

## **Образац за пријаву техничког решења<sup>1</sup>**

| <b>Техничко решење: <i>Софтвер за предикцију прозодије говора базиран на ToBI конвенцијама за прозодијску транскрипцију</i></b> |                                                                                                  |
|---------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| <b>Аутори техничког решења:</b>                                                                                                 | Милан Сечујски, Наташа Вујновић Седлар, Стеван Острогонац                                        |
| <b>Назив техничког решења:</b>                                                                                                  | Софтвер за предикцију прозодије говора базиран на ToBI конвенцијама за прозодијску транскрипцију |
| <b>Подтип техничког решења:</b>                                                                                                 | Нов производ на међународном нивоу (M81)                                                         |
| <b>Реализатори:</b>                                                                                                             | ФТН и АлфаНум д.о.о. из Новог Сада                                                               |

### **Пројекат и период реализације техничког решења:**

2011-2015: „Развој дијалошких система за српски и друге јужнословенске језике“ – ТР32035

### **За кога је техничко решење рађено:**

Техничко решење рађено је за предузеће *Speech Morphing Systems*, из Кембела у Калифорнији, Сједињене Америчке Државе, у саставу система говорних технологија за енглески језик, с перспективом проширења на друге језике, имајући у виду широку распрострањеност ToBI стандарда прозодијске транскрипције.

### **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 синтетизатора говора за енглески језик.

### **Како су резултати верификовани (од стране кога тела):**

- 1) 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 именовало рецензенте:

---

<sup>1</sup> 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).

- a. др Јернеја Жганец-Грос, *Алпинеон д.о.о*, Љубљана, Словенија,
  - b. др Зоран Шарић, *Центар за унапређење животних активности*, Београд
- 2) Употреба техничког решења у пракси потврђена је одговарајућим документима од стране предузећа *Speech Morphing Systems* из САД и *АлфаНум д.о.о*.

**На који начин се користи (кратак опис):**

Техничко решење користи се у саставу синтетизатора говора на основу текста, као компонента чији је задатак да реконструише прозодијска обележја говора који је потребно синтетизовати, како би овај звучао разумљиво и природно. Наиме, први корак синтезе говора на основу текста представља лингвистичка обрада текста на основу ког је потребно синтетизовати говор, како би се реконструисала обележја која су присутна у говору а нису експлицитно дата у тексту. Пре свега, код огромне већине светских језика фонетска транскрипција текста умногоме одступа од његовог ортографског записа, па је фонетска транскрипција главни задатак језичке обраде текста, али чак и када то није случај, потребно је реконструисати трајање појединих фонетских сегмената, као и промене основне учестаности говора у времену. Под условом да је на располагању одговарајућа прозодијска транскрипција текста, која обухвата обележја као што су положај и тип акцената, као и положај и тип граница између одговарајућих синтаксно-прозодијских целина, акустичку реализацију ових обележја могуће је одредити коришћењем система машинског учења. Задатак предложеног техничког решења је, дакле, да на основу прозодијске транскрипције полазног текста у неком од стандардних формата, на основу метода машинског учења, реконструише трајања фонетских сегмената и кретање основне учестаности гласа, као прозодијске параметре од највећег значаја за природност синтетизованог говора.

Техничко решење може се користити као самостална софтверска апликација која се ослања на оперативни систем *Windows*, или као модул у оквиру такве софтверске апликације. Првенствена намена техничког решења јесте у оквиру система за аутоматску синтезу говора на основу текста, али оно може имати и друге намене – примера ради, у евалуацији квалитета прозодијске анотације, повећању тачности аутоматске прозодијске анотације, убрзању поступка ручне прозодијске анотације говорних база података и слично.

**Опис техничког решења: Софтвер за предикцију прозодије говора базиран на ToBI конвенцијама за прозодијску транскрипцију**

Техничко решење намењено је аутоматској предикцији прозодијских обележја синтетизованог говора на основу његове прозодијске транскрипције у ToBI формату, и у овом тренутку имплементирано је за енглески језик. Транскрипција на основу

индекса тонова и граница (енг. *Tone and Break Indices – ToBI*)<sup>2</sup> представља систем за обележавање прозодијски значајних догађаја у оквиру реченице скупом ознака међу којима су најзначајније:

- индекс *границе* између две речи (B0, B1, B2, B3 или B4 – додељује се у зависности од типа границе између две речи, при чему се с једне стране налазе границе између речи које су до те мере међусобно повезане да је на граници између њих дошло и до фонолошких промена, док се с друге стране налазе границе између комплетних интонационих фраза).
- индекс *фразног тона* (L-, H-, !H- – додељује се у зависности од кретања основне учестаности на границама између потфраза од којих је састављена интонациона фраза)
- индекс *граничног тона* на крају интонационе фразе (L%, H%) и на њеном почетку (%H)
- индекс *тонског акцента* (H\*, L\*, L+H\*, H+!H\*, L\*+H...) – обележје које се додељује наглашеном слогу у оквиру речи у зависности од његовог односа са суседним слоговима, односно, утицаја на кретање основне учестаности.

ToBI транскрипција минимално обухвата индикацију намераваног прозодијског груписања речи у реченици, као и тоналних догађаја који су им придружени. Наведене прозодијске ознаке односе се на енглески језик, али је одговарајући систем развијен и за низ других језика, међу којима су и немачки, грчки, каталонски, португалски и јапански, док је делимично развијен ToBI систем прозодијске транскрипције и за српски<sup>3</sup>.

Одређеној комбинацији ToBI прозодијских ознака у одређеном фонетском контексту одговара одређена акустичка реализација, односно, одговарају јој конкретна трајања фонетских сегмената и кретање основне учестаности гласа. Предложено техничко решење реконструише ове вредности на основу система машинског учења заснованог на регресионим стаблима.

#### **Област на коју се техничко решење односи:**

Техничко решење припада области информационих, односно информационо-комуникационих технологија (ICT), и представља софтверски модул за предикцију прозодијских обележја говора који треба синтетизовати на основу текста, на основу расположиве фонетске транскрипције у ToBI формату.

---

<sup>2</sup> Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) *Prosodic Typology -- The Phonology of Intonation and Phrasing*

<sup>3</sup> Gođevac, S., "Transcribing Serbo-Croatian Intonation", in S.-A. Jun [Ed], *Prosodic Typology: The Phonology of Intonation and Phrasing*, 146-171, Oxford Linguistics, 2005.

### **Проблем који се техничким решењем решава:**

Техничко решење решава комплексан проблем у саставу аутоматске синтезе на основу текста, а то је одређивање „онога што треба да се чује“ на основу обичног текста, под условом да је на располагању његова фонетска и прозодијска транскрипција. Јаз између симболичке и акустичке репрезентације прозодије је велик, и овај проблем су раније решавали експертски системи за чији је развој био потребан концентрисан напор лингвиста, који су ручно састављали правила на основу којих се одређивало трајање појединих гласова у одређеном фонетско-прозодијском контексту, као и кретање основне учестаности (висине гласа) у току њиховог трајања.

### **Стање решености тог проблема у свету:**

Проблем одређивања прозодијских обележја говора у свету већ дуго привлачи пажњу истраживача. Постоје одређени резултати, од којих су неки и имплементирани у оквиру доступних система као што је систем за синтезу говора на основу скривених Марковљевих модела (*Hidden Markov Model Based Text-to-Speech System – HTS*), Технолошког института у Нагоји (Јапан), али је њихова имплементација у овом погледу релативно фиксна и није погодна за проширивање. С друге стране, постоје бројни веома квалитетни комерцијално расположиви системи за синтезу говора на енглеском и другим светским језицима, од којих неки постижу и веома висок степен природности говора. Међутим, решења која се користе у оквиру тих система нису комерцијално доступна, тако да постоји потреба за развојем сопственог решења.

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

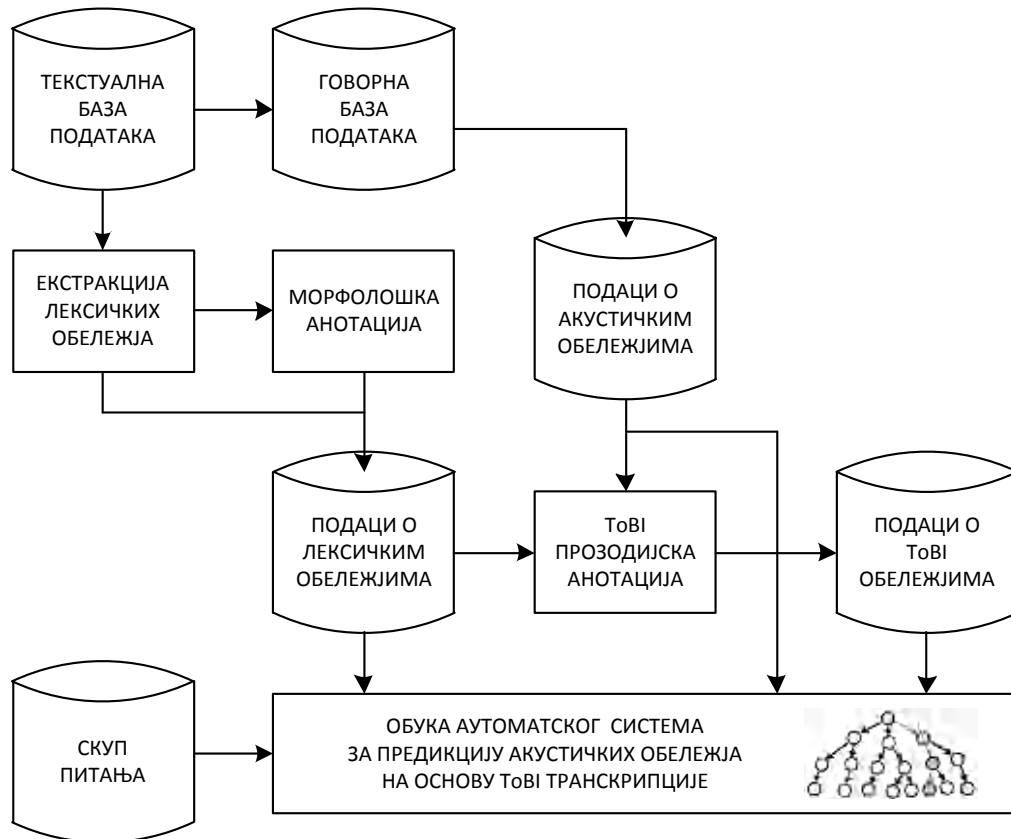
Техничко решење као улазне податке користи податке о фонетским сегментима из говорне базе података, и то:

- податке о лексичким обележјима,
- податке о ToBI обележјима,
- податке о акустичким обележјима,

а ослања се и на претходно специфицирани сет питања која се односе на одређени фонетски сегмент. Ова питања могу бити лексичког карактера, те одговор на њих може бити очигледан на основу текстуалне транскрипције говорне базе, али се могу односити и на морфолошке или прозодијске карактеристике одговарајуће лингвистичке јединице, те одговор на њих може захтевати претходну морфолошку или прозодијску анотацију говорне базе. Примери таквих питања били би:

- да ли је посматрани фонем вокал (назал, дифтонг, плонзив...)
- да ли је посматрани фонем у оквиру именице (везника, предлога...)

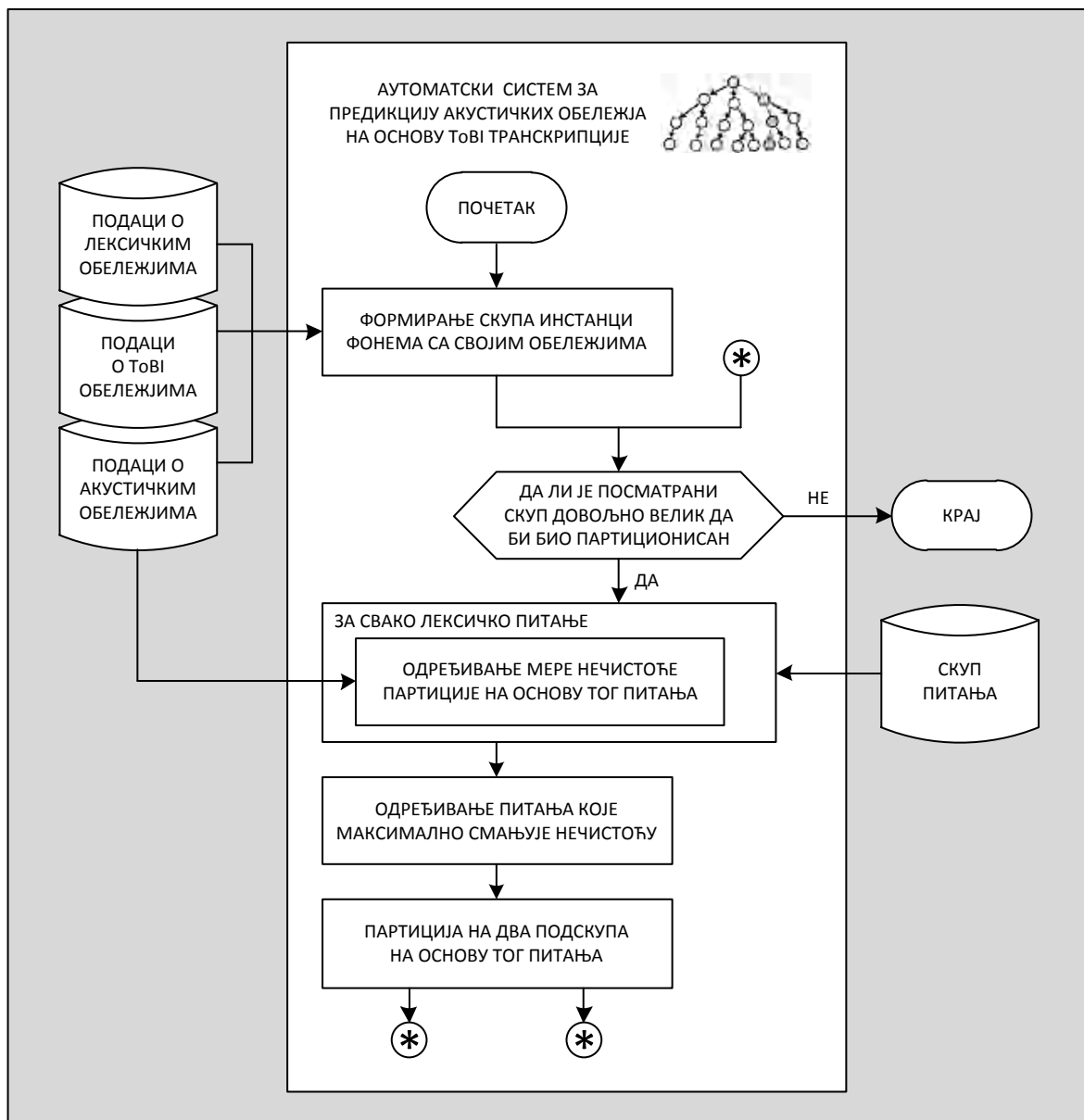
- да ли посматраном фонему непосредно претходи вокал (тишина, исти тај фонем, звучни или беззвучни парњак посматраног фонема...)
- да ли је посматрани фонем вокал под тонским акцентом  $H^*$  ( $L^*$ ,  $L+H^*$ ...)
- да ли посматраном фонему непосредно следи граница између речи типа  $B0$  ( $B1$ ,  $B2$ ...)



Слика 1. Припрема улаза аутоматског система за предикцију прозодијских обележја на основу ToBI прозодијске транскрипције

На слици 1 дат је упрошћени блок-дијаграм процеса потребних за обучавање система за аутоматску предикцију акустичких обележја. Систем за предикцију акустичких обележја представља регресионо стабло или скуп регресионих стабала на основу којих се за сваки фонетски сегмент који је потребно синтетизовати дефинише трајање и кретање основне учестаности. Упрошћен приказ начина обуке овог система, односно, формирања регресионог стабла, дат је на слици 2. Пре свега, формира се скуп инстанци фонема са својим обележјима – лексичким (укључујући и морфолошка), прозодијским (ToBI) и акустичким. Пошто говорне базе за овакве намене могу обухватати и више сати говора, у овом скупу може бити и више десетина хиљада инстанци појединих фонема. Нека, примера ради, регресионо стабло треба да прогнозира вредност основне учестаности ( $f_0$ ) на средини трајања одређеног гласа (ако се ради о звучном гласу). Нека од питања из скупа су релевантна у односу на ово

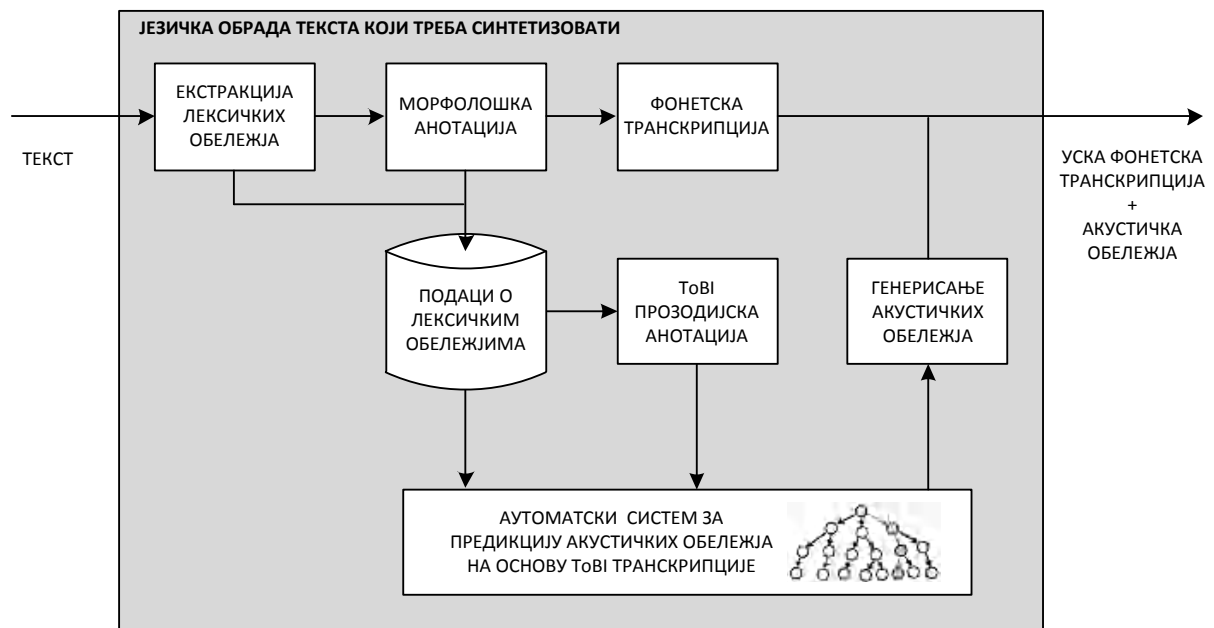
обележје, а нека не. Може се очекивати да ће питања која су у највећој мери релевантна за ово обележје (примера ради: „Да ли након речи у којој се овај фонем налази непосредно следе фразни тон L- и гранични тон L%?“) поделити посматрани скуп на гласове са претежно високом и претежно ниском  $f_0$ , за разлику од питања која у односу на исто обележје нису релевантна (примера ради: „Да ли је посматрани фонем назал?“), која ће исти скуп партиционирати тако да ће у оба подскупа и даље бити заступљени и гласови са претежно високом и гласови са претежно ниском  $f_0$ . На основу дефинисане објективне мере нечистоће партиције, могуће је идентификовати



Слика 2. Процес обуке аутоматског система за предикцију прозодијских обележја на основу ToBI прозодијске транскрипције

питање које је у највећој мери релевантно, извршити партицију на основу њега, па рекурзивно наставити исти поступак на оба добијена подскупа, све док су они довољно велики да би се партиција могла извршити. Критеријум за престанак поделе обично се везује за број инстанци у подскупу (чвору регресионог стабла), и за базе од неколико сати обично се може узети  $N = 10$ .

Што се тиче дефинисања скупа питања, ово је једини корак који захтева експертско знање из лингвистике, док је остатак процедуре у потпуности аутоматски. За обуку описаног техничког решења за енглески језик користен је сегмент стандардне радијске говорне базе Бостонског универзитета, односно, снимци који се односе на професионалну спикерку означену са F2 (правим именом Марго Мелников). Ова база једним својим делом је ручно прозодијски анотирана у складу са ToBI стандардом, а фонетска анотација је делимично ручна а делимично аутоматска. Узимајући у обзир само ToBI анотиране реченице, минутажа снимака говорнице F2 је била највећа у бази (нешто више од 2 сата говора), што је и био главни разлог за њен одабир.



Слика 3. Експлоатација аутоматског система за предикцију прозодијских обележја на основу ToBI прозодијске транскрипције

У фази експлоатације система, која је приказана блок-дијаграмом на слици 3, на располагању је само текст на основу ког је потребно синтезирати говор, и потребна акустичка обележја су непозната. Под претпоставком да су на располагању модули који ће обезбедити фонетску и прозодијску транскрипцију текста, сваки фонетски сегмент „пропушта се“ кроз регресионо стабло, путањом која је дефинисана одговорима на питања на која фонетски сегмент наилази у појединим чворовима стабла. Сегменту се на крају додељује средња вредност акустичког обележја свих фонетских сегмената који су у фази обуке формирали терминални (крајњи) чвор стабла у ком је



завршио и посматрани фонетски сегмент, и за која се, према томе, може претпоставити да имају сличан фонетски, лексички, морфолошки, односно, прозодијски контекст као и посматрани фонетски сегмент.

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

Техничко решење реализовано је као софтверска апликација, односно модул у саставу софтверске апликације, који не захтева никакав додатни хардвер за функционисање, већ му је довољан рачунар опште намене или мобилни уређај. С обзиром на развој рачунарске технологије и технологије мобилних уређаја, техничко решење може да функционише на практично свим уређајима који су данас у употреби, чак и на онима скромнијих карактеристика. Треба напоменути и да је рачунски најзахтевнији део поступка управо обука, која се врши у припремној фази, а не у фази експлоатације на уређајима крајњег корисника, тако да је рачунска захтевност поступка параметар од релативно малог значаја.

Решење је првенствено намењено реконструкцији прозодијских обележја у саставу синтезе говора на основу текста, али се може користити и за друге примене у области говорних технологија, као и лингвистичких истраживања.

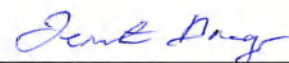
У свом тренутном облику, техничко решење се примењује између осталог у саставу синтетизатора говора за енглески језик, развијаног за предузеће *Speech Morphing Systems* из Кембела, Калифорнија, а с обзиром на постојање модификованих верзија ToBI конвенција за прозодијску анотацију и транскрипцију за низ других језика, поље потенцијалне примене техничког решења је и знатно шире од тренутног.

**Докази (прилози):**

- Писано мишљење два рецензента-експерта из области техничког решења
- Документ којим се доказује да предузеће *АлфаНум* користи ово техничко решење у развоју система говорних технологија
- Документ којим се доказује да предузеће *SpeechMorphing* користи ову технологију у развоју система говорних технологија

Нови Сад, новембар 2015. године.

Подносилац пријаве



Проф. др Владо Делић  
Руководилац пројекта TP32035





УНИВЕРЗИТЕТ  
У НОВОМ САДУ



ФАКУЛТЕТ  
ТЕХНИЧКИХ НАУКА

Трг Доситеја Обрадовића 6, 21000 Нови Сад, Република Србија  
Деканат: 021 6350-413; 021 450-810; Централна: 021 485 2000  
Рачуноводство: 021 458-220; Студентска служба: 021 6350-763  
Телефакс: 021 458-133; e-mail: [ftndeac@uns.ac.rs](mailto:ftndeac@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 17.2.9: У циљу верификације новог техничког решења предлажу се рецензенти:**

- Др Зоран Шарић (Центар за унапређење животних активности, Београд)
- Др Јернеја Жганец Грос (Alpineon, Љубљана, Словенија)

## **СОФТВЕР ЗА ПРЕДИКЦИЈУ ПРОЗОДИЈЕ ГОВОРА БАЗИРАН НА ТОБИ КОНВЕНЦИЈАМА ЗА ПРОЗОДИЈСКУ ТРАНСКРИПЦИЈУ**

Аутори: Милан Сечујски, Наташа Вујновић Седлар, Стеван Острогонац.

*-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:  | Софтвер за предикцију прозодије говора базиран на ToBI конвенцијама за прозодијску транскрипцију |
| Аутори техничког решења: | Милан Сечујски, Наташа Вујновић Седлар,<br>Стеван Острогонац                                     |
| Реализатори:             | ФТН и АлфаНум доо из Новог Сада                                                                  |
| Подтип техничког решења: | Нови производ на међународном нивоу (M81)                                                        |

### Подаци о рецензенту:

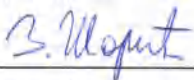
|                                                                                       |                                                                                                                                                        |
|---------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Име, презиме и звање:                                                                 | др Зоран Шарић, научни саветник                                                                                                                        |
| Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета: | Изабран 30.01.2014. у звање научни саветник на ЕТФ у Београду, у области техничко-технолошких наука - електроника, телеком. и информационе технологије |
| Установа где је запослен:                                                             | Центар за унапређење животних активности, Београд                                                                                                      |

### Стручно мишљење рецензента:

„Софтвер за предикцију прозодије говора базиран на ToBI конвенцијама за прозодијску транскрипцију“ представља **нови производ на међународном нивоу (M81)** у смислу Правилника о поступку и начину вредновања, и квантитативном исказивању научноистраживачких резултата истраживача од 21.03.2008. Образложење:

- Предложени систем решава проблем одређивања акустичких обележја говора у оквиру синтетизатора говора. Решење је имплементирано за енглески језик, док имплементација решења за друге језике за које је такође развијен ToBI систем прозодијске транскрипције треба само дефинисати релевантан сет лексичко-прозодијских питања и обезбедити потребне говорне базе за обуку.
- Решење функционише као софтверска апликација или софтверски модул, на рачунару опште намене или мобилном уређају, без посебних захтева у погледу перформанси.
- Решење је развијено без коришћења туђе патентне и лиценцне документације, претежно на сопственој опреми. Као резултат примене решења омогућена је синтеза говора на енглеском језику која звучи природно, као и даљи развој система за конверзију идентитета говорника.

У Београду, 27.11.2015. године.



др Зоран Шарић, научни саветник

## РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

### Подаци о техничком решењу:

|                          |                                                                                                  |
|--------------------------|--------------------------------------------------------------------------------------------------|
| Назив техничког решења:  | Софтвер за предикцију прозодије говора базиран на ToBI конвенцијама за прозодијску транскрипцију |
| Аутори техничког решења: | Милан Сечујски, Наташа Вујновић Седлар,<br>Стеван Острогонац                                     |
| Реализатори:             | ФТН и АлфаНум доо из Новог Сада                                                                  |
| Подтип техничког решења: | Нови производ на међународном нивоу (M81)                                                        |

### Подаци о рецензенту:

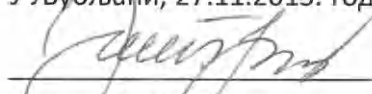
|                                                                                       |                                                                                                                                                                                  |
|---------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Име, презиме и звање:                                                                 | Др Јернеја Жганец Грос, <i>razvojni savetnik</i>                                                                                                                                 |
| Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета: | Изабрана у звање <i>razvojni savetnik za IKT</i> , 26.05.2003. године sa strane <i>Komisije za razvrščanje v raziskovalne nazive</i> за н.о. <i>Alpineon raziskave in razvoj</i> |
| Установа где је запослен:                                                             | Алпинеон д.о.о, Љубљана, Словенија                                                                                                                                               |

### Стручно мишљење рецензента:

„Софтвер за предикцију прозодије говора базиран на ToBI конвенцијама за прозодијску транскрипцију“ представља **нови производ на међународном нивоу (M81)** у смислу Правилника о поступку и начину вредновања, и квантитативном исказивању научноистраживачких резултата истраживача од 21.03.2008. Образложење:

- Софтверско решење које је изложено омогућује да се у синтетизатору говора дође до податка о томе на који начин је потребно да рачунар „изговори“ тражени текст на енглеском језику, тако што се на основу прозодијских ознака у ToBI формату (за које се претпоставља да су на располагању, нпр. одређене помоћу експертског система) генеришу одговарајућа акустичка обележја коришћењем регресионих стабала.
- Применљивост техничког решења није ограничена само на енглески језик, уз напомену да је за употребу на другим језицима за које је развијен ToBI стандард анотације потребно обезбедити одговарајуће ToBI аотиране говорне базе.
- 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 лиценцене документације.

У Љубљани, 27.11.2015. године.

  
др Јернеја Жганец-Грос



Ovim potvrđujemo da je 01.08.2015. godine u preduzeću AlfaNum d.o.o. počela upotreba tehničkog rešenja "Softver za predikciju prozodije govora baziran na ToBI konvencijama za prozodijsku transkripciju". Ova softverska komponenta u upotrebi je u sastavu razvoja govornih tehnologija za engleski jezik, prvenstveno sinteze govora, modelovanja govora i konverzije identiteta govornika, u saradnji sa preduzećem *Speech Morphing Systems* iz Kembela, Kalifornija. Radi se, takođe, i na intenzivnom testiranju tehničkog rešenja i ispitivanju mogućnosti dodavanja određenih funkcionalnosti i unapređivanja njegove tačnosti.

U Novom Sadu, 22.11.2015.

*Darko*

Darko Pekar, direktor



## **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:

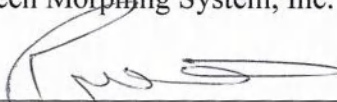
#### **Services Provided**

- 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 proceed to morphing. By "final voice model" it also doesn'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

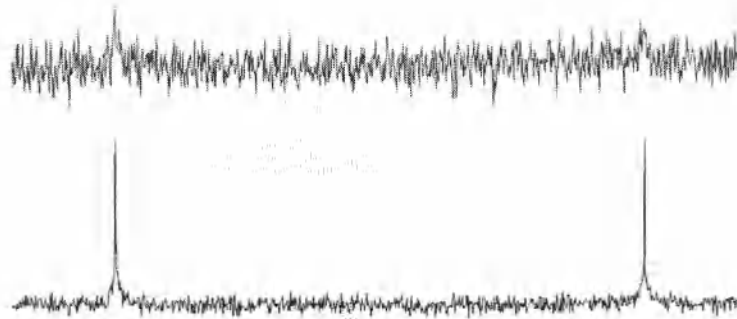
Price: [REDACTED]

**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

**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. |                |            |
| 2       | Advance payment from 3.4.2015. (Advance invoice 001/15)   | 1,00 | pcs. |                |            |

**Total**

Slovima: \_\_\_\_\_

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





УНИВЕРЗИТЕТ  
У НОВОМ САДУ



ФАКУЛТЕТ  
ТЕХНИЧКИХ НАУКА

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

ИНТЕГРИСАНИ  
СИСТЕМ  
МЕНАЏМЕНТА  
СЕРТИФИКОВАНО 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**

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

#### **СОФТВЕР ЗА ПРЕДИКЦИЈУ ПРОЗОДИЈЕ ГОВОРА БАЗИРАН НА ТОБИ КОНВЕНЦИЈАМА ЗА ПРОЗОДИЈСКУ ТРАНСКРИПЦИЈУ**

Аутори: Милан Сечујски, Наташа Вујновић Седлар, Стеван Острогонац.

*-непотребно изостављено-*

Записник водила:

Јасмина Димић, дипл. правник

Тачност података оверава:  
Секретар

Иван Нешковић, дипл. правник

Декан

  
Проф. др Раде Дорословачки