Утицај телефонског канала на аутоматско препознавање говорника и методе адаптације

Иван Д. Јокић, Стеван Д. Јокић, и Владо Д. Делић, *Члан, IEEE*

***Садржај* — Овај рад приказује кључне резултате добијене испитивањем утицаја телефонских канала на тачност аутоматског препознавања говорника и анализира могућности његове адаптације. Моделовање говорника и конструкција препознавача базирани су на примени скривених Марковљевих модела. Приликом симулирања телефонског квалитета говорних сигнала руководило се чињеницом да су главни чиниоци телефонског канала који утичу на промену изворних карактеристика говорног сигнала: тип коришћеног кодека и вероватноћа појаве грешака током преноса. Такође је извршено испитивање утицаја појаве еха у Интернет телефонији на тачност аутоматског препознавања говорника.**

***Кључне речи* — Аутоматско препознавање говорника, exo у VoIP, GMM, HMM, HTK, ITU-T STL2005.**

I. Увод

АУТОМАТСКО препознавање говорника, као примена аутоматизованог поступка у циљу разликовања односно распознавања особа на основу њиховог говора, у својој основи подразумева верификацију односно идентификацију говорника. Обзиром на то да ли је текстуална садржина посматраног говора очекивана или не, препознавање може бити зависно или независно од текста.

Говор представља сложен акустички сигнал који је резултат семантичких, лингвистичких, артикулаторних и акустичких трансформација [1], [2]. Зависно које се од њих сматрају доминантним, разликовање говорника се врши на основу обележја високог односно ниског нивоа [3]. Обележја ниског нивоа последица су акустичких трансформација унутар вокалног тракта посматраног говорног субјекта. Често се, што је примењено и у овом раду, као таква обележја користе мел-фреквенцијски кепстрални коефицијенти (енгл. ***M**el-**F**requency **C**epstral **C**oefficients*) – MFCCs. Они репрезентују спектралну обвојницу посматраног говорног сигнала, која носи информације о боји гласа припадајућег говорника. Говорна обележја високог нивоа носе информацију о прозодијским особинама говора разматраног говорника (висина тона, динамика и ритам). Вектори обележја, говора посматраних говорника, се уобличавају у одговарајуће моделе. Препознавање говорника врши се поређењем формираних модела са тест говорним узорцима. Велику примену у пракси имају стохастички модели, кроз модел мешавине Гаусових расподела (енгл. ***G**aussian **M**ixture **M**odel*) – GMM односно скривени Марковљев модел (енгл. ***H**idden **M**arkov **M**odel*) – HMM, као и одлучивачи засновани на неуронским мрежама [4].

Постојање разних видова телефонских система отвара могућност приступа великом броју услуга коришћењем телефона. У већини случајева потребно је обезбедити ауторизованост оваквог приступа. Глас представља једно од биометријских обележја тако да је у овом случају ауторизацију приступа потребно извршити анализом и препознавањем гласа на излазу из телефонског канала. Саставни делови канала - кодер, преносни систем и декодер - представљају елементе који нарушавају квалитет изворног гласа и као такви отежавају препознавање говорника на излазу система. У наведеним чињеницама видела се оправданост и потреба истраживања које је описано у наставку рада.

Ради приказа позадине извршеног експерименталног истраживања, наредна два поглавља описују реализовани препознавач говорника и начин на који је симулиран телефонски квалитет говорног сигнала над којим су вршени експерименти препознавања. Након тога дат је приказ остварених резултата препознавања, извршен осврт на поступке повећања робусности препознавања и изведена закључна разматрања.

II. Опис реализације препознавача говорника

Говорна база, коришћена приликом обуке модела и тестирања препознавача, садржи изговоре по 5 мушких и женских говорника. Она представља део говорне базе развијане у оквиру Алфанум тима на Факултету техничких наука у Новом Саду [5]. Говорници међусобно изговарају исте текстуелне садржине при чему је свака од њих једанпут изговорена од стране једног говорника. У оваквом случају погодно је применити препознавање говорника независно од

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

Иван Д. Јокић, Факултет техничких наука у Новом Саду, Трг Доситеја Обрадовића 6, 21000 Нови Сад, Србија (телефон: 381-64-3526245, e-mail: IBAHJOKIh@gmail.com).

Стеван Д. Јокић, Факултет техничких наука у Новом Саду, Трг Доситеја Обрадовића 6, 21000 Нови Сад, Србија (e-mail: stevan.jokic@gmail.com).

Владо Д. Делић, Факултет техничких наука у Новом Саду, Трг Доситеја Обрадовића 6, 21000 Нови Сад, Србија (телефон: 381-21-4852533, e-mail: vdelic@uns.ac.rs).

изговорене садржине и моделовање говора сваког од говорника као и периода пауза у говору извршити одговарајућим GMM.

Обука модела и имплементација препознавача говорника извршени су коришћењем програмског алата HTK (од енгл. ***H**idden Markov Models **T**ool**K**it*) [6] за формирање HMM. Обзиром да је у циљу моделовања било довољно користити GMM, употребом HTK то је учињено формирањем HMM-а који има једно емитујуће стање [7].

Расподела вектора обележја унутар јединог емитујућег стања, описана је као:

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

при чему μ_k и Σ_k представљају вектор средњих вредности и коваријансну матрицу *k*-те *n*-димензионалне Гаусове расподеле:

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

Обзиром на обим предвиђеног експерименталног истраживања није вршена детаљна анализа одабира параметара коришћених модела већ су за њих изабране јединствене вредности: $K = 64$ и $c_{2,k} = 1/64$, [8].

Обука модела односно тестирање тачности препознавања вршени су над дисјунктним деловима коришћене говорне базе [7]. Издвајање обележја вршено је над сегментима говорног сигнала добијеним његовим прозорирањем Хаминговим прозорским функцијама трајања 25 ms, међусобно помереним за по 10 ms. Као обележја коришћени су првих 12 MFCCs заједно са нултим мел-кепстралним коефицијентом C_0, као и њихови први односно други изводи. На овај начин моделовање говорника извршено је мешавинама Гаусових 39-димензионалних расподела.

III. Симулација телефонског квалитета говорних сигнала

Телефонски канал својим особинама, пропусним опсегом и самим преносним карактеристикама, изазива промене у говорном сигналу који се доводи на његов улаз. Са становишта преноса говорног сигнала у дигиталном облику ове промене могу се сматрати последицом примењеног кодека и вероватноће појаве грешака у преносу, односно робусности декодовања на грешке у дигиталном преносу, како је и урађено у овом раду. Ради симулирања рада одговарајућих кодека као и различитих вероватноћа појаве грешака у телефонском каналу коришћена је софтверска библиотека ITU-T STL2005 (енгл. ***ITU-T S**oftwa**T**ool **L**ibrary **2005***) [9]. Помоћу ове библиотеке програмских модула било је могуће испитати утицај следећих кодека на тачност аутоматске идентификације говорника:

G.711 – кодек на 64 kbit/s при учестаности одабирања 8 kHz. Примењује се у PSTN (од енгл. ***P**ublic **S**witched **T**elephone **N**etwork*) и VoIP (од енгл. ***V**oice **o**ver **I**nternet **P**rotocol*).

G.722 – широкопојасни кодек говора при учестаности одабирања 16 kHz. У зависности од мода рада овај кодек постиже једну од три битске брзине – 64, 56 или 48 kbit/s. Примењује се у ISDN (од енгл. ***I**ntegrated **S**ervices **D**igital **N**etwork*).

RPE-LTP (од енгл. ***R**egular **P**ulse **E**xcitation – **L**ong **T**erm **P**redictor*) – кодек на 13 kbit/s при учестаности одабирања 8 kHz. Ово је један од кодека који се користе у GSM (од енг. ***G**lobal **S**ystem for **M**obile **C**ommunications*) телефонији.

G.726 – кодек који при учестаности одабирања 8 kHz постиже битске брзине – 40, 32, 24 или 16 kbit/s. Примењује се у VoIP телефонији.

G.727 – VoIP кодек који избором броја бита језгра (енгл. *core bits* - N_c), $N_c = \{2,3,4\}$, односно бита проширења (енгл. *enhancement bits* - N_e), $N_e = \{0,1,2,3\}$, постиже једну од битских брзина – 40, 32, 24 или 16 kbit/s, при учестаности одабирања 8 kHz.

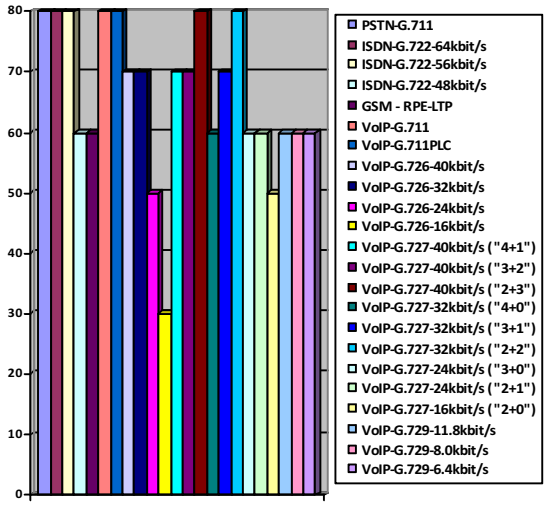
Такође, у раду је испитан и утицај кодека из Анекса C+ ITU-T препоруке G.729 [10]. Као симулација његовог рада коришћен је електронски фајл придружен овом анексу. Овај кодек се примењује у VoIP телефонији, при чему коришћењем CS-ACELP (од енгл. ***C**onjugate-**S**tructure **A**lgebraic-**C**ode-**E**xcited **L**inear **P**rediction*) поступка при учестаности одабирања улазног/излазног говорног сигнала од 8 kHz и зависно од изабраног мода рада, може постићи битске брзине 11.8, 8.0 или 6.4 kbit/s.

Симулација појаве грешака при преносу спроведена је употребом програмског модула eiddemo.exe из софтверске библиотеке ITU-T STL2005. Овим модулом извршено је симулирање појаве битских грешака и брисања говорних рамова. Вероватноће појава битских грешака односно брисања рамова вариране су у интервалима *BER* = [0, 0.001] односно *FER* = [0, 0.1] при инкрементима од 0.0001 тј. 0.01. Такође вариран је фактор спорадичности појаве ових грешака, *BER_gamma* = *FER_gamma* = {0, 0.50, 0.99}. Ове вредности редом одговарају потпуно случајној, умерено спорадичној односно потпуно спорадичној појави одговарајућих грешака. Примена овог модула при симулацији одговарајућег телефонског канала, у већини случајева вршена је након кодера а пре самог декодера. Једина одступања извршена су, по препоруци коришћене литературе [9], при тестирању RPE-LTP кодека и G.711-PLC (од енг. ***P**acket **L**oss **C**oncealment*) поступка за прикривање пакетских губитака при коришћењу G.711 кодека. Тада је симулирање грешака извршено након претходно поменутог GSM кодека односно након G.711-PLC поступка. При симулирању брисања рамова у PSTN и ISDN коришћена је величина рама од 32 ms, на излазу RPE-LTP декодера посматрани су рамови величине 20 ms, док је при испитивању VoIP канала узето да они износе 30 ms сем при испитивању G.729 кодека када је њихова величина износила 10 ms [7].

IV. Резултати препознавања говорника

Поступак испитивања процентуалне тачности аутоматске идентификације говорника започет је над изворним говорним сигналима из коришћене говорне базе, који се одликују учестаношћу одабирања 22050 Hz и резолуцијом од 16 bit. Након обуке модела говорника и тестирања препознавача забележена је тачност идентификације говорника од 100%. Ради испитивања тачности препознавања над говорним сигналима телефонског квалитета, учестаност одабирања говорних сигнала у коришћеној говорној бази је смањена у складу са пропусним опсегом телефонског канала који се испитује. Тако је за све испитиване телефонске канале сем за ISDN са примењеним G.722 кодеком, учестаност одабирања смањена на 8 kHz. Ради испитивања G.722 кодека изворни сигнали су децимирани на 16 kHz.

Као последица сужавања спектралног опсега и примене одговарајућих кодека у посматраним телефонским каналима тачност идентификације говорника на њиховим излазима показивала је максималну вредност од 90%. Ова тачност забележена је на излазима PSTN-G.711, VoIP-G.711, VoIP-G.726-40 kbit/s и VoIP-G.727-40 kbit/s (“4+1”), без присутних грешака при преносу [7].

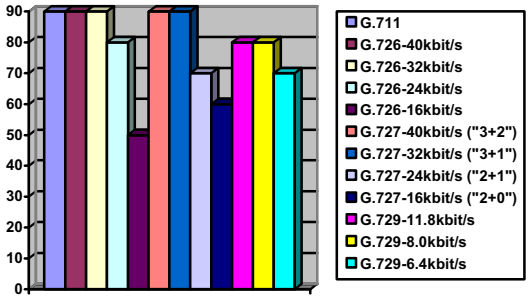


Сл. 1. Процентуална тачност идентификације говорника у зависности од типа телефонског канала и примењеног кодека при $FER = 0.05$, $BER = BER\_gamma = FER\_gamma = 0$.

Повећање вероватноће појаве случајних грешака у већини случајева резултује смањењем процента тачно идентификованих говорника, Сл. 1. Примећено је да повећана вероватноћа грешке, нарочито спорадичне појаве грешака, не резултује увек смањењем тачности препознавања [7]. У оваквим случајевима грешке могу бити лоциране у деловима говорног сигнала који су мање битни са становишта исправне идентификације говорника. Тачност идентификације на излазима PSTN-G.711, ISDN-G.722 битских брзина 64 и 56 kbit/s, VoIP-G.711, VoIP-G.726 – 40 kbit/s, VoIP-G.727 – 40 kbit/s (“4+1”), са присутним грешкама при преносу показивала је сличне вредности, 80% тачности. У односу на испитиване телефонске канале примећен је најнижи проценат тачности препознавања на излазима VoIP-G.726 и G.727 битске брзине 16 kbit/s. Тачност препознавања на излазу VoIP-G.729 у већини случајева показивала је константност од 60% са ретким одступањима од $\pm 10\%$ у односу на поменуту вредност [7].

A. Утицај појаве еха у VoIP

Појава константног временског кашњења рамова говора нарочито изражена у VoIP када достиже вредности до 400 ms [11] доводи до значајне појаве еха на излазу овог канала. Из тог разлога је било интересантно испитати утицај појаве овог ефекта на тачност идентификације говорника. Симулација појаве еха извршена је применом ефекта Delay/Echo-Simple, програмског пакета Sony Sound Forge 9.0, на излазу разматраног VoIP канала. Симулиране су вредности кашњења од 25, 100, 200 и 400 ms, при чему је задржавана и одређена вероватноћа грешке у каналу.



Сл. 2. Процентуална тачност идентификације говорника на излазу VoIP телефонског канала, у зависности од примењеног кодека при $FER = 0.05$, $BER = BER\_gamma = FER\_gamma = 0$ и временском кашњењу од 100 ms.

Изузев за поједине експерименте у којима је појава еха била проузрокавана временским кашњењима од 25 ms, утврђено је да појава овог ефекта доприноси бољој тачности препознавања у односу на канал без његовог присуства [7]. Ради илустрације утицаја еха при временском кашњењу од 100 ms, Сл.1 и Сл.2 приказују резултате препознавања за исте вероватноће појаве грешака у разматраним телефонским каналима. Евидентно је да резултати на Сл. 3, који одговарају VoIP каналима са присутним кашњењем од 100 ms, показују веће вредности односно појава еха у овом случају је допринела тачнијем препознавању.

V. Могућности адаптације препознавача

Опсежнија анализа претходно приказаних резултата захтева већу говорну базу као и више различитих тестова. У таквим условима, побољшање тачности препознавања се може остварити фонетски богатијом

говорном базом, што је у [12] резултовало већом тачношћу препознавања. На овај начин се утицај канала превазилази бољом обуком модела говорника.

Постоји низ начина побољшања препознавања одговарајућим нормализационим методама резултата препознавања, односно применом одговарајућих трансформационих поступака на векторе обележја говорних сегмената који су предмет препознавања [13]. На овај начин елиминишу се фактори који отежавају препознавање, односно покушавају се елиминисати утицаји појаве грешака у каналу.

Примена различитих нормализационих метода: *Z-norm*, *WMAP*, *T-norm* односно *D-norm* [13], подразумева одговарајућу процену расподеле погрешних резултата препознавања те на тај начин корекцију самог резултата препознавања. За разлику од њих, трансформационим поступцима: *CMS*, *RASTA*, *H-norm*, *C-norm*, *PCA*, *LDA* односно *NLDA* [13], врши се трансформација расподела вектора обележја, на основу којих се врши препознавање говорника, у циљу елиминације фактора који отежавају односно ометају поступак препознавања. У случају примене препознавања на излазима телефонских канала под тим ометајућим факторима би се могле подразумевати карактеристике преносног телефонског канала.

Стога је, конкретно за телефонски канал, развијан *C-norm* [13] поступак који пресликава векторе обележја из канално зависног простора обележја у простор обележја који је независан од посматраног телефонског канала. Над тако добијеним простором обележја независним од посматраног канала се онда врши препознавање.

VI. Закључак

Приказани резултати показују да сам кодек и појава грешака при преносу у посматраном телефонском каналу отежавају поступак аутоматске идентификације говорника на његовом излазу. Обзиром да кодек представља фиксни параметар посматраног телефонског канала то би се и његов утицај могао узети у обзир приликом обуке модела говорника. На овај начин би у будућим конструкцијама препознавача говорника за одређени телефонски канал одговарајући модели били тренирани говорним сигналима добијеним на излазу кодека примењеног у посматраном телефонском каналу.

Додатно испитивање утицаја појаве еха у VoIP показало је да у највећем броју експеримената остварена тачност препознавања говорника показује веће вредности у односу на посматрани канал без његовог присуства.

У литератури се у последње време појављују различити поступци којима се може утицати на робусност препознавача говорника. Неки од њих су укратко анализирани у овом раду, па то усмерава даља истраживања у правцу конструкције препознавача говорника који ће бити адаптиван у односу на тест окружење у ком се очекује његова примена.

Литература

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

[2] V. D. Delić, M. S. Sečujski, N. M. Jakovljević, "Akcioni model govorne komunikacije čovek-mašina," 16. Telekomunikacioni forum *TELFOR 2008*, Srbija, Beograd, novembar 25.-27., 2008., str. 680-683.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Margin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing 2004:4*, 2004, pp. 430-451.

[4] B. R. Wilderrnoth, "Text-Independent Speaker Recognition Using Source Based Features," *M. Phil. Thesis, Griffith University Brisbane, Australia*, 2001, pp. 28-37.

[5] И. Д. Јокић, Т. Н. Добријевић, Н. М. Јаковљевић, В. Д. Делић, "Опис говорне базе за препознавање говорника на српском језику," 17. Телекомуникациони форум *ТЕЛФОР 2009*, Србија, Београд, новембар 24.-26., 2009., Зборник радова, ISBN 978-86-7466-375-2, стр. 1109-1112.

[6] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershav, X. (A.) Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, "The HTK Book (for HTK Version 3.4)," ©COPYRIGHT 1995-1999 *Microsoft Corporation*, ©COPYRIGHT 2001-2009 *Cambridge University Engineering Department*.

[7] И. Јокић, "Утицај телефонских канала на аутоматско препознавање говорника," *Магистарски рад*, *Факултет техничких наука – Нови Сад*, *2010.*

[8] I. D. Jokić, V. D. Delić, N. M. Jakovljević, M. M. Dobrović and S. D. Jokić, "Accuracy of Automatic Speaker Recognition for Telephone Speech Signal Quality," in *Proc. 8th International Symposium on Intelligent Systems and Informatics*, SISY 2010, September 10-11, 2010, Subotica, Serbia, ISBN: 978-1-4244-7395-3, pp. 579-582.

[9] ITU-T User's Group on Software Tools, "ITU-T Software Tool Library 2005 User's Manual," *Geneva, August 2005*.

[10] ITU-T Recommendation G.729, "Coding of Speech at 8kbit/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

[12] P. Staroniewicz, "Speaker Recognition for VoIP Transmission Using Gaussian Mixture Models," *Proceedings of the 4th International Conference on Computer Recognition Systems, CORES'05*, Volume 30/2005, pp. 739-745, DOI: 10.1007/3-540-32390-2_87, May 22-25, 2005, Rydzyna Castle, Poland 2005.

[13] D. Wu, B. Li and H. Jiang, "Normalization and Transformation Techniques for Robust Speaker Recognition," pp. 311-330, Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert, ISBN 987-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria.

Abstract

This paper analyze the key influences of telephone channels to the accuracy of automatic speaker recognition and possibility of its adaptation. Speakers modeling and design of the speaker recognizer are based on the use of Hidden Markov Models. During the simulation of telephone quality voice signals we are guided by the fact that the main factors of telephone channel, which influencing on change the characteristics of the original speech signal, are the type of codec used and the probability of errors during transmission. Also is examined the impact of echo phenomena in Internet telephony to the accuracy of automatic speaker recognition.

IMPACT OF TELEPHONE CHANNEL TO AUTOMATIC SPEAKER IDENTIFICATION AND ADAPTATION METHODS

Ivan D. Jokić, Stevan D. Jokić, Vlado D. Delić