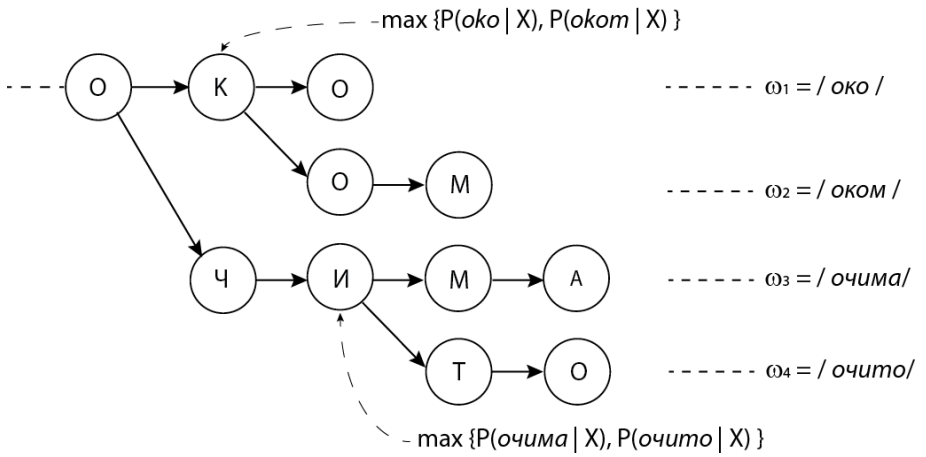
Представљање простора претраге лексичким стаблом је широко заступљено у системима за препознавање говора. Међутим, иако редукује



Слика 3.7: Организација простора претраге у виду фонетског префиксног стабла. Модели различитих речи деле заједничке секвенце фонема на почетку речи.

мрежу фонема, овај приступ је праћен сложенијим алгоритмима претраге. За разлику од мреже речи, идентификовање речи у лексичком стаблу није тривијално. Ова структура не пружа информације о посматраној речи све док се не дође до њеног последњег фонема, што онемогућава правовремену примену информација из језичког модела (н-грама). Овај проблем се превазилази применом тзв. факторисаног модела језика [39], који сваком чвору са гранањем, за дати опсервациони низ, додељује максималну вероватноћу речи која почиње префиксом одређеним посматраним чвором. Ово је илустровано на слици 3.8.

Примена н-грама изазива још један проблем у организацију претраге над лексичким стаблом. Вероватноће н-грама се не могу придружити чворовима стабла, пошто зависе од претходних $N-1$ речи у односу на текућу. У пракси, примењују се два решења за овај проблем. У првом случају, стабло претраге је статично и садржи све могуће путање између речи. Вероватноћа из језичког модела, која ће се придружити одређеној путањи се одређује у зависности од активираног чвора и речи претходно обухваћених обрадом посматране хипотезе. Друго решење подразумева експанзију стабла — поступак при ком се динамички, на



Слика 3.8: Примена факторизованог модела језика у лексичком стаблу. Свакој тачки гранања додељује се максимална вероватноћа, за дати опсервациони низ X, речи која почиње префиксом одређеним посматраним чвором.

крају сваке речи, додаје копија подстабла са речима које могу да следе. Ово значајно повећава простор претраге, због чега се у овој тези користи први приступ.

3.4.2 Реализација декодера

Проблем декодовања се своди на одређивање највероватније секвенце скривених стања посматраног модела за дати опсервациони низ. За ово се најчешће користи Витербијев алгоритам.

У секцији 3.2.1 приказан је оптимизовани поступак за прорачун вероватноће опсервационе секвенце за дати скривени Марковљев модел λ. У општем случају, пошто препознавач говора сваку реч моделује преко низа СММ, овакав поступак би се могао искористити за одређивање вероватноће опсервационе секвенце за сваку реч понаособ. Реч за коју је процењена вероватноћа опсервационе секвенце максимална би се сматрала препознатом. Међутим, овакав приступ, поред тога што је

рачунарски захтеван, није погодан за препознавање континуалног говора. Витербијев алгоритам, који представља адекватну оптимизацију поступка, резултује највероватнијом секвенцом стања СММ. Овај алгоритам користи приступ сличан процедури израчунавања унапред, с тим што се уместо сумирања по свим претходним стањима узима максимална вредност вероватноће за посматрани чвор, тј. дефинише се променљива:

$$\delta_j(t) = \max_{s_1 s_2 ... s_{t-1}} P(o_1 ... o_t, s_1 ... s_{t-1}, s_t = j) \tag{3.24}$$

која представља резултат (енгл. *Viterbi-score*) акумулисан дуж највероватније путање која се завршава у стању j и генерише секвенцу опсервација $O = o_1 o_2 ... o_t$. Одређивање највероватније путање се, слично једначини 3.13, заснива на итеративном поступку и коришћењу вредности израчунатих у претходном кораку. Додатно се чува и информација о секвенци стања која максимизује вредност израза 3.24. Ову секвенцу стања ћемо представити променљивом $\psi_j(t)$. Поступак израчунавања укључује следеће кораке:

1. Иницијализација:

$$\begin{aligned} &\delta_i(1) = \pi_i b_i(o_1), 1 \leq i < M \\ &\psi_i(1) = 0 \end{aligned} \tag{3.25}$$

2. Индуктивни корак:

$$\begin{aligned} &\delta_j(t) = \max_{1 \leq i \leq M} \delta_i(t-1) a_{ij} b_j(o_t),\ 2 \leq t \leq T,\ 1 \leq j \leq M \\ &\psi_j(t) = \operatorname*{argmax}_{1 \leq i \leq M} [\delta_i(t-1) a_{ij} b_j(o_t)] \end{aligned} \tag{3.26}$$

3. Завршни корак:

$$
\begin{aligned}
P^{\star}(\hat{S}) &= \max_{1 \leq i \leq M} \delta_i(T) \\
\hat{S} &= \operatorname*{argmax}_{1 \leq i \leq M} [\delta_i(T)]
\end{aligned}
\tag{3.27}
$$

На овај начин, Витербијев алгоритам одређује оптимално поравнање (енгл. *alignment*) између опсервационог низа акустичких обележја и секвенце стања СММ. Пошто се чувају информације о секвенци скривених стања, могуће је на крају обраде опсервационе секвенце извршити реконструкцију (енгл. *backtracking*) највероватније путање ($\hat{S}$), и идентификовати речи унутар ње.

Треба напоменути да основни облик Витербијевог алгоритма није примењив за препознавање говора над великим речницима, јер велики број стања скривених Марковљевих модела успорава процес претраге. Кључни предуслов за примену овог алгоритма за статичко стабло и велике речнике је да се врши претрага суженог обима (енгл. *beam search*). Обим претраге се редукује одбацивањем мање вероватних хипотеза (путања), чиме се број посматраних хипотеза одржава у задатим границама. Одбацивање се може вршити према различитим критеријумима, а најчешће то подразумева постављање минималне вредности за акумулисани резултат путање или дефинисање максималног броја посматраних хипотеза.

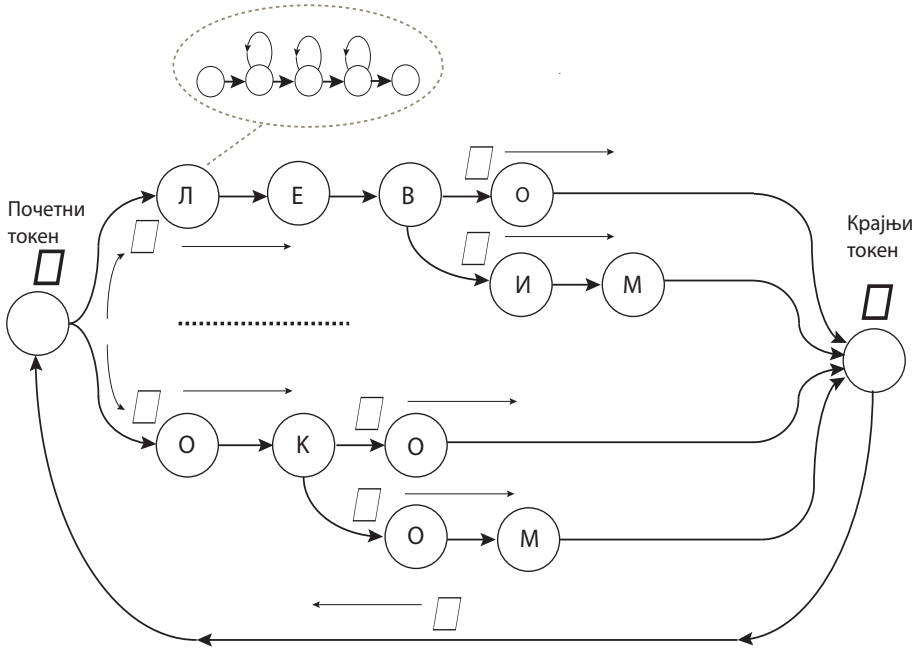
Такође, важно је рећи да директна примена овог алгоритма није погодна за укључивање различитих извора информација релевантних за декодовање (нпр. језички модел). Једини параметри који учествују у избору најбоље путање, у складу са изразом 3.26, су транзиционе вероватноће (a_{ij}) и емитујуће густине расподеле ($b_j(o_t)$).

Решење за превазилажење овог недостатка, које је коришћено и у овој тези, је претрага са прослеђивањем токена (енгл. *token passing*)

[40] [41]. Слика 3.9 илуструје овај алгоритмам. Као што се може видети, приступ задржава топологију СММ — речи се репрезентују моделима фонема, при чему сваки модел садржи низ стања СММ. На овај начин креира се транзициона мрежа декодера, а процена различитих путања се врши прослеђивањем токена — тј., структура података које поред осталог садрже информације о претходном чвору[5], акумулисаној тежини путање (која је сразмерна вероватноћи да је посматрани низ стања генерисао дати низ опсервација), и временском тренутку t. Слично приступу изворног облика Витербијевог алгоритма, акумулисана тежина представља меру слагања између опсервационе секвенце и низа стања СММ у тренутку t. У сваком кораку декодовања, копија токена из стања i се прослеђује до свих повезаних стања j, при чему се тежина путање ажурира у складу са транзиционим вероватноћама (a_{ij}) и густинама расподела емисионих вероватноћа ($b_j(o_t)$). Сужавање обима претраге се остварује пропагирањем само највероватнијих токена, док се остали одбацују.

Овде је важно напоменути да организовање простора претраге у виду лексичког стабла дозвољава примену језичког модела при пропагацији токена. У тренутку када се токен прослеђује из крајњег стања једне речи у почетно стање следеће, акумулисана тежина се ажурира у складу са вероватноћом језичког модела. У тачкама гранања примењује се факторизована вероватноћа језичког модела, а подршка за контекст шири од биграма (тј., н-грама другог реда) реализује се тако што сваки чвор чува више токена са различитим историјама [42].

[5]Важно је напоменути да при препознавању континуалног говора не посматрамо највероватнију секвенцу стања, већ највероватнију секвенцу речи. У складу са тим, токен садржи и референцу на претходну реч.



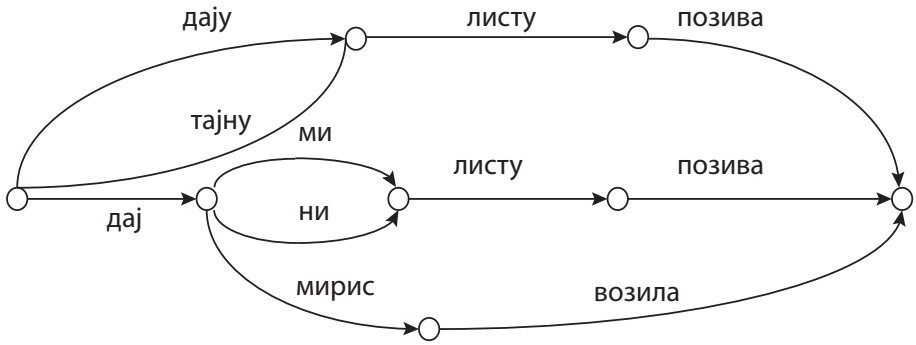
Слика 3.9: Графички приказ алгоритма за прослеђивање токена над лексичким стаблом.

3.4.3 Генерисање резултата препознавања

Као што је разматрано у претходним секцијама, задатак декодера је проналажење секвенце речи (тј., хипотезе препознавања) која по обједињеним критеријумима максимизује вероватноћу у изразу 3.24. У овиру препознате секвенце речи, свакој речи се додељују изгледност и временске границе унутар изговора.

У пракси, пожељно је уместо једне хипотезе препознавања користити више хиоптеза, нарочито у ситуацијама када препознавач, услед различитих сметњи, додели малу изгледност секвенци препознатих речи. Сужени скуп хипотеза препознавања могуће је додатно проценити, узимајући у обзир различите критеријуме [43, 44]. Коришћење контекстуалних информација које нису биле доступне у првој фази препознавања је приступ практикован у овој тези.

Овакав резултат препознавања може се представити низом који садржи N најбољих хипотеза препознавања, или мрежом речи (енгл. *word lattice*). Слика 3.10 даје графички приказ мреже речи за један пример препознавања говора. Чворови представљају временску сегментацију, а гране препознате речи различитих хипотеза препознавања.



Слика 3.10: Резултат препознавања графички представљен мрежом речи (енгл. *word lattice*).

За даљу обраду оваквих резултата постоје различити приступи. Примера ради, примењује се вишеструки поступак декодовања, при чему се у првом кораку користе н-грами нижег реда (биграм), док други корак укључује сложеније језичке моделе (триграме или четворограме). Поред тога, постоје реализације препознавача у којима се за одређивање првобитног скупа хипотеза користе акустучки модели независни од говорника, а након тога се се врши њихова валидација моделима адаптираним за говорнике.

Глава 4

Когнитивно инспирисани приступ моделовању контекста

Један од основних истраживачких проблема у реализацији дијалошких система се односи на унапређење природности интеракције између човека и система. Претпоставка за природност интеракције је да корисник не мора да улаже свесни напор да би синтаксну форму својих дијалошких чинова прилагодио унапред задатим синтаксним правилима. Дизајн система који предвиђа да корисник прилагођава свој стил изражавања у знатној мери смањује ниво природности интеракције, и не може се очекивати да ће корисници пристати да дугорочно користе овакве рестриктивне системе. Уместо тога, кориснику треба допустити да говори спонтано, а функционалност система треба да укључује могућност обраде корисничких дијалошких чинова различитих синтаксних форми које су карактеристичне за природни дијалог.

Са друге стране, уобичајени приступи развоју модула за препознавање говора укључују бројна ограничења природности интеракције.

Код препознавача над малим речницима, скуп речи и синтаксна правила за формирање корисничких дијалошких чинова се дефинишу граматикама. У случају препознавача над великим речницима, језичка ограничења нису овако експлицитна, већ последица зависности система од обучавајућег корпуса и примењеног алгоритма за обучавање (што је размотрено у секцији 3.3).

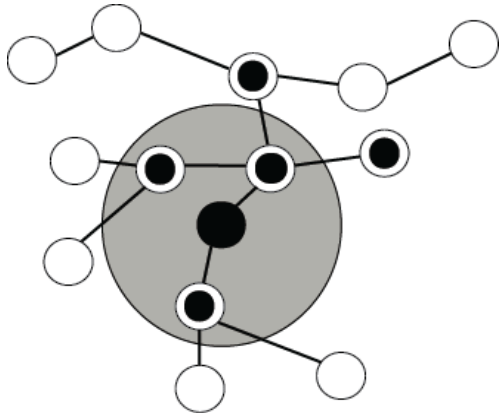
Фокусно стабло [45, 26] је симболички, когнитивно инспирисани модел фокуса пажње у интеракцији између човека и машине, који омогућава робустну интерпретацију спонтаног говора, без претходно задатих синтаксних правила. У наставку поглавља ћемо детаљније размотрити овај приступ.

4.1 Концепт модела

Модел фокусног стабла представља рачунарску концептуализацију концентричног модела радне меморије [46]. Значај радне меморије у когнитивним процесима је одавно препознат [47], и знатни број истраживања је посвећен питањима њене анатомске организације и когнитивне улоге. Почетне претпоставке о секвенцијалној организацији когнитивног апарата и о примарној, краткотрајној меморији (енгл. *short-term memory*)[48] еволуирале су временом у сложеније моделе меморије, који узимају у обзир паралелну обраду информација, селективну пажњу, итд. [49]. Може се уочити да различити когнитивни модели радне меморије деле заједничку основу:

- капацитет радне меморије је ограничен (мада различити модели претпостављају различите капацитете, нпр., до 7 ентитета који могу бити смештени у радну меморију),
- основне функционалности радне меморије укључују филтрирање

важних информација, њихову организацију и контекстно зависно приоритизовање.



Слика 4.1: Концентрични модел радне меморије. Чворови представљају семантичке ентитете присутне у меморији. У сваком тренутку, само поједини ентитети (представљени тамним чворовима) су активирани спољашњим стимулансима или унутрашњим асоцијацијама. Само један ограничени подскуп активираних чворова је доступан за тренутне когнитивне процесе (тзв. област директног приступа, чворови у сивом кругу). Притом, само један чвор из ове области носи фокус пажње (потпуно црни чвор).

У овој тези, посебну пажњу ћемо посветити концентричном моделу радне меморије, илустрованом на слици 4.1. Меморија је представљена као граф, чији чворови представљају ентитете присутне у дугорочној меморији. Семантичке везе између ових ентитета су представљене везама између чворова. Динамичка природа меморије се рефлектује кроз концептуализацију да је радна меморија функционално стање (а не одвојени анатомски ентитет) које омогућава директан приступ активираном делу дугорочне меморије. Другим речима, у сваком тренутку су

само поједини ентитети дугорочне меморије (представљени тамним чворовима) активирани спољашњим стимулансима или унутрашњим асоцијацијама. Од свих активираних чворова, само један ограничени подскуп је доступан за тренутне когнитивне процесе (ткз. област директног приступа). Притом, само један од њих се налази у фокусу пажње. Треба приметити да овај концептуални модел не специфира детаље о топологији мреже, нити о динамичком активирању чворова.

У новијим студијама [50] [51], поменути концепт радне меморије је додатно проширен. Чворови се организују по слојевима (слојеви стимуланса, контекста, кандидата, итд.) и наглашава се улога контекста као основног механизма за њихову активацију. Област директног приступа се дефинише као скуп чворова повезаних специфичним контекстом. У том смислу, и довлачење[1] појединих чворова у радну меморију започиње активирањем њиховог контекста у засебном, контекстном слоју.

Рачунарски модел фокусног стабла задржава идеју различитих нивоа приступачности појединих меморијских ентитета у зависности од тренутног фокуса пажње. Међутим, за разлику од когнитивног модела, фокусно стабло детаљно специфира:

- топологију мреже и организацију семантичких ентитета (в. секцију 4.2),
- алгоритме за активирање чворова и селектовање фокуса пажње у односу на актуелне стимулансе (в. секцију 4.3).

Са лингвистичког становишта, концепт фокусног стабла се такође заснива на теорији о структури дискурса [52], по којој комуникациони

[1]Изрази попут „довлачење ентитета из дугорочне меморије у радну меморију“ (енгл. *retrieval*) се у овој тези користе за референцирање процеса активирања ентитета у дугорочној меморији, и не подразумевају трансфер информација између засебних анатомских ентитета.

дискурс чине три повезане компоненте: лингвистичка структура, структура намере учесника у дијалогу, и фокус пажње који садржи информације о објектима, релацијама и намерама које су наглашене (експлицитно или имплицитно) у одређеном моменту. Поред тога, механизам фокуса пажње обезбеђује и акумулисање релевантних информација из претходне комуникације, чиме се елиминише потреба за чувањем њене комплетне историје. У раду [52] се такође истиче да је фокус пажње суштински важан за правилно, контексно зависно интерпретирање дијалошких чинова, и да су информације о фокусу пажње организоване хијерархијски, што је у складу са когнитивним моделом радне меморије.

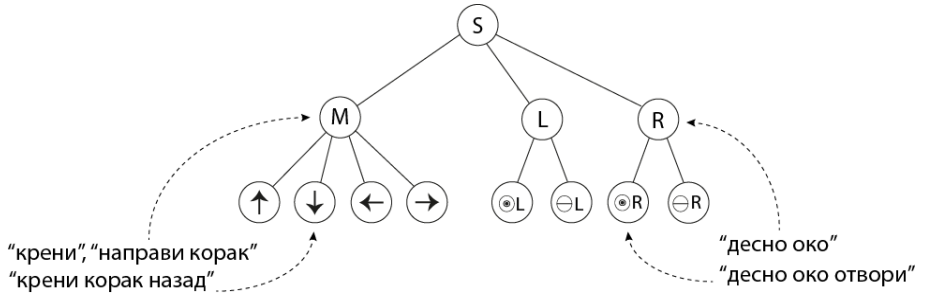
4.2 Структура фокусног стабла

Фокусно стабло представља хијерахијску структуру повезаних ентитета по узору на концентрични модел радне меморије (Слика 4.1). Чине га:

- Чворови стабла који представљају специфичне семантичке ентитете.
- Гране стабла које одређују основне промене фокуса пажње (при томе се мисли на промене фокуса пажње које су очекиване у дијалошком домену које моделује фокусно стабло).

Слика 4.2 приказује фокусно стабло које моделује изабрани домен интеракције између човека и робота. При томе, на слици је приказан само део стабла који обухвата опсег основних команди за контролу кретања робота. Табела 4.1 описује менталне представе и стимулансе везане за поједине чворове. Примера ради, чвор M је везан за кретање робота, а придружене су му кључне речи „корак“, „крени“ итд. Сваком чвору је придружена одређена ментална представа, а њихове везе

су конципиране тако да сваки чвор, изузимајући корен стабла, садржи проширену менталну представу у односу на родитељски чвор. Као што се може приметити, једино су терминалним чворовима придружене комплетне информације о одређеној радњи. Стимуланси представљају придружене кључне речи и фразе које корисник може да употреби у спонтаном говору када реферише одређени семантички ентитет представљен чвором фокусног стабла.



Слика 4.2: Приказ подстабла као дела ширег корпуса намењеног интеракцији између човека и робота. Подстабло обухвата опсег команди намењених задавању кретања и контроли очију робота.

Табела 4.1: Речник и менталне репрезентације придружене фокусном стаблу са слике 4.2.

Чвор	Ментална репрезентација	Фокусни стимуланси
S	(корен стабла)	-
M	кретање робота	„крени“, „направи“, „корак“, ...
↑	кретање робота напред	„напред“ „ка мени“, ...
↓	кретање робота назад	„назад“, „од мене“, ...
←	кретање робота лево	„лево“, „налево“, ...
→	кретање робота десно	„десно“, „надесно“, ...
L	лево око	„око“, „лево око“, „левим“, ...
$\odot_L$	отварање левог ока	„отвори“, ...
$\ominus_L$	затварање левог ока	„затвори“, ...
R	десно око	„око“, „десно око“, „десним“, ...
$\odot_R$	отварање десног ока	„отвори“, ...
$\ominus_R$	затварање десног ока	„затвори“, ...

Фокус пажње може бити придружен било ком чвору у току дијалога. При томе, свако подстабло одређено тренутним фокусом пажње као својим чвором представља област директног приступа. У конкретном случају, за случај активираног чвора „лево око“, област директног приступа чине сви чворови који садрже семантичке ентитете из команди које могу бити задате у контексту левог ока („отвори“ и „затвори“).

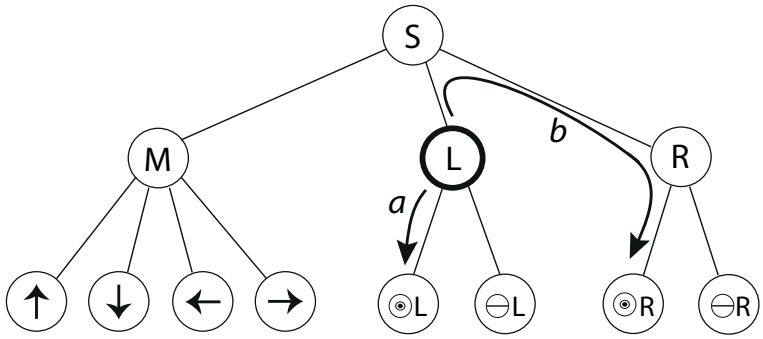
4.3 Обрада дијалошких чинова

Обрада корисничког дијалошког чина укључује екстракцију стимуланса садржаних у дијалошком чину и њихово контекстно зависно пресликавање на фокусно стабло, у зависности од тренутног фокуса пажње. Алгоритам за транзицију фокуса пажње је детаљно разматран у [45, 26], а овде су наглашени само најважнији аспекти обраде корисничких говорних команди.

Обрада команде C укључује узимање у обзир контекста интеракције, што се осликава и на постављање фокуса на нови чвор у стаблу. Претпоставимо да се команда састоји од скупа фокусних стимуланса $f_1, ..., f_n$. При томе, f_1 је стимуланс највеће општости, а f_n најмање. Поступак ажурирања фокуса пажње се може представити кроз следећа два случаја:

- Ако за сваки фокусни стимуланс из команде C постоји чвор из области директног приступа (тј., подстабла дефинисаног чвором c_n који се тренутно налази у фокусу пажње) који га представља, пресликавање фокусних стимуланса се врши у оквиру датог подстабла почевши од стимуланса највеће општости.

- У супротном, тренутни фокус пажње се помера на најближег претка који задовољава први услов. Након тога, пресликавање се изводи као у првом случају.



Слика 4.3: Прикази пресликавања фокусних стимуланса — у области директног приступа (путања ***а***) и ван ње (путања ***б***).

У циљу илустрације, посматрајмо стабло дато на слици 4.2, и претпоставимо да је процес пресликавања команде у току, а да су претходни стимуланси поставили тренутни фокус пажње на чвор са припадајућим значењем „лево око“. У тренутном контексту, нови фокусни стимуланс „отвори“ — који се, посматрано ван контекста, може односити на отварање било ког ока — ће се исправно интерпретирати као да се односи на лево око (транзиција фокуса пажње је представљена путањом ***а*** на слици 4.3).

Међутим, за случај истог тренутног фокуса пажње, команда „десно око отвори“ не може бити пресликана у тренутној области директног приступа. Ово значи да се фокус пажње прво помера у чвор вишег нивоа (у овом случају то је корен приказаног подстабла), а након тога се фокусни стимуланси пресликавају на одговарајуће чворове (путања ***b*** на слици 4.3).

У општем случају, процес обраде говорних садржаја и њиховог пресликавања на фокусно стабло разликује четири врсте дијалошких чинова:

- **Комплетни дијалошки чин** — након пресликавања оваквог дијалошког чина на фокусно стабло, нови фокус пажње се поставља

у терминални чвор (лист стабла). Примера ради, комплетни дијалошки чин за фокусно стабло на слици 4.2 је *Затвори лево око*[2].

- **Некомплетни дијалошки чин** — у овом случају постоји више могућих интерпретација датог дијалошког чина, при чему све интерпретације садрже један заједнички подскуп семантичких ентитета. Тај заједнички део се пресликава на фокусно стабло, а тренутни фокус пажње се ажурира, при чему се нови фокус пажње поставља на нетерминални чвор. Пример некомплетног дијалошког чина у посматраном домену је *Лево око*.

- **Вишемислени дијалошки чин** — слично претходном случају, и овде постоји више могућих интерпретација датог дијалошког чина, али интерпретације не садрже заједнички подскуп семантичких ентитета. Последично, фокус пажње се не ажурира. За стабло са слике 4.2, пример вишемисленог дијалошког чина би био *Затвори око*.

- **Семантички некоректни дијалошки чин** — овакви чинови не садрже фокусне стимулансе, или садрже скуп фокусних стимуланса који се не могу представити једном путањом од корена до листа стабла. У овом случају, фокус пажње се не ажурира. Пример некоректног чина у посматраном домену је *Направи лево око*.

Наредна поглавља ће размотрити аспекте модела фокусног стабла који омогућавају контекстно зависну процену комплексности обраде дијалошких чинова. Услед тога, битно је истаћи основне релације унутар стабла. Сваки чвор стабла репрезентује семантички ентитет у дуготрајној меморији. Такође, свака путања, почевши од корена стабла, је пример једне менталне представе. При томе, менталне представе

[2]Сви наведени примери подразумевају да је тренутни фокус пажње постављен на корен стабла у тренутку интерпретације.

придружене потомцима неког чвора у стаблу, интегришу и проширују менталну представу везану за њега. Фокус пажње, постављен на неки чвор n, сигнализира да је активирана његова ментална представа. То значи да су семантички ентитети придружени чвору n и његовим родитељским чворовима довучени у радну меморију. Последично, уколико је претходна историја интеракције активирала унутрашњи чвор, стабло антиципира (али не ограничава) наставак комуникације која ће активирати семантичке ентитете из области директног приступа.

Глава 5

Контекстно зависно оцењивање дијалошких чинова

Комуникација природним говором, као најважнијом манифестацијом језика, заснива се на сложеном процесу који додељује значење речима и реченицама. Услед тога, свака комуникација је везана за когнитивни напор који слушалац улаже у обраду говорних стимуланса. Другим речима, разумевање природног говора активира различите когнитивне процесе, потребне за временску и просторну детекцију сигнала, али пре свега за интерпретацију реченог у датом контексту. Са становишта когнитивне психологије [53], интерпретација говора садржи две фазе, при чему је задатак прве фазе да декодује лингвистички садржај у циљу припреме информација за наредну фазу — фазу закључивања. У другој фази, информације се интерпретирају у зависности од контекста и креирају се хипотезе о информационим намерама говорника[1].

У складу са тим, ниво когнитивног напора може послужити као индикатор тока разумевања природног говора. Значајни допринос у овом

[1]Овде је битно нагласити да декодовани лингвистички садржај не мора бити једини стимуланс за одабир хипотеза — бројни стимуланси потичу од неговорних чинова, пресупозиција итд.

смеру остварен је истраживањем електроенцефалографских осцилаторних сигнала (ЕЕГ, енгл. *electroencephalography, EEG*). Ови електрични сигнали, специфичних фреквенцијских опсега, проузроковани су активностима неурона у мозгу. Детектују се постављањем електрода на главу које региструју кумулативне флуктуације напона услед разлика потенцијала на неуронским мембранама. У односу на остале технике праћења можданих активности, ЕЕГ сигнале карактерише висока временска резолуција што омогућава брзо регистровање промена у електричној активности мозга.

У претходном поглављу, приказан је модел фокусног стабла као когнитивно инспирисани приступ за моделовање дијалога између човека и машине. У оквиру ове примене, истичу се могућности модела за:

- флексибилну обраду спонтано изговорених команди корисника, без потребе за праћењем синтаксних правила (укључујући некомплетне, вишесмислене и семантички некоректне команде),
- праћење историје интеракције,
- дизајн адаптивних дијалошких стратегија.

Поред ових карактеристика, модел нуди и могућност оцењивања комплексности дијалошких чинова [26][27]. Приступ је инспирисан електроенцефалографским истраживањима разумевања природног језика у говорној комуникацији [54] и [5].

У овом поглављу, детаљно је представљен алгоритам који обезбеђује функционалност хибридног система за аутоматско препознавање говора. Алгоритам је инспирисан истраживањима когнитивног механизама људског разумевања говора, и намењен је за обраду дијалошких чинова применом фокусног стабла. Приказана је контекстно зависна

процена хипотеза генерисаних од стране статистичког препознавача говора, а адекватност алгоритма је демонстрирана на конкретном фокусном стаблу у оквиру прототипског система.

5.1 Разумевање језика — увид у неурофизиолошка истраживања

Током говорне интеракције, након перцепције самих речи, код слушаоца започиње процес стварања, реорганизације или ажурирања менталне представе о теми комуникације. Ово, поред лингвистичких информација добијених од говорника, захтева и активирање општег знања о свету, како би процес закључивања довео до прихватљивих и примерених интерпретација говорног чина. При томе, битно је нагласити да је брзина којом се све то одвија омогућена априорним активирањем меморијских ентитета који су у одређеној семантичкој вези са ентитетима директно активираним говорним чином (семантичка примовање, енгл. *semantic priming*). Оваква предиктивна стратегија омогућава ефикаснију контекстно зависну обраду говора. Когнитивна обрада говорних стимуланса директно зависи од синтаксне и семантичке организације исказа, усклађености исказа са темом и других језичких феномена.

Током претходних неколико деценија, бројне студије у области неуронаука усмеравале су се на област разумевања језика, фокусирајући се на везе између обраде синтаксе, семантике и когнитивних процеса. Кроз примену техника ЕЕГ, истраживања се усмеравају на праћење електричних активности у мозгу које су последица различитих лингвистичких стимуланса. Основна претпоставка је да се нивои активирања различитих когнитивних процеса у току обраде говора могу детектовати праћењем одговарајућих ЕЕГ сигнала као показатеља ових биолошких активности. У том смислу, са становишта проучавања основног

механизма разумевања, веома је значајна посебна група ЕЕГ сигнала — тзв. евоцирани потенцијали (ЕП, енгл. *event related potentials, ERP*). Ово су потенцијали везани за догађај, тј. њихова појава осликава неуронску активност временски синхронизовану са одређеним догађајем (стимулансом).

За разлику од осталих електроенцефалографских сигнала који приказују осцилаторну електричну активност у оквиру различитих фреквенцијских опсега (дефинише се пет опсега за тзв. делта, тета, алфа, бета и гама таласе), ЕП-сигнали се јављају као последица сензорне, моторичке или когнитивне стимулације. При томе, поред амплитуде која описује интензитет неуронске активности, карактеристика ЕП-сигнала је кашњење у односу на појаву стимуланса, које се креће у оквирима неколико стотина милисекунди. У зависности од тога да ли је појава ЕП-сигнала обележена позитивном или негативном амплитудом, њихове ознаке добијају префикс П или Н. Поред тога, кашњење је такође укључено у ознаку. Слика 5.1 приказује типичне образце појава једног представника ЕП-сигнала. У питању је сигнал Н400, чија појава је везана за процес разумевања језика. На слици су приказане амплитудске варијације изазване различитим језичким стимулансима, што ће бити размотрено у наставку текста.

Традиционална схватања ових сигнала су била да је Н400 индикатор можданих активности везаних за семантичку интеграцију реченог, док је П600 индикатор комплексности синтаксне обраде [55] [56]. Оваква тумачења, међутим, нису могла да понуде задовољавање објашњење образаца појава сигнала ЕП код тзв. ефекта семантичке илузије (*semantic illusion*). У [56], субјекти су изложени стимулансима који садрже два дела:

- контексни део који се саопштава субјекту:

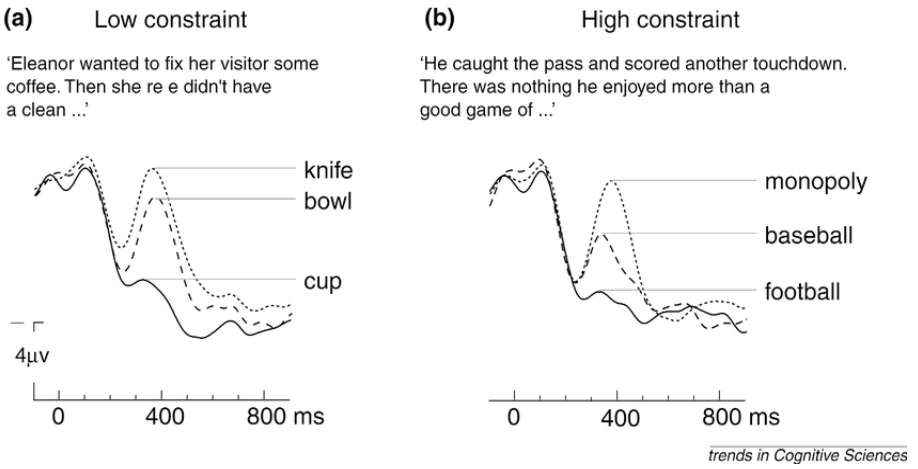
The javelin has the athletes ...

- и реч-стимуланс, у вези са чијим интерпретирањем се посматрају генерисани сигнали Н400 и П600.

Појава речи-стимуланса у виду речи *thrown* након контексног дела није изазвала појаву Н400 сигнала, док је П600 био регистрован. Пошто је целокупни стимуланс синтаксно прихватљив, али семантички неисправан, очигледна неслагања регистрованих сигнала ЕП са дефиницијама сигнала тумачена је као последица семантичке илузије — када форма реченице одлаже потпуну интерпретацију све до самог краја, јер је до тада слушалац у илузији о њеној семантичкој исправности. Закаснела појава П600 је, у складу са тим, објашњена као последица напора слушаоца да накнадном синтаксном обрадом дође до значења реченице. Слично томе, очигледна неусаглашеност поменутих тумачења са резултатима других експеримената је у појединим истраживањима превазилажена претпоставком о подели процеса обраде говора у два паралелна тока (енгл. *stream*)[57], у којима се одвојено врше синтаксна и семантичка обрада.

На основу описаног и накнадних проблема везаних за полазна тумачења сигнала ЕП, генерише се ново схватање, које на јединствени начин и без потребе за допунама објашњава резултате бројних експеримената неурофизиолошких истраживања разумевања језика. Најава савремених тумачења може се наћи већ у раду [5], док је у [54] дефинисано ново схватање ових сигнала:

- Н400 је негативно одступање сигнала, које се јавља приближно 400 милисекунди након стимуланса. Амплитуда одступања рефлектује когнитивно оптерећење које се јавља код слушаоца приликом довлачења лингвистичких информација из дуготрајне меморије у радну меморију.



Слика 5.1: Амплитуде сигнала N400 у зависности од контекстуалних ограничења у језичком исказу (преузето из [5]).

- П600 је позитивно одступање сигнала које се јавља приближно 600 милисекунди након стимуланса. Амплитуда одступања рефлектује когнитивно оптерећење које се јавља код слушаоца приликом семантичке интеграције лингвистичких информација из радне меморије.

На слици 5.1 приказани су карактеристични примери појава N400 сигнала, као индикатора менталног процеса селектовања информација о речи из дуготрајне меморије. У току експеримента, учесницима су презентовани језички стимуланси са различитим контекстуалним ограничењима у почетном делу реченице. Добијени резултати сугеришу да, поред организације семантичке меморије (тј. опште знање о свету), тренутни језички контекст игра значајну улогу у процесу активирања семантичких ентитета и њиховом довлачењу у радну меморију. Услед тога, завршетак реченице појмом за који није извршена контекстуална

припрема или који не припада очекиваној семантичкој категорији изазива појаву амплитуда сигнала Н400. Може се приметити да је интензитет амплитуде сразмеран степену одступања датог стимуланса од посматраног контекста. На слици лево 5.1**a**), приказани сигнали су детектовани на крају исказа чија форма слабије ограничава скуп појмова који се могу очекивати на крају. Слика 5.1**b** приказује сигнале Н400 за случај када контекст исказа уводи строжија ограничења за његов логичан завршетак. У оба случаја, завршетак реченица појмовима који не припадају семантичкој категорији контекстно очекиваног појма изазива појаву значајне амплитуде сигнала Н400 (нпр. појмови: *нож* и *монопол*). Ово је последица неизвршене семантичке припреме — контекст реченице није обезбедио њихово раније активирање, те је потребан додатни напор за селектовање из дуготрајне меморије. Са друге стране, завршетак реченица појмовима који нису очекивани, али припадају семантичкој категорији очекиваног појма такође изазива појаву амплитуде сигнала Н400 али мањег интензитета.

Резултати већ поменутог експеримента [56] се у светлу нове функционалне интерпретације сигнала Н400 и П600 могу тумачити на следећи начин. Излагање субјекта стимулансу у виду појма *бацити* (енгл. *thrown*) непосредно након контекста *копље је спортисту...* није изазвало генерисање амплитуде сигнала Н400 јер је појам бацања у уској семантичкој вези са појмовима копље и спортиста. Међутим, покушај семантичке интеграције ове реченице изазива повећан когнитивни напор узрокован неадекватном доделом семантичких улога различитим појмовима у реченици (копље је субјекат а спортиста објекат радње). Отежана семантичка интеграција се осликава кроз појаву амплитуде сигнала П600.

5.2 Процена комплексности дијалошког чина

Интерпретације електроенцефалографских сигнала, описане у претходној секцији, су у складу са Гибсоновом теоријом (*Syntactic Prediction Locality Theory*) [58], која приступа проблему људског разумевања говора (реченице) са аспекта ресурса потребних за њену обраду. Теорија претпоставља две компоненте неопходних ресурса:

- меморијску компоненту, као меру потребних ресурса за смештање и чување синтаксних компоненти до краја обраде реченице (до коначне интерпретације),
- интеграциону компоненту, као меру когнитивних ресурса потребних за интеграцију нове речи у постојећу семантичку структуру.

У складу са свим овим, у радовима [27, 6] је предложен приступ за процену сложености дијалошког чина. Да би се реализовала рачунарска симулација когнитивног оптерећења, у оквиру модела фокусног стабла су дефинисана два параметра инспирисана евоцираним потенцијалима Н400 и П600. Вредности ових параметара израчунавају се на основу резултата пресликавања језичких стимуланса на фокусно стабло.

У циљу њиховог дефинисања, уведимо следеће ознаке:

- $\hat{D}$ — укупни број чворова у стаблу,
- $\hat{T}$ — укупни број листова у стаблу,
- n — чвор стабла,
- n_{fp} — чвор стабла на коме је тренутни фокус пажње (пре обраде команде C),
- N — скуп свих чворова стабла који могу представљати нову менталну преставу након што се команда C обради,

- $A(n)$ — скуп који садржи чвор n и све његове претке у фокусном стаблу,
- $D(n)$ — скуп који садржи чвор n и све његове наследнике у фокусном стаблу (ово је област директног приступа за чвор n).
- $T(n)$ — скуп који садржи све терминалне чворове из скупа $D(n)$ (уколико је n терминални чвор, $T(n) = \emptyset$)

Модел фокусног стабла предвиђа два параметра за оцењивање комплексности обраде дијалошког чина:

- цена довлачења (ρ, енгл. *retrieval cost*) – параметар инспирисан евоцираним потенцијалом Н400,
- цена интеграције (η, енгл. *integration cost*) – параметар инспирисан евоцираним потенцијалом П600.

У основи формирања вредности ових параметара се налази следећа идеја. Претпоставимо да се фокус пажње налази на чвору n. Тада се сматра:

- да су семантички ентитети који одговарају чворовима из скупа $A(n)$ довучени у радну меморију,
- да менталне репрезентције које одговарају чворовима из скупа $T(n)$, представљају вероватне транзиције (тј. могуће коначне интерпретације) тренутног фокуса пажње.

На основу ових претпоставки дефинишу се параметри за евалуацију језичких исказа.

(i) Цена довлачења – ρ

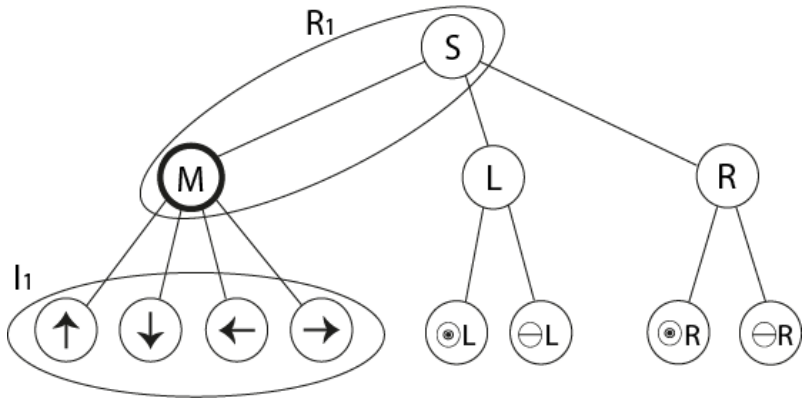
Као што је наглашено у секцији 4.3, у сваком тренутку интеракције, фокус пажње n_{fp} одговара активираној менталној представи, док скуп

$A(n_{fp})$ репрезентује семантичке ентитете довучене у радну меморију. У даљем току интеракције, сваки чвор $n \in N$ може представљати нову, активирану менталну представу. У том случају, сви чворови из скупа $\bigcup_{n \in N} A(n)$ ће бити довучени у радну меморију током обраде језичког стимуланса. Уколико из овог скупа изоставимо чворове који су већ активирани тренутним фокусом пажње (скуп $A(n_{fp})$), добијамо чворове који представљају нове семантичке ентитете довучене из дуготрајне у радну меморију. На основу тога, можемо дефинисати поступак прорачуна цене довлачења следећим изразом:

$$\rho = \begin{cases} 1, & N = \emptyset \\ \dfrac{|(\bigcup_{n \in N} A(n)) \backslash A(n_{fp})|}{\hat{D}}, & \text{у осталим случајевима} \end{cases} \tag{5.1}$$

Вредност параметра се креће у распону $\left[0, 1\right]$. Максимална вредност сигнализира семантички неисправне исказе ($N = \emptyset$), чија обрада обухвата пролажење кроз све чворове стабла и закључивање да исказ не може бити интерпретиран. У осталим случајевима добија се вредност нормализована укупним бројем чворова. Тиме се омогућава примена обрасца за произвољно фокусно стабло.

Илустрације ради, на сликама 5.2 и 5.3 приказана је промена фокуса пажње током обраде језичког исказа. У почетном тренутку (слика 5.2), фокус пажње је постављен на чвор M. Семантички ентитети који одговарају овом чвору и његовим прецима су већ довучени у радну меморију и представљени су скупом R_1. Такође, семантички ентитети представљени терминалним чворовима из скупа I_1 представљају очекивани померај фокуса пажње, тј. очекује се њихова појава у следећем исказу који би комплетирао дијалошки чин. Претпоставимо да даљи ток интеракције помери фокус пажње у чвор R (слика 5.3). Нови скуп семантичких ентитета, довучених у радну меморију, представљен



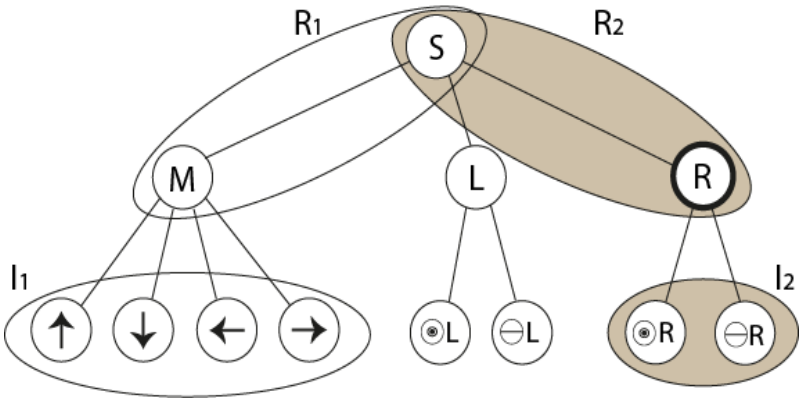
Слика 5.2: Илустрација активирања чворова фокусног стабла током обраде језичког исказа. Приказан је почетни фокус пажње, придружен чвору M.

је скупом R_2. Такође, имамо и нови скуп терминалних чворова, I_2, као очекивани померај фокуса пажње у циљу комплетирања новог исказа.

У складу са дефиницијом цене довлачења, датом формулом 5.1, вредност овог параметра за дијалошки чин који помера фокус пажње из чвора M на чвор R, израчунава се преко броја чворова у скупу R_2 који не припадају скупу R_1, подељеним укупним бројем чворова. Поступак одражава и концептуално значење самог параметра — количина нових лексичких информација из дуготрајне меморије која мора бити довучена у радну меморију током обраде дијалошког чина [59].

(ii) Цена интеграције – η

Као што је раније напоменуто, терминални чворови фокусног стабла представљају комплетне менталне представе, повезане са потпуном и једнозначном интерпретацијом дијалошког чина. Цена интеграције у овим случајевима има вредност 0 ($T(n) = \emptyset$). Насупрот томе, постављање фокуса пажње у унутрашњи чвор стабла, последица је некомплетног дијалошког чина. Одговарајућа (некомплетна) ментална представа се проширује сваким чвором из области директног приступа,



Слика 5.3: Илустрација активирања чворова фокусног стабла током обраде језичког исказа. Приказана је промена фокуса пажње, са чвора M на чвор R.

а њено комплетирање је еквивалентно померању фокуса у неки од терминалних чворова из скупа наследника (скуп $T(n)$). У складу са тим, цена интеграције за било који чвор n је сразмерна броју терминалних чворова у скупу $D(n)$, тј. величини скупа $T(n)$. У општем случају, обрада дијалошког чина дефинише скуп N (скуп чворова стабла који могу чинити нову менталну представу након обраде), а самим тим добијамо и скуп могућих комплетирања менталних представа као $\bigcup_{n \in N} T(n)$. Цена интеграције је пропорционална величини овог скупа и израчунава се према следећем обрасцу:

$$\eta = \begin{cases} 1, & |N| = 0 \\ \frac{|\bigcup_{n \in N} T(n)|}{\hat{T}}, & \text{у осталим случајевима} \end{cases} \tag{5.2}$$

Оваква дефиниција обезбеђује минималну вредност цене интеграције (0) за комплетне језичке исказе који резултују постављањем фокуса пажње у терминални чвор. Максимална вредност (1) се добија за семантички неисправне исказе, тј. оне које систем није у стању да интерпретира. Остали случајеви се односе на постављање фокуса пажње

у неки од унутрашњих чворова. Пошто се, слично приступу код претходног параметра, вредност нормализује укупним бројем терминалних чворова у стаблу, цена интеграције има у овим случајевима вредност између 0 и 1.

Уколико се поново осврнемо на слику 5.3, након постављања фокуса пажње на чвор R, скуп I_2 представља терминалне чворове у којима се очекује комплетирање менталне представе у даљем току интеракције. Дељењем броја ових чворова са укупним бројем листова, добијамо вредност цене интеграције за дати језички исказ.

Са аспекта примене ових параметара током обраде хипотеза препознавача говора, битно је нагласити следеће карактеристике:

- Оба параметра су контекстно зависна — зависе од конкретног дијалошког чина, домена интеракције (тј. структуре фокусног стабла) и историје интеракције, представљене претходном позицијом фокуса пажње.

- Вредности параметара су нормализоване тако да увек припадају опсегу $[0, 1]$, и могу се израчунавати за произвољно фокусно стабло.

- Семантички неисправни језички искази узрокују максималну вредност оба параметра.

- Вредности између 0 и 1 сигнализирају некомплетне или вишесмислене језичке исказе.

- Минимална вредност цене интеграције сигнализира комплетан језички исказ.

- Минимална вредност цене довлачења сигнализира да обрада језичког исказа не проширује постојећу менталну представу (нема

промене фокуса пажње) или је редукује (фокус пажње се помера у неки од чворова који су већ укључени у радну меморију).

5.3 Алгоритам за оцењивање хипотеза препознавања

Као што је приказано у претходној секцији, увођење контекстно зависних параметара у модел фокусно стабло омогућава симулацију H400 и П600 компоненти сигнала ЕП. Имајући у виду њихову функционалну интерпретацију, у овој секцији ће бити описан алгоритам који омогућава квалитативну анализу хипотеза препознавања говора. Поред цене довлачења и интеграције, приступ укључује још један параметар чија вредност зависи искључиво од речника енкапсулираног у фокусно стабло — лексичко подударање (λ). Његова вредност је такође у опсегу $[0-1]$ и одређује се на основу броја речи унутар хипотезе које представљају фокусни стимуланс за посматрано фокусно стабло. Нормализација се врши у односу на укупан број речи у хипотези. За фокусно стабло описано у Табели 4.1, хипотеза у виду фразе „молим те **затвори лево око**“ генерише вредност параметра 0.6 (фокусни стимуланси су задебљани).

Основна идеја приступа се може сумирати као редуковање скупа хипотеза кроз процес одбацивања оних које су мање релевантне за текући домен интеракције и тренутни фокус пажње. Редуковање се заснива на следећим критеријумима:

- **Критеријум минималне тежине семантичке интеграције**. У првом кораку алгоритам бира хипотезе које карактерише најлакша семантичка интеграција у датом моменту интеракције.

Имајући у виду карактеристике фокусног стабла, ово је задовољено за хипотезе са најмањом вредношћу цене интеграције (η). Уколико постоји само једна хипотеза са минималном вредношћу овог параметра, проглашава се за најбољу. У супротном, алгоритам прелази на следећи корак.

- **Критеријум максимизације информационог садржаја**.
  Из скупа хипотеза издвојених у претходном кораку бирају се оне са највећим информационим доприносом, тј. максималном ценом довлачења (ρ). Другим речима, из скупа хипотеза са једнаким ценама интеграције, систем бира оне које носе највише информација у датом контексту. У случају да је овај максимум везан за само једну хипотезу, она се проглашава за најбољу. У супротном, прелази се на следећи корак.

- **Критеријум максималног лексичког подударања**.
  Улазни параметар у овом кораку је подскуп хипотеза са најлакшом семантичком интеграцијом и највећим информационим доприносом. Међу њима се бирају хипотезе за које вредност лексичког подударања (λ) има највећу вредност. Као и у претходним корацима, алгоритам се завршава уколико нови подскуп садржи само једну хипотезу.

- **Критеријум максималне вероватноће појаве**.
  Овај критеријум се заснива искључиво на информацијама добијеним из система за статистичко препознавање говора. Из скупа хипотеза одабраних у претходним корацима, алгоритам бира ону којој је препознавач придружио највећу вероватноћу у процесу декодовања.

На основу датих критеријума, може се дефинисати и сам алгоритам. У циљу дефинисања појединачних корака, уведене су следеће ознаке:

- F — фокусно стабло,
- n_{FA} — фокус пажње, тренутно активирана ментална представа,
- $\{h_1, h_2, ..., h_n\}$ — скуп хипотеза препознавача сортираних по вероватноћама добијеним у процесу декодовања,
- $N(F, n_{FA}, h)$ — скуп чворова који могу представљати ажуриране менталне репрезентације након пресликавања хипотезе h на фокусно стабло F за тренутни фокус пажње n_{FA}.

Кораци алгоритма су следећи:

1. Груписање хипотеза у складу са исходом њиховог пресликавања на фокусно стабло. Обрада сваке хипотезе као резултат има ажурирану менталну репрезентацију представљену одређеним скупом чворова у фокусном стаблу. За скуп ових фокусних кандидата дефинишимо релацију еквиваленције $\backsimeq$ над скупом хипотеза $\{h_1, h_2, ..., h_n\}$:

$$h_i \backsimeq h_j \Leftrightarrow N(F, n_{FA}, h_i) = N(F, n_{FA}, h_j) \tag{5.3}$$

   Другим речима, скуп улазних хипотеза $\{h_1, h_2, ..., h_n\}$ дели се на класе еквиваленције $\Gamma = \{H_1, H_2, ..., H_k\}$, $1 \leq k \leq n$, тако да важи:

$$\begin{gathered} \bigcap_{H_i \in \Gamma} H_i = \emptyset \ \wedge \bigcup_{H_i \in \Gamma} H_i = \{h_1, ..., h_n\} \\ \wedge \ (\forall \ h_i, h_j \in \{h_1, ..., h_n\}) \\ (h_i \backsimeq h_j \Leftrightarrow (\exists \ H \in \Gamma) \ \{h_i, h_j\} \subseteq H) \end{gathered} \tag{5.4}$$

   Уколико постоји само једна класа еквиваленције $H_x \in \Gamma$, алгоритам прелази на корак 4.

2. За сваку од класа еквиваленције H, рачунају се цена интеграције

и цена довлачења. На основу дефиниције (5.3) следи да су вредности ових параметара једнаке за све хипотезе унутар класе:

$$\begin{gathered}\{h_i, h_j\} \subseteq H \Rightarrow \\ \eta(h_i) = \eta(h_j) \ \wedge \ \rho(h_i) = \rho(h_j)\end{gathered} \tag{5.5}$$

Означимо их са $\eta(H)$ и $\rho(H)$.

3. Избор класа које задовољавају први критеријум — минимална тежина семантичке интеграције:

$$I = \{H \mid H = \underset{\hat{H} \in \Gamma}{\operatorname{argmin}}\ \eta(\hat{H})\} \tag{5.6}$$

Уколико критеријум задовољава само једна класа $H_x \in I$, алгоритам прелази на корак 4. У противном, прелази се на корак 5.

4. Избор хипотеза из улазне класе H_x са максималном вредношћу лексичког подударања.

$$\Lambda = \{h \mid h = \underset{\hat{h} \in H_x}{\operatorname{argmax}}\ \lambda(\hat{h})\} \tag{5.7}$$

Након ове обраде, алгоритам прелази на корак 7.

5. Уколико постоји више класа које задовољавају први критеријум (минимална тежина интеграције), бирају се инстанце са највећим информационим доприносом (представљеним највећом ценом довлачења).

$$Q = \{H \mid H = \underset{\hat{H} \in I}{\operatorname{argmax}}\ \rho(\hat{H}\})\tag{5.8}$$

6. Из хипотеза садржаних у класама еквиваленције скупа Q бирају се оне са максималном вредношћу лексичког подударања.

$$\Lambda = \{h \mid h = \underset{\hat{h} \in \bigcup\limits_{H \in Q} H}{\operatorname{argmax}}\ \lambda(\hat{h})\} \tag{5.9}$$

7. У овом кораку се међу хипотезама h_χ из скупа Λ бира она којој је препознавач говора придружио највећу вероватноћу:

$$h_\chi = \underset{\hat{h}\in\Lambda}{\operatorname{argmax}}\ P(\hat{h}) \tag{5.10}$$

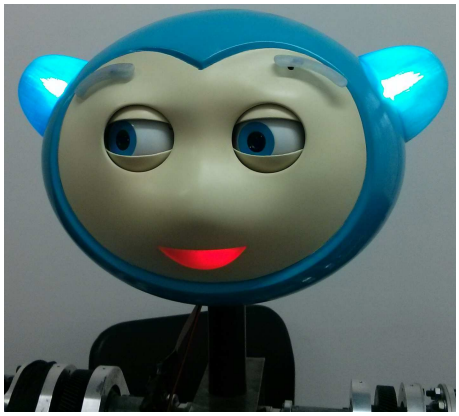
Реализацијом изложених корака алгоритма, почетни скуп хипотеза пролази кроз додатни процес оцењивања. Као резултат, систем ће у складу са контекстуално зависним критеријумима, потврдити валидност прворангиране хипотезе или изабрати нову из понуђеног скупа хипотеза аутоматског препознавача говора.

5.4 Илустрација алгоритма

У циљу демонстрације алгоритма на конкретном систему, реализован је прототип[2] конверзационог агента који рукује интеракцијом између корисника и хуманоидног робота, приказаног на слици 5.4. Због лакшег приказа, посматраћемо поједностављени домен интеракције описан у претходном поглављу. Слика 4.2 приказује фокусно стабло које моделује посматрани дијалошки домен. Речник и припадајуће менталне репрезентације за дати домен интеракције су приказани у Табели 4.1.

Хипотезе које ће бити коришћене у опису алгоритма су добијене применом статистичког препознавача говора за велике речнике, димензионираног за речник од 5000 речи. Применом оваквог препознавача желмо да демонстрирамо оправданост приступа за произвољни препознавач опште намене. Декодер препознавача је конфигурисан тако да за сваки изговор генерише по 5 највероватнијих хипотеза. Табела 5.1 приказује резултат препознавања дијалошког чина у виду сортираних

[2] Још један прототипски систем развијен у оквиру ове тезе је детаљније разматран у наредном поглављу



Слика 5.4: Глава хуманоидног робота [6]. Реализоване функционалности омогућавају задавање говорних команди за управљање положајем очију, изразима лица у циљу приказа емоција, итд.

хипотеза — h_1 је оцењена као највероватнија и представљала би резултат система без примене хибридног приступа описаног у овој тези. У овом примеру хипотеза h_3 је тачна.

Табела 5.1: Хипотезе статистичког препознавача говора за пример изговора „лево око затвори“. Хипотеза h_3 је тачна. Хипотеза h_1 је оцењена као највероватнија и представљала би резултат система без хибридног приступа описаног у овој тези. Исправно препознате речи су приказане задебљано.

Ред. број	Хипотеза
h_1	хлеб **око затвори**
h_2	лепо **око затвори**
h_3	**лево око затвори**
h_4	леву боку **затвори**
h_5	**лево** боку **затвори**

Да би се демонстрирала ефективност алгоритма, биће приказано његово извршавање за три различита момента интеракције. Сваки од њих укључује различиту историју дијалога, представљену положајем почетног фокуса пажње. Карактеристични моменти су следећи:

- Фокус пажње је постављен на чвор S, тј. корен стабла. Овај случај је карактеристичан за почетак интеракције између човека и робота.

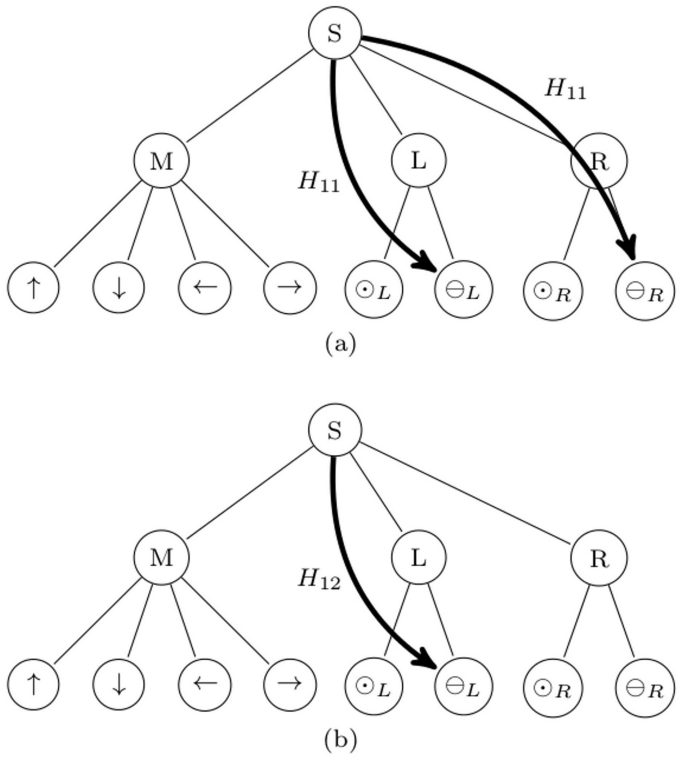
- Тренутни фокус пажње је постављен на чвор $\odot_R$. Ова позиција фокуса пажње означава да је робот успешно интерпретирао и извршио команду „отвори десно око“. Имајући у виду да анализирамо обраду хипотеза у табели 5.1, овај моменат представља случајеве у којима област директног приступа не укључује семантичке ентитете из хипотеза препознавача.

- Трећи карактеристични моменат интеракције је представљен фокусом пажње на чвору L („лево око“). Оваква ситуација репрезентује интерпретирање некомплетне команде — нпр. корисник је поменуо лево око без навођења акције коју робот треба да реализује. Овај случај посматрамо као типични случај када је фокус пажње на унутрашњем чвору стабла, а даљи ток дијалога усмерава фокус ка терминалним чворовима у области директног приступа (тј. корисник комплетира претходну команду).

У наставку ће бити илустровано извршавање алгоритма за сва три карактеристична момента. Извршавање је различито, али је у свим случајевима селектована исправна хипотеза (за разлику од резултата статистичког препознавача).

(i) Илустрација алгоритма за први карактеристични моменат интеракције — фокус пажње је у корену стабла (почетак интеракције):

1. Над скупом улазних хипотеза алгоритам креира две класе еквиваленције. Класа H_{11} садржи хипотезе(h_1, h_2 и h_4) које се пресликавају на фокусно стабло еквивалентно вишесмисленој команди



Слика 5.5: Пример обраде хипотеза када је фокус пажње постављен на корен стабла: (а) хипотезе из класе H_{11} (h_1, h_2 и h_4) се пресликавају на фокусно стабло еквивалентно вишесмисленој команди са две могуће интерпретације: „затвори лево око“ или „затвори десно око“; (б) хипотезе из класе H_{12} (h_3 и h_5) се пресликавају као комплетни дијалошки чин „затвори лево око“.

са две могуће интерпретације: *„затвори лево око“* или *„затвори десно око“*; (в. слику 5.5(а)). Класа еквиваленције H_{12} садржи хипотезе h_3 и h_5. Фокусни стимуланси из ових хипотеза активирају чворове еквивалентно комплетној команди *„затвори лево око“* (слика 5.5(б)).

2. Приступа се рачунању вредности параметара η и ρ за сваку од класа. Резултат је приказан у табели 5.2.

Табела 5.2: Оцењивање хипотеза за за први карактеристични случај — фокус пажње постављен у корен стабла, тј. чвор S.

Хипотезе	Класе еквиваленције	Вредности параметара		
		η	ρ	λ
h_1	H_{11}	0.25	0.333	0.67
h_2	H_{11}	0.25	0.333	0.67
h_4	H_{11}	0.25	0.333	0.33
h_3	H_{12}	0	0.167	1
h_5	H_{12}	0	0.167	0.67

Табела 5.3: Оцењивање хипотеза за други карактеристични случај — фокус пажње је постављен на чвор $\odot_R$.

Хипотезе	Класе еквиваленције	Вредности параметара		
		η	ρ	λ
h_1	H_{21}	0	0.083	0.67
h_2	H_{21}	0	0.083	0.67
h_4	H_{21}	0	0.083	0.33
h_3	H_{22}	0	0.167	1
h_5	H_{22}	0	0.167	0.67

3. Први критеријум за одабир хипотеза, минимална цена интеграције, (η) фаворизује класу H_{12}.

4. За хипотезе из класе H_{12}, максималну вредност лексичког подударања има h_3, тако да она постаје елемент скупа Λ.

7. Пошто је хипотеза h_3 једини елеменат скупа Λ, она постаје победничка за овај контекст интеракције.

(ii) Илустрација алгоритма за други карактеристични моменат — фокус пажње је постављен на чвор $\odot_R$:

1. На основу резулатата пресликавања хипотеза на фокусно стабло, и у овом случају се креирају две класе еквиваленције. Класа H_{21}

садржи хипотезе h_1, h_2 и h_4. Због семантичких ентитета активираних у претходној комуникацији (десно око), ове хипотезе се интерпретирају идентично комплетној команди *„затвори десно око“* (слика 5.6(а)). Друга класа еквиваленције H_{22} садржи хипотезе h_3 и h_5. Као и у претходном случају, оне одговарају комплетној команди *„затвори лево око“* (слика. 5.6(б)).

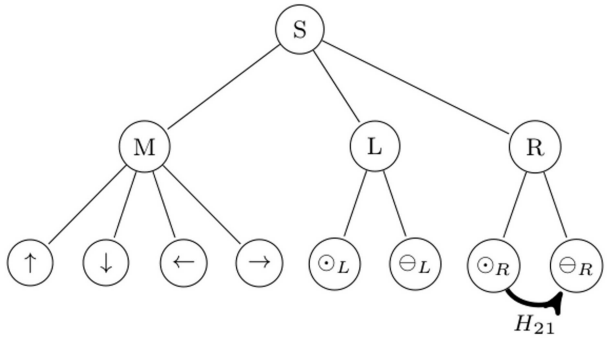
2. Рачунају се вредности цена интеграције и довлачења. Резултати су дати у табели 5.3.

3. Пошто се пресликавасе свих хипотеза завршава у терминалним чворовима, вредност параметра η је иста за све, тј. има нулту вредност. У овом случају, изостала је дискриминативна улога првог критеријума.

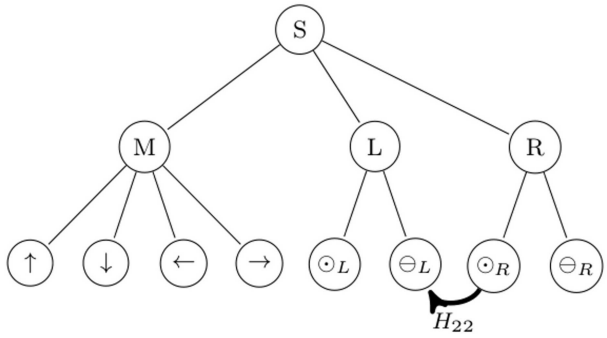
5. Алгоритам бира класе са максималном вредношћу параметра ρ (максимални информациони садржај). Класа H_{22} се прослеђује у следећи корак.

6. Вредност параметра који представља лексичко подударање уврстиће хипотезу h_3 у скуп Λ.

7. Пошто је хипотеза h_3 једини члан скупа Λ, она постаје победничка за овај контекст интеракције.



(a)



(b)

Слика 5.6: Илустрација процеса пресликавања хипотеза препознавања за фокус пажње постављен на чвор $\odot_R$: (а) Историја интеракције узрокује да се хипотезе h_1, h_2 и h_4 интерпретирају као комплетирање тренутно активне менталне репрезентације. Услед тога, померање фокуса пажње је еквивалентно пресликавању дијалошког чина „затвори десно око“. (б) Хипотезе h_3 и h_5 се пресликавају као комплетни дијалошки чин „затвори лево око“.

(iii) Илустрација алгоритма за трећи карактеристични моменат — фокус пажње је постављен на чвор L:

1. У овом примеру, фокус пажње је постављен на унутрашњи чвор стабла. Почетна обрада хипотеза, тј. генерисање класа еквиваленције на основу резултата пресликавања, резултује смештањем

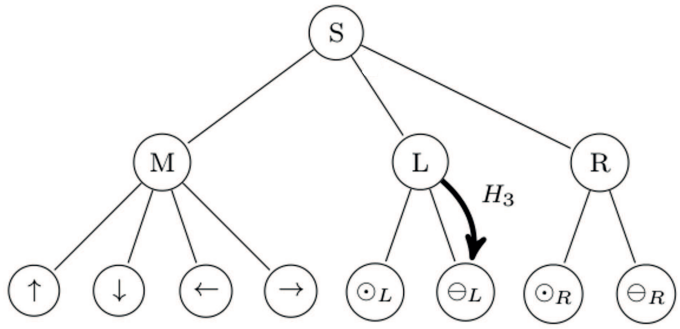
Табела 5.4: Оцењивање хипотеза за случај фокуса пажње на чвору L.

Хипотезе	Класе еквиваленције	Вредности параметара		
		η	ρ	λ
h_1	H_3	0	0.083	0.67
h_2	H_3	0	0.083	0.67
h_4	H_3	0	0.083	0.33
h_3	H_3	0	0.083	1
h_5	H_3	0	0.083	0.67

свих хипотеза у једну класу (H_3). Наиме, све хипотезе се пресликавају еквивалентно комплетној команди *„затвори лево око“*. Слика 5.7 приказује процес пресликавања.

2. Вредности параметара су приказане у табели 5.4.

4. Пошто вредности η и ρ параметара немају дискриминативну улогу у овом случају (идентични су за све хипотезе), алгоритам се ослања на проверу лексичког подударања. Хипотези h_3 је придружена максимална вредност лексичког подударања, те се она прослеђује у скуп Λ.



Слика 5.7: Пресликавање почетног скупа хипотеза препознавања у случају када је фокус пажње постављен на чвор L: све хипотезе се пресликавају еквивалентно комплетној команди „затвори лево око“.

7. Пошто је хипотеза h_3 једини члан скупа Λ, постаје победничка за

текући контекст.

5.5 Додатне напомене

Описани алгоритам за оцењивање хипотеза користи шири скуп информација. Основна идеја је инспирисина људским разумевањем говора, где се тумачење говорних исказа не завршава декодовањем језичког кода, већ накнадним процесом рационалног закључивања и извођења значења у контексту. У складу са тим, алгоритам комбинује различите оцене при избору оптималне хипотезе — поузданост резултата статистичког препознавања говора, оцену комплексности семантичке интеграције, контекстно зависну процену информационог садржаја, и лексичку усаглашеност са предвиђеним доменом интеракције.

Уколико се осврнемо на приложене примере, могао би се стећи утисак о наглашеном дискриминативном значају претпоследњег корака у процесу оцењивања хипотеза — тј. провере лексичког подударања. Међутим, битно је нагласити да се не сме преценити његова улога. У претходним корацима, хипотезе се филтрирају на основу семантичког и информационог доприноса. Ако би се изоставили ти кораци, а алгоритам свео на лексичку проверу, оцена хипотезе би се ослањала само на речник предвиђен доменом интеракције. Ово би била веома редукована концептуализација контекста. Значај провере лексичког подударања огледа се пре свега у санкционисању хипотеза са тзв. уметнутим речима (додатне речи које се појављују у резултату препознавања). Наиме, фокусно стабло омогућава флексибилну интерпретацију и индиферентно је на овакве речи. У таквим случајевима, провером лексичког подударања се потискују хипотезе које, поред исправно препознатих речи, садрже и уметнуте.

Пример добијен током тестирања прототипског система јасно илуструје горенаведено. За изговор команде „направи корак лево“, једна од хипотеза је била „направи око ка лево“. Јасно је да ова хипотеза има максималну вредност лексичког подударања за дати домен интеракције (в. слику 4.2 и табелу 4.1), али је семантички неисправна и потребно је елиминисати је у процесу оцењивања.

Глава 6

Прототипска демонстрација

У претходном поглављу је оцењивање хипотеза препознавача говора илустровано на прототипском систему за поједностављени домен интеракције. У овом поглављу ће бити демонстрирана адекватност предложеног алгоритма у оквиру другог прототипског система, за реални доменом интеракције и реалне услове употребе.

Прототипски систем разматран у овом поглављу је развијен тако да укључује двостепену обраду говорног сигнала у оквиру следећих компоненти:

- стандардни препознавач говора заснован на статистичком приступу (описаном у поглављу 3) који на основу говорног сигнала генерише скуп хипотеза,
- контекстно зависни евалуатор хипотеза (описан у поглављу 5)

Ефикасност алгоритма је илустрована кроз стандардни поступак оцењивања тачности система за аутоматско препознавање говора — израчунавањем грешака на нивоу речи и на нивоу реченице.

6.1 Карактеристике домена и прототипског система

Тестирање предложеног алгоритма извршено је на корпусу која садржи снимке дијалошких чинова за домен интеракције између корисника и мобилног телефона. Овај корпус је креиран да представи домен интеракције који подразумева брзу, поуздану и природну говорну комуникацију са мобилним телефоном, укључујући команде за позивање по имену из адресара, слање порука, додавање белешки, измене података, итд. Списак команди је дат у додатку тезе.

На основу анализе корпуса, конструисано је фокусно стабло, са 175 чворова, које моделује овај домен. Примери садржаних говорних команди укључују:

- *Отвори непрочитане поруке.*
- *Листај само одлазне позиве.*
- *Додај нови број кућног телефона.*
- *Пошаљи поруку породици Стефановић.*
- *Додај напомену.*
- *Забележи за сутра у подне састанак.*

За потребе оцењивања тачности препознавача, селектовано је 85 различитих команди које су корисници бирали на случајан начин. Статистички показатељи везани за просечан број фраза по говорнику и број речи унутар једног изговора, дати су у Табели 6.1.

База укупно садржи 772 изговорене реченице, снимљене преко фиксног телефона, и исто толико реченица снимљених преко мобилног телефона. У снимању су учествовала 64 говорника, од којих 33 мушког пола, а 31 женског.

Табела 6.1: Карактеристике селектованог корпуса.

	Средња вредност (μ)	Станд. девијација (σ)
Број команди по говорнику	24,1	7,1
Број речи по команди	3,5	1,3

За иницијално препознавање говора је коришћен статистички препознавач говора са речником од 5000 речи. Употреба овако предимензионираног речника има за циљ симулацију реалних услова — препознавач опште намене користи се за препознавање говора унутар специфичног домена. Такође, креирање корпуса који садржи снимке са мобилног и фиксног телефона има слично оправдање: симулирају се ситуације у којима акустички модели нису у потпуности у складу са амбијенталним условима присутним на месту препознавања. Ово је уобичајена појава код примене система за препознавање говора, где амбијентална бука, различити микрофони, удаљеност говорника и многи други фактори отежавају процес препознавања.

6.2 Резултати

Оцењивање тачности препознавања је вршено применом алата *sclite*. Ово је један од стандардних алата за оцену система за аутоматско препознавање говора, и део је пакета *NIST SCTK*, развијеног на Интернационалном институту за рачунарске науке (енгл. *International Computer Science Institute, UC Berkeley*) [60]. Алат врши поређење резултата препознавања са стварно изговореном секвенцом речи (задатом у референтном скупу). Након обраде, генеришу се резултати груписани по говорницима, као и збирни приказ тачности препознавања. Ове вредности се приказују као проценат у односу на број речи у референтном

скупу. Значење појединих параметара је следеће:

- #Snt — укупни број реченица у референтном скупу,
- #Wrd — укупни број речи у референтном скупу,
- Corr — тачност препознавања (удео тачно препознатих речи у односу на #Wrd).
- Sub — удео замењених речи, тј. речи које су у поступку препознавања погрешно препознате, у односу на #Wrd,
- Del — удео обрисаних речи, тј. речи које нису препознате, у односу на #Wrd,
- Ins — удео уметнутих речи, тј. речи које је препознавач додао у резултет препознавања,у односу на #Wrd,
- WER — укупна грешка на нивоу речи, једнака збиру вредности Sub, Del и Ins (енг. *word error rate — WER*),
- SER — грешка на нивоу реченице (енгл. *sentence error rate — SER*), која представља удео нетачно препознатих реченица у односу на #Snt.

Декодер статистичког препознавача говора је конфигурисан тако да за сваки изговор генерише пет најбољих хипотеза препознавања (енгл. *N-best*). Као почетно, референтно стање, изабране су прве хипотезе за сваки изговор, и за њих је извршено оцењивање тачности препознавања. Оцене резултата статистичког препознавача говора су дате у другој колони табеле 6.2. Након тога, извршено је додатно, контекстно зависно оцењивање хипотеза у складу са алгоритмом предложеним у глави 5. Оцене резултата овог препознавања су дате у трећој колони табеле 6.2.

Табела 6.2: Резултати препознавања на селектованом корпусу (в. табелу 6.1). Друга колона ове табеле садржи оцене резултата препознавања статистичког препознавача. Трећа колона садржи оцене резултата препознавања комплетног прототипског система.

Параметар	Статистичко препознавање	Хибридни приступ (контексно зависни)
#Snt	1544	1544
#Wrd	5369	5369
Corr(%)	74,7	79,3
Sub(%)	21,1	17,0
Del(%)	4,1	3,7
Ins(%)	4,2	3,4
WER(%)	29,5	24,0
SER (%)	47,0	38,6

Код ових резултата приметан је значајни напредак у перформансама препознавања — апсолутно смањење грешке на нивоу речи за $5,5\%$, и грешке на нивоу реченице за $8,4\%$. Ипак треба напоменути да постоје случајеви када скуп хипотеза који је предложио статистички препознавач не садржи тачну хипотезу. У тим случајевима, контекстно зависном евалуатору хипотеза је онемогућено да изабере тачну хипотезу. Да бисмо јасније размотрили перформансе овог прототипског система, из језичког корпуса смо издвојили 980 дијалошких чинова (од укупно 1544) за које статистички препознавач генерише скуп хипотеза које садрже тачну хипотезу[1]. Резултати препознавања над овим скупом дијалошких чинова су дати у табели 6.3.

Као што се може видети, избор адекватног скупа у правој мери истиче ефективност накнадног оцењивања хипотеза препознавача. Приметно је значајно унапређење перформанси препознавања — грешка на нивоу речи смањена је са 7,4% на 1,2%, а грешка на нивоу реченице са

[1]Овде треба приметити да је приликом тестирања прототипског система позната информација о изворним говорним чиновима, на основу које се може утврдити да ли скуп хипотеза које генерише статистички препознавач садржи тачну хипотезу.

Табела 6.3: Резултати препознавања за дијалошке чинове за које статистички препознавач нуди тачну хипотезу, не обавезно као најбоље рангирану. Друга колона ове табеле садржи оцене резултата препознавања статистичког препознавача. Трећа колона садржи оцене резултата препознавања комплетног прототипског система.

Параметар	Статистичко препознавање	Хибридни приступ (контексно зависни)
#Snt	980	980
#Wrd	3016	3016
Corr(%)	93,3	98,8
Sub(%)	5,5	0,9
Del(%)	1,1	0,2
Ins(%)	0,8	0
WER(%)	7,4	1,2
SER (%)	16,6	3,5

16,6% на 3,5%.

Поред ових показатеља, значајних са аспекта тачности препознавања говора, битно је нагласити још један допринос предложеног приступа. Он се манифестује кроз могућност система да исправно интерпретира задату команду. На основу резултата приказаних у тебели 6.3, систем је у 96,5% изабрао хипотезу која у потпуности одговара изговореној реченици (вредност SER је 3,5%). Међутим, у поглављу 4 су приказане карактеристике фокусног стабла као модела који, у оквиру модула за разумевање природног говора, омогућава флексибилну интерпретацију говорних чинова. Ово обезбеђује да успешна интерпретација реченице (исправно пресликавање на одговарајући чвор у стаблу), буде извршена и у ситуацијама када резултат препознавања не одговара у потпуности изговореној секвенци (нпр. појава уметнутих речи, изостављене поједине речи итд.). Ако са тог аспекта посматрамо процес оцењивања и избора најбоље хипотезе, хибридни поступак обезбеђује исправну интерпретацију говорних команди у 98,26% случајева за скуп

у ком нека од понуђених хипотеза садржи исправну реченицу.

Ово је још један показатељ ефективности предложеног алгоритма који узима у обзир шири контекст интеракције. Примери препознавања применом хибридног система су дати у додатку тезе.

Глава 7

Закључак

Технологија аутоматског препознавања говора је значајно напредовала последње две деценије. Услед тога, сведоци смо све шире заступљености природног говора у интеракцији између човека и машине. Ипак, тренутно стање ове технологије не задовољава бројне критеријуме за природну интеракцију. Она, без обзира на комплексност, условљену инхерентним карактеристикама језика, не сме укључивати захтеве за прилагођавање корисника систему. Са становишта машинског препознавања говора, ово захтева унапређене алгоритме, који ће омогућити ефикасно препознавање у реалним условима употребе, и без потребе за задавање специфичне форме говорне комуникације.

У овој тези је предложен нови методолошки приступ аутоматском препознавању говора у интеракцији између човека и машине. За разлику од осталих приступа овом проблему, предложени приступ укључује знатно шири опсег контекстних информација у процес избора оптималне хипотезе препознавања. На методолошком нивоу, приступ је хибридан, јер интегрише статистичке и симболичке методе, и когнитивно инспирисан, јер узима у обзир увиде у резулатате истраживања из области неурокогнитивних наука. Основни принцип је да се оцењивање

хипотеза препознавања врши на основу њихове контекстуалне усклађености, информационог садржаја и семантичке исправности. Приступ је инспирисан људским разумевањем говора — у комуникацији, људи интензивно користе контекстуалне информације за ефикасно адаптирање и интерпретирање говора у ситуацијама када не постоје идеални услови слушања.

У складу са предложеним алгоритмом, у тези је дефинисана и модификована архитектура препознавача која омогућава двостепену обраду говора. У првом степену се врши статистичко препознавање говора, које резултује скупом најбољих хипотеза, сортираних према критеријумима акустичких и језичких модела. У другом степену се врши контекстно зависно оцењивање ових хипотеза — у складу са доменом интеракције, тренутним фокусом пажње и речником система. Оваква модификација стандардне архитектуре система предвиђа да модули за препознавање говора и разумевање језика више нису независни. Модули сада деле јединствену репрезентацију домена интеракције и дијалошког контекста. Репрезентација се заснива на фокусном стаблу које моделује фокус пажње у говорној интеракцији.

Резултујући алгоритам је илустрован прототипским имплементацијама за конкретне домене интеракције. У првом случају, предложени алгоритам је примењен за препознавање говора у интеракцији са хуманоидним роботом. Други случај квантитативно демонстрира ефикасност алгоритма, и примењен је за препознавање говора у интеракцији између корисника и мобилног телефона.

Иако је демонстриран за конкретне примере, важно је напоменути да је предложени приступ независан од:

- дијалошког домена за који се примењује,
- статистичког препознавача говора који је интегрисан у конкретни

дијалошки систем.

Коначно, разматрани концепт је адаптиван у односу на стратегију контекстно зависног препознавања. Приступ у тези уводи декларативну репрезентацију доменског знања у системе за препознавање говора. За разлику од статистичких приступа, који информације извлаче из података (енгл. *data-driven*), примењена декларативна репрезентација омогућава раздвајање знања од процеса закључивања. Другим речима, начин репрезентације знања не условљава структуру алгоритама за обраду говорних стимуланса. Ово омогућава развој дијалошких стратегија које ће динамички прилагођавати критеријуме за валидацију хипотеза у односу на тренутни контекст интеракције. Ово истраживачко питање је предмет будућег рада.

Библиографија

[1] K. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[2] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," *IEICE Transactions*, vol. 88-D, no. 3, pp. 366–375, 2005.

[3] J. Daniel and H. James, "Speech and language processing: An introduction to natural language processing," *Computational Linguistics and Speech Recognition, 2nd Ed., Prentice Hall*, 2009.

[4] K. Jokinen and M. McTear, "Spoken dialogue systems," *Synthesis Lectures on Human Language Technologies*, vol. 2, no. 1, pp. 1–151, 2009.

[5] M. Kutas and K. D. Federmeier, "Electrophysiology reveals semantic memory use in language comprehension," *Trends in cognitive sciences*, vol. 4, no. 12, pp. 463–470, 2000.

[6] M. Gnjatović, "Therapist-centered design of a robot's dialogue behavior," *Cognitive Computation*, vol. 6, no. 4, pp. 775–788, 2014.

[7] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Encyclopedia of Language and Linguistics*, 2005.

[8] H. Fletcher, "The nature of speech and its interpretation," *The Bell System Technical Journal*, vol. 1, no. 1, pp. 129–144, 1922.

[9] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[10] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[11] B. T. Lowerre, *The Harpy Speech Recognition System.* PhD thesis, Pittsburgh, PA, USA, 1976. AAI7619331.

[12] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[13] B. H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of markov chains," *AT T Technical Journal*, vol. 64, pp. 1235–1249, July 1985.

[14] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, 2012.

[15] Y. Wilks, "Is there progress on talking sensibly to machines?," *Science*, vol. 318, no. 5852, pp. 927–928, 2007.

[16] C. Chelba, D. Bikel, M. Shugrina, P. Nguyen, and S. Kumar, "Large scale language modeling in automatic speech recognition," *CoRR*, vol. abs/1210.8440, 2012.

[17] R. M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 34–43, Nov 2012.

[18] G. Saon and J. T. Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *IEEE Signal Processing Magazine*, vol. 29, pp. 18–33, Nov 2012.

[19] D. Bohus and A. I. Rudnicky, *Recent Trends in Discourse and Dialogue*, ch. Sorry, I Didn't Catch That! An Investigation of Non-Understanding Errors and Recovery Strategies, pp. 123–154. Dordrecht: Springer Netherlands, 2008.

[20] M. Hacker, "Context-aware speech recognition in a robot navigation scenario," in *Proceedings of the 2nd Workshop on Context Aware Intelligent Assistance*, pp. 4–15, Citeseer, 2012.

[21] O. Lemon and I. Konstas, "User simulations for context-sensitive speech recognition in spoken dialogue systems," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, (Stroudsburg, PA, USA), pp. 505–513, Association for Computational Linguistics, 2009.

[22] V. Goel and R. A. Gopinath, "On designing context sensitive language models for spoken dialog systems," in *INTERSPEECH*, 2006.

[23] R. López-Cózar and Z. Callejas, "Asr post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information," *Speech Communication*, vol. 50, no. 8, pp. 745–766, 2008.

[24] H. Lieberman, A. Faaborg, W. Daher, and J. Espinosa, "How to wreck a nice beach you sing calm incense," in *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, (New York, NY, USA), pp. 278–280, ACM, 2005.

[25] M. Gnjatović and B. Borovac, *Toward Conscious-Like Conversational Agents*, pp. 23–45. Cham: Springer International Publishing, 2016.

[26] M. Gnjatović and V. Delić, "Cognitively-inspired representational approach to meaning in machine dialogue," *Knowledge-Based Systems*, vol. 71, pp. 25 – 33, 2014.

[27] M. Gnjatovic and V. Delic, "Electrophysiologically-inspired evaluation of dialogue act complexity," in *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pp. 167–172, Dec. 2013.

[28] H. Kurtović, *Osnovi tehničke akustike*. Naučna knjiga, 1990.

[29] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for hmm-based speaker verification," in *Third International Conference on Spoken Language Processing*, 1994.

[30] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[31] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance gaussians.," in *INTERSPEECH*, 2006.

[32] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[33] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[34] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[35] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer-Verlag New York Inc., 2006.

[36] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4.* Cambridge, UK: Cambridge University Engineering Department, 2006.

[37] S. Ostrogonac, D. Mišković, M. Sečujski, D. Pekar, and V. Delić, "A language model for highly inflective non-agglutinative languages," in *2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics*, pp. 177–181, Sept. 2012.

[38] N. Jakovljević, D. Mišković, M. Janev, and D. Pekar, "Jedno rešenje dekodera za automatsko prepoznavanje govora na velikim rečnicima," in *18. Telekomunikacioni forum TELFOR*, pp. 622–625, 2010.

[39] F. Alleva, X. Huang, and M.-Y. Hwang, "Improvements on the pronunciation prefix tree search organization," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings.,*

1996 IEEE International Conference on, vol. 1, pp. 133–136 vol. 1, May 1996.

[40] S. Young, *HMMs and Related Speech Recognition Technologies*, pp. 539–558. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[41] S. Young, N. Russell, and J. Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," tech. rep., 1989.

[42] N. Jakovljević, D. Mišković, M. Janev, and D. Pekar, "A decoder for large vocabulary speech recognition," in *Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on*, pp. 1–4, IEEE, 2011.

[43] O. Scharenborga, D. Norrisb, L. Ten Boscha, and J. M. McQueenc, "How should a speech recognizer work?," *Cognitive Science*, vol. 29, pp. 867–918, 2005.

[44] I. L. Hetherington, "A multi-pass, dynamic-vocabulary approach to real-time, large-vocabulary speech recognition," in *in Proc. of INTERSPEECH*, pp. 545–548, 2005.

[45] M. Gnjatovic, M. Janev, and V. Delic, "Focus tree: modeling attentional information in task-oriented human-machine interaction," *Appl. Intell.*, vol. 37, no. 3, pp. 305–320, 2012.

[46] K. Oberauer, "Access to information in working memory: exploring the focus of attention.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 28, no. 3, p. 411, 2002.

[47] D. Broadbent, *Perception and communication*. Pergamon Press, 1958.

[48] R. Atkinson and R. Shiffrin, "Human memory: A proposed system and its control processes," vol. 2 of *Psychology of Learning and Motivation*, pp. 89 – 195, Academic Press, 1968.

[49] N. Cowan, "Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system," *Psychological Bulletin*, vol. 104, pp. 163–191, 1988.

[50] K. Oberauer, "The focus of attention in working memory—from metaphors to mechanisms," *Frontiers in Human Neuroscience*, vol. 7, p. 673, 2013.

[51] K. Oberauer, A. S. Souza, M. D. Druey, and M. Gade, "Analogous mechanisms of selection and updating in declarative and procedural working memory: Experiments and a computational model," *Cognitive Psychology*, vol. 66, no. 2, pp. 157 – 211, 2013.

[52] B. J. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Comput. Linguist.*, vol. 12, pp. 175–204, July 1986.

[53] D. Sperber and D. Wilson, *Relevance: Communication and Cognition*. Language and Thought Series, Harvard University Press, 1986.

[54] H. Brouwer, H. Fitz, and J. Hoeks, "Getting real about semantic illusions: rethinking the functional role of the p600 in language comprehension," *Brain Res*, vol. 1446, pp. 127–143, Mar 2012.

[55] S. Coulson, *Electrophysiology and Pragmatic Language Comprehension*, pp. 187–206. London: Palgrave Macmillan UK, 2004.

[56] J. C. Hoeks, L. A. Stowe, and G. Doedens, "Seeing words in context: the interaction of lexical and sentence level information during reading," *Cognitive brain research*, vol. 19, no. 1, pp. 59–73, 2004.

[57] M. Kos, T. Vosse, D. Van Den Brink, and P. Hagoort, "About edible restaurants: Conflicts between syntax and semantics as revealed by erps," *Frontiers in Psychology*, vol. 1, p. 222, 2010.

[58] E. Gibson, "Linguistic complexity: locality of syntactic dependencies," *Cognition*, vol. 68, no. 1, pp. 1 – 76, 1998.

[59] M. Gnjatović, "Changing concepts of machine dialogue management," in *Cognitive Infocommunications (CogInfoCom), 2014 5th IEEE Conference on*, pp. 367–372, Nov. 2014.

[60] "Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech," tech. rep.

Прилог А1

Списак команди

Табела А1.1 садржи говорне чинове које су субјекти из прве групе спонтано формулисали да би илустровали говорну интеракцији између корисника и мобилног телефона. Из датог скупа, субјекти из друге групе су на случајни начин бирали команде и изговарали их преко мобилног или фиксног телефона. Корпус за тестирање прототипског система, описан у поглављу 6, формиран је на основу ових снимака.

Табела А1.1: Списак говорних чинова у корпусу за тестирање система.

	Говорни чин
1.	Отвори непрочитане поруке.
2.	Прочитај последњу Дејанову поруку.
3.	Позови Мирослава Илића.
4.	Канцеларија професора.
5.	Назови Соњу Неорчић на кућни телефон.
6.	Назови је.
7.	Убаци нов број мобилног.
8.	Прикажи пропуштене позиве.
9.	Дај ми Милицу из рачуноводства.
10.	Листај само одлазне позиве.
11.	Само пропуштени.

	Говорни чин
12.	Укључи калкулатор.
13.	Отвори именик.
14.	Иди на вести дана.
15.	Уђи у галерију.
16.	Додај нови број кућног телефона.
17.	Прикажи контакте из Теленор мреже.
18.	Излистај бројеве такси служби.
19.	Промени фиксни телефон од Жарка Смиљановића.
20.	Измени мобилни телефон ресторана Златна Птица.
21.	Покажи поруке од Драгишиног брата.
22.	Промени број Топаловића.
23.	Пошаљи поруку породици Стефановић.
24.	Зовни пекару Жеки.
25.	Сачувај све.
26.	Иди на мејлове.
27.	Врати се на почетак.
28.	Иди на слике.
29.	Ротирај фотографију.
30.	Дај ми конвертор валута.
31.	Забележи за сутра у подне састанак.
32.	Промени мелодију.
33.	Додај звук за поруку.
34.	Сними говорну поруку.
35.	Промени број Снежаниног телефона.
36.	Позови Гоцу са факултета.
37.	Сними гласовну поруку за Жељану, Анђину фризерку.
38.	Зови Беату, Драгичину сестру.
39.	Молим Аутобуску станицу.
40.	Сачувај целу преписку.
41.	Прикажи данашње поруке од Наташе.
42.	Хоћу полицију.
43.	Позови хитну помоћ.

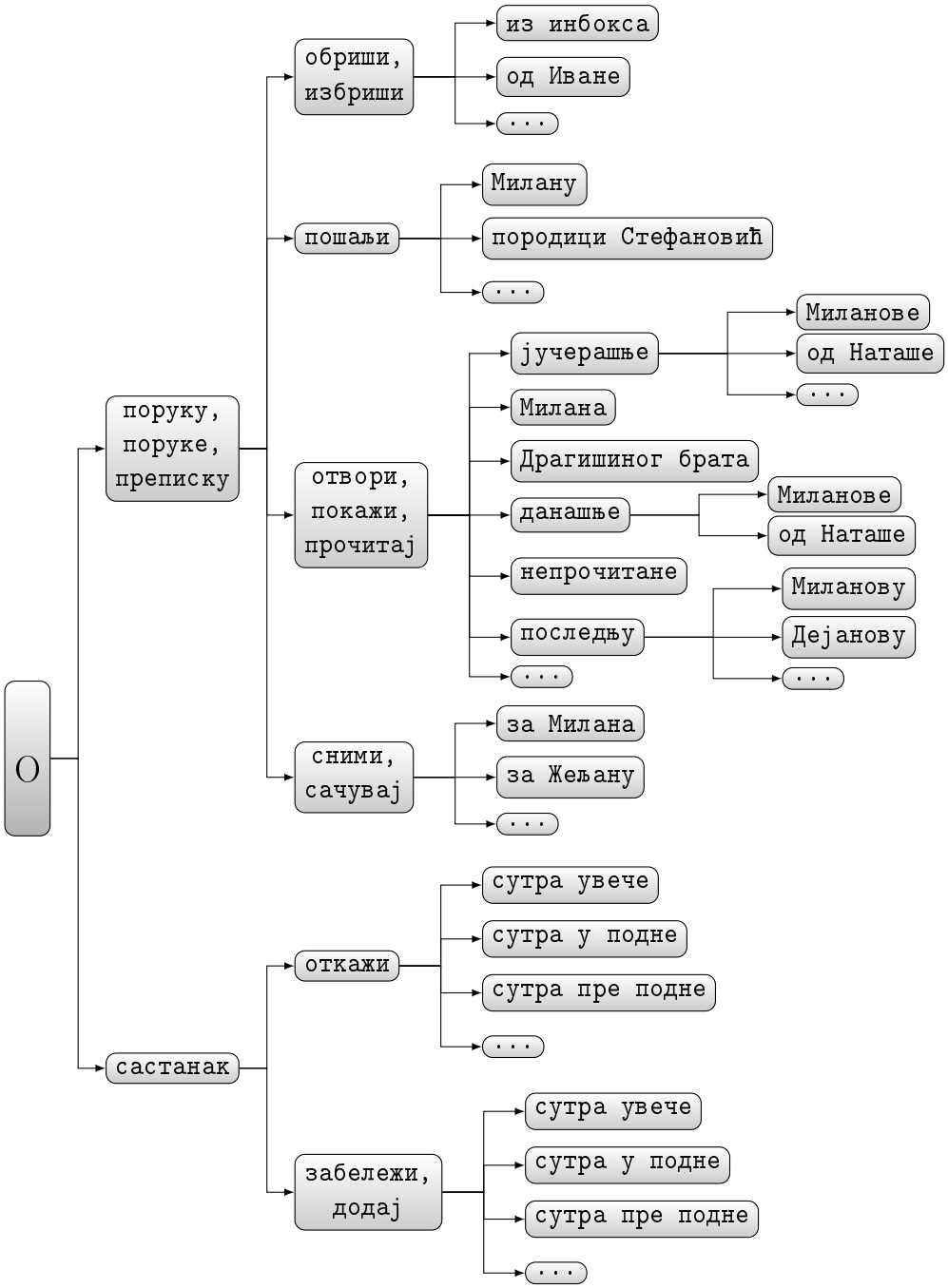
	Говорни чин
44.	Назови жену.
45.	Дај ми Јанка на мобилни.
46.	Зови шлеп службу.
47.	Назови фирму.
48.	Цимни Мирослава.
49.	Сними контакт за Ђолета.
50.	Сними у именик овај број.
51.	Прикажи биране бројеве.
52.	Отвори фотографију.
53.	Прикажи слику.
54.	Зови железничку станицу у Новом Саду.
55.	Излистај пропуштене позиве од прошле недеље.
56.	Обриши контакт Ана Тијанић.
57.	Промени слику контакта.
58.	Хајде отвори белешке.
59.	Додај напомену.
60.	Обриши све аларме.
61.	Одговори на позив.
62.	Јави се.
63.	Позови директора.
64.	Додај број за Марка Ђекића.
65.	Сними контакт.
66.	Избриши га.
67.	Додај број за Сару Кнежевић.
68.	Сачувај измене и изађи.
69.	Молим те, укључи радио.
70.	Прикажи долазне.
71.	Прикажи све.
72.	Покажи слике од Горане.
73.	Главни мени.
74.	Прикажи контакте.
75.	Сними поруку.

	Говорни чин
76.	Избриши све поруке из инбокса.
77.	Дај ми листу позива.
78.	Прикажи све конверзације.
79.	Обриши све поруке од Иване.
80.	Постави аларм.
81.	Отвори подсетник.
82.	Прикажи задњи унет контакт.
83.	Порука за Стевана Урошевића.
84.	Цимни ми зубарку.
85.	Прикажи ми данашњи подсетник.

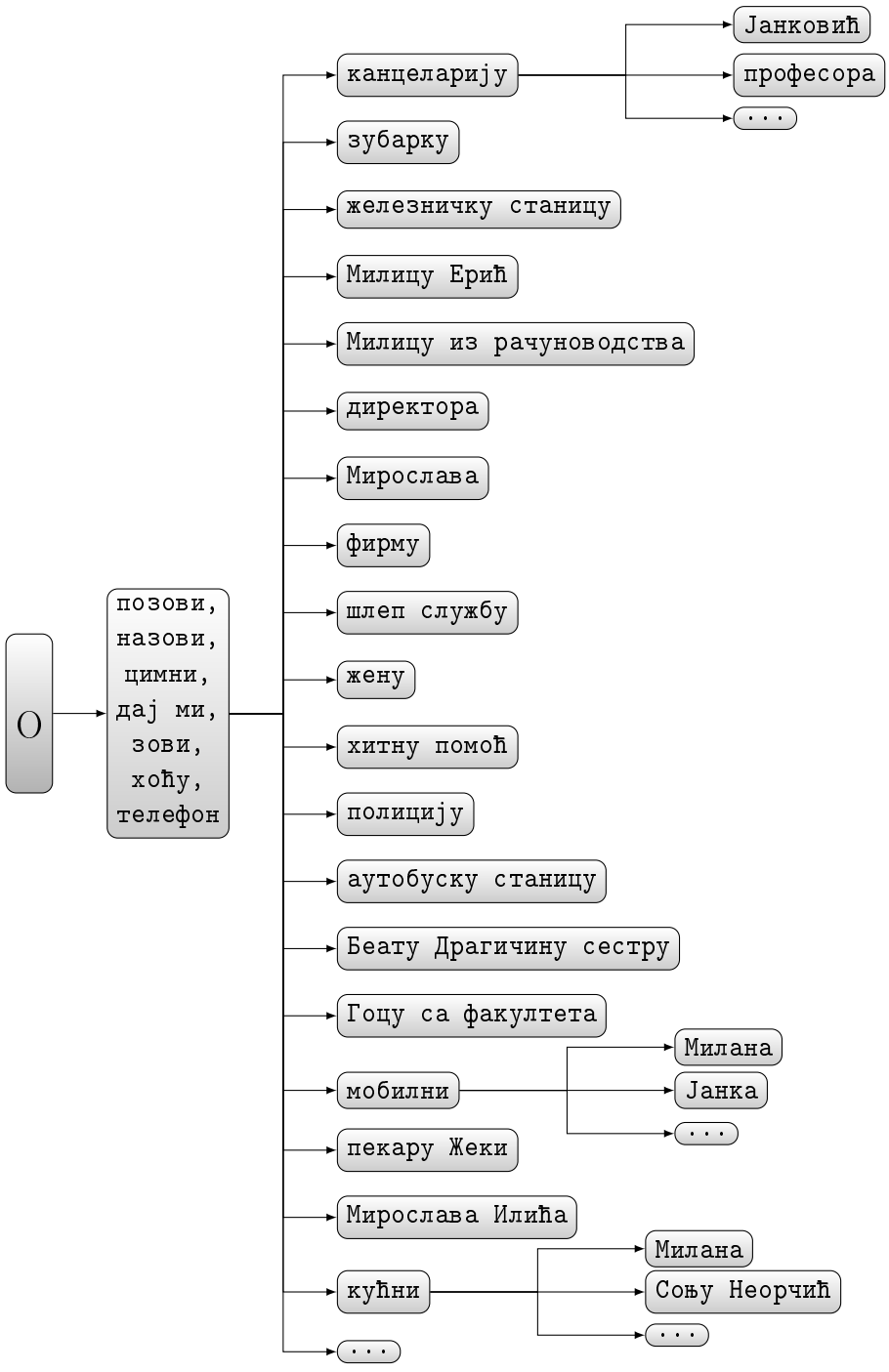
Прилог А2

Приказ фокусног стабла

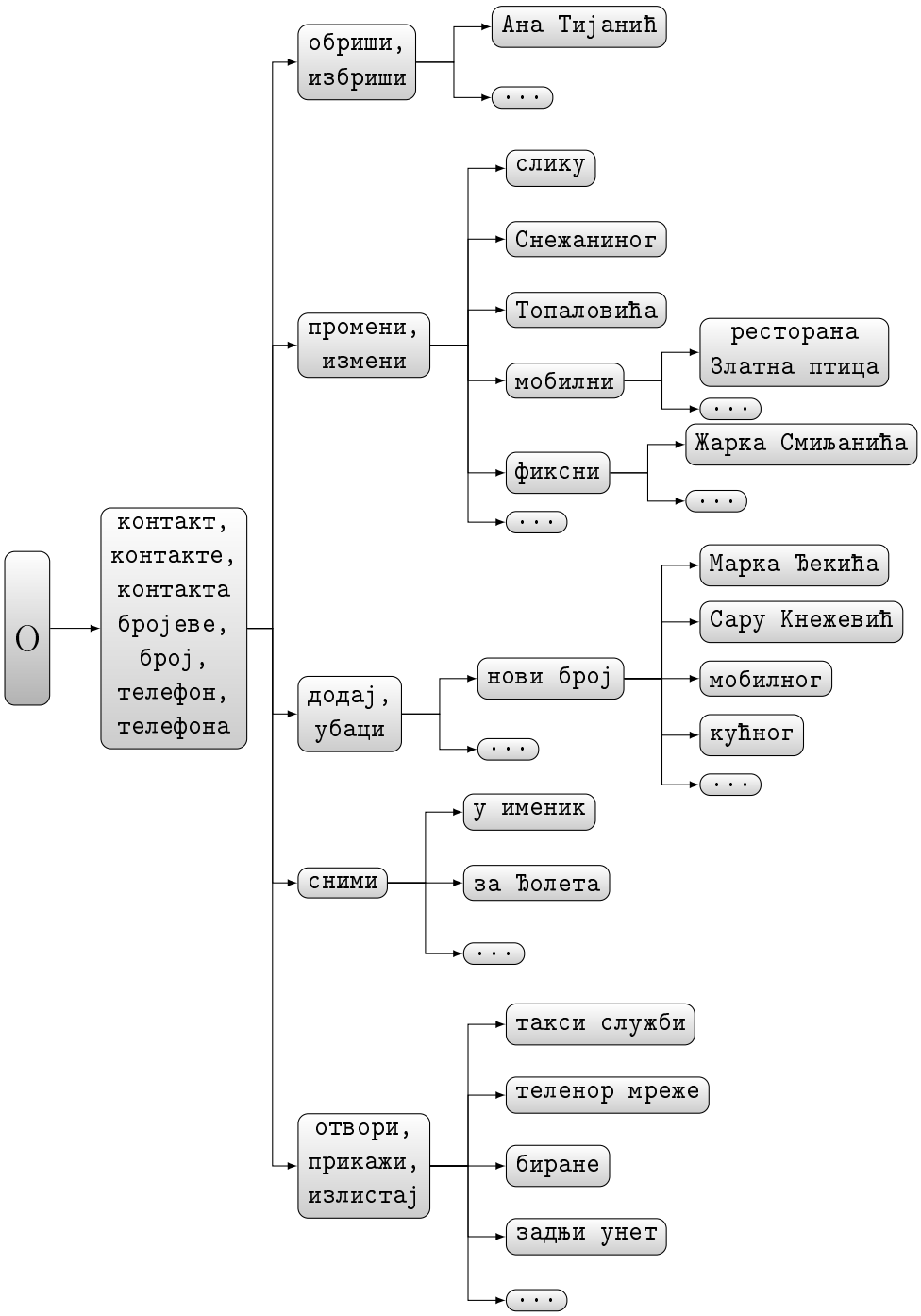
Слике А2.1—А2.5 приказују фокусно стабло коришћено при моделовању домена интеракције између корисника и мобилног телефона. Приказаан је подскуп чворова релевантних за говорне чинове заступљене у корпусу за тестирање система. Чвор O представља корен стабла.



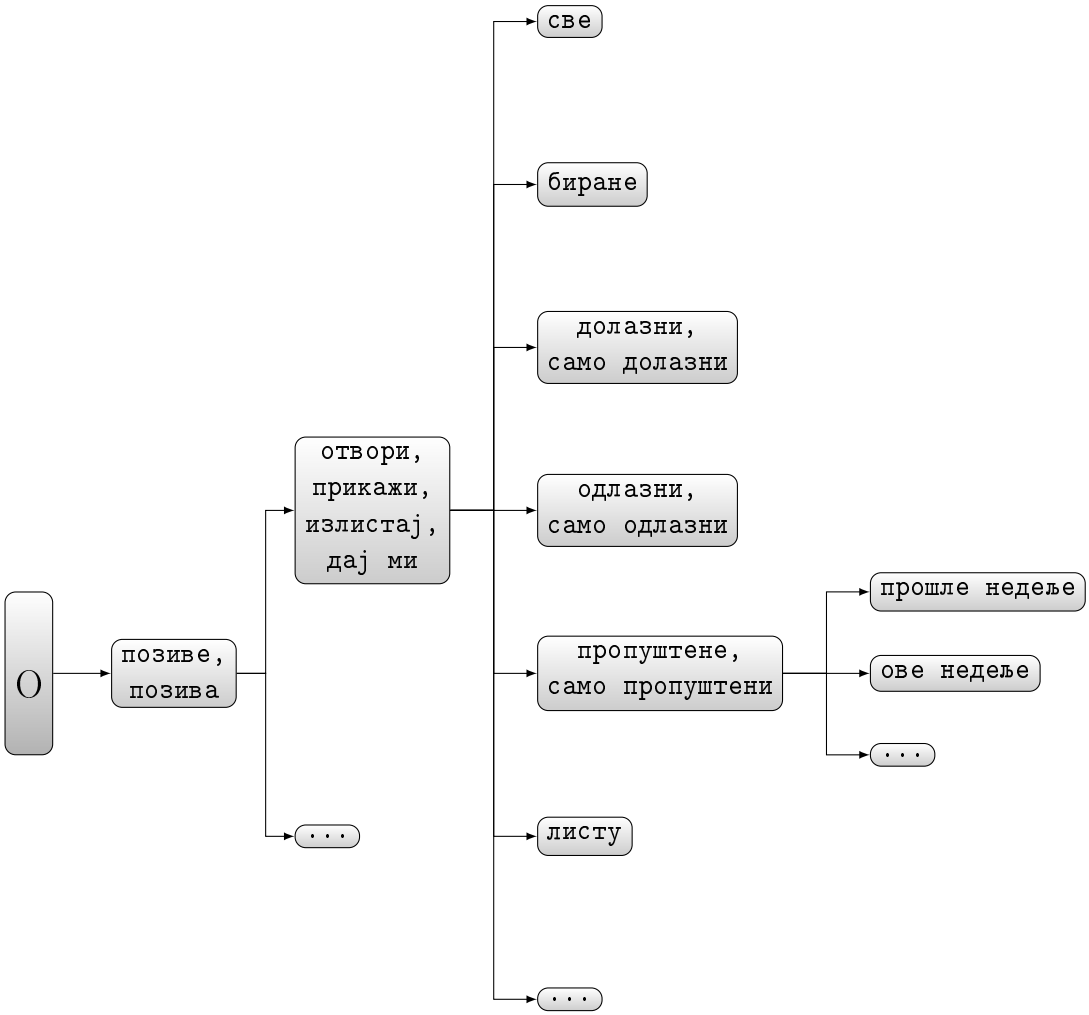
Слика А2.1: Приказ фокусног стабла намењеног моделовању домена интеракције између корисника и интегрисаног система — део I



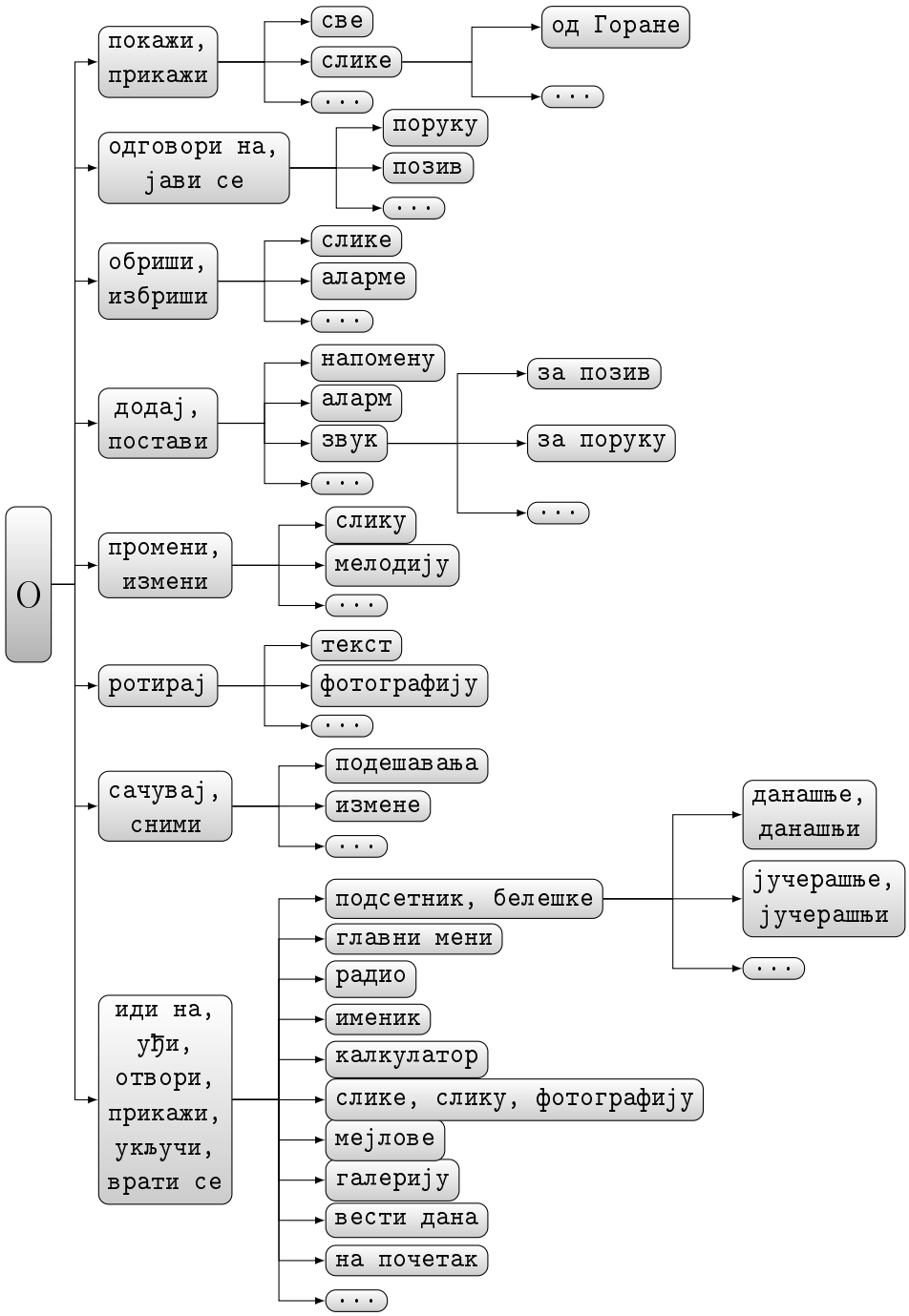
Слика А2.2: Приказ фокусног стабла намењеног моделовању домена интеракције између корисника и интегрисаног система — део II



Слика А2.3: Приказ фокусног стабла намењеног моделовању домена интеракције између корисника и интегрисаног система — део III



Слика А2.4: Приказ фокусног стабла намењеног моделовању домена интеракције између корисника и интегрисаног система — део IV



Слика А2.5: Приказ фокусног стабла намењеног моделовању домена интеракције између корисника и интегрисаног система — део V

Прилог А3

Изабрани примери препознавања

Табела А3.1 садржи примере препознавања применом предложеног алгоритма. У другој колони ове табеле су дате тачне (изговорене) реченице. У трећој колони су дати скупови који садрже по 5 најбољих хипотеза препознавања које је понудио статистички препознавач говора. Редослед одражава вероватноћу добијену током процеса декодовања, тј. прва хипотеза би представљала резултат препознавања у случају да се не користи хибридни приступ. За сваки скуп, хипотеза коју је предложени алгоритам изабрао као најбољу је написана задебљано.

Табела А3.1: Изабрани примери препознавања

	Изговорена реченица	Хипотезе препознавања
1.	хајде отвори белешке	**хајде отвори белешке** хајде отвори белешке и хајде отвори и белешке хајде о отвори белешке хајде отвори белешке
2.	назови соњу неорчић на кућни телефон	**назови соњу неорчић на кућни телефон** назови соњу неорчић напусти телефон назови соњу неорчић на кући телефон назови соњу не отићи обуци телефон назови соњу неорчић на кућни телефон

	Изговорена реченица	Хипотезе препознавања
3.	промени слику контакта	ли слику контакта ми слику контакта **промени слику контакта** ни слику контакта нини слику контакта
4.	додај напомену	додај након **додај напомену** добре након додај напорно додај на пун
5.	обриши све аларме	обриши све аларм **обриши све аларме** обриши све а ван обриши све аха ван обриши сва аларм
6.	сачувај измене и изађи	сачувај измене изађи сачувај измени изађи **сачувај измене и изађи** сачувај измене изађи их сачувај измене изађи с
7.	иди на вести дана	иди на вести дан иди на вести да иди на вести на иди на вести дао **иди на вести дана**
8.	излистај бројеве такси служби	излистај бројеве та служби излистај бројеве та луди **излистај бројеве такси служби** излистај бројеве тад служби излистај бројеве такву
9.	постави аларм	постави хари постави важан постави пар **постави аларм** постави а аларм
10.	додај нови број кућног телефона	додај нови број кућног телефон а **додај нови број кућног телефона** додај нови број кућу телефон а додај нови број кућног телефон на додај нови број кућног телефон ах

	Изговорена реченица	Хипотезе препознавања
11.	избриши све поруке из инбокса	избриши све поруке извини док сам избриши све поруке извини бог сам избриши све поруке извини баг сам **избриши све поруке из инбокса** избриши све поруке извини ок сам
12.	прикажи пропуштене позиве	прикажи пропуштене посебна прикажи пропуштене посебно **прикажи пропуштене позиве** прикажи пропуштене по седам прикажи пропуштене косе да
13.	јави се	јавити ја више јави си јави с **јави се**
14.	иди на мејлове	игра мејлове **иди на мејлове** игра милане игра не воле играм ево ове
15.	излистај пропуштене позиве од прошле недеље	излистај пропуштене позиве прошла недеља излистај пропуштене позиве прошле недеље излистај пропуштене позиве прошла недеље **излистај пропуштене позиве од прошле недеље** излистај пропуштене позиве прошло недеље