Образац за пријаву техничког решења[1]

Техничко решење: *SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база података	
Аутори техничког решења:	Драгиша Мишковић, Дарко Пекар, Милан Сечујски, Едвин Пакоци
Назив техничког решења:	*SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база података
Подтип техничког решења:	Нови производ на међународном нивоу (М81)
Реализатори:	Факултет техничких наука и *АлфаНум д.о.о.* из Новог Сада

Пројекат и период реализације техничког решења:

- Развој дијалошких система за српски и друге јужнословенске језике“ (пројекат финансиран од стране Министарства просвете, науке и технолошког развоја Републике Србије, у временском периоду 2011–2015. године, под евиденционим бројем ТР32035).

За кога је техничко решење рађено:

Техничко решење првенствено је рађено за потребе предузећа *АлфаНум д.о.о* из Новог Сада, са подршком за српски језик, да би након тога било додатно усавршено и прилагођено потребама предузећа *Speech Morphing Systems* из Кембела у Калифорнији, Сједињене Америчке Државе, увођењем подршке за енглески језик. Техничко решење коришћено је и у развоју синтезе говора за хебрејски језик у сарадњи са израелским предузећем *Aharon Speech Technologies*.

Ко техничко решење користи:

Техничко решење тренутно користи предузеће *Speech Morphing Systems*, које се бави развојем система говорних технологија (првенствено система за конверзију идентитета говорника) на више језика, а користи га и предузеће *АлфаНум д.о.о*. из Новог Сада у развоју различитих говорних технологија, почев од синтезе говора на српском, хрватском, хебрејском и енглеском језику, до препознавања говора и препознавања говорника.

[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).

Како су резултати верификовани (од стране ког тела):

1) Техничко решење базирано је на више техничких и развојних решења која су пријављивана као резултати на пројектима технолошког развоја МНТР у периоду 2005-2010. године и прво је такво техничко решење у ширем региону.
2) Писано мишљење два рецензента-експерта из области техничког решења.
3) Техничко решење верификује Наставно-научно веће Факултета техничких наука на основу позитивног мишљења два рецензента-експерта из области техничког решења. НН веће ФТН је 25.11.2015. именовало рецензенте:
    a. Др Снежана Гудурић, редовни професор на Филозофском факултету у Новом Саду
    b. Др Слободан Јовичић, редовни професор, ЕТФ у Београду и Центар за унапређење животних активности у Београду

На који начин се користи (кратак опис):

Предложено техничко решење представља апликацију за анотацију говорних база података. Основне функционалности техничког решења укључују временску сегментацију снимака, анотацију аудио-материјала у смислу уношења различитих лингвистичких информација, претрагу база према различитим критеријумима, као и манипулацију пратећим фајловима у одговарајућем интерно дефинисаном формату (*sobj*). Улаз апликације представљају звучни фајлови, односно снимци појединачних реченица изговорених од стране једног или више говорника. У току коришћења апликације овим снимцима додају се информације које се смештају у одговарајуће пратеће фајлове у интерно дефинисаном *sobj* формату, да би на крају говорна база заједно са тако припремљеним информацијама могла да се користи у различитим сегментима поступка развоја система говорних технологија – првенствено аутоматске синтезе говора на основу текста и аутоматског препознавања говора. Апликација тренутно подржава рад на српском, хрватском, енглеском и хебрејском језику.

Опис техничког решења: *SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база података

Област на коју се техничко решење односи:

Техничко решење припада области информационих, односно информационо-комуникационих технологија (ICT).

Проблем који се техничким решењем решава:

Техничко решење значајно доприноси решавању проблема развоја система говорних технологија, пре свега аутоматске синтезе говора на основу текста и аутоматског препознавања говора, али и система за препознавање говорника или препознавање емоција у говору. Развој свих ових система у великој мери се ослања на говорне базе

које морају бити анотиране у погледу различитих информација. У зависности од врсте система који се развија, може се захтевати фонетска, лексичка, и/или прозодијска анотација, и у сва три случаја реч је о комплексном проблему који најчешће захтева концентрисани напор стручних лингвиста којима је потребно обезбедити што ефикасније окружење за рад са говорним сигналима, како би анотација говорне базе података могла бити обављена што брже. Примера ради, за снимање говорне базе за синтезу говора која обухвата неколико сати говорног материјала довољно је потрошити један или два дана на студијско снимање, али је за квалитетну анотацију такве базе у погледу фонетских, лексичких и прозодијских информација некада потребно и више месеци. У овоме знатно помажу полуаутоматски и аутоматски системи за анотацију, али је ради постизања високе тачности потребно ангажовање човека, који поново мора извршити проверу тачности аутоматске анотације, те за то мора имати ефикасан алат.

Стање решености тог проблема у свету:

Интензиван развој говорних технологија диктира потребу за креирањем говорних корпуса намењених обуци и тестирању система. Нажалост, потреба за овим типом података (говорних база) није праћена и развојем стандардног скупа алата за њихово креирање, анотацију и испитивање. Додатни проблем представља и изостанак стандарда у погледу формата у ком се смештају подаци након обраде говорне базе. Поједина решења се ослањају на примену формата XML (енг. *extended markup language*) због широке подршке и могућности обраде са разним алатима. Такође, постоје и приступи код којих се алати за обраду говорних база интегришу са релационим базама података.

У наставку ће бити приказано неколико најраспрострањенијих софтверских алата који се могу користити за анотацију говорних база, и биће указано на њихове специфичне недостатке у контексту развоја система говорних технологија, због чега многе лабораторије и даље развијају сопствена решења.

Софтверски алат *Praat*[2], иако веома широко распрострањен, представља окружење првенствено намењено лингвистичким истраживањима, и користи интерни формат за смештање пратећих информација о снимцима, односно, њихових транскрипција. Овај алат, и поред изузетне ефикасности и флексибилности у обради говора, није специфично прилагођен анотацији говора у контексту развоја говорних технологија и не пружа одређене додатне могућности које су од посебног значаја за то.

[2] http://www.fon.hum.uva.nl/praat/

Софтвер *Transcriber*[3] представља јавно доступно окружење намењено *Linux* платформи. Нуди основне функционалности по питању сегментације и анотације говорних корпуса са нагласком на подршку за дугачке аудио фајлове. Развијен је у виду клијент-сервер архитектуре која подржава истовремени рад више клијената над централизованом базом. За смештање резултата обраде користи се SGML формат (енг. *standard generalized markup language*). Међутим, за разлику од предложеног решења, кориснички интерфејс *Transcriber*-а раздваја визуелни приказ аудио-сигнала и његове транскрипције, што отежава транскрипцију садржаја. Наиме, иако је овај приступ оправдан у случају дугачких аудио фајлова, при анотацији кратких може представљати проблем. Такође, ни овај софтвер не подржава многе функционалности уграђене у *SpeechLabel*.

SAPPHIRE[4] представља окружење за анализу и препознавање говора развијено на Масачусетском технолошком институту (*Massachusetts Institute of Technology – MIT*). Садржи графички интерфејс заснован на Tcl/Tk скрипт језику. Међутим, овај софтверски алат, као и многи други, није јавно доступан.

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

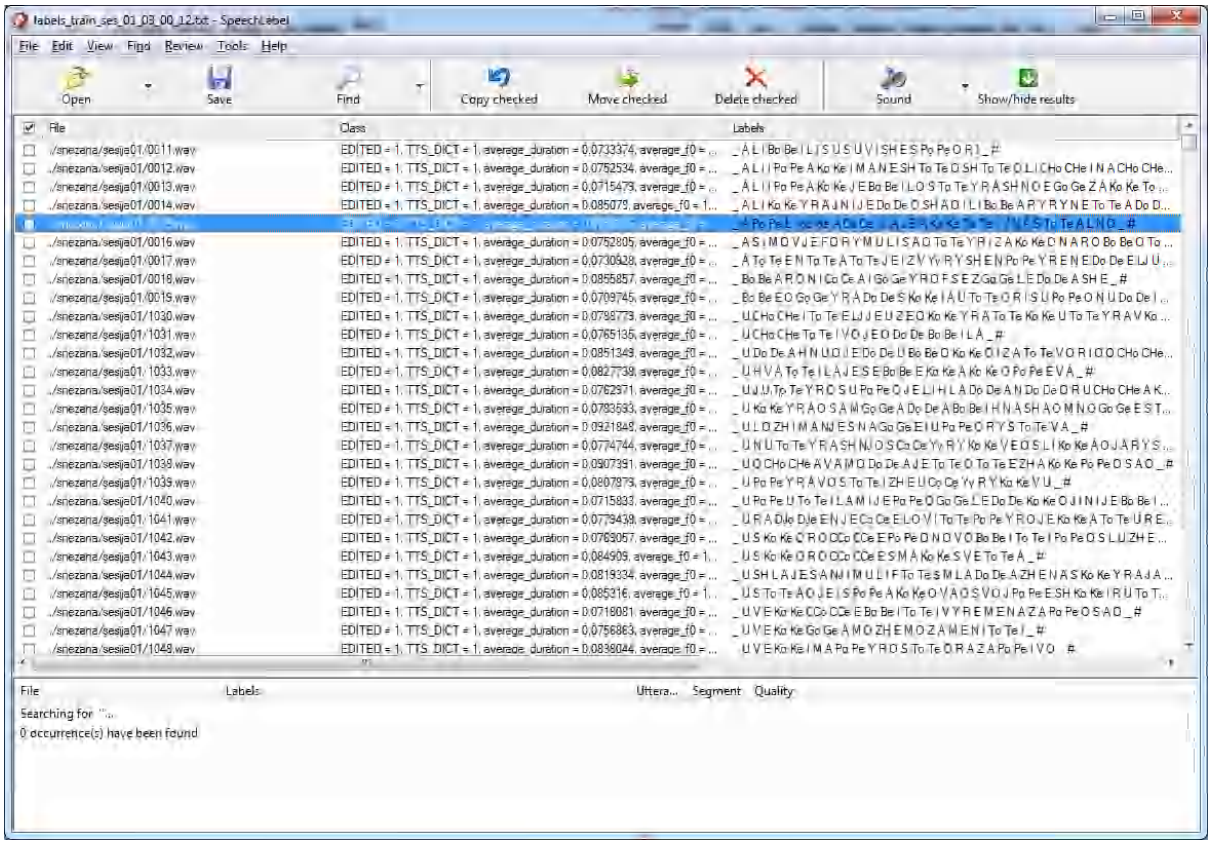
Основне функционалности предложеног техничког решења укључују временску сегментацију, анотацију аудио материјала у смислу уношења различитих лингвистичких информација, претрагу база према различитим критеријумима итд.

Резултат обраде се форматира и смешта у пратећи фајл, који се касније користи у развоју или експлоатацији система говорних технологија. Његова структура омогућава смештање лингвистичких информација за произвољан број појединачних реченица или других говорних целина репрезентованих аудио-фајловима. При томе, обрада у оквиру апликације *SpeechLabel* подразумева креирање или модификовање пратећих информација док аудио-фајлови остају неизмењени.

Апликација се састоји од графичког интерфејса, подсистема за визуелизацију параметара релевантних за аудио сигнал и пратећег лингвистичког садржаја, те рутина за обраду сигнала потребних за екстракцију акустичких обележја за потребе визуелизације. Кориснички интерфејс је реализован коришћењем окружења MFC (*Microsoft foundation classes*), а на слици 1 је приказан основни прозор апликације након учитавања пратећег фајла који се односи на одређени скуп реченица.

[3] http://transag.sourceforge.net

[4] http://www.asel.udel.edu/icslp/cdrom/vol3/222/a222.pdf



Слика 1: Изглед главног прозора апликације *SpeechLabel*. Свака од појединачних ставки у листи односи се на одређену реченицу, односно, снимак, и садржи информације о њиховом фонемском садржају, као и друге допунске информације.

Као што се може видети, листа у оквиру главног прозора садржи само грубе информације у виду идентификације аудио фајла, одређених пратећих ознака и фонемског садржаја реченице, односно, говорне целине. На овом нивоу, поред основних активности над појединачним фајловима (брисање, копирање, премештање, уношење ознака које се односе на читав снимак), апликација нуди и могућност претраге према различитим критеријумима, преслушавање појединачних снимака, постављање општих параметара окружења итд.

Избором одговарајућег снимка из листе отвара се додатни прозор за визуелизацију и приказ детаљних информација о њему (слика 2). При томе, апликација стартује рутине за обраду аудио сигнала реализоване у оквиру пратеће библиотеке *slib*[5]. Ова библиотека, поред стандардних метода у обради (филтрирање, прозорирање, Фуријеова трансформација, прилагођење учестаности одабирања итд.) нуди и рутине за естимацију енергије и основне учестаности, детекцију степена звучности и осталих параметара релевантних за говорни сигнал. Захваљујући томе, обезбеђен је приказ

[5] D. Pekar, R. Obradović: „*C++ Library for Digital Signal Processing – slib*", in Proc. TELFOR 2001.

различитих секција у оквиру прозора за визуелизацију, означених бројевима на слици 2 са десне стране.

1. *Приказ таласног облика сигнала и припадајућих лексичких информација*

   Ово је основни садржај прозора намењен уношењу лексичких информација за селектовану реченицу, односно, говорну целину. Приказ таласног облика уз могућност репродукције појединачних делова омогућава кориснику да специфицира границе између појединих гласова. Поред тога, реализовани интерфејс омогућава и унос осталих информација које се могу поделити у две групе:

   a) Лексички садржај на нивоу појединачних речи. У ову групу спадају информације о основном облику и врсти речи, положају и типу акцента, као и специфичне прозодијске информације које се односе на реченични нагласак, деакцентуацију појединих речи и поделу реченице на синтаксно-прозодијске целине.

   b) Информације о појединачном гласу. Поред постављања временских граница гласа у снимку (места почетка и завршетка његовог трајања), омогућен је и унос осталих информација, као што су атрибути, који могу репрезентовати факторе као што су специфичан начин фонације или артикулације гласа (нпр. глас изговорен са израженим *vocal fry* ефектом, назализован глас...).

2. *Приказ спектрограма*

   Приликом анализе аудио садржаја појединачних снимака, односно, реченица, врши се и естимација амплитудског DFT спектра (енг. *discrete Fourier transform*). Визуелизација спектра (спектрограма) обезбеђује додатне информације током обраде говорне базе. Тиме се олакшава тачно постављање појединих ознака (нпр. увидом у спектрограм се могу јасно детектовати промене у структури форманата што олакшава тачније одређивање граница између гласова код којих временски изглед сигнала не пружа довољно ингормација за то).

3. *Приказ естимираних вредности основне учестаности и степена звучности*

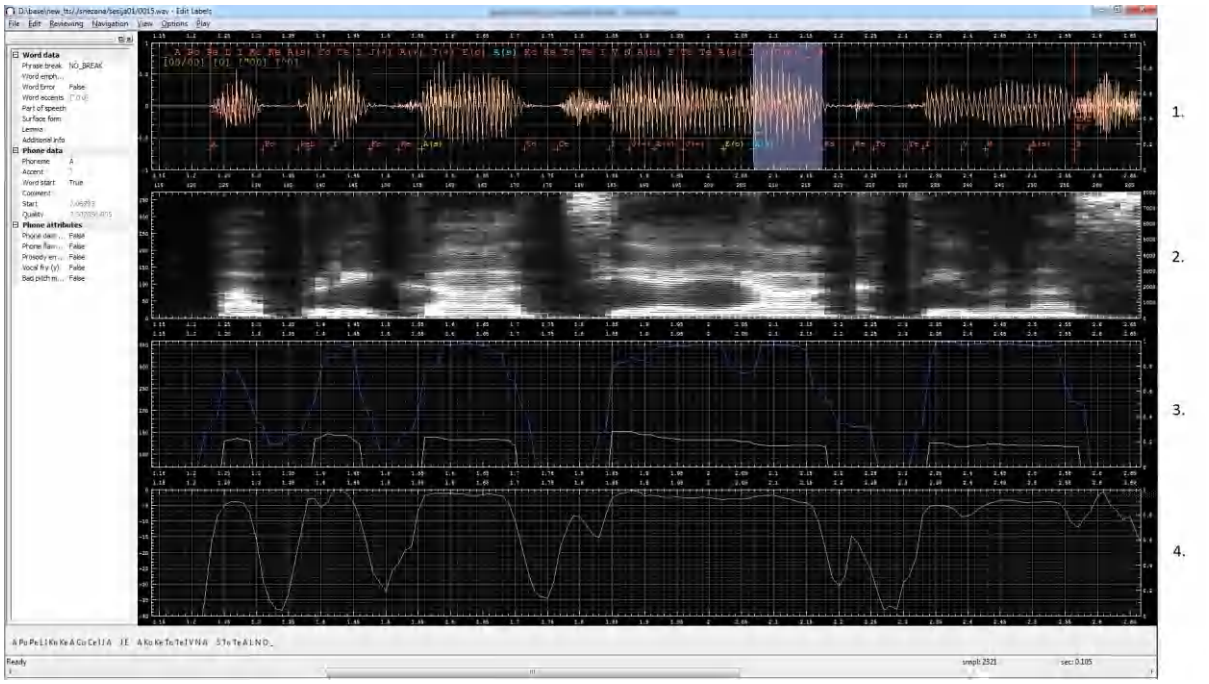
   Након процене спектра, анализом аудио-садржаја врши се и естимација степена звучности и основне учестаности за одређени говорни сегмент. Екстракција ових параметара се базира на AMDF (енг. *average magnitude difference function*) методи, која се своди на поређење сегмента сигнала са суседним сегментом помереним за одређени број одбирака. За разлику од основне верзије алгоритма, естимација основне учестаности се не ослања само на локалну информацију, већ се за сваки временски квант постави неколико хипотеза, па се потом Витербијевим алгоритмом добија оптимална f_0 крива на

неком сегменту. Визуелизација ових параметара је значајна за исправну идентификацију прозодијских обележја говорног сегмента.

Поред обележја основне учестаности које екстрахују рутине у оквиру *SpeechLabel*-а, могуће је импортовати и информације о кретању основне учестаности (f_0) из екстерног (тзв. pitch-marker) фајла, чиме се омогућује коришћење и других система и метода за одређивање f_0, односно, корисник није ограничен на коришћење метода уграђених у *SpeechLabel*.

4. *Приказ енергије аудио сигнала*

   Поред естимације спектра, анализом аудио-сигнала естимира се и његова енергија, односно, њена промена у времену, те се и она може визуелизовати у оквиру прозора за визуелизацију појединачних снимака.



Слика 2: Изглед прозора за визуелизацију појединачних снимака. Поред фонемског садржаја и временске сегментације, у оквиру приказа се визуелизују и аудио сигнал (1), спектар (2), естимирана основна учестаност (3), и енергија (4).

Поред приказа наведених секција, прозор за рад на појединачном снимку садржи и додатне елементе који олакшавају рад и навигацију унутар аудио-сигнала.

Поред могућности визуализације и претраге фајлова и пратећих информација о њима, *SpeechLabel* омогућује и ефикасну комуникацију између различитих анотатора говорне базе. Сваки анотатор може доделити одређени коментар одређеном гласу у бази, и тај коментар ће садржати и идентификацију даваоца коментара. Додељени коментар

биће смештен у пратећи фајл заједно са одговарајућом фонетском, лексичком и прозодијском информацијом, и биће видљив осталим анотаторима који касније буду радили са истим пратећим фајлом. Ефикасност овог механизма је и у томе што омогућује „преписку“ између анотатора, односно, омогућује им да размене мишљења о појединим спорним ситуацијама. Сваки спорни фонетски сегмент поменут у таквој преписци може бити идентификован јединственим идентификатором, који садржи назив снимка и временске ознаке почетка и краја сегмента. На основу оваквог идентификатора могуће је извршити и претрагу по говорној бази, како би се могло упоредити више фонетских сегмената који се у бази налазе на различитим местима, а имају сличне фонетске или прозодијске карактеристике. Корист од ефикасне комуникације између анотатора постаје јасна нарочито ако се узме у обзир варијабилност прозодијске, па чак и фонетске анотације. Наиме, при анотацији је потребно уложити изузетан напор како би се ускладили критеријуми анотације између више анотатора, и чак и тада мора се очекивати одређени степен неслагања између појединих анотатора. Просечан ниво слагања између различитих анотатора управо представља једно од репрезентативних обележја појединих система за прозодијску анотацију. Ако се анотаторима обезбеди ефикасан механизам комуникације, може се очекивати и јасније усклађивање критеријума анотације, а самим тим и већа доследност у анотацији.

Поред поменутих могућности, *SpeechLabel* подржава и интеграцију система за синтезу говора на основу постојеће фонетске, лексичке и прозодијске транскрипције. Другим речима, не само да је могуће преслушати део звучног снимка онакав какав он јесте у говорној бази, већ је могуће преслушати и синтетизовани говор, онакав какав би звучао када би имао конкретне вредности додељених фонетских, лексичких и прозодијских обележја. Примера ради, уколико анотатор није сигуран у то да ли одређеном вокалу треба доделити краткоузлазни или дугоузлазни акценат, или да ли је одређени гранични тон висок или низак, он може преслушати говор синтетизован у обе варијанте, што му може помоћи у одлучивању. Синтеза говора на основу његове транскрипције ослања се на модул за предикцију акустичких обележја говора помоћу регресионих стабала, на основу конвенција за транскрипцију које одговарају језику о коме је реч.

Како је решење реализовано и где се примењује, односно које су техничке могућности његове примене:

Техничко решење реализовано је као софтверска апликација, која може функционисати на било ком рачунару опште намене под оперативним системом *Windows*, чак и на рачунарима скромнијих карактеристика. Решење је првенствено намењено анотацији говорних снимака, односно, креирању и обради пратећих фајлова који садрже податке о фонетском, лексичком и прозодијском садржају говорних снимака. Решење је првенствено намењено развоју система говорних технологија као што су

синтетизатори говора на основу текста или аутоматски препознавачи говора, али се може користити и за друге примене у области говорних технологија, као и лингвистичких истраживања.

Докази (прилози):

- Писано мишљење два рецензента-експерта из области техничког решења
- Документа која доказују да АлфаНум и *Speech Morphing Systems* користе ово техничко решење

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 МЕНАЏМЕНТА 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 17.2.8: У циљу верификације новог техничког решења предлажу се рецензенти:

- Проф. др Снежана Гудурић (Филозофски факултет, Нови Сад)
- Др Слободан Јовичић (Центар за унапређење животних активности, Београд)

SPEECHLABEL – АЛАТ ЗА ФОНЕТСКО И ПРОЗОДИЈСКО ЛАБЕЛИРАЊЕ ГОВОРНИХ БАЗА ПОДАТАКА

Аутори: Драгиша Мишковић, Дарко Пекар, Милан Сечујски, Едвин Пакоци.

-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:	*SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база података
Аутори техничког решења:	Драгиша Мишковић, Дарко Пекар, Милан Сечујски, Едвин Пакоци
Реализатори:	ФТН и *АлфаНум д.о.о.* из Новог Сада
Подтип техничког решења:	Нови производ на међународном нивоу (М81)

Подаци о рецензенту:

Име, презиме и звање:	др Снежана Гудурић, редовни професор
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабрана 21. јули 2009. у звање редовног професора на Филозофском факултету у Новом Саду, за ужу научну област Романистика
Установа где је запослен:	Филозофски факултет у Новом Саду

Стручно мишљење рецензента:

„ *SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база података“ представља **нови производ на међународном нивоу (М81)** у смислу Правилника о поступку и начину вредновања, и квантитативном исказивању научноистраживачких резултата истраживача од 21.03.2008. Образложење:

- Предложено решење служи за придруживање фонетске, лексичке и прозодијске транскрипције реченицама у саставу говорне базе података. Овај поступак значајан је како у лингвистичким истраживањима, тако и у развоју система говорних технологија.
- Широку применљивост предложеног техничког решења потврђује постојање подршке за неколико међусобно веома различитих језика, међу којима су српски, хрватски, енглески и хебрејски језик. Увођење подршке за друге језике захтева само дефинисање стандарда одговарајућих нивоа транскрипције.
- Решење је развијено без коришћења туђе патентне и лиценцне документације, претежно на сопственој опреми. Као резултат примене решења омогућена је ефикасна анотација говорних база података, и значајно је олакшан развој система говорних технологија на више језика.

У Новом Саду, 30.11.2015. године.

Проф. др Снежана Гудурић

РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	*SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база података
Аутори техничког решења:	Драгиша Мишковић, Дарко Пекар, Милан Сечујски, Едвин Пакоци
Реализатори:	ФТН и *АлфаНум д.о.о*. из Новог Сада
Подтип техничког решења:	Нови производ на међународном нивоу (М81)

Подаци о рецензенту:

Име, презиме и звање:	др Слободан Јовичић, редовни професор
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабран у звање редовног професора 01.07.2011. на Електротехничком факултету у Београду за у.н.о. Говорна комуникација
Установа где је запослен:	Центар за унапређење животних активности, Београд

Стручно мишљење рецензента:

„ *SpeechLabel* – алат за фонетско и прозодијско лабелирање говорних база података“ представља **нови производ на међународном нивоу (М81)** у смислу Правилника о поступку и начину вредновања, и квантитативном исказивању научноистраживачких резултата истраживача од 21.03.2008. Образложење:

- Техничко решење представља апликацију помоћу које се скупу говорних снимака ручно може придружити фонетска, лексичка и/или прозодијска транскрипција, или се може кориговати транскрипција додељена од стране аутоматског система. На овај начин се уједно постиже висока тачност транскрипције уз високу ефикасност рада.
- Алтернативу коришћењу предложеног техничког решења представља коришћење других софтверских алата сличне намене, од којих су неки и јавно доступни, али од којих ниједан не испуњава све специфичне потребе анотације говорних база у саставу развоја система говорних технологија.
- Техничко решење развијено је на основу резултата научних истраживања која припадају текућем стању технике, на основу сопствених истраживања, претежно на сопственој опреми и без коришћења туђе патентне или лиценцне документације.

У Београду, 30.11.2015. године.

Проф. др Слободан Јовичић

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

Ovim potvrđujemo da je 01.10.2009. godine u preduzeću AlfaNum d.o.o. počela da se koristi osnovna verzija tehničkog rešenja "*SpeechLabel* – alat za fonetsko i prozodijsko labeliranje govornih baza podataka". Osnovnoj verziji rešenja je, tokom višegodišnjeg daljeg razvoja, dodat određeni broj naprednih mogućnosti, te je od 01.09.2015. u upotrebi najnovija verzija koja, pored vizuelizacije signala, spektrograma i energije, ima i podršku za inteligentnu pretragu audio-materijala, komunikaciju između labelatora, te sintezu govora na osnovu fonetske, leksičke i prozodijske transkripcije. Rešenje je trenutno u upotrebi u okviru razvoja sistema govornih tehnologija za srpski, hrvatski i engleski jezik, a radi se i na njegovom internom testiranju radi daljeg razvoja.

U Novom Sadu, 22.11.2015.

Darko Pekar, direktor

ALFANUM GOVORNE TEHNOLOGIJE NOVI SAD ZA PROIZVODNJU, PROMET I USLUGE

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)

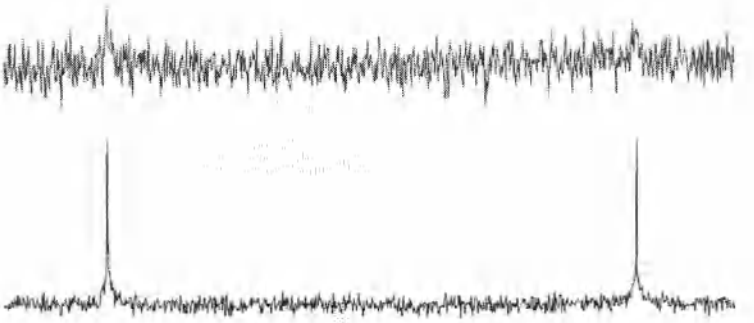
6. Development of source-filter separation tool
The result should be C/C++ library and standalone tool for source-filter separation and combination (synthesis). It should be based on some of the state-of-the-art techniques from literature. In this phase, some minor issues and spots for improvement could 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

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

SPEECHLABEL – АЛАТ ЗА ФОНЕТСКО И ПРОЗОДИЈСКО ЛАБЕЛИРАЊЕ ГОВОРНИХ БАЗА ПОДАТАКА

Аутори: Драгиша Мишковић, Дарко Пекар, Милан Сечујски, Едвин Пакоци.

-непотребно изостављено-

Записник водила:
Јасмина Димић, дипл. правник

Тачност података оверава:
Секретар
Иван Нешковић, дипл. правник

Декан
Проф. др Раде Дорословачки