Образац за пријаву техничког решења[1]

Назив техничког решења	**Параметарска синтеза говора за енглески језик на бази фонетски транскрибованог и прозодијски анотираног текста**
Аутори техничког решења	**Дарко Пекар, Синиша Сузић, Роберт Мак, Бранислав Поповић**
Категорија техничког решења	**Нови производ на међународном нивоу (М81)**

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

Техничко решење је реализовано у оквиру технолошког пројекта МПНТР "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:

Предузеће АлфаНум у Новом Саду као партиципант на пројекту ТР32035 и предузеће Speech Morphing, Inc.

Година када је техничко решење урађено:

Развој техничког решења започет је средином 2015. године, а довршен у новембру 2015. године.

Ко је прихватио-примењује техничко решење:

Поред предузећа АлфаНум у Новом Саду као партиципанта на пројекту ТР32035, техничко решење је од 2015. године у употреби од стране Speech Morphing, Inc.

Како су резултати верификовани (од стране ког тела):

1) Техничко решење је реализовано, имплементирано и испитано на развојним системима у предузећу АлфаНум.
2) Техничко решење је програмски систем реализован у оквиру пројекта ТР32035 („Развој дијалошких система за српски и друге јужнословенске језике").

[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).

Представља приступ у решавању проблема синтезе говора на основу транскрибованог и тагованог текста на енглеском језику.

3) Приложено је писано мишљење два рецензента, експерта из области техничког решења.
   a. Проф. др Снежана Гудурић, Филозофски факултет у Новом Саду,
   b. Проф. др Петар Правица, ЕТФ Београд.
4) На основу извештаја Комисије за техничка решења на Департману за енергетику, електронику и телекомуникације ФТН-а издато је Уверење о признавању техничког решења које потврђује да оно испуњава све услове да буде признато као нови производ на међународном нивоу (М81), у складу са Правилником Министарства.
5) Поједини елементи техничког решења презентирани су на већем броју домаћих и међународних научних скупова, на пример:
   - „HMM-based Speech Synthesis for the Serbian Language“, Едвин Пакоци, Роберт Мак, *Зборник радова 56. Конференције за електронику, телекомуникације, рачунарство, аутоматику и нуклеарну технику ЕТРАН*, Златибор, 2012.
   - „Use of Global Variance for the Enhancement of HMM-based Speech Synthesis“, Роберт Мак, Едвин Пакоци, *Зборник радова 56. Конф. ЕТРАН*, Златибор, 2012.
   - „Subjective Assessment of Text-to-Speech Synthesis Systems for the Serbian Language“, Едвин Пакоци, Роберт Мак, Стеван Острогонац, *Зборник радова 20. Телекомуникационог форума ТЕЛФОР*, Београд, 2012.
   - „Синтеза говора на бази скривених Марковљевих модела за српски језик“, Едвин Пакоци, мастер рад, Факултет техничких наука, Нови Сад, 2012.
   - „Примена глобалне варијансе ради унапређења квалитета говора синтетизованог на основу скривених Марковљевих модела“, Роберт Мак, мастер рад, Факултет техничких наука, Нови Сад, 2012.

На који начин се користи (кратак опис):

Апликација овог синтетизатора говора за енглески језик покреће се из командне линије оперативног система *Windows* (коришћењем програма *Cygwin*) преко позива одговарајуће скрипт датотеке написане у програмском језику *Perl*, унутар које се аутоматски позивају одговарајуће функције алата које су коришћене при синтези говора. Такође се могу у скрипту позивати функције алата који се користе за обуку модела, чиме се могу мењати параметри обуке што утиче на крајњу синтезу. Очекује се да се текст за синтезу да у виду датотеке која представља тагован текст (транскрипција са фонетским, лингвистичким и прозодијским информацијама за синтезу) на енглеском у одређеном формату. Подржан је тагован текст по ToBI[2] (енг. *tones and break indices*) конвенцијама. Апликација као излаз генерише звучне датотеке у *wav* формату.

[2] Hirschberg, J., and M. Beckman. "The ToBI Annotation Conventions." Ohio State University, 1994.

Опис техничког решења:

Параметарска синтеза говора за енглески језик на бази фонетски транскрибованог и прозодијски анотираног текста

Овај синтетизатор говора на бази скривених Марковљевих модела (енг. *hidden Markov models, HMM*) заснован је на сопственим говорним технологијама, као и на неколико савремених сличних система. Обучаван је на говорној бази на енглеском језику.

Област на коју се техничко решење односи:

Техничко решење припада области Телекомуникације и информационе технологије, и као такво представља спој неколико различитих области, попут телекомуникација и обраде сигнала, дигиталне обрада говора, синтезе говора.

Проблем који се техничким решењем решава:

Синтетизатори говора имају многобројне примене у разним областима, као што су приступ информацијама (електронска пошта, базе података, говорни портали), као помоћ особама оштећеног вида и особама са поремећајем говора, као алтернатива преносу говора преко Интернета (на економичнији, бржи и безбеднији начин), у софтверима за учење страних језика (без потребе за снимљеним репертоаром лекција или речи, онај ко учи може да чује правилан изговор било које реченице), у кућним уређајима, лингвистичким истраживањима, ванредним ситуацијама итд. Оваква технологија се може користити и у разговору двоје или више људи, где спадају нпр. системи за превођење говора, *eyes-free hands-free* системи и многи други.

Савремени системи за синтезу текста треба да испуњавају критеријуме високог квалитета, тј. да синтетизовани говор звучи природно и разумљиво, без улагања додатног напора приликом слушања. Синтетизатори на бази скривених Марковљевих модела, будући да су у основи статистички параметарски, имају додатну могућност спецификације карактеристика гласа и говорног стила. За разлику од конкатенативних, системи на бази скривених Марковљевих модела су отпорнији на недостатке у бази, заузимају мање меморијских ресурса, робуснији су, флексибилни у погледу мењања карактеристика гласа (то се лако постиже трансформацијом параметара модела).

Стање решености тог проблема у свету:

Приступи синтези говора се могу поделити у две велике групе:

- синтеза говора на основу правила (формантни синтетизатори, синтеза говора на основу скривених Марковљевих модела);
- синтеза говора повезивањем говорних сегмената (LPC(AR) модел, GAR модел, TD-PSOLA, MBR-PSOLA, LP-PSOLA и други).

Као најзначајнија врста из друге групе издваја се синтетизатор на бази *on-line* селекције и спајања говорних сегмената уз коришћење TD-PSOLA методе. Овакав је први синтетизатор говора за српски језик, такође развијен у АлфаНум-у. Он је доживео неколико верзија, а последња даје и више него прихватљиве резултате и у активној је употреби, на пример на Интернет страницама Радио телевизија Србије и Радио телевизија Војводине, али и у склопу других производа који су испоручени различитим корисницима.

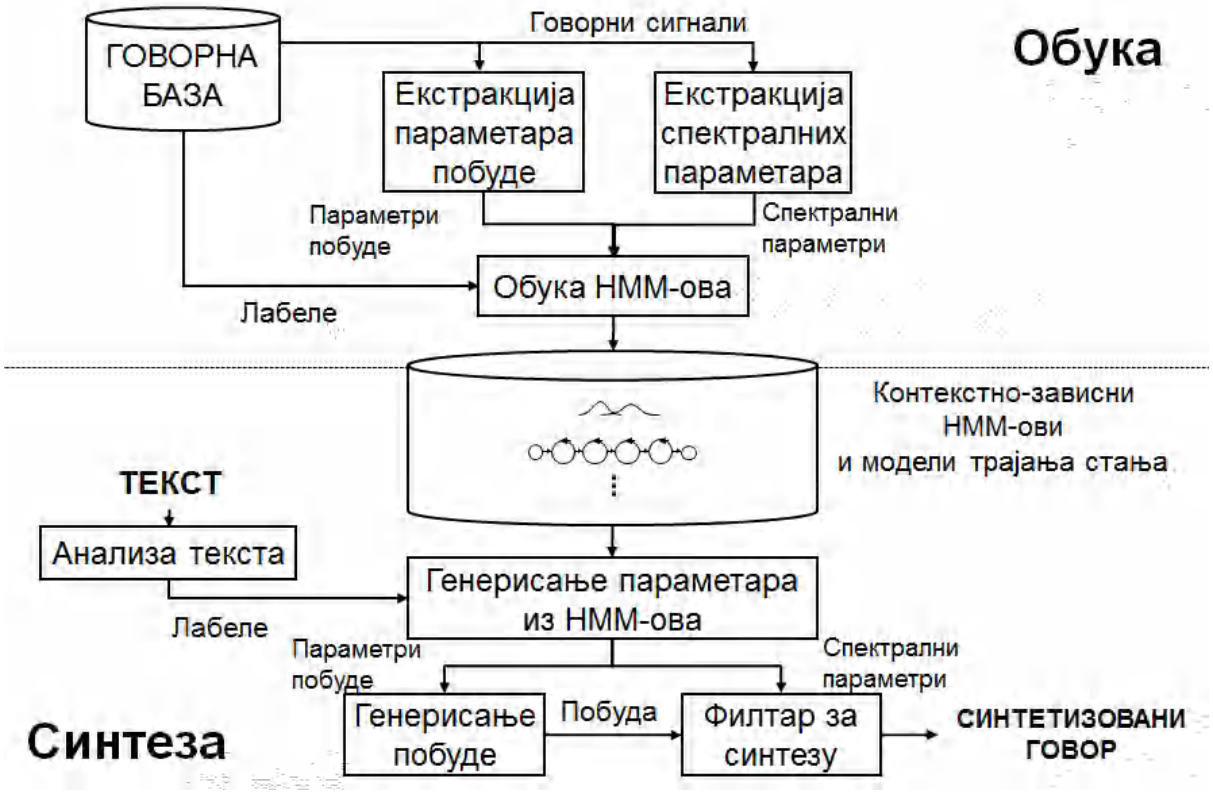
Синтеза на бази скривених Марковљевих модела тренутно је једна од најпопуларнијих метода синтезе, због раније поменутих предности у односу на конкатенативну (меморијски ресурси, флексибилост), а може да даје приближне или чак и боље резултате за исту говорну базу (стандардне величине). За њу је развијено више алата, од којих се издваја HTS (HMM-based Speech Synthesis System), који је инкорпориран у неколико апликација. Преко њега је развијен одличан синтетизатор за енглески и јапански језик, а пријављени су и синтетизатори за мандарински, арапски, корејски, немачки, португалски, шведски, фински, шпански, тајландски, грчки, мађарски, српски, хрватски, словеначки и др. Од осталих алата треба споменути Espeak, Festival, FreeTTS, Gnuspeech (на основу артикулаторних правила), који су сви бесплатни и јавно доступни.

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

Основни елементи система за синтезу говора на основу текста јесу модул за језичку обраду текста (JOT) и модул за генерисање говорног језика. Задатак модула за JOT је обезбеђивање свих потребних информација за генерисање природног и разумљивог говора. Он од текста на свом улазу формира одговарајуће транскрипције које садрже фонетске, лингвистичке (нпр. акцентуација, синтаксна структура реченице) и прозодијске информације (нпр. реченични нагласак, трајања фонема). Описан систем ради на основу датих транскрипција са потребним информацијама. Очекује се да су формиране по ToBI конвенцијама. ToBI конвенције се односе на транскрипцију и анотирање прозодије говора и широко су прихваћене као конвенције које се користе за енглески, али и за друге језике. Алат који може помоћи у креирању оваквих транскрипција предмет је другог техничког решења („Софтвер за оцену квалитета прозодијске анотације“, Дарко Пекар и др.).

Овај синтетизатор говора на бази скривених Марковљевих модела за енглески језик урађен је по угледу на систем развијен од стране HTS групе са Токијског института за технологију (слика 1). Састоји се од два главна дела – фазе обуке и фазе синтезе. Прва је фаза обуке (која се врши *off-line*, при формирању синтетизатора), у којој треба из говорне базе екстраховати параметре говорних сигнала – спектралне параметре и параметре побуде (основну учестаност), помоћу којих ће се обучити контекстно

зависни модели фонема. Процес синтезе почиње од дате транскрипције (претходно добијене након анализе текста), на основу које се одређеним критеријумом, из скупа обучених модела, добија низ контекстно зависних фонема које треба синтетизовати. Они се затим повезују у реченични модел, преко којег се уз помоћ алгоритма за генерисање параметара из скривених Марковљевих модела генеришу одговарајући параметри спектра и побуде, из којих се преко филтра за синтезу синтетише резултујући говорни сигнал.



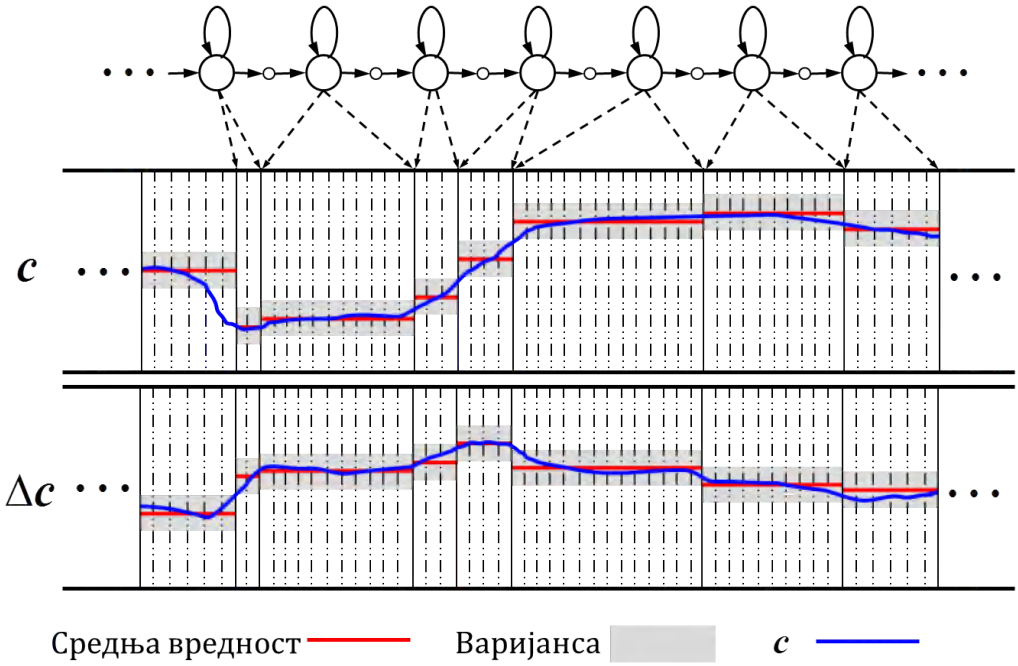
Слика 1. Систем за синтезу говора на основу текста

Говорна база за овај синтетизатор састоји се од преко 850 снимљених реченица једне радио водитељке (на енглеском), односно укупно око 1 сат говорног материјала, као и од одговарајућих транскрипција, тј. лабела са лексичким обележјима која описују контекст свих фонема и подацима о њиховом трајању, форматираних у HTS формат. База је анализирана Хеминговим прозорским функцијама са ширином фрејма од 25 милисекунди и померајем фрејма од 5 милисекунди (експериментисано је и са другим вредностима). Као спектрални параметри коришћени су мел-генерализовани кепстрални коефицијенти (укупно 35, заједно са нултим). У другој варијанти параметри су базирани на излазима из филтар банака (такође 35). Ти параметри су се показали погодним јер је уз њих екстрахован и сигнал побуде, који би се касније користио при синтези, а овакав начин издвајања обележја је погодан и за проблем конверзије говора (енг. *voice conversion*). Тај систем је детаљно описан у другом техничком решењу („Алат за декомпозицију говорног сигнала на основу модела побуда-филтар“,

Сузић и др.). У оба случаја, параметри су представљени једном континуалном вишедимензионалном Гаусовом расподелом вероватноће.

За моделовање основна учестаности потребно је проширење концепта скривених Марковљевих модела, јер она није дефинисана у безвучним сегментима говора. Једно од решења, које је примењено у овом техничком решењу, је концепт вишепросторних расподела вероватноће (енг. *multi-space probability distribution hidden Markov model – MSD-HMM)*, где се безвучни сегменти представљају бездимензионалним дискретним симболом, а звучни једнодимензионалним вредностима основне учестаности.

Осим наведених статичких параметара, користе се и динамички, који су апроксимације првог односно другог извода статичких. Узимање њих у обзир доприноси томе да као излаз алгоритма за генерисање параметара (што се извршава непосредно пре синтезе) немамо низ средњих вредности (степеничасту функцију, која резултује чујним, непријатним дисконтинуитетима), већ због уведених додатних ограничења имамо реалистичнију, глатку функцију. Ова трајекторија (заједно са трајекторијом средњих вредности) приказана је на слици 2. Укупно је коришћено 108 параметара за сваку акустичку опсервацију (3 пута 35 спектралних и 3 за основну учестаност).

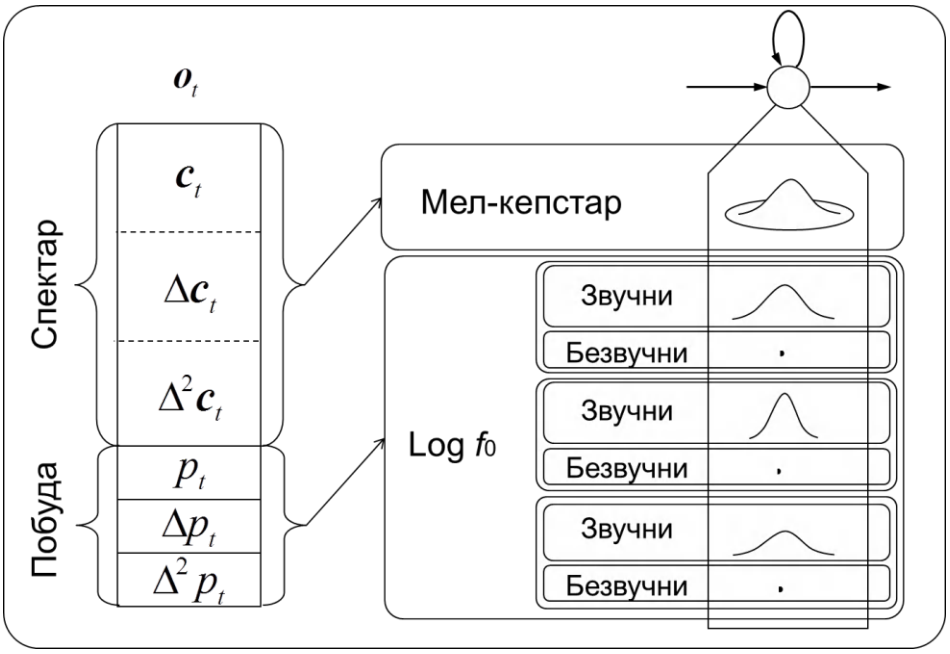


Слика 2. Трајекторија генерисаних параметара говора

Још један проблем је то што због низова дискретних симбола за основну учестаност у безвучним регијама није могуће одредити одговарајуће прелазе између стања модела, па се користи MSD-HMM са више токова (енг. *streams*) да би се укомбиновала обележја основне учестаности са спектралним, фрејм по фрејм. Раздвајање на токове (за основну учестаност) је потребно, јер би се, на границама звучних и безвучних

сегмената, за звучни фрејм добило да је његов извод димензије нула (као да је безвучан) јер је један од суседних фрејмова безвучан (за извод се узимају у обзир вредности одређеног броја суседних). Алат HTS не може да рукује оваквом опсервацијом, па се прибегава поменутом раздвајању.

Резултат је композитна опсервација, у којој један ток представљају спектрална обележја, а по један ток основна учестаност, њен први и други извод (слика 3).



Слика 3. Приказ композитне опсервације

Трајање стања модела се моделује експлицитном Гаусовом (једнодимензионалном) расподелом вероватноћа (не имплицитно преко вероватноћа прелаза), пошто то више одговара реалној ситуацији. Показује се да овакво моделовање при синтези омогућава једноставно манипулисање брзином говора.

За обуку модела користи се Baum-Welch алгоритам (субоптимална итеративна процедура). Сличан алгоритам се користи и у системима за препознавање говора, уз разлику што овде постоји више токова и узимају се у обзир, осим фонетских, и лингвистички и прозодијски контексти.

Како је је број могућих контекста који утичу на спектар, основну учестаност и трајање изузетно велик за сваки од фонема у бази, практично је немогуће имати све могуће комбинације контекстуалних фактора макар једном у бази (чак ни већину), па неки могући модели не могу бити обучени. Врло често се и постојећи модели (са контекстом) налазе само једном или неколико пута у бази. Проблем се решава контекстним кластеровањем на бази стабла одлуке. Ово се ради посебно за свако од стања модела и такође засебно за сваки тип параметара (спектар, основна учестаност

и трајање). Резултат је значајно смањена количина података, као и могућност синтезе из модела који не постоје у бази тако што се преко стабла одлуке пронађе њима одговарајући кластер, чији се параметри затим користе за синтезу.

Софтверски алати који су коришћени за обуку и синтезу су:

- HTK (Hidden Markov Model Toolkit) – садржи алате за обуку модела из базе и транскрипција. Његова надоградња за синтезу говора, HTS, омогућава генерисање параметара говора из обучених модела и адаптацију говора, као и подршку вишепросторним расподелама вероватноћа;
- SPTK (Speech Signal Processing Toolkit) – алат који је коришћен за екстракцију обележја из говорне базе и за добијање говорног сигнала из генерисаних параметара;
- Алати развијени у АлфаНум-у (ектракција обележја на бази излаза из филтар банака, екстракција основне учестаности, екстракција побуде,...).

Говор је био уједначеног квалитета, без повремених деградација. Неке уочене мане су вештачки (метални) призвук и приметна пригушеност говора, што је последица статистичког моделовања и генерисања на основу параметара. Још једна уочена мана јесте благо зујање на вокалима и фрикативима због прејединоставног модела побуде (код варијанте где се она не издваја), што је могуће решити реалистичнијим моделом побуде на бази глоталних импулса.

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

У оквиру овог синтетизатора говора за енглески језик примењене су неке од савремених техника моделовања и генерисања говора. Систем је развијен за потребе предузећа Speech Morphing, и у плану је да се користи као део већег система за конверзију говора, који је још увек предмет интензивних истраживања у свету.

Корисницима су доступне могућности:

- *Синтеза говора на енглеском језику, уз значајну редукцију потребних хардверских ресурса (процесорске снаге и меморијског простора) у поређењу са конкатенативном верзијом синтетизатора*.

- *Могућности адаптације карактеристика генерисаног говора у погледу различитих параметара.*

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 на основу извештаја Комисије за техничка решења 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 17.2.5: У циљу верификације новог техничког решења предлажу се рецензенти:

- Проф. др Снежана Гудурић, Филозофски факултет у Новом Саду
- Проф. др Петар Правица, Електротехнички факултет у Београду

ПАРАМЕТАРСКА СИНТЕЗА ГОВОРА ЗА ЕНГЛЕСКИ ЈЕЗИК НА БАЗИ ФОНЕТСКИ ТРАНСКРИБОВАНОГ И ПРОЗОДИЈСКИ АНОТИРАНОГ ТЕКСТА

Аутори: Дарко Пекар, Синиша Сузић, Роберт Мак, Бранислав Поповић.

-непотребно изостављено-

Записник водила:	Тачност података оверава: Секретар	Декан
Јасмина Димић, дипл. правник	Иван Нешковић, дипл. правник	Проф. др Раде Дорословачки

РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	Параметарска синтеза говора за енглески језик на бази фонетски транскрибованог и прозодијски анотираног текста
Аутори техничког решења:	Дарко Пекар, Синиша Сузић, Роберт Мак, Бранислав Поповић
Реализатори:	ФТН и АлфаНум доо из Новог Сада
Подтип техничког решења:	Нови производ на међународном нивоу (М81)

Подаци о рецензенту:

Име, презиме и звање:	Др Петар Правица, редовни професор у пензији
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабран у звање редовни професор (**1986**) на Електротехничком факултету у Београду, у.н.о. Аудиотехника и обрада говорног сигнала
Установа где је запослен:	Електротехнички факултет, Београд

Стручно мишљење рецензента:

„ Параметарска синтеза говора за енглески језик на бази фонетски транскрибованог и прозодијски анотираног текста " 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 обуку и синтезу говорног сигнала на енглеском језику, коришћењем скривених Марковљевих модела
- Техничко решење развијено је на основу резултата научних истраживања која припадају текућем стању технике.
- Омогућено је коришћење система позивом одговарајућег скрипта, без додатних захтева у погледу рачунарских ресурса.
- На бази добијених резултата могуће је радити на развоју система за конверзију гласа говорника.

У Београду, 30.11.2015. године.

Проф. др Петар Правица

ИНЖЕЊЕРСКА КОМОРА СРБИЈЕ
Петар
Б. Правица
дипл. инж. ел.
353 6193 03
ОДГОВОРНИ ПРОЈЕКТАНТ

РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	Параметарска синтеза говора за енглески језик на бази фонетски транскрибованог и прозодијски анотираног текста
Аутори техничког решења:	Дарко Пекар, Синиша Сузић, Роберт Мак, Бранислав Поповић
Реализатори:	ФТН и АлфаНум доо 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. за ужу научну област Романистика, Филозофски факултет, Нови Сад
Установа где је запослен:	Филозофски факултет, Нови Сад

Стручно мишљење рецензента:

„Параметарска синтеза говора за енглески језик на бази фонетски транскрибованог и прозодијски анотираног текста“ 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 врши обуку параметара и синтезу на основу њих, на енглеском језику. Решење је могуће имплементирати и за друге језике уз одговарајуће транскрипције, базу и скуп лексичко-прозодијских питања.
- Решење је развијено коришћењем углавном постојећих програмских алата, са сопственом базом и транскрипцијама. Коришћен је сопствени алгоритам за издвајање основне учестаности, а, у једној варијанти, и сопствена обележја и алгоритам за раздвајање побуде и вокалног тракта.
- Као резултат примене решења омогућена је параметарска синтеза говора на енглеском језику, која се може користити у даљем развоју система за конверзију идентитета говорника.

У Новом Саду, 30.11.2015. године.

Проф. др Снежана Гудурић

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

Novi Sad
22.11.2015.

POTVRDA

Ovim potvrdujemo da je 01.10.2015. godine u preduzeću AlfaNum d.o.o. počelo da se koristi tehničko rešenje *Parametarska sinteza govora za engleski jezik na bazi fonetski transkribovanog i prozodijski anotiranog teksta,* koje je zajednički razvijeno od strane Fakulteta tehničkih nauka Novi Sad i preduzeća AlfaNum, a koristi se za potrebe preduzeća Alfanum i SpeechMorphing Inc.

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)

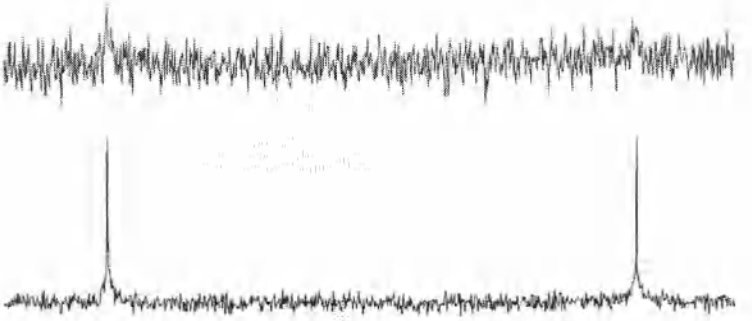
6. Development of source-filter separation tool
The result should be C/C++ library and standalone tool for source-filter separation and combination (synthesis). It should be based on some of the state-of-the-art techniques from literature. In this phase, some minor issues and spots for improvement could 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:
TÜV NORD ISO 9001
TÜV NORD ISO 14001
TÜV NORD OHSAS 18001

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

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

ПАРАМЕТАРСКА СИНТЕЗА ГОВОРА ЗА ЕНГЛЕСКИ ЈЕЗИК НА БАЗИ ФОНЕТСКИ ТРАНСКРИБОВАНОГ И ПРОЗОДИЈСКИ АНОТИРАНОГ ТЕКСТА

Аутори: Дарко Пекар, Синиша Сузић, Роберт Мак, Бранислав Поповић.

-непотребно изостављено-

Записник водила:	Тачност података оверава: Секретар	Декан
Јасмина Димић, дипл. правник	Иван Нешковић, дипл. правник	Проф. др Раде Дорословачки