Опис говорне базе за препознавање говорника на српском језику

Иван Д. Јокић, Томислав Н. Добријевић, Никша М. Јаковљевић, Владо Д. Делић *Member, IEEE*

***Садржај* — У овом раду је дат приказ говорне базе која је намењена обуци и тестирању система за аутоматско препознавање (верификацију) говорника. Говорна база је развијена на Факултету техничких наука у Новом Саду. Базу чине спонтано изговорене секвенце цифара српског језика од стране четрдесетак говорника, снимљене у периоду од шест месеци. Снимање је вршено у канцеларијским условима, коришћењем просечног рачунарског микрофона (Genius MIC-01) и интегрисане звучне картице (SoundMax на MB ASUS-A7V8K).**

***Кључне речи* — Аутоматско препознавање (верификација) говорника, говорна база.**

I. Увод

Аутоматско препознавање говорника представља област дигиталне обраде говора која има за циљ да се на основу гласа говорника утврди његов идентитет, односно „ко је рекао?“. Разликују се две групе система: системи за идентификацију и верификацију говорника. Код система за идентификацију циљ је одредити идентитет, за разлику од система за верификацију који потврђују или одбацују тврдњу о идентитету говорника.

Један од битних елемената који утичу на квалитет и поузданост експеримената представља и говорна база. Говорне базе које постоје и које су намењене за потребе препознавања говора [1], нису адекватне пошто не обухватају промене гласа говорника током времена. Из тог разлога снимања говорних база намењених аутоматском препознавању говорника се врше у више временски померених сесија.

Препознавање говорника за разлику од других говорних технологија није у тој мери зависно од језика, стога је могуће користити и говорне базе снимљене на другим језицима. Одлучили смо се на снимање говорне базе за препознавање говорника на српском језику пошто сваки језик има своје фонетске и прозодијске посебности, а које у мањој или већој мери варирају између говорника. Са друге стране постоји тренд развоја говорних база и за мање језике у циљу формирања система који је независан од говорног подручја [2]. Сажет преглед постојећих база за европске језике дат је у одељку II.

Процес формирања говорне базе обухвата неколико корака:

- снимање аудио материјала,
- формирање транскрипција,
- временско поравнавање транскрипција (текстуалног садржаја) са аудио садржајем,
- верификација базе.

Детаљан опис услова и начина на који је снимана база дат је у делу III. Садржај базе је описан у делу IV. Начин поравнавања транскрипција и аудио материјала је наведен у делу V, након ког следи закључак.

II. Преглед постојећих база

Ради развоја и процене квалитета препознавача говорника у литератури се могу пронаћи разне јавно доступне стандардизоване говорне базе. Најчешће коришћене су [3]:

SIVA – италијанска говорна база снимана преко јавне телефонске мреже. Садржи исказе 840 говорника, при чему је свега 40 говорника снимљено у 18 сесија, а преосталих 800 у по једној сесији. Најмање време које је протекло између две узастопне сесије је 3 дана. Садржава изговорене речи, цифре, кратка питања и читан текст.

PolyVar – говорна база на француском језику (говорници су из Француске и Швајцарске) снимана преко јавне телефонске мреже. Садржи гласове 143 говорника (85 мушких и 58 женских) снимљених у просеку са 25 сесија по говорнику (код неких говорника постоји и 229 сесија). Интервал снимања зависи од говорника и износио је од једног дана до једног месеца. Говорници су читали секвенце цифара, реченице, али постоје и сегменти са спонтаним говором.

POLYCOST – развијена у оквиру Европског пројекта за верификацију говорника COST 250, који обухвата 13 европских држава. Снимана је преко међународних ISDN телефонских линија. Чине је снимци 133 говорника (74 мушкараца и 59 жена) снимљених у бар 5 сесија. Интервал између сесија варира од једног дана до неколико недеља. Говорници су читали цифре, секвенце цифара и реченице, али и произвољан спонтан монолог. Доминантан језик је енглески, али су присутни и снимци на другим европским језицима.

Овај рад је подржан од стране Министарства за науку и технолошки развој републике Србије у оквиру пројекта “ Говорна комуникација човек-машина” (ТР 11001).

И. Д. Јокић, Факултет техничких наука у Новом Саду, Србија (телефон: 381-64-3526245; факс: 381-21-4752997; e-mail: IBAHJOKIh@gmail.com).

Т. Н. Добријевић, Телеком Србија а.д., Таковска 2, 11000 Београд, Србија; (e-mail: tomislavd@telekom.rs).

Н. М. Јаковљевић, Факултет техничких наука у Новом Саду, Трг Доситеја Обрадовића 6, 21000 Нови Сад, Србија; (e-mail: jakovnik@uns.ac.rs).

В. Д. Делић, Факултет техничких наука у Новом Саду, Трг Доситеја Обрадовића 6, 21000 Нови Сад, Србија; (e-mail: vdelic@uns.ac.rs).

KING – развијана по истраживачком уговору са владом Сједињених Држава при чему је један део говорне базе сниман преко висококвалитетног микрофона док је други сниман преко јавне телефонске мреже. Садржи снимке 51 говорника (искључиво мушкарци) који су снимани у 10 сесија, а интервал између две сесије је био недељу дана. Говорници су требали својим речима да опишу фотографије које виде. Језик је амерички енглески.

YOHO – дизајнирана ради подршке верификацији говорника зависној од текста за безбедан приступ владиним апликацијама. Чине је снимци 138 говорника (106 мушкараца и 32 жене) снимљених у 14 сесија од којих су 4 намењене обуци, а 10 верификацији. Временски интервал између две узастопне сесије снимања је бар 3 дана. Садржи искључиво изговорене секвенце цифара. Снимање је вршено и преко лоших и преко квалитетних микрофона. Језик је амерички енглески.

Switchboard I-II – база телефонског квалитета, а чине је снимци спонтаних телефонских разговора. Снимљено је 1200, с приближно подједнаким бројем мушких и женских, говорника. Сваки од говорника је снимљен у бар 5 сесија. Интервал између снимања варира од неколико дана до неколико недеља. Језик је амерички енглески. Део овог корпуса је Амерички институт за стандарде и технологију **NIST** издвојио и формирао тзв. евалуационе базе: **NIST Evaluation Corpus 1996 ('97,'98)**.

OGI Speaker Recognition Corpus – снимана преко јавне телефонске мреже, а чине је снимци 91 говорника (~50% мушкараца и жена) сниманих у 12 сесија, при чему је минимално време између две узастопне сесије месец дана. Говорници су читали фразе, цифре и реченице. Језик је амерички енглески.

Применом ових говорних база може се испитати и зависност тачности коришћеног препознавача говорника од језика који се користи, као и могућности његове вишејезичне примене.

III. Начин снимања базе

Велики број система за препознавање и верификацију говорника од корисника захтева да изговори на случај изабрану кратку секвенцу цифара (обично од 4 до 6 цифара у низу), на основу које доноси одлуку о идентитету корисника. Овакав приступ је уједно и основни разлог зашто базу чине изговори секвенци од по 4 цифре српског језика. Мали лексикон, који чини 10 цифара српског језика, не обезбеђује одговарајућу фонетску варијабилност тако да ова база није најпогоднија за експерименте са системима за препознавање говорника који су независни од текста.

Секвенцу од 4 цифре особа која се снима може лако запамтити и изговорити спонтано, што обично није случај са реченицама или дужим фразама. Приликом снимања говорници су инструисани да изговоре задате секвенце на начин како би их изговарали да користе неки реални систем за идентификацију на основу гласа. Оваквим приступом је постигнуто да се у контролисаним условима у снимљеним исказима добију ефекти који постоје код спонтаног говора, као што су непотпуна артикулација, гутање појединих гласова који додатно могу да укажу на идентитет говорника.

Сви говорници су снимани у истим канцеларијским условима коришћењем просечног рачунарског микрофона (Genius MIC-01) и интегрисане звучне картице (SoundMax на MB ASUS-A7V8K). Снимци су сачувани у wav формату (учестаност одабирања 22050 Hz и 16 бита по одмерку).

Због непостојања одговарајућих услова није вршено симултано снимање помоћу више различитих микрофона. Релативно скромна и лако доступна опрема којом је вршено снимање омогућила је да разлике које обично постоје између говорне базе и снимака који би се добили у реалним условима, на кућном рачунару у канцеларијском или собном окружењу, буду релативно мале. Говорна база намењена препознавању и верификацији говорника треба да омогући и праћење промена карактеристика гласа једног говорника током времена. Да би се ово обезбедило снимање се врши у неколико сесија. Минималан временски период између две узастопне сесије је недељу дана. Ове препоруке су испоштоване и при снимању ове базе.

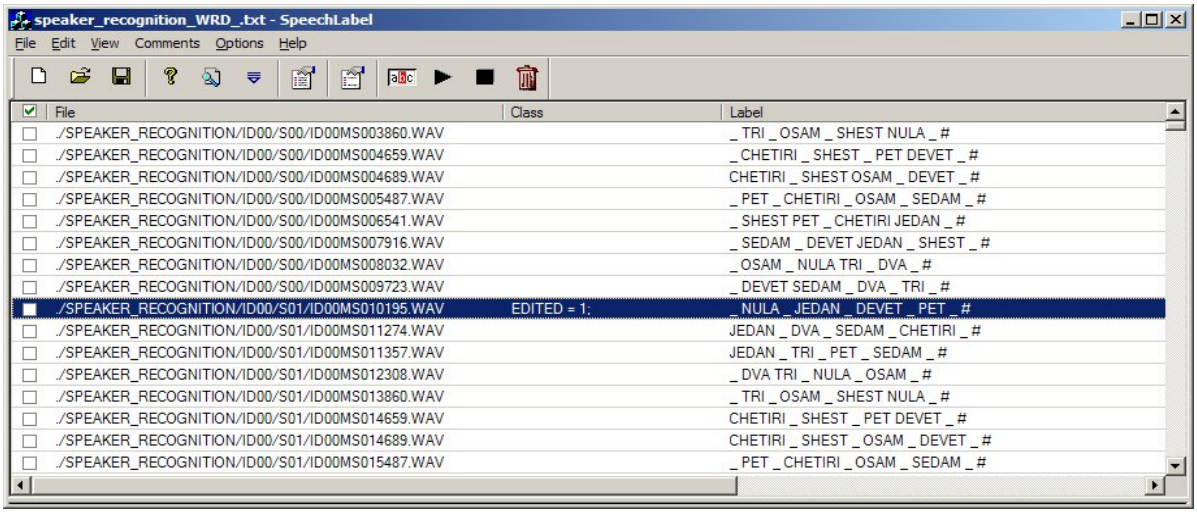
IV. Садржај базе

Говорну базу чине искази 44 говорника (29 мушкараца и 15 жена) животне доби од 20 до 40 година. Сваки исказ садржава секвенцу од четири спонтано изговорене цифре српског језика.

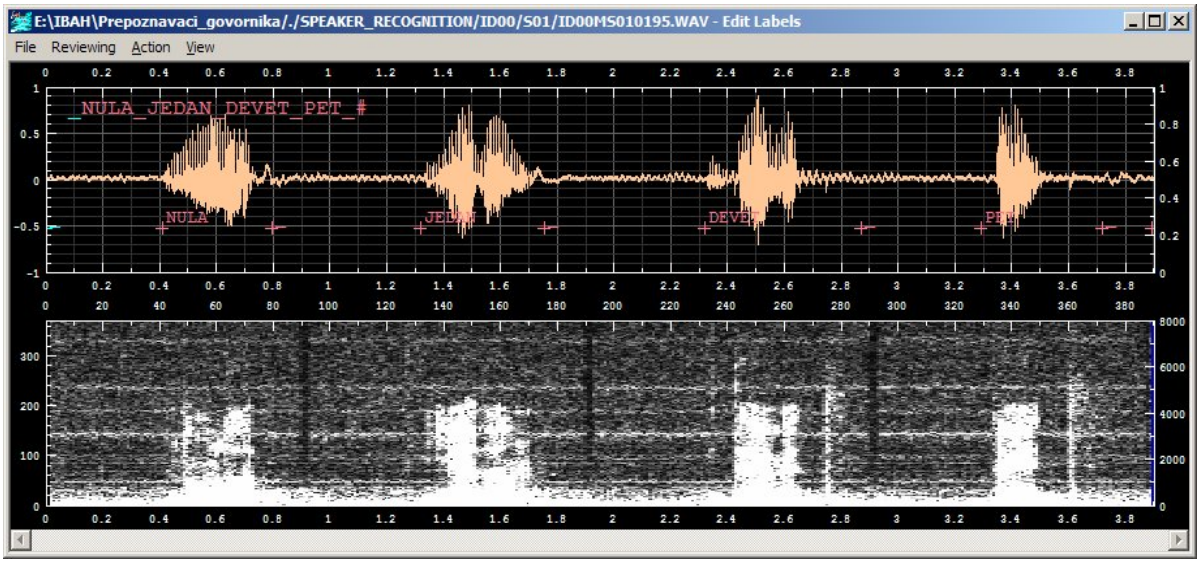
Планирано је да се снимање сваког од говорника реализује у 10 сесија. У свакој од 10 сесија говорник треба да изговори 12 различитих секвенци од по 4 цифре. Иако је временски период у коме је реализовано снимање био релативно кратак (6 месеци), поједини говорници су одустали тако да за њих постоји нешто мањи број снимака. У табели 1 су наведене ознаке њихових идентитета. Узорци гласа говорника који су снимљени у мање од 5 сесија се не могу искористити за анализу промена карактеристика гласа говорника током времена нити за обуку одговарајућих модела говорника, али могу послужити као снимци потенцијалних уљеза (уљези су говорници које систем за препознавање говорника није моделовао и не очекује на улазу).

У свакој од сесија говорници су изговарали по две исте секвенце (1-3-5-7 и 4-6-8-9), док су остале секвенце цифара биране тако да се свака од цифара нађе бар једном на првој, другој, трећој или четвртој позицији у измењеном контексту. Основна намена ове две секвенце цифара које су заједничке за све говорнике је при тестирању система.

Од правила да говорник у различитим сесијама изговори различите секвенце цифара се одступило код



Сл. 1а. Маска апликације за надзор говорних фајлова.



Сл. 1б. Оперативни прозор за кориговање граница.

10 говорника (носе ознаке ID00 ÷ ID09), да би се на већем броју секвенци могле проучити промене карактеристика гласа говорника током времена.

Дефинисан је јединствен начин обележавања снимака чији је формат: IDzzpSyyxxxx, а значење појединих поља је дато у табели 2. Усвојени формат омогућава да се на основу самог имена фајла закључи и његов садржај.

У циљу будуће обуке и тестирања препознавача потребно је поделити говорну базу у две дисјунктне целине, скуп за обуку и скуп за тестирање. Ако ово не би било испуњено тестови не би били фер, пошто би се систем обучавао и тестирао на истим снимцима што у реалним условима није могуће, а добијене перформансе би биле боље од оних које би систем имао у реалним условима. У зависности од тога да ли се тестира систем зависан односно независан од текста разликујемо два тест скупа. У случају система зависног од текста, секвенце цифара 1-3-5-7 и 4-6-8-9 које су снимљене у првих 5 сесија чине тест скуп, што чини приближно 1/10 целокупне базе.

ТАБЕЛА 1: ЛИСТА ГОВОРНИКА КОД КОЈИХ ПОСТОЈЕ ОДСТУПАЊА У БРОЈУ СЕСИЈА.

ID34, ID35, ID36, ID38, ID41, ID43	мање од 5 сесија
ID01, ID37, ID40	9 сесија
ID21	11 сесија
ID39	6 сесија
ID08	10 сесија, с тим што сесија S06 садржи 11 снимака

ТАБЕЛА 2: ОБЈАШЊЕЊЕ ОЗНАКА ФАЈЛОВА.

zz	Двоцифрена идентификациона ознака говорника
p	Ознака пола говорника (**m**-мушки, **f**-женски)
yy	Двоцифрен редни број сесије снимања
xxxx	Четири изговорене цифре

Са друге стране, у случају система независног од текста, тест скуп чине снимци свих говорника снимљених у првој сесији проширен са снимцима говорника код којих постоји мање од 6 сесија (види табелу 1). Напоменимо и то да су снимци говорника код којих не постоји бар шест сесија изостављени из скупа за обуку, без обзира на то да ли се обучава систем зависан или независан од текста.

Скуп за обуку не мора да садржи део који није обухваћен скупом за тестирање, већ само један његов део, те се смањивањем скупа за обуку може испитивати утицај величине скупа за обуку на перформансе препознавача.

V. Транскрипције

За сваки исказ који се налази у бази формирана је транскрипција. Још у фази снимања извршена је контрола да ли се оно што је речено и оно што треба да буде речено поклапало. У случајевима када је дошло до овог одступања, име фајла (последње четири цифре у имену) је кориговано у складу са оним што је речено.

Као последица спонтаног говора, код појединих говорника, при изговору неких цифара дошло је до гутања као и изобличења појединих фонема. Ове модификације се могу приписати карактеристикама посматраних говорника. Ово у великој мери отежава поступак формирања транскрипција на нивоу фонема, што је уједно и разлог зашто у овој фази развоја базе постоје само транскрипције на нивоу речи.

Поред информације о томе шта је речено потребно је поставити и границе између речи. Иницијално постављање граница између речи се обично врши аутоматски помоћу система за аутоматско препознавање говора. У те сврхе искоришћен је систем за препознавање говора развијен на Факултету техничких наука у Новом Саду [4][5].

Након аутоматски постављених граница неопходно је извршити њихово контролисање и евентуалну корекцију, што је урађено софтверским алатом развијеним у оквиру Алфанум пројекта [6]. Маска апликације за надзор посматраних говорних фајлова и сам изглед оперативног прозора за кориговање граница приказани су на Сл. 1а и Сл. 1б. Преслушавањем говорних узорака и праћењем таласног облика сигнала и/или спектрограма, у горњем односно доњем делу оперативног прозора, контролише се исправност претходно постављених граница и по потреби врши њихова корекција. Подаци о транскрипцијама говорних фајлова и временским границама у секундама аутоматски се чувају у одговарајућем .txt фајлу.

VI. Закључак

Дат је приказ прве верзије Говорне базе намењене препознавању говорника на српском језику. Тренутно броји 44 говорника са 4643 снимка. У поређењу са стандардним светским базама нешто је мања по свом обиму како по броју говорника тако и по свом фонетском садржају. База је тренутно у фази верификације, односно проверавања да ли су аутоматски постављене границе исправне.

Пошто је снимана у канцеларијским условима са лако доступном опремом, ова база је идеална за формирање система за ограничавање приступа на кућним рачунарима, што јој је била иницијална намена.

У наредном периоду планира се даље проширење базе, како бројем говорника тако и фонетским садржајем увођењем фонетски богатих реченица. Поред тога снимање ће се вршити и помоћу нешто квалитетнијих уређаја за снимање.

Тренутна верзија је првенствено намењена системима зависним од текста, али може да послужи за развој система независних од текста, при чему се фонд речи ограничава на цифре српског језика.

База тренутно није јавно доступна, али након завршене верификације ће то постати, у циљу лакшег поређења перформанси различитих система за препознавање говорника.

Захвалница

Захваљујемо се свим колегама и пријатељима који су учествовали у снимању ове говорне базе, нема потребе посебно наводити да без њиховог учешћа ова говорна база не би постојала.

Литература

[1] В. Делић, “Говорне базе на српском језику снимљене у оквиру пројекта АлфаНум,” *III ДОГС*, стр. 29-32, Нови Сад, 2000.

[2] R. Jones and J. Mason and R. Jones and L. Helliker and M. Pawlewski, “SpeechDat Cymru: A large-scale Welsh telephony database,” *LREC Workshop: Language Resources for European Minority Lang.,* 1998.

[3] J. Campbell, D. Reynolds, “Corpora for the evaluation of speaker recognition systems”, *ICASSP’99*, pp. 829-832, 1999.

[4] Д. Пекар, Р. Обрадовић, В. Делић, “Програмски пакет AlfaNumCASR систем за препознавање континуалног говора,” *IV ДОГС*, стр. 49–56, Бечеј, мај 2002.

[5] М. Јанев, Н. Јаковљевић, Д. Пекар, “Поређење система за препознавање говора на српском језику базираних на пуним и дијагоналним коваријансним матрицама” *Телфор 2007*, стр. 342–345, Београд, 2007.

[6] http://alfanum.ftn.ns.ac.yu/speech_label/speech_label_bin.zip.

Abstract

The paper presents a speech database intended for training and testing systems for the automatic speaker recognition (verification). The speech database is developed at the Faculty of Technical Sciences in Novi Sad. The database contains utterances with sequences of digits which are spontaneously spoken by about forty speakers in Serbian, and recorded in the period of six months in an office environment. An average computer microphone (Genius MIC-01) and integrated sound card (SoundMax ASUS on MB-A7V8K) were used for all recordings.

DESCRIPTION OF SPEECH DATABASES FOR SPEAKER RECOGNITION IN SERBIAN

Ivan D. Jokić, T. N. Dobrijević, N. M. Jakovljević, V. D. Delić