UDC: 621.39:534.87 INFO M: str. 37-41

ИСПИТИВАЊЕ ТАЧНОСТИ АУТОМАТСКОГ ПРЕПОЗНАВАЊА ГОВОРНИКА НА ГОВОРНИМ СИГНАЛИМА ТЕЛЕФОНСКОГ КВАЛИТЕТА
ACCURACY INVESTIGATION OF AUTOMATIC SPEAKER RECOGNITION FOR TELEPHONE SPEECH SIGNAL QUALITY

Иван Јокић, Владо Делић, Никша Јаковљевић, Милан Добровић, Стеван Јокић

РЕЗИМЕ: У овом раду извршено је испитивање тачности идентификације говорника на говорним сигналима телефонског квалитета. Имплементација коришћеног препознавача говорника извршена је употребом HTK (енгл. Hidden Markov models ToolKit – HTK). Утицај разматраних телефонских канала на преношени говорни сигнал посматран је кроз њихове основне особености, типове примењених кодека и ефеката који су последица самог стања преносног канала. Поменути ефекти су посматрани кроз фактор вероватноће грешке приликом преноса, док је за VoIP телефонске канале анализирана и појава еха. Симулације рада одговарајућих кодека као и различитих вероватноћа појаве грешака приликом преноса извршене су коришћењем јавно доступне софтверске библиотеке алата, ITU-T STL2005, док је појава еха симулирана применом ефекта Delay/Echo-Simple програмског пакета Sony Sound Forge 9.0.
КЉУЧНЕ РЕЧИ: аутоматско препознавање говорника, мел-фреквенцијски кепстрални коефицијенти, модел мешавине Гаусових расподела, скривени Марковљев модел, HTK, ITU-T STL2005, ITU-T препорука G.729, ехо у VoIP.
ABSTRACT: This work was performed by examining the accuracy of speaker identification on telephone quality voice signals. Implementation of the used speaker recognizer was performed using HTK. Influence of the considered telephone channels on transmitted voice signal is seen through its basic characteristics, types of the applied codecs and the effects caused by the condition of the transmission channel. These effects were observed by a factor of transmission error probability, while the VoIP telephone channels were analyzed and the appearance of echo. Simulation of the appropriate codecs and the probability of various errors made during transmission by using publicly available library of software tools, ITU-T STL2005, while the echo phenomenon was simulated using effect Delay / Echo-Simple suite Sony Sound Forge 9.0.
KEY WORDS: Automatic Speaker Recognition, Mel – Frequency Cepstral Coefficients, Gaussian Mixture Models, Hidden Markov Model, HTK, ITU-T STL2005, ITU-T Recommendation G.729, echo in VoIP.

1. УВОД

Аутоматско препознавање говорника представља скуп процедура и поступака које имају за циљ утврђивање идентитета човека на основу његовог гласа, а које се извршавају на рачунарима односно рачунарским системима опште или специјализоване намене. У зависности од задатка препознавања оно може бити сведено на верификацију говорника или и на његову идентификацију међу више говорника. Верификација подразумева потврду или одбацивање тврђеног идентитета препознаваног говорника, док идентификација има за циљ утврђивање идентитета посматраног говорника.

Глас односно говор представља сложени сигнал добијен као резултат неколико трансформација, семантичких, лингвистичких, артикулаторних и акустичких, које се дешавају на различитим нивоима [1,2]. У циљу ефикасне рачунарске обраде посматраних говорних сигнала потребно је на одговарајући начин у коначном домену извршити њихову представу. Тим се сваки од њих репрезентује одговарајућим скупом обележја. У зависности од тога да ли се циљана обележја говора првенствено сматрају последицом анатомских особености посматраног говорног субјекта, или се приликом њиховог добијања у обзир узима њему својствен речник као и прозодијске особености његовог говора, често се врши њихова категоризација на обележја ниског односно високог нивоа респективно [3]. У овом раду су коришћена обележја ниског нивоа, односно говор посматраног говорника сматран је последицом дејства органа вокалног тракта на ваздушну струју покренуту радом његових плућа. Од три основне карактеристике гласа – интензитет, висина и боја – овде је акценат дат на његовој боји. Она се огледа у спектралној обвојници говорног сигнала, а као погодна обележја за њену параметризацију коришћени су мел-фреквенцијски кепстрални коефицијенти – *MFCC* (енгл. *Mel–Frequency Cepstral Coefficients*). Дељењем сваког од посматраних говорних сигнала на сегменте те одређивањем њима одговарајућих вектора изабраних обележја добија се потребан сет на основу чијег једног дела је извршена обука модела посматраних говорника. Други део претходно поменутог скупа, дисјунктан делу коришћеном за обуку а посматран кроз исти скуп обележја као и приликом обуке, употребљаван је приликом тестирања препознавача. Прегледом литературе може се закључити да постоји широк спектар потенцијалних начина моделовања говорника који се у општем случају могу применити, од детерминистичких преко стохастичких ка употреби неуронских мрежа [1,4], те комбинацији дискриминативних, *SVM* (енгл. *Support Vectors Machines – SVM*), и генеративних модела [5,6], у смислу добијања језгра односно кернела дискриминативног модела помоћу генеративних модела који одговарају класама секвенци од интереса.

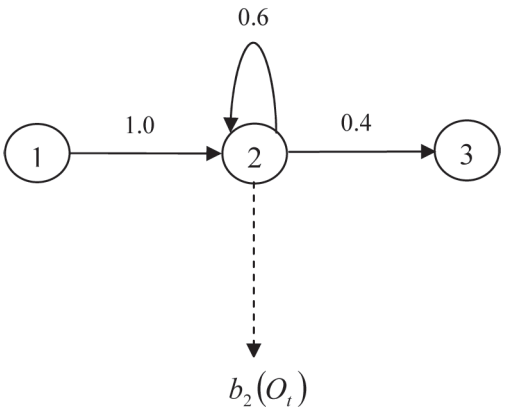
Обзиром на карактеристике говорне базе [8] која је коришћена приликом тестирања, по пет мушких и женских говорника при чему су изговорене садржине снимака на нивоу говорника међусобно фонетски различите, при-

мењено је препознавање независно од изговорене садржине. Препорука литературе је да се у оваквим околностима препознавања, када је сваки фонетски садржај изговорен само једанпут, користи модел мешавине Гаусових расподела - *GMM* (енгл. *Gaussian Mixture Model*). Руководећи се овим сазнањем у овом раду је говор сваког од посматраних говорника у говорној бази као и интервали пауза присутни у посматраним говорним узорцима моделован по једним оваквим моделом.

У наредним поглављима извршен је опис реализације препознавача говорника као и примењеног начина добијања говорних сигнала телефонског квалитета, потребних ради испитивања њиховог утицаја на тачност препознавања говорника. Након тога дат је приказ остварених резултата препознавања, извршена је њихова анализа и дата закључна разматрања.

2. ОПИС РЕАЛИЗОВАНОГ ПРЕПОЗНАВАЧА

У циљу конструкције препознавача говорника у овом раду је коришћен програмски алат за формирање скривених Марковљевих модела – *HTK* (енгл. ***H**idden Markov models **T**ool**K**it*) [7]. Пошто је овим алатом потребно конструисати препознавач који врши препознавање говорника независно од изговорене садржине, односно користити моделе мешавина Гаусових расподела за њихово моделовање, у овом раду је то урађено формирањем скривеног Марковљевог модела – *HMM* (енгл. *Hidden Markov Model*) са једним емитујућим стањем (слика 1).



Слика 1. – *Приказ коришћеног HMM-а.*

Расподела вектора обележја, O_t, унутар једног емитујућег, стања 2, што је у ознаци поменуте назначено еквивалентним индексом у једнакости која следи, описана је *GMM*-ом:

$$b_2(O_t)=\sum_{k=1}^{K} c_{2,k}\cdot N(O_t,\mu_k,\Sigma_k), \qquad (1)$$

при чему μ_k представља вектор средњих вредности, док Σ_k одговара коваријансној матрици посматране *k*-те *n* димензионалне Гаусове расподеле:

$$\mathbf{N}(O,\mu,\Sigma)=\frac{1}{\sqrt{(2\cdot\pi)^n\cdot|\Sigma|}}\cdot e^{-\frac{1}{2}(O-\mu)^T\cdot\Sigma^{-1}\cdot(O-\mu)}. \qquad (2)$$

Изабране су следеће вредности параметара коришћеног *GMM-a*:

- K=64, број коришћених Гаусових n димензионалних расподела,
- $c_{2,k}=\frac{1}{64}=0.015625$, јединствен тежински фактор за било коју од 64 Гаусове расподеле,
- n=39, димензионалност вектора обележја O_t.

Обзиром да већина говорних језика садржи од 30 до 40 фонема, одабрани број K је могао бити и мањи. Његова вредност се обично усклађује са бројем акустичких целина које се посматрају приликом обуке модела сваког од посматраних говорника. У овом раду је лабелирање коришћене говорне базе вршено на нивоу фонема односно субфонема што узимајући у обзир њен српски фонетски садржај чини укупно 42 акустичке целине за сваког од говорника. Очекујући финију статистичку представу поменутих акустичких целина у оквиру модела сваког од говорника, изабран је нешто већи број мешавина, K, од поменутог броја акустичких целина. Као мера статистичког доприноса сваке од акустичких целина моделу посматраног говорника у расподелу приказану изразом 1 уведен је одговарајући тежински фактор. Сходно постављеном циљу у оквиру овог рада нису вршена опсежнија испитивања претходно поменутих параметара у оквиру коришћеног *GMM*-а, но су за њих узете вредности како је претходно наведено. Дефинисањем претходно приказаних параметара (једнакости 1 и 2, слика1) у светлу коришћеног *HTK* те коришћењем адекватног конфигурационог фајла који садржи параметарске описе дигиталне обраде посматраних говорних узорака ради добијања репрезентативних вектора обележја, извршена је поставка прототипа коришћених модела. Приликом формирања вектора обележја који садржи првих 12 *MFCC* заједно са тзв. нултим мел-кепстралним коефицијентом, као и њихове прве односно друге изводе, посматрани су сегменти говорног сигнала који су прозорирани Хеминговом прозорском функцијом величине 25ms. При томе је прозорирање вршено на сваких 10ms. Обука модела извршена је говорним узорцима чија је учестаност одабирања усклађена са пропусним опсегом испитиваних телефонских канала, што резултује већом тачношћу препознавања говорних сигнала телефонског квалитета у односу на случај када је обука вршена говорним сигналима из изворне говорне базе чије је снимање вршено при учестаности одабирања од 22050Hz [8].

Иницијализација коришћених модела извршена је употребом *HTK* функције *HCompV*, након чега је извршена њихова процена применом функције *HERest*. Приликом тестирања тачности препознавача, препознавање говорника вршено је помоћу функције *Hvite*, а естимација перформанси помоћу функције *HResults*.

3. СИМУЛАЦИЈА ГОВОРНИХ СИГНАЛА ТЕЛЕФОНСКОГ КВАЛИТЕТА

Посматран из угла препознавача говорника, говорни сигнал представља предмет препознавања на основу ког

je потребно извршити идентификацију говорника. Стога је битно да карактеристике посматраних говорних сигнала буду што мање нарушене преносним каналом. У случају говорних сигнала телефонског квалитета поменуто нарушавање карактеристика је последица два општа узрока, и то:

– коришћеног кодека, те самим тим и наметнутог ограничења у погледу пропусног опсега посматраног телефонског канала и
– вероватноће појаве грешака у преносном каналу.

Сходно поменутим чињеницама, у овом раду је апстракција посматраних телефонских канала извршена путем симулације рада одговарајућих кодека и различитих квалитативних стања самог преносног канала у смислу вероватноће појаве грешака у њему. Ради тога је коришћена софтверска библиотека међународне телекомуникационе уније – *ITU-T STL2005* (енгл. ''*ITU-T Software Tool Library*'') [9].

Ова софтверска библиотека омогућава симулацију рада следећих кодека:

– *G.711*, битске брзине 64kbit/s, коришћен у *PSTN* (енгл. *Public Switched Telephone Network*) као и у *VoIP* (енгл. *Voice over Internet Protocol*), при чему су учестаности одабирања улазног односно излазног сигнала 8kHz,
– *G.722*, широкопојасни кодек говора код кога је учестаност одабирања улазног/излазног сигнала 16kHz, за примене у *ISDN* (енгл. *Integrated Services Digital Network*) са подржаним битским брзинама од 64, 56 односно 48kbit/s,
– *RPE-LTP* (енгл. *Regular Pulse Excitation – Long Term Predictor*), битске брзине 13kbit/s при учестаности одабирања улазног/излазног сигнала од 8kHz, *GSM* (енгл. *Global System for Mobile communications,* прекло од фр. *Groupe Spécial Mobile*) кодек,
– *G.726* кодек примењен у *VoIP*, који омогућава битске брзине од 40, 32, 24 и 16kbit/s при учестаности одабирања улазног/излазног сигнала од 8kHz,
– *G.727* кодек примењен такође у *VoIP* који се може сматрати унапређеном верзијом претходно поменутог кодека по препоруци *G.726* пошто обезбеђује идентичне битске брзине, с тим да се одабиром броја бита језгра (енгл. *core bits* - N_c), $N_c = \{2,3,4\}$, односно бита проширења (енгл. *enhancement bits* - N_e), $N_e = \{0,1,2,3\}$ може бирати кодна структура којом ће се циљана битска брзина постићи.

Поред наведених типова кодека у раду је испитан и утицај кодека по *ITU-T* препоруци *G.729* из Анекса ***C***+, при чему је коришћена његова софтверска имплементација доступна у електронском фајлу придруженом овом анексу [10]. Користећи *CS-ACELP* (енгл. *Conjugate-Structure Algebraic-Code-Excited Linear Prediction*) поступак при учестаности одабирања улазног/излазног сигнала од 8kHz овим кодеком могуће је постићи једну од три битске брзине – 11.8, 8.0 или 6.4kbit/s.

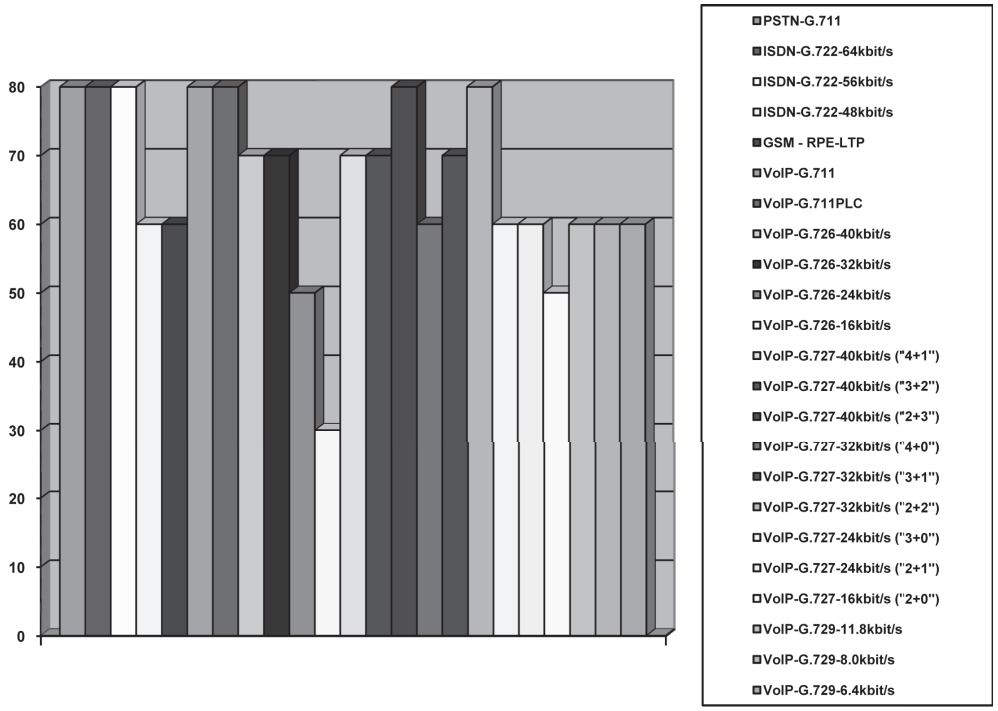
Ради симулације грешака при преносу коришћен је програм ***eiddemo.exe*** из *STL2005*, који омогућава генерисање битских грешака и појаве брисања рамова у улазном битском низу. Адекватним одабиром параметара при позиву овог програма контролисана је вероватноћа појаве ових грешака као и ниво њиховог спорадичног појављивања.

4. ПРИКАЗ РЕЗУЛТАТА ПРЕПОЗНАВАЊА ГОВОРНИКА

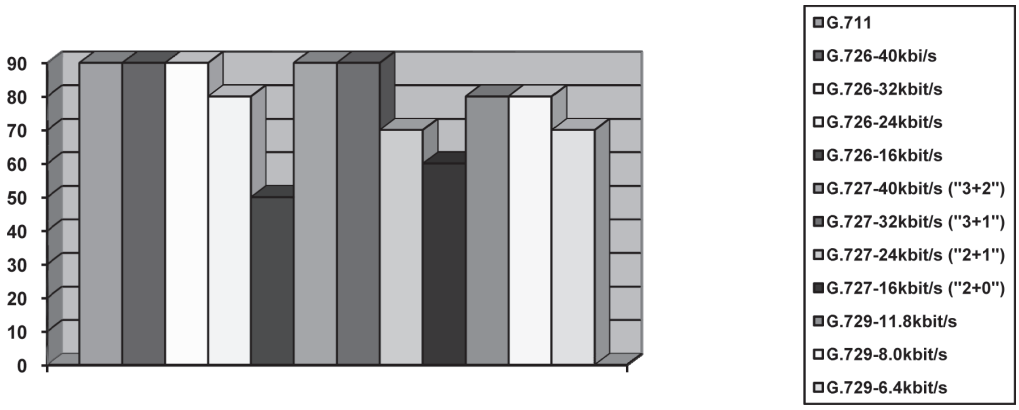
Располажући широком палетом кодека доступних у коришћеној софтверској библиотеци *STL2005,* те варирањем вероватноће појаве грешака при преносу испитан је утицај различитих стања посматраних телефонских канала на тачност аутоматског препознавања говорника. При томе је појава случајних грешака контролисана параметрима вероватноће појаве битских грешака, *BER*=[0,0.001], односно вероватноће појаве брисања рамова, *FER*=[0,0.1], при инкрементима 0.0001 тј. 0.01. Фактори који описују спорадичну појаву битских грешака односно брисања рамова варирани су кроз три вредности, *BER_gamma=FER_gamma*={0, 0.50, 0.99}, које одговарају каналима са потпуно случајном, умерено спорадичном односно потпуно спорадичном појавом одговарајућих грешака.

У већини случајева, сем при тестирању *RPE-LTP* кодека и поступка за прикривање пакетских губитака (енгл. *Packet Loss Concealment* – ***PLC***) при употреби *G.711* кодека, *G.711-PLC*, уметање грешака је симулирано након примене одговарајућег кодера. За поменута два случаја то је у складу са препоруком из коришћене литературе [9] вршено након декодовања.

Изворни снимци у коришћеној говорној бази одликују се учестаношћу одабирања од 22050Hz при резолуцији од 16 бита. Обука и тестирање препознавача над оваквом говорном базом резултује 100%-ом оствареном идентификацијом говорника. Сходно типу кодека који карактерише испитивани телефонски канал, вршено је смањење учестаности одабирања у употребљеној говорној бази. Као последица сужавања посматраног спектралног опсега у односу на изворне говорне сигнале као и примене одговарајућих кодека евидентирано је смањење у оствареној тачности препознавања. Сада је тачност идентификације имала максималну вредност од 90% [8]. Овај максимум остварене процентуалне тачности евидентиран је на излазима *PSTN-G.711*, *VoIP-G.711*, *VoIP-G.726-40*kbit/s и *VoIP-G.727-40*kbit/s (''4+1'') симулираних телефонских канала без присутних грешака приликом преноса. Повећавање вероватноће појаве случајних грешака углавном је резултовало смањењем процента тачно препознатих говорника (слика 2).



Слика 2. – *Процентуална тачност препознавања при FER=0.05, BER=BER_gamma=FER_gamma=0, у зависности од типа кодека посматраног телефонског канала.*



Слика 3. – *Процентуална тачност препознавања говорника у зависности од примењеног кодека посматраних VoIP канала при BER=BER_gamma=FER_gamma=0, FER=0.05 и временском кашњењу од 100ms.*

Резултати препознавања такође показују [8] да постоје ситуације када повећани број грешака не резултује у смањењу тачности препознавања. Ово се може тумачити тиме да тачност препознавања зависи такође и од места дејства грешака у преношеном говорном сигналу. Уколико се грешке догоде у периодима пауза, што може бити последица њихове спорадичне појаве, или у деловима односно на фонемима унутар посматраног говорног узорка који су мање битни са становишта исправног препознавања у односу на посматрану говорну базу, препознавач може показивати исту тачност без обзира на повећан број преносних грешака.

Интересантно је приметити да *VoIP-G.729* телефонски канал показује, у извесној мери, отпорност на појаву грешака при преносу. Наиме тачност препознавања на излазу оваквог канала без присутних грешака износи 60%, док даље повећавање нивоа грешака у само малом броју случајева доводи до њених апсолутних варијација од ±10% [8].

Појава константног временског кашњења говорних рамова нарочито може доћи до изражаја у *VoIP* телефонским каналима када достиже вредност до 400ms [11]. Оваква изражена кашњења доводе до значајне појаве еха на пријемној страни ових телефонских канала, што је ову појаву учинило интересантном са становишта њеног утицаја на тачност препознавања говорника. У експериментима су применом ефекта *Delay*/*Echo*-*Simple* програмског пакета *Sony Sound Forge 9.0*, вршене симулације

кашњења од 25, 100, 200 и 400ms на излазима разматраних *VoIP* канала са присутним одређеним степеном грешака у преносу. Вршено је поређење утицаја поменутих вредности кашњења, у односу на случај без присутног кашњења, на тачност препознавања говорника на њиховом излазу. Утврђено је да појава еха, сем у ретким случајевима кашњења од 25ms, доприноси бољој тачности препознавања у односу на посматрани канал без присутног еха (слика 3).

5. ЗАКЉУЧАК

Коришћењем софтверског алата *STL2005* експериментално је утврђено да како сама примена кодека тако и појава грешака у преносу, опште посматрано доприноси деградацији процента тачности препознавања говорника на излазима симулираних телефонских канала. Ово смањење тачности израженије је при већим вероватноћама појаве грешака при преносу, као и при коришћењу кодека мањих битских брзина. У циљу унапређења тачности препознавања један од праваца даљих истраживања огледа се и у конструкцији препознавача прилагођеног телефонском каналу на чијем излазу се врши препознавање. Ово би подразумевало не само прилагођеност препознавача на учестаност одабирања која је у складу са пропусним опсегом посматраног телефонског канала, како је урађено у овом раду, него и на кодек примењен у њему. На овај начин би се препознавач обучавао говорним сигналима који се добијају на излазу кодека посматраног телефонског канала.

Додатна анализа утицаја појаве еха у *VoIP* телефонији показала је да у већини случајева појава временског кашњења поправља тачност препознавања говорника. Природа настанка еха има за последицу преклапање директног и временски закаснелог сигнала што у зависности од величине кашњења резултује већим или мањим позитивним односно негативним истицањем дистинктивних особина фонема односно субфонема посматраних говорних узорака у односу на коришћену говорну базу.

Детаљнија анализа утицаја величине временског кашњења при појави еха као и одабир параметара (једнакост 1) приликом обуке модела посматраних говорника у циљу њихове што боље обучености на посматрани телефонски канал представљају могуће правце даљих истраживања у оквиру претходно описаног испитивања.

Овај рад је резултат истраживања на пројекту "Говорна комуникација човек машина", ТР-11001, који се одвија под покровитељством Министарства за науку и технолошки развој Републике Србије.

Коришћена говорна база за потребе испитивања у овом раду представља део говорне базе развијене у оквиру АлфаНум тима са Факултета техничких наука у Новом Саду.

ЛИТЕРАТУРА

[1] Campbell P. Joseph, Jr., (1997). "Speaker recognition: a tutorial", *Proceedings of IEEE,* Vol. 85, No. 9, pp.1437-1462.

[2] Делић Д. Владо, Сечујски С. Милан, Јаковљевић М. Никша, (2008). ‘’Акциони модел говорне комуникације човек-машина’’, 16. Телекомуникациони форум *ТЕЛФОР 2008*, pp. 680-683, Србија, Београд, новембар 25.-27., 2008.

[3] Bimbot Frédéric, Bonastre Jean-François, Fredouille Corinne, Gravier Guillaume, Magrin- Chagnolleau Ivan, Meignier Sylvain, Merlin Teva, Ortega-Garcia Javier, Petrovska-Delacrétaz Dijana, and Reynolds A. Douglas, (2004). "A Tutorial on Text-Independent Speaker Verification", *EURASIP Journal on Applied Signal Processing 2004:4*, pp. 430-451.

[4] Wildermoth Richard Brett, (2001). "Text-Independent Speaker Recognition Using Source Based Features", *M. Phil. Thesis, Griffith University, Brisbane, Australia*, 2001.

[5] Moreno J. Pedro, Ho P. Purdy, (2003). "A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels", *Published in Eurospeech 2003, 1-4 September 2003, Geneva, Switzerland*.

[6] Quan Le, Bengio Samy, (2002). «Hybrid generative-discriminative models for speech and speaker recognition», Technical Report IDIAP-RR 02-06, *IDIAP*, March 2002.

[7] Young Steve, Evermann Gunnar, Gales Mark, Hain Thomas, Kershav Dan, Liu Xunying (Andrew), Moore Gareth, Odell Julian, Ollason Dave, Povey Dan, Valtchev Valtcho, Woodland Phil, (2009). ‘’The HTK Book (for HTK Version 3.4)’’, ©COPYRIGHT 1995-1999 *Microsoft Corporation*, ©COPYRIGHT 2001-2009 *Cambridge University Engineering Department*.

[8] Јокић Иван, (2010). ‘’Утицај телефонских канала на аутоматско препознавање говорника’’, *Магистарски рад, Факултет техничких наука – Нови Сад, јун 2010*.

[9] ITU-T User’s Group on Software Tools, (2005). ‘’ITU-T Software Tool Library 2005 User’s Manual’’, *Geneva, August 2005*.

[10] ITU-T Recommendation G.729, (2007). ‘’Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)’’, *ITU-T Rec. G.729 (01/2007)*.

[11] http://voip.about.com/od/glossary/g/echo.htm

Иван Јокић, Факултет техничких наука, Универзитет у Новом Саду
Контакт: e-mail: IBAHJOKIh@gmail.com
Области интересовања: обрада сигнала, препознавање, говорне технологије, аутоматско препознавање говорника.

Владо Делић, Факултет техничких наука, Универзитет у Новом Саду
Контакт: e-mail: vdelic@uns.ac.rs
Области интересовања: акустика, обрада аудио сигнала, говорне технологије, комуникација човек-рачунар.

Никша Јаковљевић, Факултет техничких наука, Универзитет у Новом Саду
Контакт: e-mail: jakovnik@uns.ac.rs
Области интересовања: обрада сигнала, препознавање говора, машинско учење.

Милан Добровић, Телеком Србија
Контакт: e-mail: milando@telekom.rs
Области професионалног интересовања: дигитална обрада слике (филтрирање слике), дигитална обрада говорног сигнала (препознавање говорника), рачунарске мреже.

Стеван Јокић, Факултет техничких наука, Универзитет у Новом Саду
Контакт: e-mail: stevan.jokic@gmail.com
Области професионалног интересовања: обрада сигнала (говора и биомедицинских), мреже сензора.