Образац за пријаву техничког решења[1]

Назив техничког решења	**Алат за декомпозицију говорног сигнала на основу модела побуда-филтар**
Аутори техничког решења	**Синиша Сузић, Роберт Мак, Дарко Пекар, Никша Јаковљевић**
Категорија техничког решења	**Нови производ на међународном нивоу (М81)**

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

Ово техничко решење урађено је у оквиру технолошког пројекта "Развој дијалошких система за српски и друге јужнословенске језике" (ТР32035, 2011-15) на Факултету техничких наука и у предузећу „АлфаНум" у Новом Саду.

Ко користи техничко решење:

Техничко решење рађено је за предузеће *Speech Morphing Systems*, из Кембела у Калифорнији, Сједињене Америчке Државе, као део њиховог система за конверзију гласа једног говорника у глас другог.

Година када је техничко решење урађено:

Развој техничког решења започет је у априлу 2015. године, а довршен у јуну 2015. године.

Ко је прихватио-примењује техничко решење:

Техничко решење тренутно користи предузеће *Speech Morphing Systems*, које се бави развојем говорних технологија (првенствено система за конверзију гласа једног говорника у глас другог) на више језика, а користи га и предузеће АлфаНум д.о.о. из Новог Сада у развоју параметарског синтетизатора за српски и енглески језик.

Како су резултати верификовани (од стране ког тела):

1) Употреба техничког решења у пракси потврђена је одговарајућим документима од стране предузећа *Speech Morphing Systems* і *АлфаНум д.о.о*.

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

2) Дато је писано мишљење два рецензента-експерта из области техничког решења:
   a. Проф. др Зоран Ивановски, Факултет електротехнике и информационих технологија у Скопљу
   b. Проф. др Зденка Бабић, Електротехнички факултет у Бањој Луци

3) За ово техничко решење спроведена је предвиђена процедура на Департману за енергетику, електронику и телекомуникације и ФТН-у који издаје Уверење о признавању техничког решења, а које потврђује да оно испуњава све услове да буде признато као нови производ на међународном тржишту, у складу са Правилником Министарства.

На који начин се користи (кратак опис):

Техничко решење је реализовано у виду статичке C++ библиотеке намењене за употребу у програмском окружењу Microsoft Visual Studio 2013. Поред библиотеке реализована је и конзолна апликација намењена Windows оперативном систему која користи дату библиотеку. Библиотека као и сама конзолна апликација предвиђени су за рад са једноканалним аудио фајловима снимљеним у *wav*[2] формату. Сви улазни параметри осим у командној линији могу да се дефинишу и у одговарајућим конфигурационим фајловима. За добијање тачнијих резултата неопходно је као улазни параметар проследити и фајл са временским положајем пичмаркера. Резултат коришћења програма су сигнал побуде снимљен у *wav* формату као и параметри филтра снимљени у бинарном или текстуалном формату. Добијени резултати могу да се користе како у параметарској синтези говора тако и у поступку конверзије гласа једног говорника у глас другог.

Опис техничког решења: Алат за декомпозицију говорног сигнала на основу модела побуда-филтар

Област на коју се техничко решење односи:

Техничко решење припада области дигиталне обраде говорних сигнала.

Проблем који се техничким решењем решава:

Постоји више различитих приступа моделовању продукције говорног сигнала. Најбољи резултати се постижу употребом модела који директно описују функционисање артикулаторних органа. Међутим, такви модели су по правилу веома сложени и њихово решавање је исувише комплексно да би могли да се примењују у реалним апликацијама. Компромисно решење је моделовање процеса продукције говора помоћу линеарних система. Такав модел се још назива и модел побуда-филтар. Овај модел полази од претпоставке да се сигнал побуде $\boldsymbol{u}(\boldsymbol{n})$ користи као улаз у линеарни систем $\boldsymbol{v}(\boldsymbol{n})$ који представља вокални тракт, односно говорни сигнал може да се представи у облику конволуције:

[2] http://soundfile.sapp.org/doc/WaveFormat/

$$\boldsymbol{s(n) = u(n) * v(n)}\ ,$$

или у спектралном домену:

$$S(f) = U(f)V(f).$$

Моделовање сигнала $u(n)$ зависи од типа побуде: за звучне гласове користи се периодични сигнал, а у случају безвучних гласова користи се шум. Звучни гласови се обично моделују као линеарни систем који описује карактеристике глоталног импулса $G(f)$, који се побуђује периодичном поворком импулса чија је преносна карактеристика $E(f)$, а основна учестаност f_0.

Последњи корак у моделовању продукције говора је укључивање зрачења на уснама. Овај ефекат се обично описује помоћу VF филтра преносне карактеристике $L(f)$.

Узимајући у обзир претходне наводе потпуни модел продукције говора добија следећи облик:

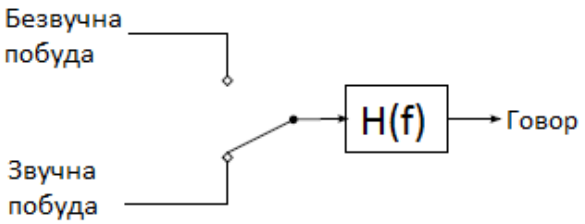
$$S(f) = E(f)G(f)V(f)L(f),$$

при чему је $G(f) = 1$ за безвучне сигнале, а $E(f)$ у том случају моделује сигнал шума.

Засебно моделовање делова система описаних преносним карактеристикама $G(f)$, $V(f)$, $L(f)$ је компликовано и због тога се они често обједињују у јединствену компоненту означену са $H(f)$, односно:

$$S(f) = E(f)H(f).$$

Блок дијаграм оваквог система дат је на Слици 1.



Слика 1. Блок дијаграм модела продукције говора

У овом техничком решењу описан је предлог једног поступка који се користи да би се на основу сигнала $s(n)$ одредио сигнал e(n) и параметри блока $H(f)$.

Стање решености тог проблема у свету:

Издвајање побуде и коефицијената филтра вокалног тракта предмет је научног истраживања већ неколико деценија будући да резултати имају различите примене: кодовање говорног сигнала, синтеза говора на основу текста, конверзија гласа једног говорника у други. Стога данас постоји више различитих модела, а нови модели се и даље развијају.

Један од најзаступљенијих приступа је линеарно предиктивно кодовање (LPC) говорног сигнала које је засновано на представљању преносне карактеристике вокалног тракта *all-pole* филтром (Atal, Bishnu S. "The history of linear prediction." *IEEE Signal Processing Magazine* 23.2 (2006): 154-161).

Поред LPC анализе значајну примену у обради говора имају и мел-фреквенцијски кепстрални коефицијенти (MFCC) који се добијају применом дискретне косинусне трансформације на логаритмованом спектру сигнала. (Tychtl, Zbynek, and Josef Psutka. "Speech production based on the mel-frequency cepstral coefficients." *EuroSpeech*. Vol. 99. 1999).

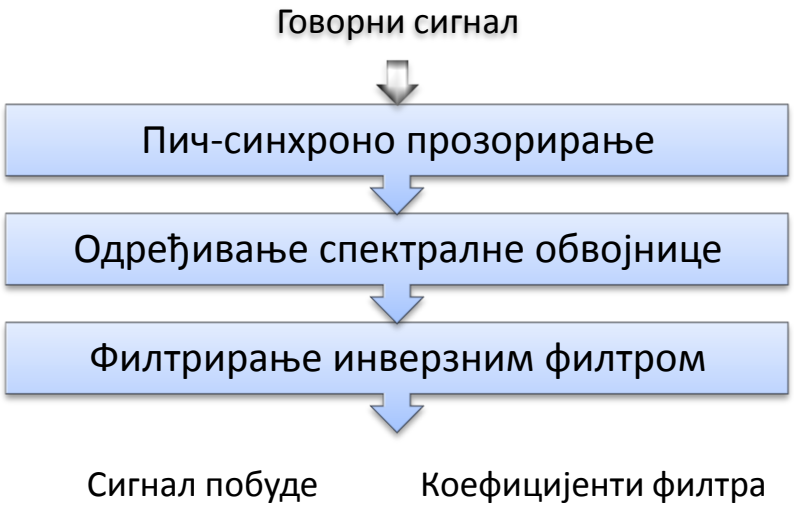
Поступци засновани на издвајању глоталних импулса из говорног сигнала такође су доста коришћени (Walker, Jacqueline, and Peter Murphy. "A review of glottal waveform analysis. "*Progress in nonlinear speech processing*. Springer Berlin Heidelberg, 2007. 1-21).

Међутим, ниједан од наведених поступака није савршен и има недостатке. На пример, LPC анализа не функционише исувише добро у условима повећаног шума, не представља добро звучне гласове. И то је главни разлог због чега је проблем издвајања побуде и параметара вокалног тракта и даље интересантан за истраживање.

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

Метод за издвајање побуде описан о овом раду заснован је на већ неким познатим поступцима у процесу обраде говорног сигнала уз одређене модификације. Поред процеса анализе, тј. добијања сигнала побуде и параметара филтра, у овом документу је описано и како се врши синтеза полазног сигнала.

Основни кораци у анализи сигнала приказани су на Слици 2.



Слика 2. Дијаграм предложеног поступка анализе говорног сигнала

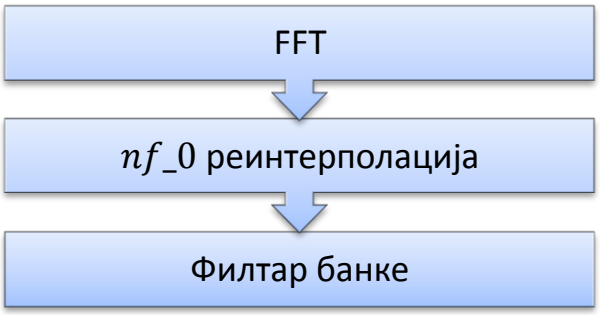
Први корак у процесу синтезе је пич-синхроно прозорирање. Фајл се пичмаркерима један је од улазних параметара нашег алгоритма. Уколико пичмаркер фајл није дефинисан

користи се еквидистантна апроксимација положаја пичмаркера на основу основне учестаности дефинисане у конфигурационом фајлу. Треба ипак напоменути да се употребом еквидистантних пичмаркера постижу нешто лошији резултати. За издвајање прозорираног сигнала користи се Хемингова прозорска функција дефинисана на следећи начин:

$$h(n) = 0.54 - 0.45\cos\left(\frac{2\pi n}{N-1}\right)$$

Дужина прозора је такође дефинисана у конфигурационом фајлу.

Следећи корак у процесу анализе јесте одређивање спектралне обвојнице. Овај поступак представљен је на Слици 3. Над прозорираним сигналом прво се израчунава дискретна Фуријеова трансформација применом брзих алгоритама (FFT). FFT се рачуна у најмање 1024 тачке. Добијене вредности FFT коефицијената се потом користе да би се одредиле вредности спектра у тачкама које одговарају целобројним умношцима основне учестаности, nf_0. За добијање тражених вредности користи се линеарна интерполација. На основу добијених вредности спектра у тачкама nf_0 врши се реинтерполација вредности у тачкама које одговарају полажају FFT коефицијената на фреквенцијској оси.



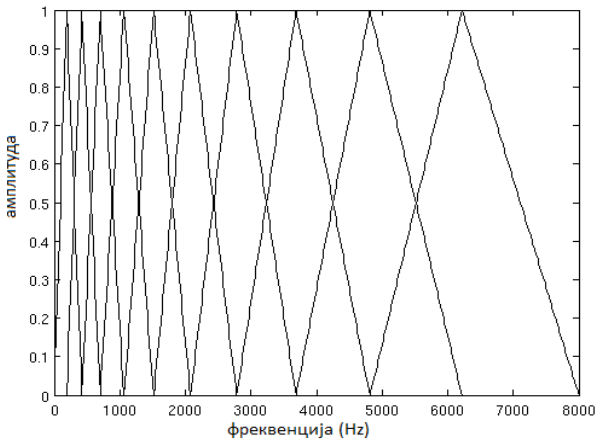
Слика 3. Одређивање спектралне обвојнице

У случају када се користе еквидистантни пичмаркери процес реинтерполације се прескаче, управо због непознавања тачне учестности f_0 и могућности да дође до озбиљнијих грешака. Тада се првобитно израчунати FFT коефицијенти користе као излаз из тренутног блока.

Вредности коефицијената из блока за интерполацију користе се као улази у филтар банке које имају приближно једнаку резолуцију на мел фреквенцијској скали која је дефинисана следећом једначином:

$$mel(f) = 2595\log_{10}(1 + \frac{f}{700})$$

Филтар банке су троугаоног облика, а сам процес је имплементиран тако да се фреквенцијске компоненте множе одговарајућим коефицијентима и потом се за дату банку акумулирају. Изглед филтар банака дат је на Слици 4.



Слика 4. Десет мел филтар банака

Број филтар банака се такође дефинише у конфигурационим фајловима.

Следећи корак у предложеном алгоритму је одређивање вредности филтар банака у еквидистантим временским тренуцима. Вредност временског помераја је још један од конфигурационих параметара. Коефицијенти филтар банака се рачунају као линеарна комбинација коефицијената филтар банака који одговарају претходно издвојеним пич-синхроним прозорима између којих се налази посматрани временски тренутак,

$$IF(t_0) = c_1 FB_{n-1} + c_2 FB_n \ ,$$

Коефицијенти c_1 и c_2 рачунају се на следећи начин:

$$c_1 = \frac{d_1}{T0}, \quad c_1 = \frac{d_2}{T0}$$

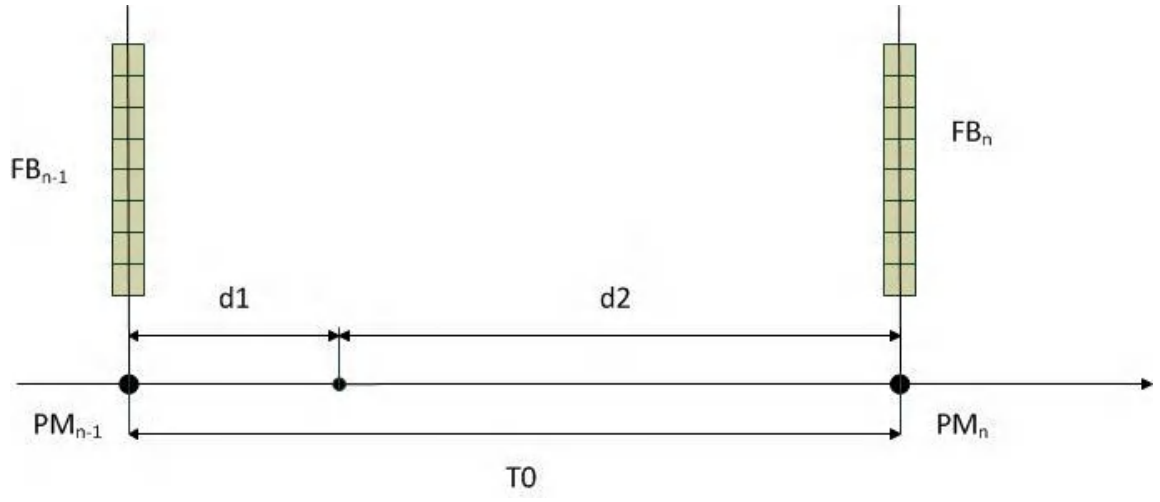
где d_1 и d_2 представљају удаљеност тренутне тачке од најближег левог и десног пичмаркера, респективно, а $T0$ растојање између два посматрана пичмаркера. Графичка илустрација дата је на Слици 5.

Вредности који се добијају на излазу овог блока се сматрају тачкама које се могу искористити за добијање преносне карактеристике вокалног тракта. Међутим, у последњој операцији у овом блоку израчунава се и нормализациони коефицијент који ће бити примењен на коефицијентима филтар банака. Прво се на основу филтар банака врши интерполација спектра у броју тачака у ком ће се рачунати инверзни филтар. На основу тих вредности нормализациони коефицијент се рачуна према следећој једначини:

$$norm = \sqrt{\frac{\sum_{k=0}^{N_1} X(k)^2}{N_1 N_2}},$$

при чему су $X(k)$ интерполиране вредности спектра, $N1$ је број тачака у којим је спектар интерполиран, а $N2$ је дужина пич-синхроног прозора.

Овако добијене вредности уз коефицијенте филтар банака чине први излаз алгоритма.



Слика 5. Илустрација израчунавања коефицијената филтра

Последњи корак у фази анализе је одређивање сигнала побуде. Полазећи од модела којим је раније дефинисано да је спектар говорног сигнала дефинисан као

$$S(f) = E(f)H(f)$$

следи да је спектар побуде дефинисан на следећи начин:

$$E(f) = \frac{S(f)}{H(f)} = S(f)I(f)$$

Добијање сигнала побуде у описаном методу управо је дефинисано на основу претходне једначине. $I(f)$ се добије инверзијом преносне карактеристике која одговара преносној карактеристици вокалног тракта. Преносна карактеристика инверзног филтра адаптивно се мења и одређена је на основу формуле:

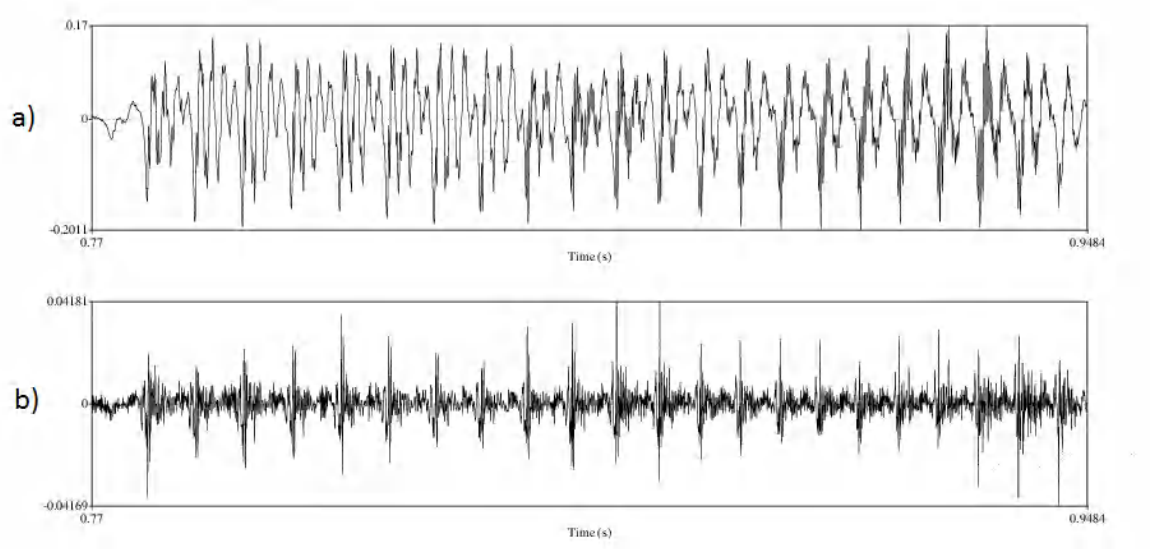
$$I(f) = \frac{1}{c_1 H_{n-1}(f) + c_2 H_n(f)}$$

Коефицијенти c_1 и c_2 одређени су тренутним положајем посматране тачке у односу на два најближа фрејма (који су еквидистантни). Одређивање коефицијената c_1 и c_2 је аналогно одређивању коефицијената филтар банака који одговарају фрејмовима. Вредности спектралних компоненти $H_{n-1}(f)$ и $H_n(f)$ добијају се интерполацијом израчунатих вредности филтар банака. Примењује се линеарна интерполација на основу познатих вредности излаза филтар банака и њиховог положаја на фреквенцијској оси. Коефицијенти филтра у временском домену се добијају рачунањем инверзне дискретне Фуријеове трансформације применом брзих алгоритама, тј.

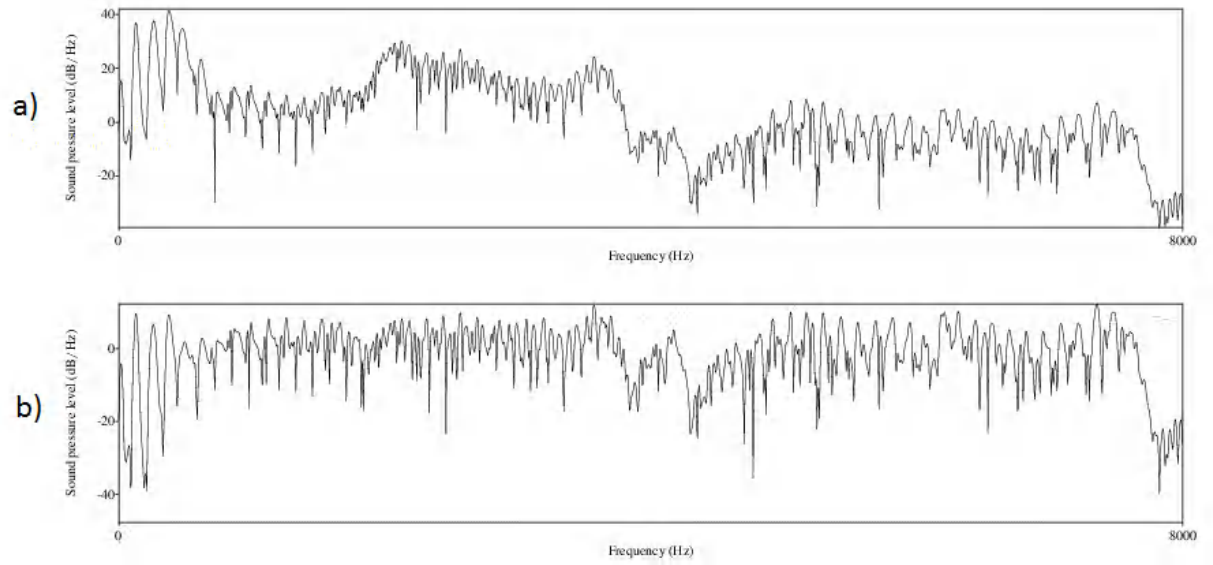
$$i(n) = IFFT\{I(f)\}$$

Због потребе да се инверзни филтар рачуна у свакој тачки описани поступак је преспор за употребу у реалном времену. Међутим, имајући на уму чињеницу да су добијена обележја намењена у процесу обуке за параметарску синтезу говора или у процесу обуке за конверзија говорника, ова чињеница и не представља превелики проблем.

Резултати примене описаног алгоритма могу се видети на Сликама 6 и 7. На Слици 6 приказани су изглед оригиналног сигнала и сигнала побуде, док су на Слици 7 приказани спектар оригиналног сигнала и спектар добијеног сигнала побуде. Са Слике 6 је видљиво и шта је главна одлика предложеног алгоритма у спектралном домену. Наиме, алгоритам тежи да већину спектралих компоненти доведе на исте вредности.



Слика 6. Упоредни приказ оригиналног (а) и сигнала побуде (б)



Слика 7. Упоредни приказ спекатара оригиналног (а) и сигнала побуде (б)

Овим методом је такође предвиђена и синтеза полазног сигнала на основу сигнала побуде и коефицијената филтар банака. С обзиром на раније дату једначину за израчунавање сигнала побуде јасно је да у спектралном домену реконструкција полазног сигнала има следећи облик:

$$S(f) = E(f)H(f),$$

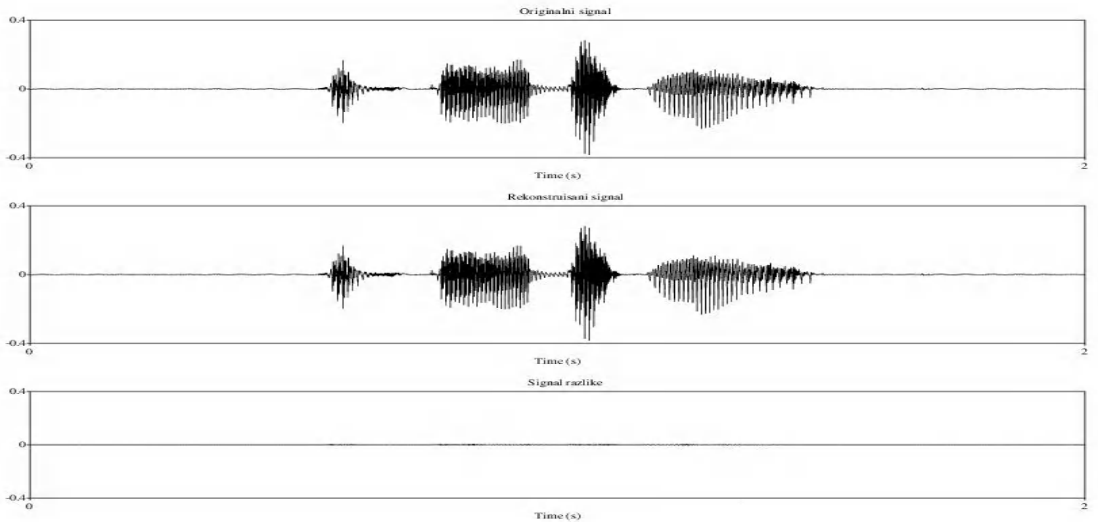
односно у временском домену се рачуна конволуција раније добијеног сигнала побуде и филтра вокалног тракта. Преносна каратеристика филтра се такође адаптивно мења зависно од посматране улазне тачке и одређена је на следећи начин:

$$H(f) = c_1 H_{n-1}(f) + c_2 H_{n-1}(f)$$

Као и у случају инверзног филтра коефицијенти c_1 и c_2 одређени су тренутним положајем посматране тачке у односу на два најближа еквидистантна фрејма, док се вредности фреквенцијских компоненти $H_{n-1}(f)$ и $H_{n-1}(f)$ добијају на основу израчунатих вредности филтар банака. Коефицијенти филтра у временском домену добијају се инверзном дискретном Фуријеовом трансформацијом на основу вредности компоненти $H(f)$.

$$h(n) = IFFT\{H(f)\}$$

Добијени сигнал није у потпуности идентичан полазном сигналу, што се може видети на Слици 8. Међутим, добијене разлике су минималне, а тестови слушања су такође потврдили да су и слушаоцу неприметне.



Слика 8. Приказ оригиналног, реконструисаног и сигнала побуде

За разлику од процеса анализе процес синтезе нема проблем са брзином извршавања. Наиме, могуће је одредите инверзне филтре које одговарају преносним карактеристикама $H(f)$ само једанпут, а потом све процесе интерполације радити у временском домену без потребе за додатним израчунавањем инверзних трансформација.

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

Техничко решење је реализовано у виду статичке C++ библиотеке коју је могуће користити у имплементацији нових поступака. Поред тога развијена је и конзолна апликација која користи дату библиотеку и омогућава израчунавање и чување сигнала побуде у *wav* формату, а коефицијената филтар банака у бинараном или текстуалном формату.

Резултати примене описане методе коришћени су у параметарској синтези на основу скривених Марковљевих модела. Иако за сада не постоји опсежнија анализа први тестови слушања показују да су добијени резултати упоредиви са резултатима добијених на основу мел-генерализованих кепстралних (MGC) коефицијената („Параметарска синтеза говора за енглески језик на бази фонетски транскрибованог и прозодијски анотираног текста", Пекар и други).

Такође су вршени и одређени експерименти са трансформацијом карактеристика једног говорника у другог. Идеја је заснована на прилагођавању прозодије побуде полазног сигнала на прозодију циљног говорника, као и пресликавању коефицијената.

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
- Уверење о признавању техничког решења од стране Департмана 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 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 17.2.6: У циљу верификације новог техничког решења предлажу се рецензенти:

- Проф. др Зоран Ивановски (Факултет за електротехнику и информационе технологије, Скопље, Македонија)
- Проф. др Зденка Бабић (Електротехнички факултет, Бања Лука, БиХ)

АЛАТ ЗА ДЕКОМПОЗИЦИЈУ ГОВОРНОГ СИГНАЛА НА ОСНОВУ МОДЕЛА ПОБУДА-ФИЛТАР

Аутори: Синиша Сузић, Роберт Мак, 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:	Алат за декомпозицију говорног сигнала на основу модела побуда-филтар
Аутори техничког решења:	Синиша Сузић, Роберт Мак, Дарко Пекар, Никша Јаковљевић
Реализатори:	ФТН и АлфаНум доо из Новог Сада
Подтип техничког решења:	Нови производ на међународном нивоу (М81)

Подаци о рецензенту:

Име, презиме и звање:	Проф др Зденка Бабић, редовни професор
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабрана 05.12.2012. у звање редовног професора Универзитета у Бањој Луци, у области опште електротехнике
Установа где је запослен:	Електротехнички факултет, Бања Лука

Стручно мишљење рецензента:

„Алат за декомпозицију говорног сигнала на основу модела побуда-филтар“ 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:

- Предложени систем представља једно решење проблема издвајања побуде и параметара филтра који описује вокални тракт на основу говорног сигнала, настало коришћењем научних метода на пројекту МПНТР
- Решење функционише као софтверска апликација или софтверски модул, на рачунару опште намене, без посебних захтева у погледу перформанси.
- Описани поступак може да се користи у процесу синтезе говора на основу текста или приликом конверзије гласа једног говорника у глас другог

У Бањој Луци, 27.11.2015. године.

Проф. др Зденка Бабић

РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	Алат за декомпозицију говорног сигнала на основу модела побуда-филтар
Аутори техничког решења:	Синиша Сузић, Роберт Мак, Дарко Пекар, Никша Јаковљевић
Реализатори:	ФТН и АлфаНум доо из Новог Сада
Подтип техничког решења:	Нови производ на међународном нивоу (М81)

Подаци о рецензенту:

Име, презиме и звање:	Проф. д-р Зоран Ивановски
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабран у октобру 2011. године у звање ванредног професора на ФЕИТ у Скопљу, за област електронике
Установа где је запослен:	Факултет за електротехнику и информационе технологије, Скопље

Стручно мишљење рецензента:

„Алат за декомпозицију говорног сигнала на основу модела побуда-филтар“ 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:

- Софтверско решење описује поступак за декомпозицију говорног сигнала на сигнал побуде и параметре филтра коришћењем инверзне преносне карактеристике естимиране спектралне обвојнице
- Техничко решење развијено је на основу резултата научних истраживања која припадају текућем стању технике, али и на на основу сопствених истраживања, при чему није коришћена туђа патентна или лиценцна документација
- Омогућено је коришћење као самосталне апликације, као и библиотеке за развој нових метода
- На бази добијених резултата могуће је радити на развоју параметарске синтезе говора,али и конверзију гласа једног говорника у глас другог

У Скопљу, 27.11.2015. године.

Проф. д-р Зоран Ивановски

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 15.07.2015. godine u preduzeću AlfaNum d.o.o. počelo da se koristi tehničko rešenje *Alat za dekompoziciju govornog signala na osnovu modela pobuda-filtar,* koje je zajednički razvijeno od strane Fakulteta tehničkih nauka Novi Sad i preduzeća Alfanum, a koristi se za potrebe preduzeća Alfanum i SpeechMorphing Inc. Dato tehničko rešenje se koristi u razvoju parametarske sinteze govora na osnovu teksta kao i za razvoj sistema za konverziju glasa jednog govornika u glas drugog.

U Novom Sadu, 22.11.2015.

Darko Pekar, direktor

DRUŠTVO SA OGRANIČENOM ODGOVORNOŠĆU 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)

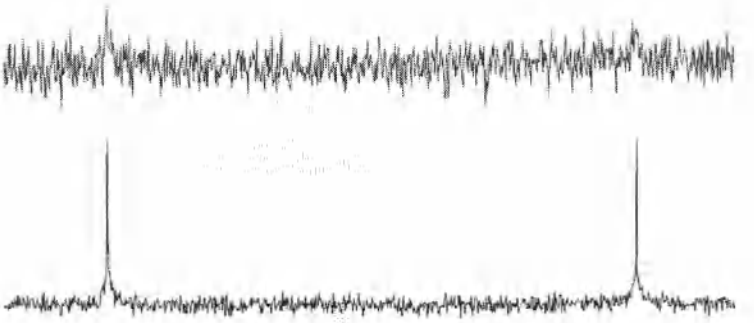
6. Development of source-filter separation tool
The result should be C/C++ library and standalone tool for source-filter separation and combination (synthesis). It should be based on some of the state-of-the-art techniques from literature. In this phase, some minor issues and spots for improvement could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

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

АЛАТ ЗА ДЕКОМПОЗИЦИЈУ ГОВОРНОГ СИГНАЛА НА ОСНОВУ МОДЕЛА ПОБУДА-ФИЛТАР

Аутори: Синиша Сузић, Роберт Мак, 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