Elektrotehniški vestnik 74(4): 229-235, 2007
Electrotechnical Review: Ljubljana, Slovenija

Metoda signalnega podprostora s sprotno oceno šuma in njena uspešnost pri robustnem avtomatskem razpoznavanju govora

Bojan Jarc, Rudolf Babič

Univerza v Mariboru, Fakulteta za elektrotehniko računalništvo in informatiko,
Smetanova ul. 17, 2000 Maribor, Slovenija
E-pošta: bojan.jarc@uni-mb.si, rudolf.babic@uni-mb.si

Povzetek. V prispevku predstavljamo metodo zmanjševanja nivoja šuma na podlagi teorije signalnega podprostora in njeno uspešnost pri izboljšanju avtomatskega razpoznavanja govora. V [1] predstavljeno metodo, ki je v osnovi primerna za zmanjševanje nivoja belega šuma, smo nadgradili s sprotno oceno šuma brez neposredne detekcije intervalov aktivnosti govora. Podali smo novo oceno lastnih vrednosti govora s pomočjo razmerja signal-šum in nov postopek detekcije intervalov aktivnosti govora na podlagi stopnje koreliranosti. Uspešnost metode smo ovrednotili z rezultati avtomatskega razpoznavanja govora v eksperimentalnih okoljih Aurora 2 in 3 ter dosegli skupno absolutno uspešnost razpoznavanja besed 83,90 in 78,29 odstotno.

Ključne besede: razpoznavanje govora v šumnem okolju, procesiranje signalov, signalni podprostor, detekcija aktivnosti govora

Signal subspace method with continuous noise estimation and its efficiency in robust automatic speech recognition

Extended abstract. In most of automatic speech-recognition (ASR) systems, recognition performance significantly decreases when moving from the studio to a real-world environment. A noisy environment and echo are the most common reasons for ASR performance degradation. New trends in the area of mobile communications demand development of efficient recognition and pre-processing methods in order to improve noise robustness.

This paper presents a signal subspace-based method for noise reduction and its efficiency for ASR improvement in a real noisy environment. The signal subspace method was first presented in [1] presuming the white noise as an interfering signal. According to [1], the clean signal is estimated by using noisy-signal covariance matrix eigenvalues. Since the calculation of eigenvalues with the Karhunen-Loève transformation (KLT) is a computationally intensive task, they can be approximated with the use of fast discrete cosine transformation (FDCT) [4]. They are called approximate eigenvalues. To achieve the method suitability for real-world environments, we propose a minima tracking-based approach for noise covariance matrix eigenvalues estimation. Since it is presumed that the noise and speech are uncorrelated zero mean signals, covariance coefficients can be estimated with autocorrelation coefficients and additive relation given in (13). According to (14), the additive relation is also preserved between speech and noise approximate eigenvalues $\hat{\lambda}_s$ and $\hat{\lambda}_d$. This is the basis for the use of the minima tracking-based approach for estimation of $\hat{\lambda}_d$ (see Eq. 15 and Fig. 1). To reduce overestimation of $\hat{\lambda}_d$ in areas of speech presence, we propose a signal-to-noise ratio (SNR)-dependent estimation of $\hat{\lambda}_s$ by (17). For clean speech estimation, a spectral domain-constraint estimator (SDC) is used by (10).

The SDC estimator wrongly presumes that the speech is always present in a noisy signal. Since speech is a correlated signal, we propose a voice-activity detection (VAD) method based upon the level of autocorrelation (see Eq. 18-19). Presumption that the noise is a more weakly correlated signal than speech allows us to use the minima tracking-based approach for determination of the noise-correlation level (see Fig. 3 b). A novel VAD function based on the ratio of speech and noise correlation levels is defined by (19). The clean speech is then estimated in the time domain using an SDC estimator and VAD gain function by (20).

The proposed method efficiency is confirmed with ASR results in Aurora 2 and 3 experimental frameworks comprising the noisy speech of connected digits with train and test schemes for ASR. The mel-cepstrum feature extraction algorithm is applied with 12 mel cepstrum coefficients and the energy coefficient. The

Prejet 12. september, 2006
Odobren 10. maj, 2007

absolute recognition performance for the Aurora 2 and 3 ASR tasks are shown in Tables 1 and 2, respectively. The best overall word-recognition accuracy of 83.90% and 78.29% respectively are achieved. Relatively to the baseline results, this stands for a 35.49 % and 10.86% improvement.

Keywords: speech recognition in a noisy environment, signal processing, signal subspace, voice-activity detection

1 Uvod

Zadnjih nekaj let so opazna velika prizadevanja tako izdelovalcev telekomunikacijskih naprav kot tudi širše strokovne javnosti za izboljšanje uspešnosti avtomatskega razpoznavanja govora (ARG) v različnih šumnih okoljih. Da bi spodbudili razvoj in dosegli standardizacijo čim uspešnejšega algoritma, je skupina Aurora, ki deluje v okviru evropskega inštituta za standardizacijo v telekomunikacijah (ang.: "*European Telecommunications Standard Institute - ETSI*"), izdala eksperimentalna okolja Aurora 2, 3 in 4. Kljub dolgotrajnim prizadevanjem na tem področju trenutna uspešnost razpoznavanja govora v šumnem okolju še vedno ne zadovoljuje vseh vidikov, potrebnih za uspešno komercialno rabo.

V naslednjem prispevku predstavljamo metodo zmanjševanja nivoja šuma na podlagi teorije signalnega podprostora s sprotno oceno šuma. Različne metode na podlagi teorije signalnega podprostora so predstavljene v [1, 2, 3, 4]. V osnovi temeljijo na izračunu lastnih vrednosti Toeplitzove avtokorelacijske matrike signala šumnega govora in obravnavajo primere, ko je signalu govora dodan beli šum. Njihova uspešnost v sistemih ARG ni znana. V našem prispevku bomo predstavili metodo na podlagi teorije signalnega podprostora, primerno za poljubne šume okolja, ter podali njeno uspešnost v sistemu ARG.

Prispevek je organiziran kot sledi. V drugem poglavju je opisana teorija signalnega podprostora in v tretjem postopek ocenjevanja lastnih vrednost za poljubne šume. Metoda detekcije intervalov aktivnosti govora je predstavljena v četrtem poglavju, v petem pa so predstavljeni rezultati.

2 Teorija signalnega podprostora

V tem poglavju je na kratko povzeta teorija signalnega podprostora predstavljena v [1]. Govorni signal je predstavljen z linearnim modelom. Predpostavljeno je, da sta signala govora in šuma nekorelirana in aditivna.

Zaporedje otipkov signala šumnega govora v vektorski obliki zapišemo z enačbo:

$$\mathbf{y} = \mathbf{s} + \mathbf{d} . \tag{1}$$

Pri tem so **y**, **s** in **d** vektorji dimenzij *K*, in sicer: šumnega govornega signala, signala govora in signala šuma. Vektor **y** je v prostoru $\mathbb{R}^K$.

Glede na predpostavljeni, linearni model je v [1] vektor **s** definiran z enačbo:

$$\mathbf{s} = \mathbf{V}\mathbf{x} . \tag{2}$$

Pri tem je $\mathbf{V} = [\mathbf{v}, ..., \mathbf{v}_M]_{K\times M}$ matrika linearnih neodvisnih baznih vektorjev, $\mathbf{x} = (x[1], ..., x[M])^T$ pa vektor naključnih spremenljivk s srednjo vrednostjo nič. Rang matrike **V** je *M* in velja, da je $M \le K$. Kadar je $M < K$, leži poljubna množica vektorjev {**s**} v podprostoru prostora $\mathbb{R}^K$. Podprostor imenujemo signalni podprostor.

Avtorja v [1] predvidevata, da je srednja vrednost vektorja **s** enaka nič ($\mu_s = 0$) oz. da je kovariančna matrika enaka avtokorelacijski matriki:

$$\mathbf{R}_s = E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{V}\mathbf{R}_x\mathbf{V}^T . \tag{3}$$

Pri tem je "^{T}" operator transponiranje, $E\{.\}$ matematično upanje in $\mathbf{R}_x$ kovariančna oz. avtokorelacijska matrika vektorja **x**. Ker je rang matrike $rang(\mathbf{R}_s) = M$, ima matrika K-M ničelnih lastnih vrednosti.

Za vektor signala šuma **d** sta avtorja v [1] predvidela Gaussovo porazdelitev s srednjo vrednostjo $\mu_d = 0$ in varianco λ_d. Kovariančna matrika šuma, izračunana z avtokorelacijsko, je naslednja:

$$\mathbf{R}_d = E\{\mathbf{d}\mathbf{d}^T\} = \lambda_d\mathbf{I} . \tag{4}$$

Pri tem je **I** enotina matrika in λ_d varianca šuma. Rang matrike $\mathbf{R}_d$ je *K*, kar pomeni, da se šum nahaja v celotnem prostoru $\mathbb{R}^K$. Iz enačbe (1) ob upoštevanju enačbe (2) sledi, da je:

$$\mathbf{y} = \mathbf{V}\mathbf{x} + \mathbf{d} . \tag{5}$$

Kovariančno matriko vektorja **y** lahko zapišemo tudi v naslednji obliki:

$$\mathbf{R}_y = E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{V}\mathbf{R}_x\mathbf{V}^T + \mathbf{R}_d . \tag{6}$$

Matrika $\mathbf{R}_y$ je Hermitna. Upoštevajoč spektralni teorem [6] obstaja dekompozicija matrike $\mathbf{R}_y$ na lastne vrednosti in lastne vektorje. Dekompozicijo zapišemo z enačbo:

$$\mathbf{R}_y = \mathbf{U}\mathbf{\Lambda}_y\mathbf{U}^T . \tag{7}$$

Pri tem je $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_K]_{K\times K}$ ortonormalna matrika lastnih vektorjev in $\mathbf{\Lambda}_y$ diagonalna matrika lastnih vrednosti $\mathbf{\Lambda}_y = diag(\lambda_y[1], ..., \lambda_y[K])$. Ker je šum beli oz. $\mathbf{R}_d$ diagonalna matrika, so lastni vektorji matrike $\mathbf{R}_y$ hkrati tudi lastni vektorji matrik $\mathbf{R}_s$ in $\mathbf{R}_d$, lastne

vrednosti pa so vsota lastnih vrednosti matrik $\mathbf{R}_s$ in $\mathbf{R}_d$ [1]. Zato velja, da je:

$$\begin{aligned}&\mathbf{V}\mathbf{R}_x\mathbf{V}^T = \mathbf{U}\mathbf{\Lambda}_y\mathbf{U}^T - \lambda_d\mathbf{U}\mathbf{U}^T \text{ oz.}\\ &\mathbf{U}\mathbf{\Lambda}_s\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}_y\mathbf{U}^T - \mathbf{U}\lambda_d\mathbf{U}^T.\end{aligned} \tag{8}$$

Lastne vrednosti matrike $\mathbf{R}_s$ sedaj zapišemo z:

$$\lambda_s[k] = \begin{cases}\lambda_y[k] - \lambda_d & \text{za } k = 1, ..., M\\ 0 & \text{za } k = M+1, ..., K.\end{cases} \tag{9}$$

Zmanjševanje nivoja šuma realiziramo z modifikacijo $\lambda_y[k]$. V [1] je podanih več optimalnih linearnih cenilk. Mi smo se osredotočili na cenilko SDC, ki je optimalna v smislu minimalizacije energije popačenj govornega signala glede na omejeno energijo preostalega šuma posamezne spektralne komponente. Takrat posamezno lastno vrednost $\lambda_y[k]$ modificiramo s pomočjo faktorja:

$$g_{SDC}[k] = \left(\frac{\lambda_s[k]}{\lambda_s[k] + \lambda_d}\right)^{\gamma}, \qquad k = 1, ..., M\ . \tag{10}$$

Pri tem je $\gamma \geq 0{,}5$ eksperimentalno določena konstanta, s katero spreminjamo nivo preostalega šuma in popačenja govornega signala. Z naraščanjem konstante γ se energija preostalega šuma zmanjšuje, povečujejo pa se popačenja govornega signala.

Nekorelirane lastne vrednosti λ_y izračunamo s transformacijo Karhunen-Loève (KLT) oz. z analizo glavnih komponent (PCA). Dimenziji prostora K in podprostora M na splošno nista znani, zato ju izberemo sami. Glede na izbrano dimenzijo K dobimo s transformacijo KLT optimalno rešitev, v smislu minimalne srednje kvadratne napake, iz $M<K$ lastnih vrednosti rekonstruiranega signala [5].

2.1 Uporaba hitre transformacije DCT

Računska zahtevnost transformacije KLT narašča s četrto potenco dolžine transformiranega vektorja **y**. Velika računska zahtevnost metod zmanjševanja nivoja šuma je nezaželena, saj slabo vpliva na odzivni čas sistemov ARG.

Znano je, da lahko tvorjenje govornega signala v govornem traktu modeliramo z avtoregresivnim procesom [7]. Tak model imenujemo vir-filter model [7]. Prav tako je dokazano [4], da lahko lastne vrednosti kovariančne matrike avtoregresivnega procesa aproksimiramo s pomočjo transformacije DCT. Pri tem ne gre za direktno uporabo transformacije DCT, pač pa je s koeficienti transformacije DCT, definirana nova transformacijska matrika. Ker obstaja možnost izračuna približnih lastnih vrednosti in zaradi bistveno manjše računske zahtevnosti smo se odločili za uporabo transformacije DCT oz. v [4] predlaganega postopka s hitro transformacijo DCT (metoda FDCT). Z metodo FDCT vektor približnih lastnih vrednosti izračunamo po enačbi:

$$\hat{\boldsymbol{\lambda}}_y = \mathbf{B}\mathbf{r}_y, \tag{11}$$

pri tem je $\hat{\boldsymbol{\lambda}}_y = (\hat{\lambda}_y[1], \hat{\lambda}_y[2], ..., \hat{\lambda}_y[K])^T$ vektor približnih lastnih vrednosti, $\mathbf{r}_y = (r_y[0], r_y[1], \ldots, r_y[K\text{-}1])^T$ avtokorelacijski vektor in $\mathbf{B} = [b_{ij}]_{K\times K}$ matrika, katere elementi so:

$$b_{ij} = \begin{cases}\sum_{k=1}^{K} c_{i,k}^2, & j = 1\\ 2\sum_{k=1}^{K-j+1} c_{i,k}c_{i,k+j-1}, & 2 \leq j \leq K.\end{cases} \tag{12}$$

Pri tem je b_{ij} j-ti element i-te vrstice matrike **B** in $c_{i,j}$ j-ti element i-te vrstice matrike DCT. Računska kompleksnost metode FDCT narašča z drugo potenco števila lastnih vrednosti K [4].

3 Ocenjevanje lastnih vrednosti

V naslednjem poglavju predstavljamo predlagan postopek ocenjevanja lastnih vrednosti poljubnih signalov šuma in govora brez neposredne detekcije intervalov aktivnosti govora.

V realnem okolju so situacije s stacionarnim šumom izredno redke. Še redkeje imamo opravka z belim šumom. Zato je smiselno poiskati rešitev, ki je primerna za poljubne časovno spremenljive šume. Podobno kot za signal govora tudi za signal šuma predpostavimo kvazi-stacionarnost. To pomeni, da lahko avtokorelacijske koeficiente šumnega signala ocenimo v dovolj kratkih časovnih intervalih. Ob predpostavki nekoreliranosti med signaloma govora in šuma lahko zapišemo enačbo:

$$\mathbf{r}_y = \mathbf{r}_s + \mathbf{r}_d, \tag{13}$$

pri čemer sta $\mathbf{r}_s$ in $\mathbf{r}_d$ avtokorelacijska vektorja signalov govora in šuma. Iz enačbe (11) sedaj sledi:

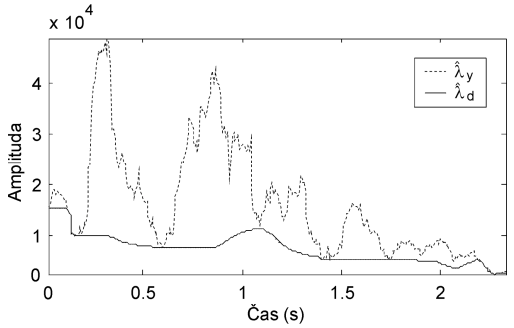
$$\begin{aligned}\hat{\boldsymbol{\lambda}}_y &= \mathbf{B}(\mathbf{r}_s + \mathbf{r}_d)\\ &= \hat{\boldsymbol{\lambda}}_s + \hat{\boldsymbol{\lambda}}_d.\end{aligned} \tag{14}$$

$\hat{\boldsymbol{\lambda}}_s$ in $\hat{\boldsymbol{\lambda}}_d$ sta približna lastna vektorja matrik $\mathbf{R}_s$ in $\mathbf{R}_d$. Na voljo imamo samo šumni govorni signal, zato določitev $\hat{\boldsymbol{\lambda}}_s$ in $\hat{\boldsymbol{\lambda}}_d$ ni trivialna.

3.1 Lastne vrednosti šumnega signala

Na podlagi opazovanja smo ugotovili, da ima spekter približnih lastnih vrednosti $\hat{\lambda}_y$ značilno obliko lokalnih minimumov in maksimumov (slika 1) in da lokalni

minimumi sovpadajo z intervali aktivnosti govora. Zato lahko s pomočjo sledenja minimumov $\hat{\boldsymbol{\lambda}}_y^{(w)}$ med sosednjimi časovnimi intervali ocenimo $\hat{\boldsymbol{\lambda}}_d^{(w)}$. Pri tem smo z *w* označili indeks trenutnega časovnega intervala. V [8] je predstavljena ocena močnostnega spektra šumnega signala na podlagi statističnega minimuma. Tako je močnostni spekter šumnega signala ocenjen z amplitudo močnostnega spektra šumnega govornega signala v območjih lokalnih minimumov. Za ocenjevanje lastnih vrednosti $\hat{\boldsymbol{\lambda}}_d^{(w)}$ predlagamo izboljšan pristop.



Slika 1: Ocena $\hat{\lambda}_d$ na podlagi sledenja minimumov $\hat{\lambda}_y$ (k=40, β=0,94, časovni interval iskanja minimuma je 0,5s)
Figure 1. Minimum tracking-based estimation of $\hat{\lambda}_d$ from $\hat{\lambda}_y$ (k=40 β=0,94, interval for minimum searching is 0,5s).

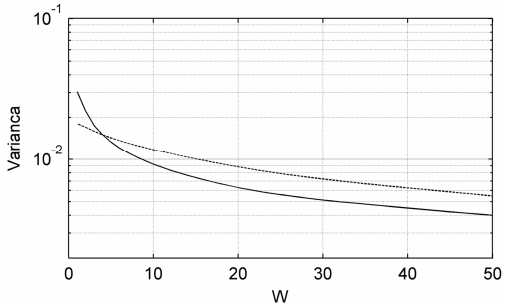
Ker obstaja koreliranost med posameznimi komponentami časovno sosednjih lastnih vektorjev, predlagamo iskanje minimuma *k*-te lastne vrednosti tudi v prihodnjih in ne le v predhodnih časovnih intervalih ter glajenje časovno zaporednih lastnih vrednosti z rekurzivno enačbo prvega reda. Posamezno komponento vektorja $\hat{\boldsymbol{\lambda}}_d^{(w)}$ tako ocenimo z enačbo:

$$\hat{\lambda}_d^{(w)} = \beta\hat{\lambda}_d^{(w-1)} + (1-\beta)\min(\hat{\lambda}_y^{(n)}), \quad n = w-W,\ldots,w+W. \tag{15}$$

Pri tem je β rekurzivni koeficient v mejah $\beta \in [0, 1)$, *w* je indeks časovnega intervala, $\min(\hat{\lambda}_y^{(n)})$ je minimalna vrednost $\hat{\lambda}_y^{(n)}$ in $2W+1$ je število sosednjih časovnih intervalov, uporabljenih pri iskanju minimalne vrednosti. Zaradi preglednosti smo v enačbi (15) izpustili indeks komponent vektorja *k*.

Koeficient β določa časovno konstanto oz. stopnjo pomnjenja predhodne lastne vrednosti. Ker smo na začetku prispevka predpostavili, da imamo opravka s poljubnim šumom, analitična določitev β ni mogoča. Koeficient β smo določili eksperimentalno s poslušanjem ocenjenega govornega signala. Najboljše rezultate smo dosegali z vrednostmi v območju β = 0,9 ÷ 0,95. Izbira števila intervalov $2W+1$ oz. ustreznega časovnega intervala iskanja minimuma je kompromis med primernostjo metode za zelo spremenljive šumne signale in verjetnostjo, da bomo izbrali lastne vrednosti šumnega signala brez govora. Slednje je močno odvisno od narave govornega signala. Testi nad zaporedji števk v različnih šumnih okoljih so pokazali dobre rezultate, pri iskanju minimuma v intervalih dolžine od 0,3s ÷ 1s. Primer ocene $\hat{\lambda}_d$ iz $\hat{\lambda}_y$ po enačbi (15) prikazuje slika 1. Izbrali smo časovni interval iskanja minimuma 0,5s in rekurzivni faktor β = 0,94. Predlagan postopek ocenjevanja $\hat{\lambda}_d$ vnaša v proces ARG končno zakasnitev (0,25s) odziva razpoznavalnika in je slabost našega postopka.

Uspešnost predlaganega postopka ocenjevanja $\hat{\lambda}_d$ smo primerjali s postopkom, podanim v [8]. Podobno kot v [8] smo generirali naključni časovno diskretni signal z varianco σ^2 = 1 in primerjali varianco lastnih vrednosti $\hat{\lambda}_d$. Pri tem smo avtokorelacijske vektorje $\mathbf{r}_y$ dimenzije 50 ocenjevali v intervalih z 800 otipki in s prekrivanjem med intervali 750 otipkov. $\hat{\boldsymbol{\lambda}}_y^{(w)}$ smo izračunali z enačbo (14) in $\hat{\boldsymbol{\lambda}}_d^{(w)}$ z enačbo (15). Izbrali smo rekurzivni faktor β = 0,94, število intervalov *W*, pa smo spreminjali v mejah od 1 do 50. Iz rezultatov na sliki 2 je razvidna manjša varianca s predlaganim postopkom ocenjenih lastnih vrednosti za $W > 5$, kar potrjuje večjo uspešnost predlaganega pristopa. To so potrdili tudi rezultati ARG.



Slika 2: Varianca lastnih vrednosti $\hat{\lambda}_d$ ocenjenih s (15) (polna črta) in po postopku, predlaganem v [8] (prekinjena črta)
Figure 2. Variance of eigenvalues $\hat{\lambda}_d$ estimated with (15) (solid line) and using the approach proposed in [8] (dashed line).

3.2 Lastne vrednosti govornega signala

Glede na enačbo (14) izračunamo $\hat{\boldsymbol{\lambda}}_s$ z razliko $\hat{\boldsymbol{\lambda}}_y$ in $\hat{\boldsymbol{\lambda}}_d$. Komponente $\hat{\boldsymbol{\lambda}}_s$ ne morejo biti negativne, zato tak izračun kombiniramo s funkcijo polvalnega usmerjanja oz. s funkcijo praga, s pragom pri vrednosti nič. Eksperimentalni rezultati so pokazali, da opisan subtraktivni izračun $\hat{\boldsymbol{\lambda}}_s$, povzroči nastanek motenj podobnih tako imenovanemu "glasbenemu šumu" (pojav tonalnih komponent zaradi variabilnosti $\hat{\boldsymbol{\lambda}}_y$ in odštevanja glajenega spektra $\hat{\lambda}_d$, ang. *musical noise*). Pojav je dobro poznan iz metod zmanjševanja nivoja šuma na podlagi spektralnega odštevanja [7, 9] in je najbolj moteč v področjih spektra z majhnim razmerjem signal-šum (razmerje SNR). Klasičen pristop k maskiranju glasbenega šuma je podan v [9]. Moteč pojav naključnih spektralnih vrhov pri nizkih razmerjih

SNR je zmanjšan z odštevanjem tudi do petkrat večje amplitude šuma od ocenjene. Eksperimenti z ARG so pokazali, da tak pristop k oceni $\hat{\boldsymbol{\lambda}}_s$ ne zagotavlja največjega števila pravilno razpoznanih besed.

K temu pripomore tudi omejena uspešnost postopka ocenjevanja $\hat{\boldsymbol{\lambda}}_d$, opisanega v prejšnjem podpoglavju. Zaradi zahteve, da je postopek primeren tudi za spremenljive šume, smo izbrali kratek interval iskanja minimuma (0,5s), ki ne zagotavlja detekcije lastnih vrednosti šumnega signala izključno v intervalih brez govora. Posledično lahko v energijsko šibkih intervalih začetkov in koncev besed lastne vrednosti šumnega govornega signala napačno opredelimo kot lastne vrednosti šuma oz. precenimo komponente $\hat{\boldsymbol{\lambda}}_d$. Zato predlagamo izračun $\hat{\boldsymbol{\lambda}}_s$, kjer z razmerjem SNR zmanjšamo vpliv netočne ocene $\hat{\boldsymbol{\lambda}}_d$ na vrednost $\hat{\boldsymbol{\lambda}}_s$.

Definirajmo razmerje *SNR* z enačbo:

$$SNR^{(w)}[k] = (\hat{\lambda}_y^{(w)}[k] - \hat{\lambda}_d^{(w)}[k]) \big/ \hat{\lambda}_d^{(w)}[k], \quad k = 1,..., K, \quad (16)$$

pri čemer je *k* indeks komponent vektorja **SNR** in *w* je indeks časovnega intervala. Zaradi precenjenosti $\hat{\boldsymbol{\lambda}}_d$ lahko imajo komponente vektorja **SNR** vrednost nič tudi, ko so dejanske lastne vrednosti govornega signala od nič različne oz. je govor aktiven. Ker pri ničelnem razmerju SNR informacije o dejanskih lastnih vrednostih govora nimamo, smo jih ocenili kar z lastnimi vrednostmi šumnega govora. Z naraščanjem razmerja SNR je vpliv netočne vrednosti $\hat{\lambda}_d$ na subtraktivno izračunano vrednost $\hat{\lambda}_s$ manjši, zato smo delež odštevanca $\hat{\lambda}_d$ z razmerjem SNR povečevali od vrednosti nič do celotne vrednosti $\hat{\lambda}_d$. Linearno povečevanje odštevanca ni dalo želenih rezultatov. Zato smo izbrali eksponentno povečevanje odštevanca, kot opisuje enačba:

$$\hat{\lambda}_s^{(w)}[k] = \hat{\lambda}_y^{(w)}[k] - \hat{\lambda}_d^{(w)}[k](1 - e^{-SNR^{(w)}[k]}), \quad k = 1, ..., K. \quad (17)$$

Pri tako ocenjenem spektru $\hat{\lambda}_s$ ni bilo težav z nastankom glasbenemu šumu podobnih motenj. Z uporabo enačbe (17) dobimo iz enačbe (10) množitelj za modifikacijo lastnih vrednosti $\hat{\lambda}_y$.

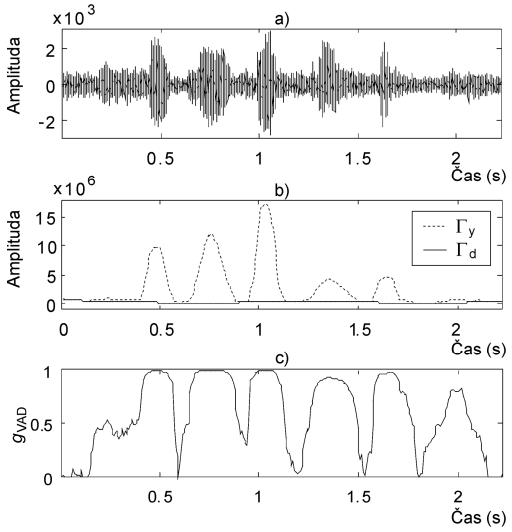
4 Detekcija aktivnosti govora

Večina metod za detekcijo aktivnosti govora (VAD), kot odločitveno funkcijo, govor je oz. ni prisoten, uporablja funkcijo na podlagi razmerja *SNR* [7]. Takšne so tudi statistične metode [10, 12]. V naslednjem poglavju je predstavljena metoda VAD na podlagi avtokorelacijskih koeficientov $\mathbf{r}_y$.

Na splošno lahko govor delimo na zvenečega in nezvenečega. Njegovo tvorjenje opisujemo kot filtriranje zračnega toka s prenosno funkcijo govornega trakta. Pri zvenečem govoru moduliramo zračni tok z nihanjem glasilk. Posledično sta zveneči in nezveneči govor korelirana signala. Zato je smotrno stopnjo koreliranosti uporabiti za detekcijo aktivnosti govora. Stopnjo koreliranosti **y** bomo ovrednotili s prvo normo vektorja $\mathbf{r}_y$, kar zapišemo z enačbo:

$$\Gamma_y^{(w)} = \| \mathbf{r}_y^{(w)} \|_1 . \quad (18)$$

Predpostavimo, da je šumni signal šibkeje koreliran od govornega signala in da je v točkah minimumov $\Gamma_y^{(w)}$ aktiven samo šum oz. da velja $\mathbf{r}_y = \mathbf{r}_d$. Stopnjo koreliranosti šumnega signala $\Gamma_d^{(w)}$ lahko takrat ocenimo s sledenjem minimumov funkcije $\Gamma_y^{(w)}$ oz. z enačbo, analogno enačbi (15). Primer ocene $\Gamma_d^{(w)}$ iz $\Gamma_y^{(w)}$ prikazuje slika 3 b). Izbrali smo časovni interval iskanja minimuma 0,5s in faktor $\beta = 0{,}94$.



Slika 3: Metoda VAD na podlagi avtokorelacije: a) Šumni govor, b) Ocena Γ_d na podlagi sledenja minimumov Γ_y, c) Funkcija ojačenja g_{VAD}

Figure 3. Autocorrelation-based VAD method: a) Noisy speech, b) Minimum tracking-based estimation of Γ_d from Γ_y, c) VAD gain function.

Definirajmo funkcijo za detekcijo aktivnosti govora z enačbo:

$$g_{VAD}^{(w)} = 1 - \frac{\mu \Gamma_d^{(w)}}{\Gamma_y^{(w)}} . \quad (19)$$

Pri tem je μ empirično izbrano realno število v mejah $0 < \mu \leq 1$, s katerim zmanjšamo vpliv precenitve $\Gamma_d^{(w)}$ oz. določimo spodnjo mejo funkcije $g_{VAD}^{(w)}$. Če je $\mu = 1$, vpliv precenitve $\Gamma_d^{(w)}$ ni kompenziran oz. je spodnja meja g_{VAD} enaka nič. Časovni potek g_{VAD} pri $\mu = 1$ prikazuje slika 3 c).

5 Eksperimentalni rezultati

Uspešnost predlagane metode smo potrdili z rezultati avtomatskega razpoznavanja govora v eksperimentalnih okoljih Aurora 2 [13] in 3 [16]. Okolje Aurora 2 sestavljajo zaporedja angleških števk različnih govorcev. Vsebuje čiste govorne signale, signale govora z dodanimi različnimi šumnimi signali pri različnih razmerjih *SNR* ter modele za učenje in testiranje. Govorno gradivo je razdeljeno na tri dele: A, B in C. Glede na učno gradivo so eksperimenti razdeljeni na dve učno-testni skupini: učenje na čistem govoru (UČG) ter učenje na čistem in šumnem govoru (UŠG). Govorno gradivo Aurore 3 je del večje multijezikovne baze SDC (ang.: "*SpeechDat Car*"). Zajema števke štirih evropskih jezikov: nemški, španski, danski in finski. Vsako zaporedje števk je posneto z bližnjim in oddaljenim mikrofonom. Glede na ujemanje učnega in testnega okolja so eksperimenti v okolju Aurora 3 razdeljeni na tri učno-testne skupine: dobro ujemanje (DU), srednje neujemanje (SN) in veliko neujemanje (VN). V obeh okoljih je izračun kepstralnih vektorjev realiziran po standardni Aurora WI007 predlogi [15] in razpoznavanje izvedeno z razpoznavalnikom HTK [14].

Predprocesiranje šumnega govornega signala smo izvajali z metodo signalnega podprostora s pomočjo transformacije FDCT. Avtokorelacijske vektorje $\mathbf{r}_y$ dimenzije 50 smo ocenjevali v intervalih z 800 otipki in s prekrivanjem med intervali 750 otipkov. Glede na podatek o frekvenci vzorčenja govornega gradiva f_s = 8kHz je bil čas trajanja intervalov 100ms in prekrivanje med sosednjimi intervali 93,75ms. V vsakem intervalu smo po enačbi (11) izračunali vektor $\hat{\boldsymbol{\lambda}}_y$ in z rekurzivno enačbo (15) komponente vektorja $\hat{\boldsymbol{\lambda}}_d$ pri K = 50. Pri tem smo izbrali rekurzivni koeficient β = 0,94 in število sosednjih intervalov za iskanje minimuma $2W$+1 = 80. Komponente vektorja $\hat{\boldsymbol{\lambda}}_s$ smo izračunali z enačbo (17) in komponente vektorja $\mathbf{g}_{SDC}$ z enačbo (10) pri K=M=50. Aktivnost govora smo upoštevali po enačbi (19). Koeficienta γ = 4 in μ = 0,5 smo določili empirično na podlagi rezultatov ARG. Čisti govorni signal smo ocenili s cenilko:

$$\mathbf{s}^{(w)} = \mathcal{T}\{\mathbf{B}^{-1} g_{VAD}^{(w)} \mathbf{g}_{SDC}^{(w)}\}\mathbf{y}^{(w)}. \qquad (20)$$

Pri tem je w indeks časovnega intervala, $\mathbf{B}^{-1}$ inverzna matrika $\mathbf{B}$ in $\mathcal{T}\{.\}$ Toeplitzov operator.

Absolutne vrednosti uspešnosti razpoznavanja besed v eksperimentalnem okolju Aurora 2 prikazuje tabela 1. Vrednosti so podane v odstotkih. Dosegli smo vrednosti 89,38% in 78,41% za primera učenja na šumnem (UŠG) ter na čistem govoru (UČG) oz. relativno izboljšanje rezultatov za 15,29% in 55,68% glede na referenco v [13]. Predvsem za učenje na šumnem govoru smo tako izboljšali rezultate, objavljene v [11] in [12], kjer je bilo doseženo relativno izboljšanje 30,57% oz. 35,42%.

Učenje	Del A	Del B	Del C	Vsota [a]
UŠG [b]	90,94	88,76	87,51	89,38
UČG [c]	79,60	74,73	83,40	78,41
Povprečje	85,27	81,75	85,46	83,90

Tabela 1: Absolutna uspešnost razpoznavanja besed v eksperimentalnem okolju Aurora 2. [a] Utežna vsota z utežmi 0,4, 0,4 in 0,2. [b] Učenje na šumnem govoru. [c] Učenje na čistem govoru.
Table 1. Absolute word accuracy results in Aurora 2 experimental framework. [a] Weighted sum with weights 0.4, 0.4 and 0.2. [b] Multicondition training. [c] Clean only training.

Absolutne vrednosti uspešnosti razpoznavanja besed v okolju Aurora 3 prikazuje tabela 2. Dosegli smo skupno absolutno uspešnost razpoznavanja besed 78,29% oz. relativno izboljšanje 10,86% glede na [16].

Jezik	Ujemanje učno-test. okolja			Vsota [d]
	DU[a]	SN[b]	VN[c]	
Finščina	90,53	72,50	30,35	69,17
Španščina	94,13	86,68	70,17	85,53
Nemščina	93,05	87,63	84,00	88,89
Danščina	85,89	64,41	50,59	69,55
Povprečje	90,90	77,81	58,78	78,29

Tabela 2: Absolutna uspešnost razpoznavanja besed v eksperimentalnem okolju Aurora 3. [a] Dobro ujemanje. [b] Srednje neujemanje. [c] Veliko neujemanje. [d] Utežna vsota z utežmi 0,4, 0,35 in 0,25.
Table 2. Absolute word accuracy results in Aurora 3 experimental framework. [a] Well matched. [b] Medium mismatch. [c] High mismatch. [d] Weighted sum with weights 0.4, 0.35 and 0.25.

Rezultati skupne absolutne uspešnosti razpoznavanja besed v okoljih Aurora 2 in 3, ki so 83,90% in 78,29%, ne presegajo rezultatov metode, podane v [17] (89,29% in 90,77%). V [17] podana metoda je vrhunec skupnih prizadevanj podjetij Motorola, Francoski telekom in Alcatel ter je vključena v standardni algoritem robustne parametrizacije govora [18].

Iz tabel 1 in 2 vidimo, da je predlagana metoda uspešna v obeh eksperimentalnih okoljih. Zmerno relativno izboljšanje v okolju Aurora 3 je posledica rezultatov ARG finskih števk, kjer nismo dosegli izboljšanja glede na rezultate v [16]. Predvidevamo, da je vzrok v hitri zaporedni izgovarjavi števk, kjer daje metoda ocene šuma s sledenjem minimumov slabše rezultate. Drugi vzrok je različno šumno okolje. Velik del števk finske baze je posnet z glasbo v ozadju. Glasba je močno koreliran signal in jo posledično predlagana metoda napačno opredeli kot govor. To potrjujejo rezultati v okolju Aurora 2, kjer je bila uspešnost metode odvisna od šumnega okolja in najslabša pri govoru iz ozadja (ang.: *babble*).

6 Sklep

V prispevku smo prestavili metodo zmanjševanja nivoja šuma na podlagi teorije signalnega podprostora s sprotno oceno šuma. Metodo, ki je primerna le za beli šum, smo posplošili za primere poljubnega, časovno spremenljivega šuma. Predlagali smo postopek ocenjevanja lastnih vrednosti šumnega in govornega signala na podlagi razmerja *SNR* ter postopek detekcije intervalov aktivnosti govora z uporabo avtokorelacijskih koeficientov. Uspešnost metode zmanjševanja nivoja šuma smo potrdili z rezultati ARG v eksperimentalnih okoljih Aurora 2 in Aurora 3. Dosegli smo skupno relativno izboljšanje razpoznavanja besed 35,49% za eksperimentalno okolje Aurora 2 in 10,86% za eksperimentalno okolje Aurora 3 glede na referenčne rezultate v [13, 16].

7 Literatura

[1] Y. Ephraim, H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, Volume: 3 Issue: 4 , July 1995, Page(s): 251 -266.

[2] S. H. Jensen, P. C. Hansen, S. D. Hansen, J. A. Sörensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 439-448, Nov. 1995.

[3] P. S. K. Hansen, "*Signal Subspace Methods for Speech Enhancement*", Ph.D. Thesis, Technical Univ. of Denmark, Lyngby, Denmark, Sept. 1997.

[4] J. Huang, Y. Zhao, "A DCT-Based Fast Signal Subspace Technique for Robust Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 6, pp. 747-751, Nov. 2000.

[5] N. Pavešić, "*Razpoznavanje vzorcev: Uvod v analizo in razumevanje vidnih in slušnih signalov"*, Fakulteta za elektrotehniko, Ljubljana, 2000.

[6] M. H. Hayes, "*Statistical digital signal processing and modeling*," John Wiley & sons, inc., New York, 1996.

[7] Deller, R. J., Proakis J. G. and Hansen, J. H. L., "*Discrete-Time Processing of Speech Signal*", Macmillan Publishing Company, 1993.

[8] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. of the 7th European Signal Proc. Conf.*, pp. 1182-1185, Sept. 1994.

[9] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *ICASSP '79*, vol. 4, pp. 208–211, Apr. 1979.

[10] J. Sohn, N. S. Kim, W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, Jan. 1999.

[11] B. Jarc, R. Babič, "Second Order Statistics Spectrum Estimation Method for Robust Speech Recognition," *Eurospeech 2001, Proceedings*, pp. 229-232, Sep. 2001.

[12] B. Jarc, R. Babič, "Izboljšanje natančnosti razpoznavanja govora z določanjem njegove aktivnosti na podlagi statističnega modela," *Elektroteh. vestn.*, 2002, zvez. 69, št. 1, str. 75-81.

[13] H. G. Hirsch, D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*; Paris, France, Sept. 18-20, 2000.

[14] S. Young, J. Odell, D. Ollason, V. Valtchev, P. Woodland, "*The HTK Book* (*for HTK Version 3.0*)", July 2000, Microsoft Corporation.

[15] ETSI standard document - ETSI ES 201 108 v1.1.1, "*Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm*", Feb. 2000.

[16] Aurora documents, AU/225/00, AU/271/00, AU/273/00, AU/378/01, Finnish, Spanish, German, Danish databases for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: description and baseline results, 2000.

[17] Aurora documents, "*Motorola - France Télécom - Alcatel Advanced Front End Proposal*," Adopted by ETSI for DSR advanced front-end evaluation, Jan 2002.

[18] ETSI standard document - ETSI ES 202 050 v1.1.5, *"Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithm*", Jan. 2007.

Bojan Jarc je diplomiral leta 1992, magistriral leta 1999 in doktoriral leta 2003 na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru. Zaposlen je na Fakulteti za elektrotehniko, računalništvo in informatiko, Inštitut za elektroniko, kot asistent. Njegovo raziskovalno področje so robustno avtomatsko razpoznavanje govora, obdelava signalov in digitalna sita.

Rudolf Babič je diplomiral leta 1970 in magistriral leta 1980 na Fakulteti za elektrotehniko v Ljubljani. Doktoriral je leta 1991 na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru. Zaposlen je na Fakulteti za elektrotehniko, računalništvo in informatiko, Inštitut za elektroniko, kot visokošolski predavatelj in je vodja Laboratorija za elektronske sisteme. Njegovo raziskovalno področje so obdelava signalov, načrtovanje in izdelava elektronskih vezij, sistemov in naprav ter načrtovanje in izvedba analognih in digitalnih sit.