Образац за пријаву техничког решења[1]

Назив техничког решења	Скуп модула и процедура за висококвалитетну анотацију говорних база података
Аутори техничког решења	Едвин Пакоци, Бранислав Поповић, Дарко Пекар, Никша Јаковљевић
Категорија техничког решења	Нови производ примењен на међународном нивоу (М81)

За кога је рађено техничко решење и у оквиру ког пројекта МПНТР:

Техничко решење је реализовано у склопу пројекта технолошког развоја "Развој дијалошких система за српски и друге јужнословенске језике" (ТР32035, 2011-2015) на Факултету техничких наука и у предузећу АлфаНум у Новом Саду.

Ко користи техничко решење:

Предузеће АлфаНум у Новом Саду као партиципант на пројекту ТР32035, као и предузеће Speech Morphing System, Inc. (Campbell, California) из Сједињених Америчких Држава.

Година када је техничко решење урађено:

Техничко решење је реализовано током 2015. године. Део програмских алата који се користе у ове сврхе примењиван је у склопу разних задатака везаних за аутоматско препознавање говора од средине 2014. године, конкретно је почето са анотирањем говорних база почетком 2015. године, а описани скуп модула и процедура је финализиран до средине 2015. године. Цела главна процедура је прилагођена раду на *Linux* оперативним системима због захтева неких од коришћених програмских алата, уз неколико корака који се могу обавити у *Windows* окружењу. Техничко решење је настало услед потребе за ефикасним, прецизним и што више аутоматизованим процесом за анотирање нових говорних база.

[1] У складу са одредбама *Правилника о поступку и начину вредновања, и квантитавном исказивању научноистраживачких резултата истраживача*, који је 21.03.2008. године донео Национални савет за научни и технолошки развој Републике Србије («Службени гласник РС», бр. 38/2008).

Ко је прихватио-примењује техничко решење:

Ово техничко решење иницијално је реализовано за потребе развоја говорних и језичких ресурса у предузећу АлфаНум, а користи се и у развојном тиму на Факултету техничких наука. Процедура која је описана у техничком решењу је затим почела да се користи и у америчком предузећу „Speech Morphing System“ за развој говорних технологија, где је изабрана као најбоље решење за анотирање енглеских говорних база за њихове потребе.

Како су резултати верификовани (од стране ког тела):

1) Техничко решење је реализовано у Лабораторији за акустику и говорне технологије на Факултету техничких наука у Новом Саду. Имплементирано је и испитано на развојним системима у предузећу АлфаНум. Од средине 2015. године техничко решење је у свакодневној експлоатацији на серверима овог предузећа. Такође је независно испитано у предузећу „*Speech Morphing System*“, где је детаљном анализом резултата анотације на датој енглеској говорној бази изабрано као процедура која је дала најбоље резултате, па се активно користи и тамо за њихове намене.
2) Техничко решење је базирано на програмским алатима и скриптовима у оквиру софтверског пакета за аутоматско препознавање говора *Kaldi*, који је један од најчешће коришћених модерних алата који се користи у овој области. Поједини елементи техничког решења се ослањају на резултате остварене у оквиру више техничких и развојних решења и научних радова, која представљају верификоване резултате на технолошким пројектима МПНТР у периоду од 2005. до 2015. године. Основни начин на који је *Kaldi* коришћен у оквиру МПНТР пројеката описан је у раду:
    - *Бранислав Поповић, Едвин Пакоци, Стеван Острогонац, Дарко Пекар, „Large Vocabulary Continuous Speech Recognition for Serbian Using the Kaldi Toolkit“, 10. конференција Дигитална обрада говора и слике, DOGS 2014, Нови Сад, Србија: Факултет техничких наука Нови Сад, 5-9. октобар 2014, Милан Сечујски, Никша Јаковљевић, Владо Делић (eds.), 31-34, ISBN: 978-86-7892-633-4.*

    Техничко решење је успешно искоришћено за анотирање неколико прикупљених говорних база, и оне су затим такве коришћене за обучавање система за препознавање говора, од којих је један формиран за потребе апликације за управљање мобилним телефоном путем гласа на српском језику, која је описана у раду:
    - *Бранислав Поповић, Едвин Пакоци, Никша Јаковљевић, Горан Кочиш, Дарко Пекар, „Voice Assistant Application for the Serbian Language“, 23rd Telecommunications Forum, TELFOR 2015, Београд, Србија: Друштво за телекомуникације, 24-26. новембар 2015. (прихваћено за објављивање)*

3) Техничко решење верификује Наставно-научно веће Факултета техничких наука на основу позитивног мишљења два рецензента-експерта из области техничког решења. Наставно-научно веће ФТН је 25.11.2015. именовало рецензенте:
    - Проф. др Зорица Николић, Електронски факултет Ниш (Ниш, Србија)
    - Доц. др Бранислав Геразов, Факултет за електротехнику и информационе технологије (Скопље, Македонија)
4) Наставно-научно веће Департмана за енергетику, електронику и телекомуникације и Факултета техничких наука, на основу мишљења рецензената и приложених доказа, издаје Уверење о признавању техничког решења које потврђује да оно испуњава све услове да буде признато као нови производ примењен на међународном нивоу (М81), у складу са Правилником Министарства.

На који начин се користи (кратак опис):

Техничко решење омогућава анотацију новоприкупљене говорне базе у Waveform аудио формату, односно креирање лабела са изговореним фонемима за сваку аудио датотеку из базе, у којима су забележене границе између свака два фонема, као и оцена веродостојности сваке постављене границе. Излазне лабеле су у АлфаНум-овом текстуалном формату. Могуће је задавање низа конфигурационих параметара од којих се већина односи на издвајање акустичких обележја из аудио датотека и моделовање фонема, а неки и на начин формирања оцене фонема. Техничко решење се користи позивањем низа готових апликација које примењују развијене модуле под *Windows* окружењем (из командне линије), односно постављањем конфигурационих параметара и позивањем готовог скрипта под *Linux* окружењем.

Опис техничког решења:
Скуп модула и процедура за висококвалитетну анотацију говорних база података

C++ програмски модули, *Shell* скриптови и *Windows* и *Linux* апликације у склопу процедуре за аноритање говорне базе представљају нови производ заснован на сопственим говорним технологијама које су развијене на пројектима Министарства од 2005. до 2015. године, као и готовим програмским алатима у оквиру *Kaldi* софтверског пакета. Производ је реализован средином 2015. и од тада је у редовној употреби на серверима предузећа АлфаНум и у лабораторији развојног тима на Факултету техничких наука у Новом Саду, као и на серверима предузећа „Speech Morphing System“ у САД. Говорне базе анотиране применом овог техничког решења детаљно су тестиране, при чему је показано да ова процедура даје конзистентно добре границе између фонема, боље од неколико компетитивних метода, а омогућава и релативно лако и брзо детектовање проблематичких делова базе на основу оцене веродостојности.

Област на коју се техничко решење односи:

Техничко решење припада области Електроника, телекомуникације и информационе технологије, и као такво налази примену у већем броју различитих области, попут телекомуникација и обраде сигнала, обраде аудио сигнала, говорне комуникације човек-машина и разних других сродних области – заправо практично свуда где се користе говорне базе.

Проблем који се техничким решењем решава:

Где год се користе говорне базе, поготово велике, од изузетног је значаја да се оне што брже и што боље анотирају, да би се омогућило њихово успешно коришћење приликом развоја одговарајућих крајњих апликација. Један проблем је прецизност анотације – циљ јесте да цела база буде конзистентно добро лабелирана, а други је брзина – базу треба спремити за коришћење што брже, лакше и уз колико год је то могуће минимално ручно прегледање (да што више буде аутоматизовано). Лоше анотирана база у било којој апликацији везаној за системе за препознавање говора може довести до значајне деградације перформанси, јер се модели неће обучити на одговарајућим говорним сегментима. Са друге стране, пошто величина говорне базе може такође допринети боље обученим моделима говорних јединица, значајно је да се то одради и што је ефикасније могуће, а без у великој мери аутоматизованих процедура то није изводљиво.

Стање решености тог проблема у свету:

Проблем анотације, у смислу постављања граница фонема у аудио датотекама задате говорне базе, није често директно адресиран у научним радовима. Често је фокус на самој обуци модела фонема, која или почиње већ ручно прегледаним лабелама или од равномерно распоређеног низа фонема, у ком случају се конвергенцијом током одређеног броја итерација обуке долази до коначних вредности за границе између фонема. Два софтверска пакета за аутоматско препознавање говора која се често примењују у модерним апликацијама су *Kaldi* – „*The Kaldi Speech Recognition Toolkit*“ (2011, Povey D. et al.) и *HTK* - „*The HTK Book (For HTK Version 3.4)*“ (2006, Young S. et al.), од којих се први користи и у овом техничком решењу. Постоје и многи други јавно доступни алати за обуку система за препознавање говора, базирани на различитим начинима моделовања говорних јединица. Наш алат за анотирање говорних база омогућава да се преко генерализације жељених конвенција везаних за границе између појединих фонема на одређеном, релативно малом броју ручно задатих граница изграде иницијални модели и затим одреде конзистентне границе за читаву говорну базу. Осим тога, омогућава прегледање анотираног материјала по оценама одређених фонема, чиме се лако може детектовати ако у бази постоје проблематични сегменти које треба ручно кориговати или избацити (рецимо јер транскрипција неке реченице

не одговара ономе што је заиста речено, ако постоје оштећења у аудио сигналу или ако је нека реч изговорена на другачији начин у односу на онај предвиђен задатим лексиконом). Предложено решење је и једноставно за примену и омогућује кориговање параметара процедуре зависно од карактеристика базе (величине, разноликости, броја говорника, типа реченица...).

Објашњење суштине техничког решења и детаљан опис са карактеристикама, укључујући и пратеће илустрације и техничке цртеже (техничке карактеристике):

Техничко решење, дакле, решава проблем анотирања (лабелирања) произвољне дате говорне базе. Све што захтева на улазу су две датотеке. Једна је датотека са транскрипцијом (односно, низом изговорених речи) за сваку од аудио датотека. Друга је лексикон, односно речник изговора сваке од речи (низ фонема за сваку од речи). Свака реч која се појављује у транскрипцијама мора да се нађе и у лексикону, са тим што једна реч може имати и више алтернативних изговора у лексикону. У току процедуре обуке и постављања граница, систем ће се одлучити за одговарајући изговор. Уколико је у питању говорна база са више говорника, могуће је у датотеци са транскрипцијама задати и идентификаторе говорника и пола, као и низ других карактеристика говорника на основу којих се касније може једноставно направити подскуп говорне базе за неку следећу намену. Пример улазних транскрипција и лексикона дат је на слици 1.

```
\BLOCK00\SES001 AP001BA01.wav   BA      SPK001  M       28      New_York_City   four six seven seven
\BLOCK00\SES001 AP001BA02.wav   BA      SPK001  M       28      New_York_City   nine eight three three
...


EIGHT EY1 T
FOUR F AO1 R
NINE N AY1 N
...
```

Слика 1. Улазне транскрипције и лексикон (сегменти)

Комплетна процедура анотације говорне базе састоји се од следећих корака:

Конверзија транскрипција и лексикона у иницијалне лабеле (*Windows*)

Апликација претвара улазне транскрипције и изговоре појединих речи из транскрипција у иницијалне лабеле за сваку аудио датотеку из говорне базе. У овим лабелама нису маркиране границе између фонема нити постоје оцене, већ су ту само идентитети фонема, називи аудио датотека и описни подаци за говорнике који су наведени у датотеци са транскрипцијама. Између сваке две речи поставља се тишина (она може нестати током процеса обуке). У конфигурационим опцијама може се специфицирати коришћење одређених атрибута фонема (наглашеност вокала, оштећеност фонема), као и конверзије одређених фонема у друге у случајима када је

то пожељно. Пример позива апликације дат је испод, а део резултујућих лабела приказан је на слици 2. Лабеле су у АлфаНум-овом *sobj* текстуалном формату.

```
label_tool --kaldi-align-apps init-labs-eng –i transcripts_file;lexicon_file
           -o out_labels_file [-p config_file]
```

```
{
       files =
       [
              {
                     class = { ACCENT='New_York_City', AGE='28', GENDER='M',
                               SPK_ID='SPK001', UTT_TYPE='BA' },
                     file_name = './BLOCK00/SES001/AP001BA01.wav',
                     phonemes =
                     [
                            {
                                   phoneme = 'F',
                                   start = 0.0,
                                   surface_form = 'FOUR',
                                   word_start = 1
                            },
                            {
                                   phoneme = 'AO1',
                                   start = 0.0
                            },
                            {
                                   phoneme = 'R',
                                   start = 0.0
                            },
                            ...
```

Слика 2. Креиране иницијалне лабеле (сегмент)

*Конверзија лабела у **Kaldi data** датотеке* (*Windows*)

Апликација на основу улазних иницијалних лабела формира све датотеке са подацима о говорној бази потребне за стандардну *Kaldi* обуку. Ту спада 6 датотека, које се након извршавања апликације нађу у специфицираном излазном директоријуму: *text* – садржи идентификаторе реченица и уз сваки од њих њену транскрипцију; *utt2spk* – мапира сваки идентификатор реченице на одговарајућег говорника; *wav.scp* – мапира идентификатор реченице на путању до одговарајуће аудио датотеке; *spk2gender* – упарује сваког говорника са полом; *lexicon.txt* – речник изговора у *Kaldi* формату, који садржи изговоре речи из транскрипција; и *lm.arpa* – базични униграм модел језика са једнаким вероватноћама сваке од речи у стандардном ARPA формату. Улазни конфигурациони параметри омогућавају спецификацију основног директоријума аудио базе (ако су у лабелама релативне путање), улазног лексикона (ако желимо омогућити још алтернативних изговора неких речи), начина генерисања идентификатора реченице и разних других параметара. Све остале датотеке потребне за обуку модела су углавном стандардне и не обезбеђују се преко ове апликације.

```
label_tool --kaldi-align-apps kaldi-files –i in_labels_list_file
           -o out_files_directory [-p config_file]
```

Ручна анотација референтног скупа аудио датотека (*Windows*)

Следећи корак је издвајање извесног броја аудио датотека, односно њихових лабела као референтних. У овом релативно малом скупу лабела треба ручно поставити

границе између фонема, а он ће након тога служити за обучавање иницијалног скупа модела. Овај скуп треба да је довољно велик да садржи што више различитих говорних сегмената (фонема у различитим контекстима), да би се на тај начин „зацртале“ основне карактеристике модела фонема. Међутим, нема потребе да буде превелик – свакако, што више аудио датотека се издвоји у овај референтни скуп, то боље, али је сасвим довољно да то буде свега око 100-200 реченица које довољно добро покривају све говорне јединице, што не доводи до предугачког мануелног рада. Прегледање и ручна анотација референтних датотека је предвиђена да се врши преко АлфаНум-овог алата *SpeechLabel* („*SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база“ (2015, Мишковић Д. и др, техничко решење)).

*Конверзија референтних лабела у облик потребан за **Kaldi** обуку* (*Windows*)

Подскуп лабела који је издвојен у референтни скуп се затим конвертује у формат који одговара апликацијама које *Kaldi* користи приликом обучавања модела. *Kaldi* као начин моделовања говорних јединица користи тип аутомата са коначним бројем стања – FST (*Finite State Transducer*), у оквиру којих сваком моделу придружује целобројни идентификатор. За рад ове апликације потребно је преко улазних аргумената обезбедити број стања за сваки од модела (тј. идентификатора) – типично се тишине (не-говорни сегменти) моделују са по 5 стања, а све не-тишине са по 3, затим пресликавања појединих фонема у идентификаторе, путање до аудио датотека, као и пресликавања аудио датотека у идентификаторе реченице. Због ових потреба, типично се пре овакве процедуре анотирања ради једна чиста, стандардна *flat-start Kaldi* обука (без референтног дела базе), која аутоматски генерише све потребне идентификаторе. Апликација, на основу улазних лабела, за сваку аудио датотеку избацује низ идентификатора фонема за сваки фрејм, све у одговарајућем *Kaldi* формату. Такође се формира и подскуп датотеке са идентификаторима реченица које одговарају само референтним аудио датотекама – ово се користи у иницијалној обуци модела као специфичан списак датотека за ту обуку.

```
label_tool --labs-to-ali –i in_ref_labels_file -o out_ali_file -p config_file
```

***Kaldi** обука и аутоматско одређивање граница фонема* (*Linux*)

Процес обуке и одређивања граница фонема је у потпуности аутоматизован, и покреће се стартом *Shell* скрипта *run.sh* из *Linux* командног терминала. У оквиру главног скрипта позива се низ помоћних. Након почетног анализирања свих *data* датотека и генерисања помоћних датотека и идентификатора, први корак јесте издвајање обележја (сви параметри за то се задају у засебној конфигурационој датотеци) за целу базу, а затим следи обука модела монофона на основу ручно анотираног дела аудио базе. Све спецификације се задају на почетку скрипта (жељени укупни број гаусијана, итерација, итд). Аутоматски се издвоји подскуп свих потребних датотека везан за референтни скуп, а затим почиње обука модела у задатом броју итерација користећи

ручно задате границе између фонема. Након тога се одреде границе фонема на целој бази користећи обучене моделе монофона. Затим следи генерисање стабла повезаних трифона (на основу добијених граница), а потом и обука трифона. Све спецификације се опет задају на почетку скрипта *run.sh* (број листова у стаблу, број гаусијана, итерације...). Обука трифона се ради на целој бази, а раде се и повремена ажурирања граница у току обуке. После тога се одређују границе на основу резултујућих модела трифона, а затим се цео поступак обуке трифона понавља још једанпут, са новогенерисаним стаблом (на основу нових, бољих граница). Сви параметри су конфигурабилни. Након крајњег одређивања граница, формирају се излазне лабеле у *Kaldi* формату. Све споменуте целине су имплементиране у оквиру засебних скриптова, који пак позивају одговарајуће програме који су у оквиру *Kaldi* пакета алата (*compute-mfcc-feats*, *build-tree*, *gmm-init-model*, *gmm-mixup*, *gmm-est*, *compile-train-graphs*, *gmm-align-compiled*, *show-alignments*, као и многи други помоћни програми – последња два наведена програма су модификована тако да омогуће приказ и акустичких скорова фонема, односно израчунавање оцене веродостојности постављене границе на основу тих скорова). Начин позивања главног скрипта дат је испод, а пример задавања параметара у њему дат је на слици 3.

```
./run.sh -l 2>&1 | tee run_log.txt
```

```
# Number of states for silence phones
num_sil_states=5
# Number of states for non-silence phones
num_nonsil_states=3

# Fixed alignment training specifics
# complexity
ref_num_gauss=500
# iteration count
ref_num_iters=10
# last iteration to increase Gaussians count
ref_max_iter_inc=8

# First triphone pass training specifics
# complexity
tri1_num_states=1800
tri1_num_gauss=2500
# iteration count
tri1_num_iters=35
# iterations for realignment
tri1_realign_iters="10 20 30"
# last iteration to increase Gaussians count
tri1_max_iter_inc=25
```

Слика 3. Постављање параметара у скрипту *run.sh*

Ажурирање граница и оцена у почетним лабелама (*Windows*)

Ово је апликација која на основу информација из *Kaldi* лабела, мапе идентификатора реченица на путање до аудио датотека (*wav.scp*) и неколицине других параметара (дужина фрејма, померај фрејма) ажурира границе између фонема (тј. израчунава и уписује оне из *Kaldi* лабела) и уписује нове оцене. Овим се завршава прва итерација анотирања аудио базе. Начин позива је дат у наставку (путања до *Kaldi* лабела се наводи у конфигурационој датотеци), а пример добијених лабела на слици 4.

```
label_tool --kaldi-align-apps new-align –i in_labels_file -o out_labels_file
           [-p config_file]
```

```
{
        files =
        [
                {
                        class = { ACCENT='New_York_City', AGE='28', GENDER='M',
                                  SPK_ID='SPK001', UTT_TYPE='BA' },
                        file_name = './BLOCK00/SES001/AP001BA01.wav',
                        phonemes =
                        [
                                {
                                        phoneme = 'F',
                                        quality = 0.768,
                                        start = 0.01,
                                        surface_form = 'FOUR',
                                        word_start = 1
                                },
                                {
                                        phoneme = 'AO1',
                                        quality = 0.722,
                                        start = 0.09
                                },
                                {
                                        phoneme = 'R',
                                        quality = 0.806,
                                        start = 0.18
                                },
                                ...
```

Слика 4. Добијене лабеле са новим границама (сегмент)

Апликација за анализу граница – поређење референтног скупа лабела (*Windows*)

Апликација учитава лабеле са новим границама и референтне лабеле, па пореди границе истих фонема у аудио датотекама које су у референтном скупу. На основу разлике у позицији границе и оцене веродостојности уписује ажурирану оцену – што је већа разлика, то се досадашња оцена више квари. Такође, апликација исписује и детаљну статистику одступања од ручно задатих граница по сваком прелазу било која два фонема, као и проценат граница које су у оквиру задатог броја милисекунди од ручно постављене (типично се ставља 5, 10, 20, 30 и 50ms, и на основу тога се могу брзо поредити резултати две обуке са различитим параметрима). Такође, користећи нове оцене граница фонема, могу се лако пронаћи и прегледати проблематични сегменти у целој бази, како у референтном скупу, тако и ван њега (тамо је остала оцена коју је израчунао *Kaldi* програм на крају обуке), користећи *SpeechLabel* алат.

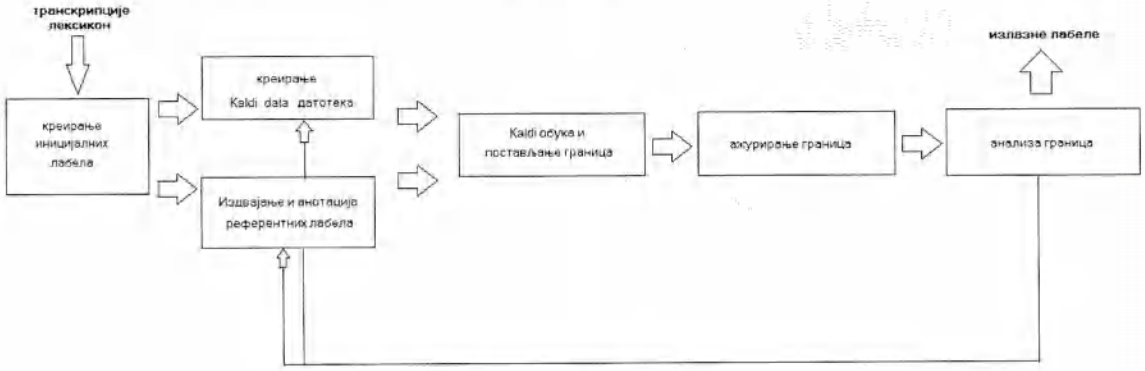
```
label_tool --kaldi-align-apps compare-boundaries
           –i aligned_labels_file;ref_labels_file -o out_labels_file
           [-p config_file]
```

Корекције и нова итерација

Ако су анализом пронађене грешке у референтном скупу (нпр. због грубе грешке постављача границе) или у остатку базе (нпр. због грешке у транскрипцији неке реченице), оне се могу исправити, а затим се може пустити наредна итерација обуке и постављања граница. Такође је могуће променити неки параметар обуке, ако се закључи да је то потребно. Овај поступак се може понављати све док анализа постављених граница не произведе задовољавајући резултат.

Како је реализовано и где се примењује, односно које су могућности примене (техничке могућности):

Техничко решење се, дакле, примењује сукцесивним позивом низа апликација у *Windows*, односно скрипта у *Linux* окружењу, из командне линије, односно терминала. За исправан рад апликација обезбеђен је низ модула и функција написаних и C++ програмском језику (у оба окружења). На приказан начин може се применити за скоро потпуно аутоматску анотација потпуно нове говорне базе, на било ком језику. Блок шема целе процедуре приказана је на слици 5.



Слика 5. Блок шема процедуре за анотирање говорне базе

Докази (прилози):

- Писано мишљење два рецензента - експерта из области техничког решења.
- Потврда предузећа АлфаНум.
- Потврда предузећа Speech Morphing System.
- Уверење о признавању техничког решења Наставно-научног већа Факултета техничких наука, на основу мишљења рецензената и приложених доказа, издато након одговарајуће процедуре на Департману за енергетику, електронику и телекомуникације.

Нови Сад, новембар 2015. године.

Подносилац пријаве

Проф. др Владо Делић
Руководилац пројекта ТР32035

УНИВЕРЗИТЕТ У НОВОМ САДУ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА
Трг Доситеја Обрадовића 6, 21000 Нови Сад, Република Србија
Деканат: 021 6350-413; 021 450-810; Централа: 021 485 2000
Рачуноводство: 021 458-220; Студентска служба: 021 6350-763
Телефакс: 021 458-133; e-mail: ftndean@uns.ac.rs
ИНТЕГРИСАНИ СИСТЕМ МЕНАЦМЕНТА СЕРТИФИКОВАН ОД:

Наш број: _____
Ваш број: _____
Датум: 2015-11-26

ИЗВОД ИЗ ЗАПИСНИКА

Наставно-научног већа Факултета техничких наука у Новом Саду, на 5. редовној седници одржаној дана 25.11.2015. године, донело је следећу одлуку:

-непотребно изостављено-

Тачка 17.2.10: У циљу верификације новог техничког решења предлажу се рецензенти:

- Проф. др Зорица Николић (Електронски Факултет, Ниш)
- Доц. др Бранислав Геразов (Факултет за електротехнику и информационе технологије, Скопље, Македонија)

СКУП МОДУЛА И ПРОЦЕДУРА ЗА ВИСОКОКВАЛИТЕТНУ АНОТАЦИЈУ ГОВОРНИХ БАЗА ПОДАТАКА

Аутори: Едвин Пакоци, Бранислав Поповић, Дарко Пекар, Никша Јаковљевић.

-непотребно изостављено-

Записник водила:
Јасмина Димић, дипл. правник

Тачност података оверава:
Секретар
Иван Нешковић, дипл. правник

Декан
Проф. др Раде Дорословачки

РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	**Скуп модула и процедура за висококвалитетну анотацију говорних база података**
Аутори техничког решења:	Едвин Пакоци, Бранислав Поповић, Дарко Пекар, Никша Јаковљевић
Реализатори:	Факултет техничких наука и АлфаНум у Новом Саду
Пројекти на којима је развијено:	"Развој дијалошких система за српски и друге јужно-словенске језике", ТР32035 код МПНТР (2011-2015)
Област на коју се односи:	Електроника, телекомуникације и информационе технологије
Корисници:	Предузеће АлфаНум у Новом Саду, од 2015. године, Предузеће Speech Morphing System, Inc. од 2015. године
Категорија техничког решења:	Нови производ примењен на међународном нивоу (М81)

Подаци о рецензенту:

Име, презиме и звање:	**Доц. др Бранислав Геразов**
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабран у звање насловни доцент 22.04.2015. год. на Факултету за електротехнику и информационе технологије, Скопље, Македонија за у.н.о. електроника
Установа где је запослен:	Факултет за електротехнику и информационе технологије, Скопље, Македонија

Стручно мишљење рецензента:

Резултат научно-истраживачког рада "Скуп модула и процедура за висококвалитетну анотацију говорних база података" представља **Нови производ примењен на међународном нивоу (М81)** у смислу Правилника о поступку и начину вредновања, и квантитативном исказивању научноистраживачких резултата истраживача од 21.03.2008.

Образложење за техничко решење (ТР):

- Техничко решење као резултат даје потпуно анотирану нову говорну базу, при чему не захтева много мануелног рада.
- Процедура описана у ТР је јасна и прецизна, а омогућава и подешавање бројних параметара ради даље оптимизације резултата.
- ТР је специфично по томе што омогућава да се жељене конвенције дефинисане ручном анотацијом веома малог дела говорне базе генерализују на целу базу. Оваква процедура даје конзистентније и квалитетније резултате од других сличних метода.
- ТР се већ активно примењује и у предузећу АлфаНум и у америчком предузећу Speech Morphing System, Inc, што је показатељ да описана процедура функционише успешно.

У Скопљу, 30.11.2015. године.

Доц. др Бранислав Геразов

РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	**Скуп модула и процедура за висококвалитетну анотацију говорних база података**
Аутори техничког решења:	Едвин Пакоци, Бранислав Поповић, Дарко Пекар, Никша Јаковљевић
Реализатори:	Факултет техничких наука и АлфаНум у Новом Саду
Пројекти на којима је развијено:	"Развој дијалошких система за српски и друге јужнословенске језике", ТР32035 код МПНТР (2011-2015)
Област на коју се односи:	Електроника, телекомуникације и информационе технологије
Корисници:	Предузеће АлфаНум у Новом Саду, од 2015. године, Предузеће Speech Morphing System, Inc. од 2015. године
Категорија техничког решења:	Нови производ примењен на међународном нивоу (М81)

Подаци о рецензенту:

Име, презиме и звање:	Проф. др Зорица Николић, редовни професор
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабрана у звање редовног професора 07.03.2000.на Електронском факултету у Нишу за у.н.о. Телекомуникације.
Установа где је запослен:	Електронски факултет, Ниш

Стручно мишљење рецензента:

Резултат научно-истраживачког рада "Скуп модула и процедура за висококвалитетну анотацију говорних база података" представља **Нови производ примењен на међународном нивоу (М81)** у смислу Правилника о поступку и начину вредновања, и квантитативном исказивању научноистраживачких резултата истраживача од 21.03.2008.

Образложење за техничко решење (ТР):

- Техничко решење омогућава готово потпуно аутоматско анотирање произољне говорне базе, на било ком језику.
- ТР се користи позивом низа готових апликација из командне линије у *Windows* окружењу, односно позивом *Shell* скрипта из терминала у *Linux* окружењу.
- ТР представља нови производ највише заснован на сопственим говорним технологијама које су развијене на пројектима МПНТР у претходном периоду. Такође се базира и на најмодернијим јавно доступним програмским алатима у научној области из које је ТР.
- ТР се примењује у предузећу АлфаНум и на Факултету техничких наука у Новом Саду, али и у америчком предузећу Speech Morphing System, Inc. Сама ова сарадња је настала као резултат квалитета ТР. Очекује се и писање више научних радова на тему резултата ТР.

У Нишу, 28.11.2015. године.

Зорица Николић

Проф. др Зорица Николић

ALFANUM
GOVORNE TEHNOLOGIJE

AlfaNum d.o.o., Novi Sad
Trg Dositeja Obradovića 6
021/4750204
PIB: 102948614
Mat.br. 08790949
Tekući račun: 335-6188-21
Registarski broj: 002890219
Šifra delatnosti: 72400

U Novom Sadu, 22.11.2015.

POTVRDA

Ovim potvrđujemo da je 01.07.2015. godine u preduzeću AlfaNum d.o.o. počelo da se koristi tehničko rešenje *Skup modula i procedura za visokokvalitetnu anotaciju govornih baza podataka*, koje je razvijeno od strane Fakulteta tehničkih nauka Novi Sad i preduzeća AlfaNum, a koristi se za potrebe pripreme i modifikacije govornih baza podataka kako za interne potrebe, tako i za potrebe saradnika. Preduzeće AlfaNum ga koristi u sklopu razvoja više proizvoda koji će promovisati govorne tehnologije za srpski jezik.

U ime firme:

Direktor
Darko Pekar, dipl. ing.

Darko Pekar, direktor

DRUŠTVO SA OGRANIČENOM ODGOVORNOŠĆU ZA PROIZVODNJU, PROMET I USLUGE ALFANUM GOVORNE TEHNOLOGIJE NOVI SAD

SOFTWARE DEVELOPMENT SERVICES AGREEMENT

THIS SOFTWARE DEVELOPMENT SERVICES AGREEMENT (the "Agreement") dated this 15th day of March, 2015

BETWEEN

Speech Morphing System, Inc of 1320 White Oaks Road, Campbell, California
(SMI or the "Customer")

- AND -

AlfaNum Ltd, Trg Dositeja Obradovića 6, Novi Sad, Serbia
Phone Number: +381 21 475 0204
(the "Alfanum or the Software Development Services Provider").

BACKGROUND:

A. SMI is of the opinion that the Alfanum has the necessary qualifications, experience and abilities to provide services to SMI.

B. Alfanum is agreeable to providing such services to SMI on the terms and conditions set out in this Agreement.

C. Alfanum has signed and is agreeable to SMI NDA and IP policies

IN CONSIDERATION OF the matters described above and of the mutual benefits and obligations set forth in this Agreement, the receipt and sufficiency of which consideration is hereby acknowledged, SMI and Alfanum (individually the "Party" and collectively the "Parties" to this Agreement) agree as follows:

Services Provided

1. SMI hereby agrees to engage Alfanum to provide SMI with services with their respective costs (the "Services" or the "Project") pursuant to the attached Statement of Work (The SoW or Attachment A).

2. The Services will also include other tasks which the Parties may agree on. Alfanum hereby agrees to provide such Services to SMI.

Term of Agreement

3. The term of this Agreement (the "Term") will begin on the date of this Agreement and will remain in full force and effect until the completion of the Services, subject to earlier termination as provided in this Agreement. The Term and scope of this Agreement may

Speech Morphing System, Inc. (Customer)

Per: _____
Meri Friedlander
Title: EVP Product, Operations & BD

AlfaNum Ltd (Software Development Services Provider)

Per: _____
Darko Pekar
Title: CEO

Statement of Work

1. Familiarize with existing source-filter separation tool and customize it for our purposes
This tool is made for Linux, and is ready as command line application. Since AlfaNum's preferred environment is Visual Studio and Windows, AlfaNum will try to use it via Cygwin. AlfaNum will try to use it as-is, without trying to change anything in the code (it would take significant time to get familiar with the source). If needed, AlfaNum will make some additional tool, which will convert its input-output to something more convenient for AlfaNum's other tools.

Milestone 1.1: Set of tools which will enable AlfaNum to extract features for training phase, and to synthesize voice when appropriate files are provided. Getting familiar with this tool, and writing additional documentation is included.
Required time (for M1.1): 2 weeks
Estimated finish date: T+2w
Price: [redacted]

2a. Implement training algorithm for hybrid synthesis
This step will include customization of open source toolkit so it can work with filter parameters obtained from tool described in 1 ("SFS" tool), instead of MFCC. It should also preserve source parts of the training samples, since they will be used during synthesis. As an input it will take, besides output from SFS tool for each recording, phone boundaries, text with pronunciation and prosodic tags (if provided). After training, its output will be acoustic model (AM) of speaker A. AM will comprise of several thousand of states, each consisting of:

- filter features, as well as their time derivatives (first and second order)
- prosodic features (duration, pitch and power)
- source parts for several pitch values - in this phase algorithm will probably preserve arbitrary representative of source part (for specific pitch), since there will be many. In later phases, algorithm could try to find "centroid" of these source parts, which will be more appropriate representative. It may turn out that this step is not necessary, although, in the worst case scenario, it may turn out that differences between source parts of one state are so big that even this procedure won't provide satisfactory results. According to information available so far, and conducted experiments, this should not be the case.
- context information, i.e. in which context it should be used. By "context" it is meant surrounding phones, but also some wider information: proximity to other phones, prosodic tags, position in syllable, word, sentence, etc.

2b. Implement hybrid TTS back-end
Implementation of customized HMM TTS, which uses source parts from the original signal, instead of generic excitation (pulse train + white noise). It will do the following:
- Input will be phonetized text with (optional) prosodic tags. Prosodic tags will be handcrafted, until the appropriate front-end of TTS is developed.
- For each phoneme in sentence it will generate the sequence of states, which will be the basis for the calculation of:
  - Filter parameters. Will be calculated based on static and dynamic values of these parameters for this and surrounding states.
  - Prosodic parameters (pitch, duration and power).
  - Source signal for this part of phone.
- Provided with source signal, pitch and duration for each state, PSOLA (or other appropriate algorithm) will generate source signal on the sentence level.
- Filter parameters will be applied to generated source signal, by using SFS tool from 1.

For the first milestone AlfaNum will try to reduce the amount of work, so the first results could be visible ASAP. It will be decided in the process. It will also work with only one speaker, not just for the first milestone, but for the whole phase.

Milestone 2.1: First working version of the training algorithm is finished. Training of the first voice model is finished.
First working version of the synthesizing algorithm is finished. Some features may not be implemented or optimized, but it should be audible that generated voice is natural enough and sounds like the original.
Required time: 7 weeks
Estimated finish date: T+12w
Price: [redacted]

Milestone 2.2: Optimized version is finished. Training of the final voice model is finished. By "optimized" it does not mean that it will be fully optimized, market-ready, version, but it will be sufficiently comfortable to work with, so it can proceeded to morphing. By "final voice model" it also don't mean that there won't be any room for improvement, especially in source-filter separation part.
Required time: 17 weeks
Estimated finish date: T+22w
Price: [redacted]

3. Perform phonetic alignment on one database
It will be done (semi)automatically, by using ASR tools. Some manual corrections might be required. If successful, AlfaNum will negotiate transferring this tool and process to SMI.

Required time: 2 weeks
Estimated finish date: T+2w
Price: [redacted]

4. Prosodically annotate several hours of one talent voice
It will be done by AlfaNum's semi-automatic annotation tool and process. This tool is not part of the offer, just this annotation. If successful, AlfaNum will negotiate transferring this tool and process to SMI.

Required time: 4 weeks
Estimated finish date: T+6w
Price: [redacted]

5a. Produce prosody, given the tagged text - standalone tools for testing purposes
In the training phase, input to this module will be prosodically annotated speech database (obtained in task 4). It will build Classification And Regression Trees (CARTs), which are actually "models" of exact prosody, or the means of converting tagged text into prosody (pitch, duration, and optional energy).
Actually, there will be two modules, one for extraction of acoustic features (pitch, duration, energy) from the database in appropriate format, and the mentioned one for building CARTs.
In the usage (or test) phase, input will be file with tagged text and pronunciation, and output will be file with the prosody for that text. Each phone in the text will have its duration, and pitch if applicable (for vowels). Pitch can be given for two points in a vowel (e.g. 1/4 and 3/4), or as required. Exact format of output file can be defined later.

Although SMI will be able to train CARTs on its own, AlfaNum will do that as well, so SMI can just use them as input in the usage phase.

Milestone 5a.1: Set of standalone tools for acoustic features extraction, building CARTs and prosody prediction. Documentation included.
Required time: 3 weeks (includes training of initial CARTs to be at SMI's disposal)
Estimated finish date: T+9w (since it is dependent on task 4)
Price: [redacted]

5b. (optional) Produce prosody, given the tagged text - libraries and APIs with licenses for unlimited use
Same as 5a, but in the form of defined APIs and source code. SMI will have the right of unlimited commercial use.

Milestone 5b.1: Set of APIs and source code for acoustic features extraction, building CARTs and prosody prediction. Documentation included.
Required time: 5 weeks
Estimated finish date: 5w after start (currently unknown)
Price: [redacted] + [redacted] yearly maintenance (6 months warrantee)

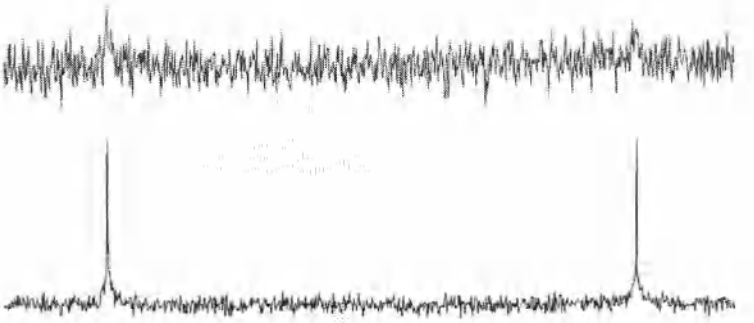
6. Development of source-filter separation tool
The result should be C/C++ library and standalone tool for source-filter separation and combination (synthesis). It should be based on some of the state-of-the-art techniques from literature. In this phase, some minor issues and spots for improvement could remain.
In analysis phase, the tool should output glottal source of the signal, and corresponding filter parameters. In synthesis phase, the tool should be able to combine glottal source and filter parameters in order to produce the original voice.
As a standalone tool, it will be available from command line, with appropriate options. As a library it will implement appropriate APIs (interface), which will be defined.
Since this task demands initial research and trial phase, it will be divided into two subtasks:
- Exploration of state-of-the-art and deciding on the method
- Implementation of chosen method

Milestone 6.1: Presentation about SFS approaches, with reference to the code availability and patent protection. Recommendation based on exploration and initial trials.
Required time: 2 weeks
Estimated finish date: 2w after start
Price: [redacted]

Milestone 6.2: Developed and tested SFS library and standalone tool. In absence of full TTS, tests will be based on its ability to reconstruct the original signal, as well as visualizations of spectral envelope and source.
Required time: 4 weeks
Estimated finish date: 4w after start
Price: [redacted]

6.a Additional costs and improved SFS tool
Due to unexpected complications in R&D process mostly related to TD-PSOLA issues (which is actually not a part of these tasks), there was more than 1.5 person-month of additional work to improve the prosody quality to an acceptable level.
Besides having additional work related to agreed version of SFS tool (task 6), AlfaNum will provide new SFS tool which generates source which should be more robust to prosody changes induced by

TD-PSOLA. The tool will achieve this by synchronizing phases of all harmonics in input signal. Visually, source signal should look more like the red signal on the figure below, as opposed to the green one:



AlfaNum cannot guarantee that this will resolve all the problems regarding changes in prosody (it will not), but numerous experiments and spectrum analysis in the past showed that it should bring significant improvement.

Price: [redacted]

6.b (Optional) Further improvements of prosody changing algorithm
Currently, prosody (pitch, duration and power) changes in source signal during alignment process (and later during synthesis), are done by using TD-PSOLA algorithm. There are numerous, more advanced, approaches to this problem (e.g. HNM, STRAIGHT), which can provide more natural output, even when changes in prosody are substantial. AlfaNum will further investigate state-of-the-art in this area and propose optimal solution.

Since this task demands initial research and trial phase, it will be divided into two subtasks:
- Exploration of state-of-the-art and deciding on the method
- Implementation of chosen method

Milestone 6.b.1: Presentation about prosody changing approaches, with reference to the code availability and patent protection. Recommendation based on exploration and initial trials.
Required time: 3 weeks
Estimated finish date: 3w after start
Price: [redacted]

Milestone 6.b.2: Developed and tested library and standalone tool. Tests will be based on comparison of this method with TD-PSOLA, both objective and subjective.
Required time: 8-12 weeks (depends on 6.b.1)
Price: depends on 6.b.1

7. Adapting one sequence to another, pitch marking and pitch marking mapping
The task is to take filter parameters from one sequence (speaker A) and apply them to the source of the speaker B, by using the prosody of the speaker C (which also could be A or B). All speakers must have uttered exactly the same sentence (same labels, including pauses, otherwise results will be suboptimal).
Inputs for the tool are:
- Filter parameters and label file of speaker A.
- Source and label file of speaker B.
- Wav and label file of speaker C.

Note: label file will have the same phonetic content, so only the boundaries will be different. Hence, the process of generating these files should not take too long, even if done manually.
Labels in SpeechLabel format will be required. Those are easy to make even from scratch, and if you already have them for one file, then it's really easy to make them for another file with the same content.
Output are resulting (aligned) source and filter which are then fed to SFS tool. It will generate the resulting signal. All this will be a part of one batch process.

Price: [redacted]
Time required: 2w

8. SpeechLabel tool
The tool has the following features (among others):
- Accepts input in two label formats.
- Accepts specific AM
- Support for numerous audio formats.
- Enables visualization of:
    - Speech signal.
    - Spectrum.
    - Pitch.
    - Pitch marks.
    - Power.
    - Phone data (phone, LOC, comment, attribute).
    - Word data (stress, type of word...).
- Intuitive navigation:
    - Moving through phonemes and words.
    - Zoom in/out, zoom to selection/all/phone.
    - Control by mouse, using both buttons and net-scroll.
    - Playing signal in various ways.
- Editing options:
    - Edit any phone, word, attribute or comment.
    - Change position of phones (could be restricted not to go outside adjacent phones).
    - Undo changes.
    - Auto save.
- Prosodic annotation:
    - Included support for standard ToBI tags.
    - Included TTS which synthesizes sequence (or its part) in accordance to current ToBI tags, so the agent can compare that to recorded audio. Very helpful feature for prosodic annotation and control. For this to work, we need to have one TTS already working (even poorly).
- Search options:
    - Search by filename.
    - Search by phone or sequence of phones.
    - Search by some attribute.
    - Search by quality (LOC or something else).
    - Advanced search options.
- Reviewing mode:
    - Starts by setting search options and selecting pop-up result.
    - Search results keep popping up, until reviewing mode is turned off.
- Comment features include:

- Adding comment to any phone.
- Search through commented phones, by numerous criteria including reviewing date or user.
- Editing comments, which virtually enables chat on that phone.

Requested additions:
- Support for JSON format.
- Support for wildcards in search.
- Visualization of all phones that were altered during editing of one sequence.
- Comparison of two label files by generating comments which describe those changes.

Software is intended only for changing and reviewing of label files. It is not intended for any kind of signal and speech processing (although it performs some signal operations internally).
Software is for Windows platform only.
Source code is not included.

Pricing:
- [redacted] for 5 licenses
- [redacted] for additional 5 licenses
- [redacted] for unlimited number of licenses.

Annual maintenance and support after 12 months of warranty: [redacted] (regardless # of users)
Additional significant changes in this software: upon request

Delivery: initial version already delivered. Delivery of the version above with requested additions for acceptance tests by 7th of September.

Optional: Source code price: [redacted]

9. Phonetic labeling service

First database price (not including task 3): [redacted]
Every other database: [redacted]
Additional cost per corrected phoneme: [redacted] - this is where we will lose most of the time, if the database is dirty. If there are only few hundred corrections, then this additional cost will be symbolic. If there are few thousand, then not so symbolic, but in that case our manual effort will also be substantial.

Output:
- Fully labeled database ready to use for voice build
- Language and voice optimized AM (pending conclusions of 13.a)

Time required: 2-4w, depending on the database.

9.a Set of tools for AM building
This set of tools and scripts will accept as input:
- Speech database
- Labels (identity and position of each phone + optional LOC)

Output will be model file which can be used in 12.
Source code and all the scripts will be provided.

Notes:
- Output will be ASR-like AM model, in terms that it can be used for alignment, but not for TTS

ALFANUM
GOVORNE TEHNOLOGIJE

Trg Dositeja Obradovića 6
Novi Sad, Srbija
+381(0)21 4750204
Reg.No. 08790949
IBAN: RS35340000001100269766
SWIFT:GIBARS22
SWIFT CODE FOR CORRESPONDENT BANK:GIBAATWG

Invoice No. 083/15

Buyer

Company	Speech Morphing System, Inc
Address	White Oaks Road
City	Campbell
State	California

Invoice Date:	24.sept.15
Town:	Novi Sad

Art. No	Description	QTY	Unit	Unit price($)	Total ($)
1	Software Development Services (Agreement from 15.03.2015	1,00	pcs.	[redacted]	[redacted]
2	Advance payment from 3.4.2015. (Advance invoice 001/15)	1,00	pcs.	[redacted]	[redacted]

Total [redacted]

Slovima: [redacted]

Service intended for export. VAT not chargeable according to the article 12 of the Law, section 3, paragraph 4, subparagraph 7

DRUŠTVO SA OGRANIČENOM ODGOVORNOŠĆU ALFANUM GOVORNE TEHNOLOGIJE ZA PROIZVODNJU, PROMET I USLUGE NOVI SAD

УНИВЕРЗИТЕТ У НОВОМ САДУ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА

Трг Доситеја Обрадовића 6, 21000 Нови Сад, Република Србија
Деканат: 021 6350-413; 021 450-810; Централа: 021 485 2000
Рачуноводство: 021 458-220; Студентска служба: 021 6350-763
Телефакс: 021 458-133; e-mail: ftndean@uns.ac.rs

ИНТЕГРИСАНИ СИСТЕМ МЕНАЏМЕНТА СЕРТИФИКОВАН ОД:

Наш број: 01.сл
Ваш број:
Датум: 2015-12-25

ИЗВОД ИЗ ЗАПИСНИКА

Наставно-научно веће Факултета техничких наука у Новом Саду, на 6. редовној седници одржаној дана 23.12.2015. године, донело је следећу одлуку:

-непотребно изостављено-

ТАЧКА 24. Питања научноистраживачког рада и међународне сарадње

24.2.13. На основу позитивног извештаја рецензената верификује се техничко решење (М 81) под називом:

СКУП МОДУЛА И ПРОЦЕДУРА ЗА ВИСОКОКВАЛИТЕТНУ АНОТАЦИЈУ ГОВОРНИХ БАЗА ПОДАТАКА

Аутори: Едвин Пакоци, Бранислав Поповић, Дарко Пекар, Никша Јаковљевић.

-непотребно изостављено-

Записник водила:
Јасмина Димић, дипл. правник

Тачност података оверава:
Секретар
Иван Нешковић, дипл. правник

Декан
Проф. др Раде Дорословачки