Поређење карактеристика неколико метода за детекцију говорне активности

Никша Јаковљевић, Драгиша Мишковић, Жељен Трповски
Факултет техничких наука
Универзитет у Новом Саду
Нови Сад, Србија
jakovnik@uns.ac.rs, dragisa@uns.ac.rs, zeljen@uns.ac.rs

Сажетак — **Овај рад даје преглед перформанси неколико метода за детекцију говорне активности које су процењене на говорној бази која садржи реченице српског језика на које је додат реалан шум. Разматране су следеће врсте шума: звук у аутомобилу, аутобусу и возу који се креће, звук у кухињи, канцеларији и фабрици, на метро станици, улици и терену, звук поред потока и веш машине која ради. Најбољи резултати добијен су за систем где се проблем детекције третира као проблем препознавања облика и где се користе обележја која одвојено моделују побуду и вокални тракт.**

Кључне речи- Обрада говора, Детекција говорне активности; Реални шумови; DEMAND;

I. Увод

Под детекцијом говорне активности (ДГА) подразумева се лоцирање сегмената аудио сигнала у којима постоји говор. ДГА заузима значајно место у кодовању говора [1], аутоматском препознавању говора [2], аутоматском препознавању говорника [3] и унапређењу квалитета говорног сигнала [4] где се обично појављује као почетни блок ових система. Због своје широке примене развијен је велики број различитих алгоритама којима се практично реализује ДГА.

Један од најједноставнијих приступа ДГА јесте детекција на основу нивоа енергије сигнала. У случају аудио сигнала који су снимани у студијским условима ово представља веома ефикасан метод, пошто сегменти са малом енергијом одговарају тишини, а сегменти са већом енергијом говору. У овом раду је анализиран приступ који је описан у [3] где се при израчунавању енергије над једним анализаторском прозором из даље анализе изоставља енергија коју даје једносмерна компонента у оквиру тог прозора, а праг се израчунава у односу на максималну енергију посматраног сигнала. У наставку овог рада овај приступ носи ознаку А1.

Поред нивоа укупне енергије на сегменту, често се посматра и енергија на појединачним подопсезима, амплитудски спектар као и број пресека са нулом. На пример у оквиру ИТУ-Т стандарда са ознаком G.729B дефинисан је ДГА који укупну енергију, енергију на ниским учестаностима, линијски спектар и број пресека са нулом тренутног сегмента пореди са одговарајућим вредностима тих параметара процењеним на неговорним сегментима. У зависности од вредности ових разлика доноси се одлука да ли дати сегмент садржи говор или не [1], [5]. У наставку овог рада овај алгоритам носи ознаку А2.

Трећи алгоритам који је разматран у оквиру овог рада јесте статистички метод који је предложен у [6]. За сваки одбирак дискретне Фуријеове трансформације (ДФТ) посматраног сегмента израчунава се количник вероватноће да он садржи само шумну компоненту и вероватноће да садржи и говорну и шумну компоненту. Потом се за дати сегмент сигнала израчунава количник вероватноћа као геометријска средина количника који су добијени за појединачне одбирке ДФТ. Овај количник придружен сегменту пореди се са унапред задатим прагом на основу ког се одређује коју је хипотезу потребно прихватити. Да би се избегло одбацивање сегмената у којима је говор тих (на крајевима реченица и речи) имплементирана је тзв. *hang-over* процедура којом се помоћу скривеног Марковљевог модела омогућава постепен прелаз их хипотезе да постоји зашумљен говор у хипотезу да постоји само шум. Успех ове методе у великој мери зависи од методе која је коришћена за процену шума. У овом раду су искоришћене две методе: метода базирана на минималним статистикама [7] и метода базирана на непристрасној процени минималне средње квадратне грешке [8]. Прва верзија алгоритма у овом раду носи ознаку А3, а друга А4.

Четврти алгоритам који је разматран у оквиру овог рада предложен је у [3]. Алгоритам започиње уклањањем шума из зашумљеног говорног сигнала помоћу Винеровог (*Wiener*) филтра. Основни циљ овог корака јесте да се повећа разлика између говорних и неговорних сегмената, те при избору параметара којима се регулише понашање Винеровог филтра није вођена пажња о степену оштећења говорног сигнала. Овај сигнал се дели на сегменте и за сваки од њих се израчунава енергија и мел фреквенцијски кепстрални коефицијенти (МФКК). Потом се МФКК за 10% сегмената са највишом енергијом користе за обуку модела говорног сигнала, а 10% сегмената са најмањом енергијом за обуку модела „тишине". За моделовање искоришћене су мешавине Гаусових расподела, а за процену параметара модела кластеризација методом *k*-средњих вредности (*k-means*). Ови модели се потом користе за одређивање вероватноћа да ли сегмент садржи говор или не. На основу поређења количника ове две

вероватноће и унапред задатог прага доноси се одлука да ли је сегмент говорни или не. У овом раду овај приступ носи ознаку A5.

Последњи алгоритам који је разматран у овом раду представљен је у раду [9]. Оно по чему се издваја од претходних алгоритама јесте по врсти обележја која се користе као и начину одлучивања да ли је сегмент говорни или не. За описивање вокалног тракта користили су 13 МФКК и њихове прве и друге изводе по времену. За описивање побуде користили су следећа обележја: хармоничност, јасноћу, грешку друге предикције, спектар хармонијског производа, израженост вршне вредности у кепстру, и две мере хармоничности које се изводе из алгоритма за сумирање резидуалних хармоника, као и њихове прве и друге изводе по времену. За класификацију су користили вештачке неуронске мреже са једним скривеним слојем. Задовољавајуће резултате добили су са 32 неурона. За активациону функцију неурона изабрали су хиперболичну тангентну сигмоид функцију. Обуку неуронских мрежа реализовали су на *TIMIT* и *Noisex*-92 бази. Варијанта ДГА у којој се користе само обележја за описивање вокалног тракта (МФКК обележја и њихови први и други изводи по времену) биће означена са A6. Варијанта у којој се користе само прва четири набројана обележја за описивање побуде A7, а последња три A8. Последња варијанта која обједињује одлуке које су дале претходне три варијанте, тако што израчунава геометријску средину њихових одлука, биће означена са A9.

II. Говорна база

Да би се оцениле перформансе појединих алгоритама за ДГА за различите типове шума као и за различите односе снага сигнала и шума формирана је посебна база од неколико постојећих. Из говорне базе која носи ознаку S70W100s120 [10] на случај су изабране по четири реченице српског језика за пет женских и пет мушких говорника. Ову базу чине снимци студијског квалитета, који су дигитализовани са учестаношћу одабирања 22050 Hz и са по 16 бита по одбирку. Изабрана је ова база пошто су реченице биране тако да буду фонетски богате (да се са што мање реченица покрије што више фонема српског језика). У исказима се тишина појављивала на почетку и крају реченице, а врло ретко у средини реченице, а да би се покрио и овај случај извршено је спајање по два исказа истог говорника. Спајање је вршено на исказима истог говорника, пошто је примећено да снимци различитих говорника могу да имају различит ниво шума (који је последица индукције од електричних инсталација).

Снимци шумова узети су из неколико база и то:

- *DEMAND* [11] – Ову базу чине снимци различитих акустичких окружења снимљених са шеснаесто каналним микрофонским низом. Оригинално су снимљени са 48 kHz, али је обезбеђена и верзија са 16 kHz која је и коришћена у овом раду. Искоришћени су снимци звукова: *i*) веш машине у току прања, *ii*) у кухињи у току припреме хране, *iii*) на спортском терену када постоји нека активност, *iv*) поред потока, *v*) у канцеларији где запослени активно користе тастатуре, *vi*) на прометној метро станици и *vii*) на градској саобраћајници. Преостали типови шума који постоје у овој бази изостављени су пошто се у њима јавља јасан говор, што би могло да референтне информација о границама појединих сегмената учини инвалидним и неупотребљивим.
- *Noisex*-92 [12] – Ова база представља једну од првих база која је садржавала снимке реалних врста шума. Развијена је за потребе војске тако да углавном садржи снимке звукова који су везани за војску (звукови унутар кокпита авиона, оклопног возила, и сл.). Из ове базе (односно из једног њеног дела који је бесплатно доступан на Интернету) преузет је само шум који постоји у фабрици, пошто су остали шумови ван области интересовања или су вештачки генерисани. Учестаност одабирања је 19.98 kHz, а одбирци су представљени са по 16 бита.
- *FreeSound* [13] – Ово је страница са које је могуће преузети снимке различитих природиних и вештачки генерисаних звукова високог квалитета. За потребе овог рада преузети су снимци са реалним звуковима који се јављају у аутобусу и у возу који су у покрету и код којих не постоји говорна активност. Учестаност одабирања у преузетим снимцима је 44.1 kHz, а одбирци су представљени са барем 16 бита. Сви снимци су добијени помоћу професионалних ручних снимача.

Да би се реализовало комбиновање говорног сигнала са сигналом шума било је неопходно у свим снимцима поставити исту учестаност одабирања. За учестаност одабирања је изабрана вредност од 16 kHz, пошто она обезбеђује довољно висок квалитет говорног сигнала. За комбиновање говорног сигнала и шума са жељеним односом сигнал шум искоришћен је јавно доступан алат са отвореним кодом *Filtering and Noise Adding Tool* (*FaNT*) [14] који је развијен за наведене потребе у оквиру Аурора-2 и Аурора-4 пројекта. При одређивању нивоа енергије говорног сигнала и шума коришћен је непондерисан пун спектар (до 8 kHz), али су код говорног сигнала изузети делови који одговарају паузама што је у складу са ИТУ препорукама П.56.

Говорна база са пратећим транскрипцијама која је коришћена у овом раду јавно је доступна на адреси: https://drive.google.com/open?id=0B6Ith1idiU6IaG0zQ0F1ekNCTG8.

III. Оцена квалитета ДГА

При ДГА постоје 4 могућа исхода:

- говорни сегмент је класификован као говорни, тзв. коректни позитивни исход;
- неговорни сегмент је класификован као говорни, тзв. лажно позитивни исход;

- говорни сегмент је класификован као неговорни, тзв. лажно негативни исход;
- неговорни сегмент је класификован као неговорни, тзв. коректни негативни исход.

На основу ових исхода дефинишу се следећи параметри:

$$TPR = TP/(TP + FN) \quad (1)$$

$$FNR = FN/(TP + FN) \quad (2)$$

$$FPR = FP/(FP + TN) \quad (3)$$

$$TNR = TN/(FP + TN) \quad (4)$$

где је: *TP* – број коректних позитивних исхода, *FN* – број лажних негативних исхода, *FP* – број лажних позитивних исхода, *TN* – број коректних негативних исхода, *TPR* – сензитивност, *FNR* – удео промашаја, *FPR* – удео лажних аларма и *TNR* – специфичност.

Стандардни начин поређења различитих система за детекцију је помоћу криве радне карактеристике пријемника (*ROC* енгл. *Receiver operating characteristic*) која приказује зависност сензитивности и удела лажних аларма за различите вредности параметра који дефинише понашање алгоритма (нпр. прага одлучивања). Овакав начин поређења постаје непрактичан уколико је потребно упоредити велики број различитих алгоритама у различитим условима и тада се као мера користи тзв. уједначени удео грешака (*EER* енгл. *Equal error rate*). *EER* представља вредност *FNR* (или *FPR*) када је вредност променљиве којом се контролише понашање алгоритма таква да је *FNR* = *FPR*.

IV. Резултати

Вредности *EER* за различите алгоритме и различите врсте шума као и односе снага сигнала и шума (*SNR*) дате су на Сл. 1 и 2. За реализацију појединих алгоритама искоришћена су јавно доступна решења која су обезбедили аутори радова [3] и [9]. Иако је основни циљ овог рада анализа ДТА алгоритама у случају када у сигналу постоји шум, урађени су експерименти са чистим говорним сигналом, да би измерили релативну деградацију тачности која је последица додавања шума. Посматрањем резултата тестова са чистим говорним сигналом може се уочити да је најнижи *EER* добијен за варијанту А9. Треба приметити да се и за остале варијанте алгоритама које су предложене у [9] добијају вредности *EER* приближне вредности *EER* која се добија за А9. Интересантно је да је допринос обележја која описују вокални тракт (А6) и побуду (А7 и А8) приближно исти, а да њихова комбинација доводи до смањења *EER* (А9).

Колико је у статистичком приступу ДГА битна тачност процене нивоа шума показује разлика између алгоритама А3 и А4 која износи скоро 3%. Ова предност методе А4 над методом А3 готово да је конзистентна за све анализиране врсте шума и све вредности *SNR* (једино за фабрички шум при *SNR* од -5 dB добијено је да је А3 незнатно боље од А4 41.8% наспрам 41.9%). Једноставни алгоритми А1 и А2 који се често користе у случајевима када у сигналу није присутан шум дали су знатно лошије перформансе од осталих алгоритама, пошто су оклузије безвучних фрикатива и африката често детектовали као тишине. Накнадном обрадом којом би се елиминисале кратке тишине (што постоји у алгоритмима А6-А9) значајније би се поправиле перформансе ових алгоритама.

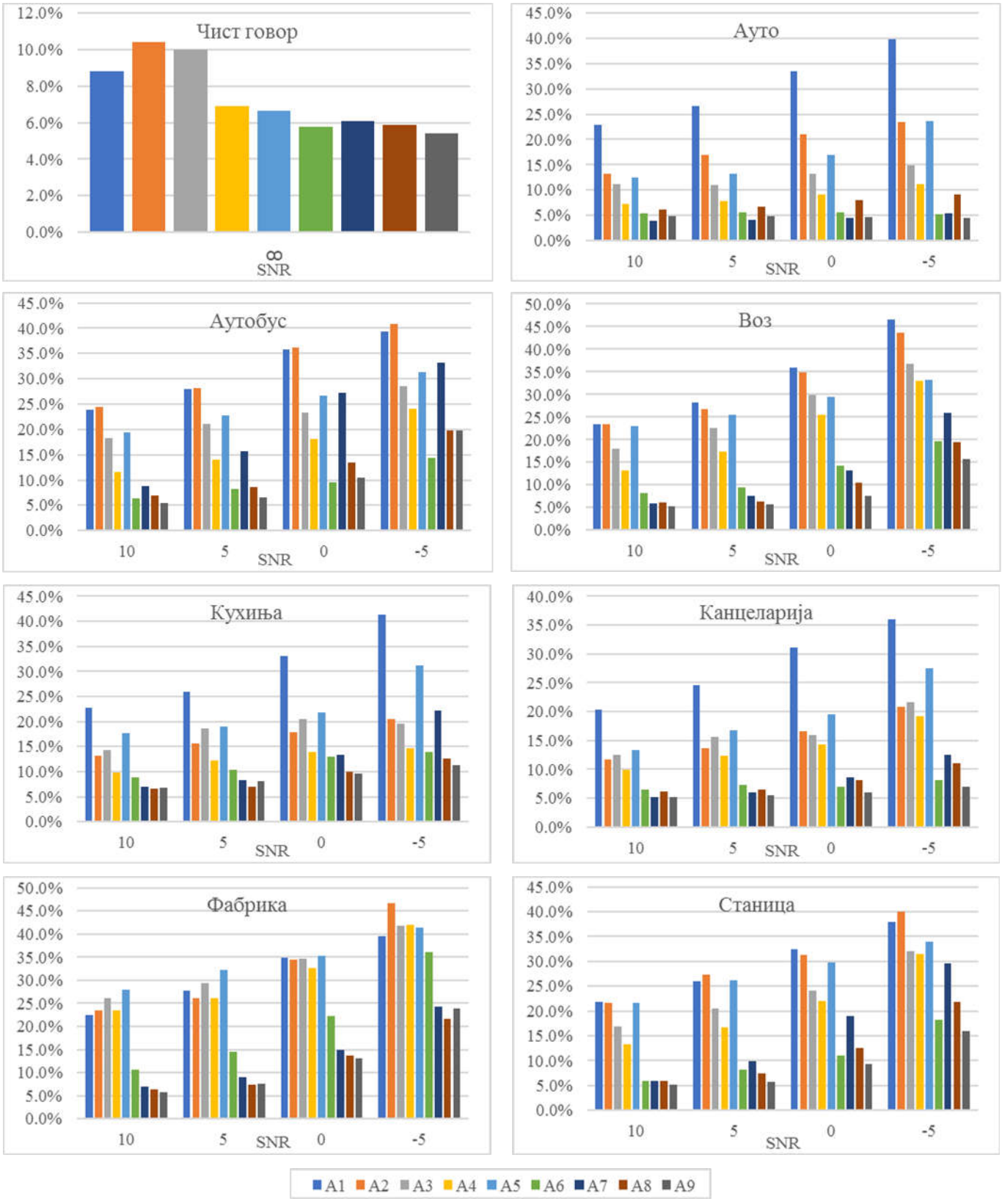
Додавање шума говорном сигналу је у просеку деградирало перформансе свих алгоритама. Највећа просечна деградација је добијена за А1 и износи око 14% (минимално 11% и максимално 15%) за *SNR* од 10 dB односно око 32% (минимално 27% и максимално 38%) за *SNR* од –5 dB, док су вредности за остале *SNR* између њих.

Прилично велика деградација је приметна и у случају А5 где је за *SNR* од 10 dB око 12% (минимално 5% и максимално 21%), односно за *SNR* од –5 dB око 25% (минимално 17% и максимално 35%), док су вредности за остале *SNR* између њих.

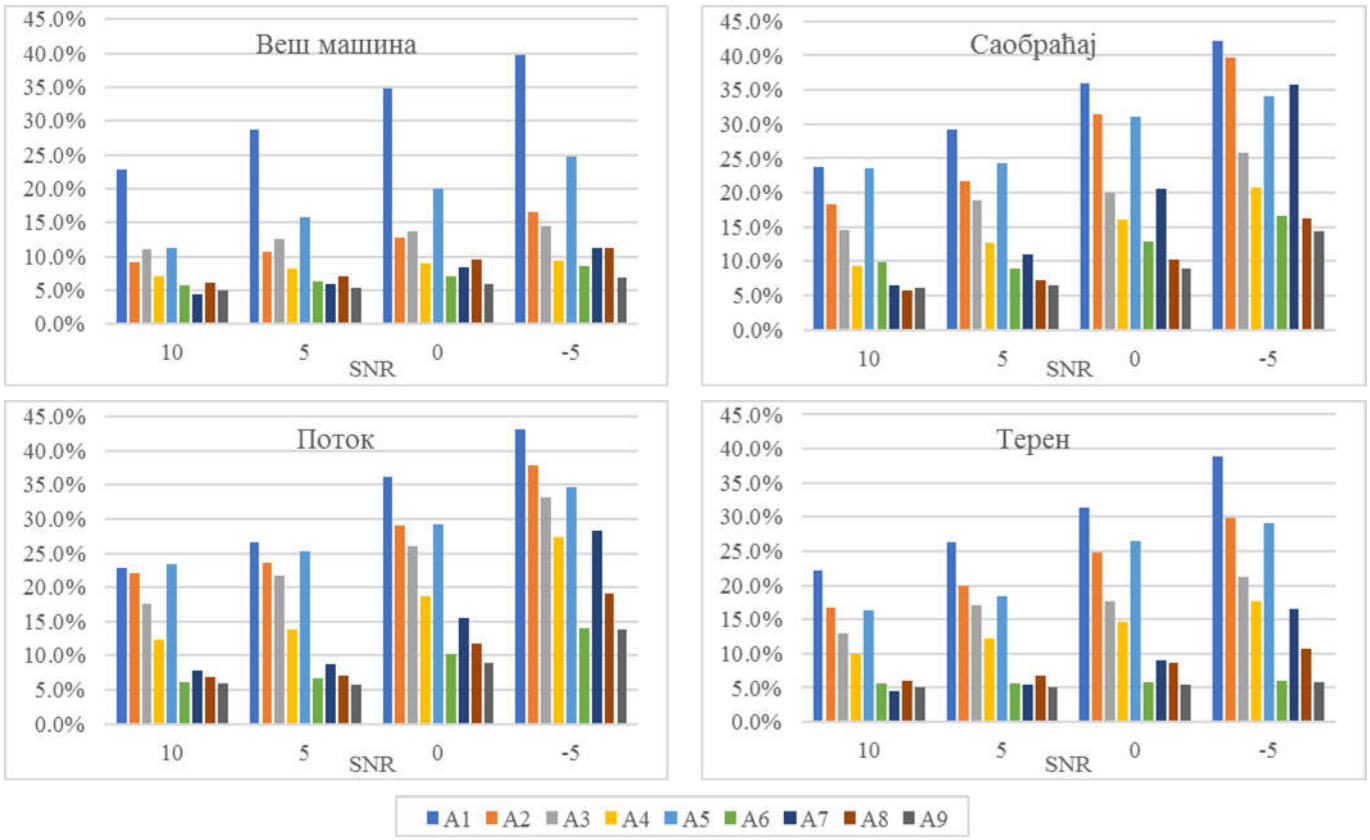
Деградација перформанси код алгоритма А2, А3 и А4 приближно је иста и у просеку износи око 6% (максимално 16%) за *SNR* од 10 dB, и око 17% (минимално 5%, максимално 35%) за *SNR* од –5 dB. Интересантно је да је за А2 при *SNR* од 10 dB добијено побољшање перформанси од 1% за случај веш машине, што је последица смањења броја лажних негативних исхода које генеришу сегменти са оклузијама. Статистичке методе А3 и А4 показале су се прилично отпорне на шум који генерише ауто у покрету и веш машина.

Методе предложене у [9] (А6–А9) показале су се као отпорне на шум ниског интензитета (*SNR* већи од 0 dB) где је просечна деградација мања од 3%. Најмање добра се показала метода А6 где је просечна деградација око 1.4% (максимално 5%) за *SNR* од 10 dB, односно око 9% (максимално 30%) за *SNR* од –5 dB. Иако методе А7 и А8 користе обележја којима се описује глотална побуда, А8 се показала као нешто робустнија за мање вредности *SNR*. Као најбоља метода се показала А9 где је просечна деградација око 0.1% (максимално 1%) за *SNR* од 10 dB, односно око 9% (максимално 30%) за *SNR* од –5 dB. Можда је интересантно да поступак А9 (слично важи и за А6 и А7) даје нешто боље резултате у присуству шума који представља звукове у аутомобилу. Претпостављамо да је то последица обуке модела који је коришћен за класификацију где су за обуку користили шумове из *Noisex*-92 базе, односно да су се услови тестирања много боље поклопили са условима у којима су рађени тестови. Зашто се то није десило са шумом који представља звуке из фабрике, који је за потребе овог теста директно преузет из *Noisex*-92 врло вероватно је последица тога да та врста шума није коришћена при обуци неуронске мреже.

Нисмо очекивали овако низак ниво грешке за методе од А6-А7, пошто је за обуку класификатора коришћена енглеска *TIMIT* говорна база. Овакво понашање алгоритама указује да су неуронске мреже у току обуке научиле карактеристике говорног сигнала у целини, а не појединачних фонема, јер се скуп фонема енглеског и српског језика знатно разликује. Идеја да покушамо директно са моделима које су они створили последица је чињенице да су они тестове вршили на бази која је садржавала реченице јапанског језика.



Слика 1: Вредности уједначеног удела грешака (*EER*) за различите алгоритме ДГА (представљени су различитим бојама) и за различите врсте шума (засебни подграфици) у зависности од односа сигнал шум (*SNR*).



Слика 2: Вредности уједначеног удела грешака (*EER*) за различите алгоритме ДГА (представљени су различитим бојама) и за различите врсте шума (засебни подграфици) у зависности од односа сигнал шум (*SNR*).

Што се тиче шумова, највећи проблем су представљали звукове из фабричке хале, потом аутобуса у покрету, станице и саобраћајнице. Претпостављамо да је проблем што шум у овим случајевима има велику динамику коју анализирани алгоритми не могу да уклоне.

V. Закључак

У овом раду је упоређено неколико алгоритама за детекцију говорне активности који издвајају различита обележја и користе различите принципе да би детектовали говорне сегменте. У експериментима су се као најбољи показали системи за које постоји модел говора и шума, а који као обележја користи како она која моделују изглед вокалног тракта (у овом случају мел фреквенцијски кепстрални коефицијенти) тако и облежја која моделују побуду (хармоничност, јасноћу и друга слична обележја). Добијен је помало изненађујући резултат да систем који је обучен коришћењем исказа енглеског језика ради изузетно добро и на исказима српског језика. Ово је вероватно последица чињенице да је вештачка неуронска мрежа научила карактеристике говора које су независне од језика. Још једном је потврђена хипотеза да се комбиновањем одлука класификатора скромнијих перформанси може добити класификатор који има боље перформансе од свих њих појединачно. Задовољавајуће перформансе могу се постићи и помоћу статистичког приступа, уколико се изабере добар модел за процену нивоа шума што је у овом случају метода базирана на непристрасној процени минималне средње квадратне грешке.

Захвалница

Овај рад је делимично финансиран од стране Министарства просвете, науке и технолошког развоја Републике Србије кроз пројекте ТР32035 и ИИИ47020.

Литература

[1] A. M. Kondoz, Digital speech: Coding for Low Bit Rate Communication Systems, 2nd ed., Chichester, England, John Wiley & Sons Ltd, 2004.

[2] B. Popović, E. Pakoci and D. Pekar, "Advanced voice activity detection on mobile phones by using microphone array and phoneme-specific Gaussian mixture models," in Proc. 14th Intern. Symp. on Intelligent Systems and Informatics (SISY 2016), 2016, pp. 45-48.

[3] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," *IEEE Inter. Conf. on Acoustics, Speech and Signal Process.* (ICASSP 2013), 2013, pp. 7229-7233.

[4] P. C. Loizou, *Speech enhancement*: *Theory and Practice,* 2nd ed., Boca Raton, FL, CRC Press, 2013.

[5] ITU-T (1996) A silence compression scheme for G.729 optimised for terminals conforming to ITU-T V.70, ITU-T Rec. G.729 Annex B.

[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6(1), pp. 1-3, Jan. 1999.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech & Audio Process.*, vol. 9(5), pp. 504-512, July 2001.

[8] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," I*EEE Trans. Audio, Speech & Language Processing*, vol. 20(4), pp. 1383-1393, May 2012.

[9] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Process. Lett.*, vol. 23(2), pp. 252-256, Feb. 2016.

[10] В. Делић, "Говорне базе на српском језику снимљене у оквиру пројекта АлфаНум," *Зборник радова 3. конф. Дигитална обрада говора и слике* (ДОГС 2000), 2000, стр. 29-32.

[11] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *Proceedings of Meetings on Acoustics* (ICA2013), 2013, pp. 3591.

[12] Јавно доступан део Noisex-92 базе. Доступан: http://spib.linse.ufsc.br/noise.html (датум 8.2.2017), 1992.

[13] Јавна база звукова FreeSound. Доступан: http://www.freesound.org (датум 8.2.2017).

[14] H. Hirsch, FaNT: Filtering and Noise Adding Tool. Доступан: http://dnt.kr.hs-niederrhein.de/index964b.html (датум 30.1.2017), 2005.

ABSTRACT

The paper obtains an overview of the performances of the several algorithms for voice activity detection estimated on a Serbian speech database with added real noise samples. The following noise types were taken into consideration: sounds in moving car, bus and train, sounds in a kitchen, office and factory, sounds on the subway station, street and sport field, sounds of creek and washing machine. The best results are obtained with the algorithm where detection is treated as pattern recognition problem using features which separately model source and vocal tract.

A Comparison of Several Voice Activity Detection Methods

Nikša Jakovljević, Dragiša Mišković, Željen Trpovski