UDC: 004.934.8'1 INFO M: str. 24-33

ZAVISNOST TAČNOSTI PREPOZNAVANJA GOVORNIKA OD IZBORA OBELEŽJA
COMPARISON OF THE AUTOMATIC SPEAKER RECOGNITION PERFORMANCE OVER STANDARD FEATURES

Milan Dobrović - Telekom Srbija, Vlado Delić - FTN Novi Sad,
Nikša Jakovljević - FTN Novi Sad, Ivan Jokić - FTN Novi Sad

REZIME: U ovom radu je dat pregled rezultata eksperimenata koji su imali za cilj utvrđivanje zavisnosti tačnosti prepoznavanja govornika od izbora obeležja. Razmatrana su standardna obeležja poput linearnih i perceptualnih prediktivnih koeficijenta (LPC i PLP), kao i mel-frekvencijski kepstralni koeficijenti (MFCC). Pored toga ispitana je mogućnost primene heteroscedastičke linearne diskriminativne analize (HLDA) kojom bi se povećale razlike između modela govornika. Govornici su modelovani pomoću modela mešavina Gausovih raspodela uz pomoć HTK alata. Uticaj složenosti ovih modela na tačnost prepoznavanja je takođe razmotren u ovom radu. Za obuku i testiranje je korišćeno 30 govornika iz studijske baze S70W100s120. Nešto bolje performanse su pokazali sistemi koji koriste MFCC i PLP obeležja. Primena HLDA je u većini slučajeva doprinela poboljšanju tačnosti pri čemu je to poboljšanje bilo manje što je bila veća tačnost referentnog sistema (sa istim obeležjima bez primene HLDA).

KLJUČNE REČI: Automatic Speaker Recognition, Gaussian Mixture Models, Mel-Frequency Cepstral Coefficients, Linear Prediction Coefficients, Perceptual Linear Prediction, Hidden Markov Model, HTK

ABSTRACT: This paper presents a study of speaker recognition accuracy depending on the choice of features, window width and model complexity. The standard features were considered, such as linear and perceptual prediction coefficients (LPC and PLP) and mel-frequency cepstral coefficients (MFCC). In addition, the application of Heteroscedastic Linear Discriminant Analysis (HLDA) was examined, in order to increase the difference between speaker models. Gaussian mixture model (GMM), with the use of HTK tools, was chosen for speaker modelling. Thirty speakers from the speech database S70W100s120 were used for system training and testing. It showed better system performance using MFCC and PLP features. Application of HLDA in most cases helped improve the accuracy, while that improvement was less as the accuracy of the reference system (the one with the same features without the use of HLDA) was higher.

KEY WORDS: Automatic Speaker Recognition, Gaussian Mixture Models, Mel-Frequency Cepstral Coefficients, Linear Prediction Coefficients, Perceptual Linear Prediction, Hidden Markov Model, HTK

1. UVOD

Govor je ljudima najprirodniji način za prenošenje informacija. Karakteristike ljudskog glasa jedinstvene su za svakog čoveka, pa je sasvim prirodno da ljude prepoznajemo po njihovom glasu. Automatsko prepoznavanje govornika predstavlja oblast digitalne obrade signala koja se odnosi na prepoznavanje ljudi od strane mašina na osnovu njihovog govora. U zavisnosti od namene prepoznavanje govornika može se podeliti na dva posebna slučaja: identifikaciju i verifikaciju. Zadatak verifikacije govornika je potvrđivanje ili odbacivanje tvrđenog identiteta na osnovu njegovog govora. Za razliku od verifikacije, u procesu identifikacije govornika odluka nije binarna jer sistem odlučuje ko je mogući govornik, kojoj grupi govornika pripada ili je nepoznata osoba [1].

U zavisnosti od toga da li je sistemu poznat tekst koji govornik izgovara razlikuje se prepoznavanje zavisno od teksta (tekst je zadat od strane sistema ili ga je u fazi obuke izabrao govornik) i nezavisno od teksta (tekst je proizvoljan i bira ga govornik u toku prepoznavanja). U ovom radu predstavljen je sistem za prepoznavanje govornika nezavisan od teksta.

Iz perspektive korisnika, sve aplikacije zasnovane na prepoznavanju govornika mogu se podeliti u dve grupe [2]:

1) one koje koristi sam govornik, i

2) one kod kojih su govornik i korisnik odvojeni.

U prvu grupu spadaju aplikacije koje uključuju kontrolu pristupa do informacija ili fizičkog pristupa zaštićenim (autorizovanim) zonama. U drugu grupu spadaju aplikacije koje sakupljaju informacije ili prate određenog govornika – tipično za forenzičke aplikacije. Osnovna razlika između ove dve grupe je i u ponašanju govornika: u prvoj grupi aplikacija govornik je kooperativan u procesu prepoznavanja dok u drugoj grupi govornik nije kooperativan ili je čak namerno opstruktivan.

U odeljku 2 predstavljene su pogodnosti korišćenja govora kao biometrijskog obeležja, kao i metodi koji se koriste u automatskom prepoznavanju govornika i modelovanju govornika. U odeljku 3 opisana je govorna baza, način realizacije sistema za prepoznavanje govornika kao i postupak obučavanja sistema. Dobijeni rezultati predstavljeni su u odeljku 4, nakon kog sledi zaključak.

2. TEORIJSKE OSNOVE

Do danas je predstavljen i istražen veliki broj postupaka koji se koriste u biometrijskim sistemima za prepoznavanje. Među najpopularnijim biometrijskim obeležjima su otisci prstiju, oblik lica i glas [3]. Svako biometrijsko obeležje ima svoje prednosti i mane, postoji više faktora zbog kojih se govorni signal koristi u biometriji [4]:

- Ne postoji ugrožavanje privatnosti, pošto zahtev za izgovaranje neke sekvence reči ljudi ne smatraju ugrožavanjem privatnosti.
- Postojanje velikog broja aplikacija u kojima je govor glavni (ako ne i jedini) signal koji je na raspolaganju – na primer telefonija.
- Jednostavno dobijanje podataka i prenos, što je posledica široke rasprostranjenosti telefonske mreže.
- Jeftini i lako dostupni uređaji za prikupljanje informacija. Kod aplikacija vezanih za telefonsku mrežu, u pristupnim tačkama ne moraju se instalirati posebni prenosnici ili mreže, pošto mobilni telefoni pružaju pristup gotovo svuda. Čak i za aplikacije koje nisu vezane za telefoniju, zvučne kartice i mikrofoni su jeftini i lako dostupni.

Ljudski govor pored lingvističkih informacija (informacije šta je rečeno) sadrži i čitav niz drugih informacija kao što su emocionalno stanje govornika, geografsko poreklo govornika, obrazovanje govornika, pol govornika ali i njegov identitet. Iako se sve ove informacije prenose u paraleli, ljudi lako ili bez mnogo problema raspoznaju svaku od njih. U zavisnosti od konkretne namene sistema zavisi koje informacije govornog signala su relevantne. Na primer, lingvističke informacije su relevantne ukoliko je cilj prepoznati sekvencu reči koja je izgovorena. Prisustvo irelevantnih informacija (poput okruženja i identiteta govornika) može negativno uticati na performanse sistema [1,4].

Osnovni principi koji se koriste kod automatskog prepoznavanja govornika su vrlo slični principima kod automatskog prepoznavanja govora. Parametrizacija govora predstavlja proces u kome se vrši transformacija odbiraka govora u sekvencu skupa karakteristika, tj. u sekvencu vektora obeležja. Karakteristike koje su esencijalne za prepoznavanje govornika menjaju se relativno sporo. Prema tome, izdvajanje obeležja je proces kompresije podataka pri čemu se zadržavaju informacije bitne za razlikovanje govornika [5].

U procesu automatskog prepoznavanja govornika izdvaja se nekoliko vrsta obeležja govora koja se danas koriste. Obeležja se mogu podeliti na [2]:

(1) obeležja niskog nivoa, odnosno obeležja zasnovana na: (a) spektralnim karakteristikama govora, (b) obliku vokalnog trakta i

(2) obeležja višeg i visokog nivoa koja se zasnivaju na: (a) prozodijskim karakteristikama, ritmu i brzini (b) semantičkim karakteristikama i dikciji.

Spektralne karakteristike i karakteristike vokalnog trakta predstavljaju obeležja niskog nivoa. Takva obeležja se mogu automatski izdvajati iz govora, pa se mogu koristiti kod automatizovanih sistema za prepoznavanje govornika. U pogledu pouzdanosti sistema koji koriste ove parametre za prepoznavanje govornika, uticaj imitacije (lažnog govornika) nije od većeg značaja. Najčešće korišćena obeležja u prepoznavanju govornika su LPC (*Linear Prediction Coefficients*), PLP (*Perceptual Linear Prediction*) i MFCC (*Mel-Frequency Cepstral Coefficients*). Ovo su ujedno i obeležja koja se koriste i pri prepoznavanju govora, što je posledica činjenice da oblik vokalnog trakta i glotalne pobude određuje kako identitet govornika tako i identitet fonema. Oblik vokalnog trakta određuje vrednosti rezonantnih učestanosti, a time i oblik obvojnice spektra, dok glotalna pobuda stepen i oblik harmonijske strukture u spektru. Iako se karakteristike govornika menjaju sporo tokom vremena, oblik vokalnog trakta i glotalna pobuda variraju sa promenom glasa stoga je i dalje potrebno govorni signal izdeliti na manje segmente za koje se može smatrati da su oblik vokalnog trakta i pobuda konstantni.

Obeležja višeg nivoa nose prozodijske karakteristike, ritam, značenje i dikciju i konverzacione karakteristike govornika, poput reči. Prozodijska obeležja kao i obeležja visokog nivoa karakterišu govorni iskaz kao celinu, na primer, osnovna učestanost i energija po intervalu vremena od nekoliko desetina do nekoliko stotina milisekundi. Ove karakteristike sadrže informacije o stilu govora neke osobe, pa se mogu koristiti pri prepoznavanju govornika. Dok se energija po intervalu računa na standardan način i ne predstavlja problem, računanje osnovne učestanosti svojevrstan je problem u digitalnoj obradi govora, naročito u praktičnoj primeni kada je govorni signal pod uticajem prenosnog kanala i okoline, a naročito u uslovima kada je odnos signal šum mali. Prozodijska obeležja se često dodaju parametrima baziranim na spektru te se tako dobija jedan složeni vektor.

Ipak, obzirom da su prozodijski parametri pod jakim uticajem samog lingvističkog konteksta rečenice, nisu našli široku primenu u savremenim sistemima za prepoznavanje govornika. Osim toga, ovi parametri su mnogo lakši za imitiranje od karakteristika vokalnog trakta. Zato je upotreba prozodijskih parametara u praksi veoma ograničena. Naročito ako je potrebno obezbediti veću pouzdanost, a postoji veliki rizik od zlonamernih korisnika koji bi mogli imitirajući da učine svoj glas sličan ciljnom glasu.

Klasični modeli govornika mogu se podeliti na šablonske i stohastičke modele [6], odnosno na neparametarske i parametarske modele, respektivno. Kod šablonskih modela karakteristično je da se trening i test vektori obeležja međusobno porede pod pretpostavkom da su oni međusobno nesavršene replike. Stepen izobličenja između njih predstavlja meru sličnosti, odnosno različitosti. Vektorska kvantizacija (VQ - *Vector quantization*) [7] i dinamičko vremensko slaganje (DTW - *dynamic time warping*) [8] predstavljaju primere šablonskih modela u prepoznavanju govornika nezavisno ili zavisno od teksta. Kod stohastičkih modela, svaki govornik je modelovan kao stohastički izvor nepoznate, ali nepromenljive funkcije gustine raspodele. U fazi obuke, iz trening podataka vrši se estimacija parametara funkcije gustine raspodele. Provera slaganja test iskaza govornika vrši se računanjem izglednosti modela govornika. U savremenim primenama prepoznavanja govornika zavisno od teksta, koriste se statistički modeli zasnovani na skrivenim Markovljevim modelima (HMM - *Hidden Markov Models*) baš kao i kod prepoznavanja govora. Kod prepoznavanja govornika nezavisno od teksta, koristi se poseban slučaj skrivenih Markovljevih modela sa jednim stanjem. Takvi modeli se nazivaju modeli mešavina Gausovih raspodela (GMM - *Gaussian Mixture Models*).

Zbog svoje efikasnosti i lakoće primene, trenutno dominantan pristup modelovanju govornika predstavlja model mešavina Gausovih raspodela [9], kod koje je gustina raspodele data izrazom:

$$p(x|\lambda)=\sum_{k=1}^{K} w_k \cdot p(x|\lambda_k)=\sum_{k=1}^{K} w_k \cdot \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right] \quad (1)$$

gde je x n-dimenzionalni vektor obeležja, w_k, μ_k i Σ_k težina, srednja vrednosti i kovarijansna matrica k-te Gausove raspodele respektivno, a K broj Gausovih raspodela koje čine mešavinu. Za težine Gausovih raspodela koje čine jednu mešavinu važi ograničenje da je $0 \leq w_k \leq 1$ i $\Sigma_{k=1}^{K} w_k = 1$. Prostor obeležja svakog od govornika (uključujući i univerzalnog govornika) je opisan sa po jednom mešavinom Gausovih raspodela.

Kovarijansne matrice GMM komponenti mogu biti: pune matrice, dijagonalne matrice i sferične matrice. Pune matrice sastoje se od dijagonalnih elemenata koji predstavljaju varijansu svake dimenzije u prostoru obeležja. Nedijagonalni elementi ukazuju na korelaciju između dimenzija. Dijagonalne kovarijansne matrice imaju samo dijagonalne elemente različite od nule. Dijagonalna kovarijansna matrica se primenjuje ako ne postoji korelacija između različitih dimenzija u prostoru obeležja, ali se mogu koristiti i u situacijama kada postoji izvesna korelacija između različitih dimenzija prostora obeležja. Sferične kovarijansne matrice [10] predstavljene su istom vrednošću varijanse. Ova vrednost je ista za elemente na dijagonali matrice, pri čemu su nedijagonalni elementi su jednaki nuli. Ova vrsta kovarijansne matrice može se koristiti kada ne postoji korelacija između dimenzija u prostoru obeležja i varijanse su iste za sve dimenzije. Izbor tipa kovarijansne matrice zavisi prvenstveno od potrebne preciznosti i raspoložive memorije, kao i same primene. Kovarijanse Gausovih raspodela u ovom radu aproksimirane su dijagonalnim kovarijansnim matricama da bi se smanjila kompleksnost modela.

GMM se koristi u prepoznavanju govornika iz dva razloga [11]. Realno je pretpostaviti da se glas osobe može okarakterisati skupom akustičkih klasa, na primer različitih fonetskih klasa. Pojedinačne komponente u GMM treba da modeluju ishodišne akustičke klase. Akustičke klase ukazuju na određene karakteristike vokalnog trakta koje mogu biti zavisne od govornika. Drugi razlog za korišćenje GMM je to što može da pruži dobre aproksimacije gustine verovatnoće proizvoljnih oblika. GMM je pogodan u situacijama kada podaci datog modela imaju veliki broj lokalnih maksimuma. Lokalni maksimumi su oblasti prostora obeležja gde su vektori obeležja „gušće“ raspoređeni. Svaka komponenta može se predstaviti Gausovom raspodelom.

Pri implementaciji sistema za prepoznavanje govornika vrlo često se nameće potreba obučavanja pozadinskog modela - UBM (*Universal Background Model*), koji je zapravo veoma veliki GMM obučavan da predstavlja raspodelu obeležja govora [12] nezavisno od govornika, odnosno svih govornika uopšte. UBM se koristi kao mogući alternativni model govornika tokom procesa verifikacije. Takođe, koristi se i u sistemima za prepoznavanje govornika u otvorenom skupu. U procesu obučavanja UBM različiti parametri imaju uticaja. Njih možemo podeliti u dve kategorije: a) parametri algoritma i b) parametri podataka [13]. Parametri algoritma odnose se na promene u procesu obuke, poput broja mešavina, metoda obuke, broja iteracija, metoda inicijalizacije itd. Parametri podataka obuhvataju različite načine definisanja podskupa podataka koji su na raspolaganju za obučavanje modela, poput govorne baze, količine podataka, broja govornika, količine podataka po govorniku, metoda izbora govornika, načina korišćenja vektora obeležja, balansiranja podataka prema kanalu, mikrofonu, jeziku ili nekoj drugoj promenljivoj.

Za obuku modela je obično na raspolaganju relativno mali broj opservacija, tako da se javlja problem efikasnosti estimacije parametara statističkih modela. ($\{w_k, \mu_k$ i $\Sigma_k\}, k = 1...K$). Problem postaje izraženiji što je dimenzija vektora obeležja veća. Pored problema efikasnosti estimacije parametara, pojedine procedure za obuku koje su efikasne u slučaju male dimenzionalnosti prostora obeležja postaju netraktabilne s povećanjem dimenzionalnosti prostora obeležja. Ovaj problem se često naziva prokletstvom velike dimenzionalnosti.

Jedan od načina da se prevaziđe problem velike dimenzionalnosti prostora obeležja, jeste smanjenje broja parametara koje treba proceniti, što se u slučaju mešavine Gausovih raspodela obično postiže aproksimacijom pune kovarijansne matrice dijagonalnom. Dijagonalizacija kovarijansne matrice odgovara uvođenju pretpostavke da su obeležja koja čine opservaciju međusobno nekorelisana. Jedan od načina da se obeležja dekorelišu i da se smanji dimenzionalnost prostora jeste primenom raščlanjivanja na osnovne komponente (PCA – *Principal Component Analysis*) ili primenom linearne diskriminantne analize (LDA – *Linear Discriminant Analysis*). U ovom radu je iskorišćena heteroscedastička LDA [14] koja transformiše prostor tako da se maksimizuje rastojanje između različitih klasa i mimimizuje rasipanje opservacija unutar pojedinačne klase.

3. REALIZACIJA SISTEMA

Za obuku i testiranje sistema korišćen je deo govorne baze S70W100s120 [15] - baza sadrži iskaze 120 govornika koji su izgovarali po 70 rečenica i oko 100 izolovano izgovorenih reči. Baza je originalno snimljena na magnetofonsku traku, u gluvoj sobi ETF-a u Beogradu, ali je naknadno u okviru AlfaNum projekta konvertovana u digitalni zapis (16 bita po odmerku i učestanošću odabiranja 22050*Hz*). Iz ove baze je na slučaj izabrano 30 govornika. Da bi se sa relativno malim brojem govornika dobila adekvatna procena performansi realnog sistema, svi govornici su istog (muškog) pola.

Tokom obuke sistema formirano je 10 modela govornika, UBM, kao i model tišine. Za obuku svakog pojedinačnog modela govornika korišćeno je 11 iskaza u kojima govornik izgovara pojedinačno reči. Za obuku jednog govornika u proseku je bilo na raspolaganju 40 sekundi samog govora (sam govorni signal bez pauza). UBM je obučen na osnovu iskaza drugih 10 govornika (ukupno 110 iskaza u kojima su takođe izgovarane izolovane reči) koji su predstavljeni kolektivnim identitetom kao uljezi. Količina podataka korišćena za obučavanje UBM nakon uk-

lanjanja tišine iznosila je oko 400 sekundi. Pošto tišina ne nosi nikakvu informaciju o identitetu govornika, a može značajno da naruši stvarne karakteristike modela govornika, pristupilo se labeliranju podataka korišćenih za obučavanje modela govornika, UBM i modela tišine. U fazi testiranja sistema korišćeno je 100 iskaza, po 5 iskaza svakog od 20 govornika. Test skup činili su iskazi 10 govornika čiji su individualni modeli bili formirani u toku obuke sistema, kao i iskazi 10 govornika za koje nije generisan poseban model, niti su njihovi iskazi korišćeni za obuku UBM. Na ovaj način je obezbeđeno da su skup podataka za obuku modela i skup podataka za testiranje sistema bili disjunktni, kao i da uslovi testiranja budu bliski stvarnim gde postoje uljezi koji nisu poznati sistemu. Prosečno trajanje iskaza za testiranje sistema iznosilo je 3,5*s* (odnosno 1,8*s* ako se izuzme tišina između izgovorenih reči), što daje ukupno trajanje test skupa od oko 6 minuta (odnosno 3 minuta).

Za potrebe obuke i testiranja sistema korišćen je softverski alat *HTK* (*Hidden Markov Models ToolKit*) koji je prvenstveno namenjen sistemima za prepoznavanje govora zasnovanim na skrivenim Markovljevim modelima [16]. Pošto smo se u radu ograničili da govornika modelujemo pomoću mešavine Gausovih raspodela, skriveni Markovljev model ima samo jedno emitujuće stanje.

Za obuku sistema za prepoznavanje govornika korišćeni su HTK alati: HCopy, HCompV, HERest, HHed.

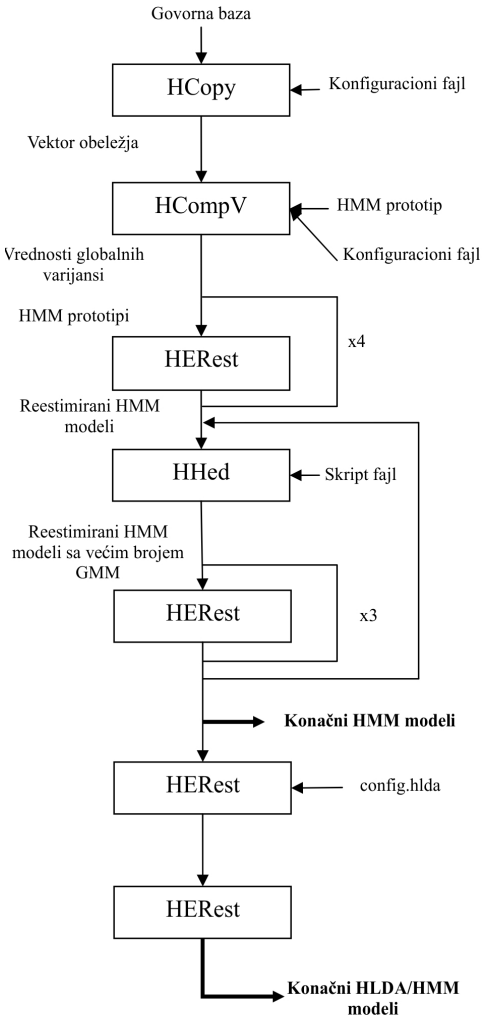
Na slici 1 je prikazan proces obuke sistema u HTK. Izdvajanje vektora obeležja izvršeno pomoću alata HCopy. U laboratorijskim uslovima pošto su svi audio fajlovi (kako iz skupa za obuku tako i iz skupa za testiranje) unapred dostupni obično se izdvajanje obeležja za sve fajlove (kako one koji se koriste za obuku tako i one koji se koriste za testiranje) vrši na početku obuke sistema, što je i ovde bio slučaj. Ulaz predstavljaju govorna baza za obuku u *.wav* formatu, i konfiguracioni fajl u kome su dati parametri na osnovu kojih se vrši konverzija govora u parametarski oblik. Izlaz predstavlja baza za obuku u parametarskom obliku.

Inicijalizacija korišćenih modela izvršena je upotrebom HTK alata HCompV. Ova funkcija učitava ulazni HMM model i govornu bazu za obuku, a kao izlaz daje novi HMM model čije srednje vrednosti i varijanse su jednake globalnoj srednjoj vrednosti i varijansi. Vrednosti globalne varijanse u narednim koracima koriste se kao prag, čime se sprečava da pojedini modeli usled malog broja observacija u skupu za obuku imaju suviše male vrednosti varijansi.

Procena modela govornika izvršena je primenom funkcije HERest. HERest istovremeno ažurira sve HMM modele koristeći celokupnu govornu bazu za obuku. Ukratko, HERest radi na sledeći način. HERest učitava sve HMM modele i uz svaki govorni fajl mora postojati njemu pridružena odgovarajuća transkripcija u obliku *label* fajla. Nakon učitavanja govornog fajla u memoriju, HERest koristi pridruženu transkripciju i generiše kombinovani HMM. Ovaj model se zapravo dobija spajanjem pojedinih modela koji odgovaraju labelama navedenim u transkripciji. Nakon toga primenjuje se *Forward-Backward* algoritam nad čitavim kombinovanim modelom. Za procenu modela govornika funkcija HERest je korišćena u pet iteracija.

Za povećanje broja Gausovih raspodela po stanju korišćena je funkcija HHed i odgovarajući skript fajl. Konverzija iz HMM sa jednom Gausovom raspodelom u HMM sa više Gausovih raspodela obično je jedan od poslednjih koraka u obučavanju sistema. Mehanizam povećanja Gausovih raspodela po stanju se naziva podela komponenti. Ovaj postupak je veoma fleksibilan jer omogućava postepeno povećanje Gausovih komponenti. Pri obučavanju sistema broj Gausovih raspodela je povećavan postepeno za jedan. Svako povećanje broja Gausovih raspodela praćeno je sa četiri iteracije HERest.

Pored toga što se koristi za estimaciju standardnih parametara modela, alat HERest je korišćen i za estimaciju HLDA projekcija. Za konačnu estimaciju HLDA, HERest je korišćena u dve iteracije: u prvoj iteraciji uz korišćenje konfiguracionog fajla *config.hlda* u kome se definisani parametri HLDA transformacije.



Slika 1:*Obuka sistema u HTK*

Postupak testiranja performansi sistema za prepoznavanje govornika izvršen je pozivima funkcija HVite i HResults. Pre toga, alatom HParse kreirana je mreža na osnovu gramatike u kojoj su navedeni dozvoljeni modeli govornika. Mreža je bila ograničena

na strukturu u kojoj je bilo moguće da se u okviru jednog fajla pojavi samo jedan govornik. Nakon što je formirana mreža koja definiše moguće realizacije i prelaze stanja moglo se pristupiti prepoznavanju govornika pomoću funkcije HVite. Zadatak prepoznavanja govornika je da nađe najverovatniju putanju kroz mrežu za dat nepoznat govorni fajl i HMM modele. U ovom koraku nije bila postavljena margina pouzdanosti kojom bi se tražilo da akumulisana izglednost bude za neku unapred zadatu vrednost veća od mogućih kompetitivnih modela. Alat HResult služi za analizu dobijenih rezultata i generiše odgovarajuće statistike vezane za performanse sistema.

4. REZULTATI

Kao što je u uvodu napomenuto razmotrene su tri vrste standardnih obeležja koja se koriste za opisivanje govornog signala i to;MFCC, PLP i LPC [17]. Pored samih obeležja varirana je i širina analizatorskog prozora (frejma). Za širinu su izabrane standardne vrednosti 20*ms*, 25*ms* i 30*ms*, ali i nešto veće vrednosti 40*ms*, 50*ms* i 100*ms*. Motivacija za ispitivanje sa nešto širim analizatorskim prozorom leži u činjenici da nije bitan tačan oblik vokalnog trakta već njegov prosek, odnosno da su karakteristike govornika sporopromenljive. Bez obzira na širinu uvek je korišćena Hamingova prozorska funkcija. Vrednost pomeraja analizatorskog prozora iznosila je 10*ms* i nije menjana sa promenom veličine prozora da bi broj opservacija koji se koristi pri obuci modela ostao isti. Složenost modela nije bila unapred fiksirana tako da su modeli govornika i UBM bili opisani sa brojem Gausovih raspodela, *K*, od 1 do 64. Dimenzija prostora obeležja je zavisila od konkretnih obeležja koja su korišćena i iznosila je 39 u slučaju MFCC i PLP, odnosno 30 u slučaju LPC. U slučaju LPC obeležja za potrebe ovog rada ispitani su rezultati prepoznavanja govornika i za nestandardne vrednosti, njihov broj je povećan na 39 odnosno 48.

Radi preglednosti u tabelama 1 - 4 prikazana je zavisnost tačnosti prepoznavanja govornika od veličine analizatorskog prozora i broja Gausovih raspodela po stanju (izdvojene sledeće vrednosti K= 1, 2, 4, 8, 16, 32, 64) za različite vektore obeležja. Tabela 1 se odnosi na eksperimente u kojima je korišćeno 12 MFCC koeficijenata i tzv. nulti mel-frekvencijski kepstralni koeficijent, kao i njihove prve i druge izvode (MFCC_0_D_A). Kao što se iz priloženog može videti, povećanjem broja Gausovih raspodela koje se koriste za modelovanje generalno raste i tačnost prepoznavanja govornika. Za prozore koji ne obuhvataju stacionarne segmente govora (čije je trajanje veće od 30*ms*) ovaj trend nije monoton. Pretpostavljamo da je to posledica veće korelisanosti opservacija, koja je u ovim slučajevima značajna. Interesantno je da se samo u slučaju kad se koristi analizatorski prozor dužine 100*ms*, ni za relativno velik broj Gausovih raspodela u modelu ne dobija sistem dovoljno visoke tačnosti (preko 95%). Treba primetiti da je već oko 30-ak Gausovih raspodela dovoljno za kvalitetno modelovanje jednog govornika, a da se duplo većim brojem Gausovih raspodela ne dobijaju značajno bolje performanse.

Tabela 1: *Uspešnost prepoznavanja govornika [%] za vektor obeležja MFCC_0_D_A*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	41	40	37	39	36	36
2GMM	35	32	27	24	26	39
4GMM	35	38	37	37	37	50
8GMM	54	53	50	24	25	46
16GMM	93	94	82	58	40	67
32GMM	97	96	96	96	95	75
64GMM	97	97	94	93	94	53

U slučaju PLP obeležja analizirane su dve varijante koje se najčešće sreću u literaturi. U prvoj varijanti ispitane su performanse sistema za prepoznavanje govornika za vektor obeležja koji sadrži prvih 12 PLP koeficijenata i nulti kepstralni koeficijent, kao i njihove prve i druge izvode (PLP_0_D_A) čiji su rezultati predstavljeni u tabeli 2. Kao i u slučaju MFCC obeležja može se primetiti da duži analizatorski prozor znači i lošije performanse. Pored toga s povećanjem broja Gausovih raspodela raste i tačnost prepoznavača. Analizatorski prozor dužine 100*ms* i za ovu kombinaciju obeležja ne rezultuje sistemom zadovoljavajuće tačnosti bez obzira na složenost modela. I u ovoj varijanti optimalan broj Gausovih raspodela je oko 30, a duplo više Gausovih raspodela po stanju ne donosi značajno poboljšanje tačnosti.

Tabela 2: *Uspešnost prepoznavanja govornika [%] za vektor obeležja PLP_0_D_A*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	40	38	39	39	41	42
2GMM	39	57	60	60	58	54
4GMM	64	64	63	74	58	17
8GMM	79	82	51	47	46	49
16GMM	86	88	87	74	46	83
32GMM	95	97	99	94	86	56
64GMM	96	97	96	96	91	70

Performanse sistema koje su dobijene za drugu varijantu PLP obeležja (prvih 13 PLP koeficijenata i njihove prve i druge izvode (PLP_D_A) prikazane su u tabeli 3. Ovi rezultati su takođe na nivou rezultata postignutih za MFCC. Interesantno je zapaziti da za sve veličine Hamingovog prozora postoji značajan pad uspešnosti prepoznavanja govornika za 64 GMM u odnosu na slučaj 32 GMM, što je verovatno posledica preobučenosti sistema.

Tabela 3: *Uspešnost prepoznavanja govornika [%] za vektor obeležja PLP_D_A*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	32	32	35	38	39	38
2GMM	50	46	43	38	38	42
4GMM	54	25	23	23	22	62
8GMM	28	39	39	63	51	30
16GMM	87	85	84	70	82	62
32GMM	99	97	92	95	95	81
64GMM	86	89	86	83	78	79

Na kraju, ispitane su performanse sistema za prepoznavanje govornika za vektor obeležja koji sadrži prvih 10 LPC koeficijenata i njihove prve i druge izvode (LPC_D_A). Rezultati, prikazani u tabeli 4 pokazuju da LPC obeležja daju od 20% do 30% lošije rezultate prepoznavanja govornika u odnosu na PLP i MFCC obeležja, i tačnost je suviše niska da bi sistem bio uspešno korišćen. Ispitana je tačnost prepoznavanja govornika i za slučaj vektora obeležja prvih 13, odnosno 16, LPC koeficijenata i njihovih prvih i drugih izvoda. Međutim, tačnost prepoznavanja govornika nije se značajno povećala (poboljšanje u odnosu na rezultate prikazane u tabeli 4 je iznosilo 5% za vektor obeležja dimenzije 39, odnosno 7 % za vektor obeležja dimenzije 48).

Tabela 4: *Uspešnost prepoznavanja govornika [%] za vektor obeležja LPC_D_A*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	11	11	12	12	12	15
2GMM	13	13	12	12	11	9
4GMM	10	8	8	6	6	13
8GMM	56	42	59	65	67	23
16GMM	37	40	39	25	42	35
32GMM	66	66	65	51	66	57
64GMM	71	70	38	73	70	20

Slika 2 prikazuje uporedne rezultate za različite vektore obeležja (MFCC_0_D_A, PLP_0_D_A, PLP_D_A, LPC_D_A) i veličinu prozora 25*ms* i broj Gausovih raspodela od 1 do 64. Rezultati pokazuju da primena LPC obeležja daje znatno lošije rezultate prepoznavanja govornika u odnosu na PLP i MFCC obeležja.

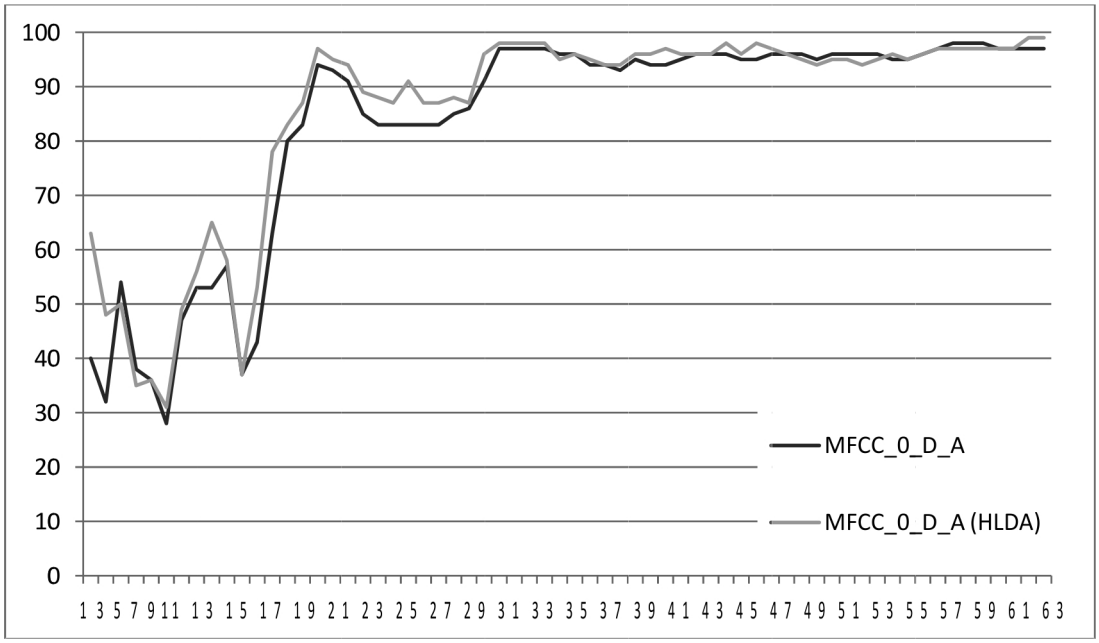
Pri korišćenju HLDA transformacije nad standardnim vektorima obeležja dimenzionalnost prostora vektora obeležja je smanjena za 5. Nakon HLDA transformacije, dimenzionalnost vektora obeležja je n=34 za MFCC i PLP i n=25 za LPC. U tabelama 5 - 8 prikazana je, nakon primene HLDA, zavisnost tačnosti prepoznavanja govornika od veličine analizatorskog prozora i broja Gausovih raspodela po stanju (izdvojene sledeće vrednosti *K= 1,2,4,8, 16, 32, 64*) za različite vektore obeležja.

U tabeli 5 prikazana je uspešnost prepoznavanja govornika za MFCC_0_D_A vektor obeležja nakon HLDA transformacije za različite širine analizatorskog prozora. Slika 3 prikazuje uporedne rezultate uspešnosti prepoznavanja govornika za vektor obeležja MFCC_0_D_A pre i nakon primene HLDA za analizatorski prozor širine 25*ms*. Ispitivanje je pokazalo da nakon primene HLDA, za mali broj Gausovih raspodela po stanju (1GMM i 2GMM) evidentno je poboljšanje performansi za oko 20%. Može se primetiti da se već za modele sa po 16 raspodela postiže dovoljna visoka tačnost, prepoznavanja, naravno u slučajevima kada je analizatorski prozor dovoljno mali.

Tabela 5: *Uspešnost prepoznavanja govornika [%] za vektor obeležja MFCC_0_D_A nakon primene HLDA*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	53	63	59	63	60	54
2GMM	52	48	48	50	52	70
4GMM	37	35	37	45	44	81
8GMM	56	56	55	33	29	64
16GMM	95	97	93	71	39	61
32GMM	100	95	96	93	95	75
64GMM	99	99	97	96	96	71

Slika 2: *Uspešnost prepoznavanja govornika [%] za različite vektore obeležja i analizatorski prozor od 25ms*



Slika 3: *Uspešnost prepoznavanja govornika [%] pre i nakon primene HLDA na MFCC_0_D_A za analizatorski prozor od 25ms*

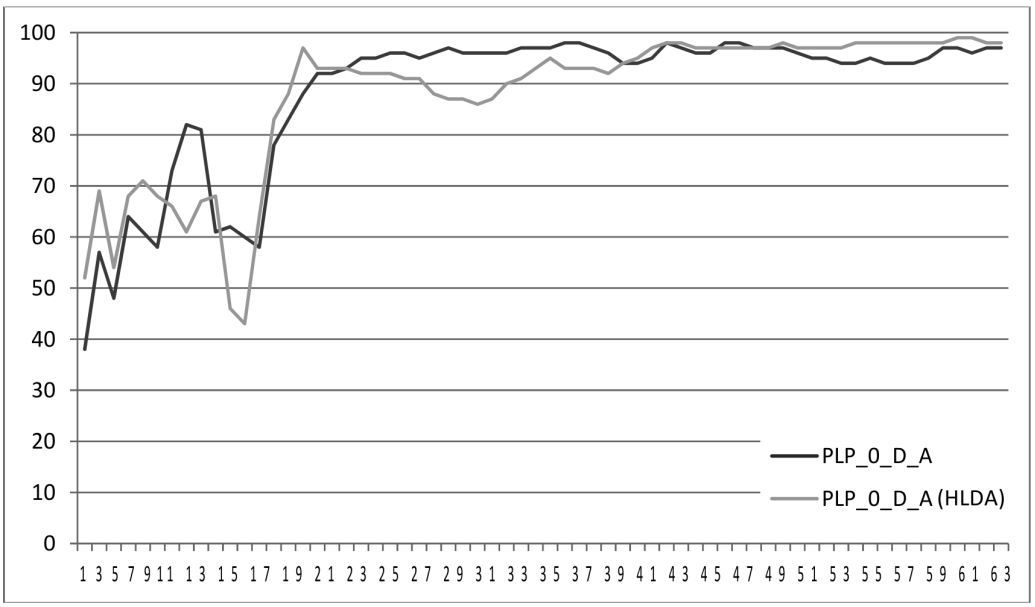
U tabeli 6 prikazana je uspešnost prepoznavanja govornika za vektor obeležja PLP_0_D_A nakon HLDA transformacije za različite širine analizatorskog prozora. Eksperimenti pokazuju slične rezultate kao i za MFCC obeležja; za mali broj Gausovih raspodela po stanju (1GMM i 2GMM) evidentno je poboljšanje performansi za oko 15% ali i dalje nedovoljno visoku tačnost. Sa druge strane i u ovom slučaju se za oko 16 raspodela po modelu dobija sistem zadovoljavajuće tačnosti. Za veći broj Gausovih raspodela HLDA ne pruža nikakvo ili minimalno poboljšanje.

Slika 4 prikazuje uporedne rezultate uspešnosti prepoznavanja govornika za vektor obeležja PLP_0_D_A pre i nakon primene HLDA za analizatorski prozor širine 25*ms*. Sličan odnos važi i za druge širine analizatorskih prozora koji su manji od 40*ms*.

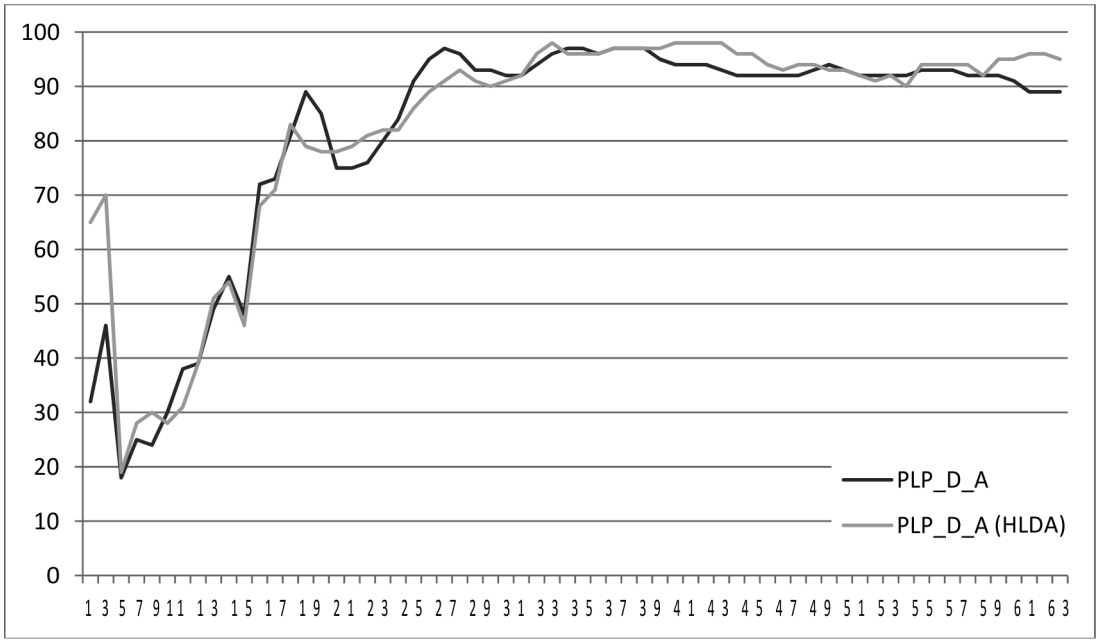
Tabela 6: *Uspešnost prepoznavanja govornika [%] za vektor obeležja PLP_0_D_A nakon primene HLDA*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	51	52	52	56	58	59
2GMM	66	69	62	64	60	59
4GMM	66	68	67	69	62	48
8GMM	59	61	47	49	49	56
16GMM	97	97	91	75	42	85
32GMM	96	93	96	92	88	74
64GMM	98	98	98	97	98	78

U tabeli 7 prikazana je uspešnost prepoznavanja govornika za vektor obeležja PLP_D_A nakon HLDA transformacije za



Slika 4: *Uspešnost prepoznavanja govornika [%] pre i nakon primene HLDA na PLP_0_D_A za analizatorski prozor od 25ms*

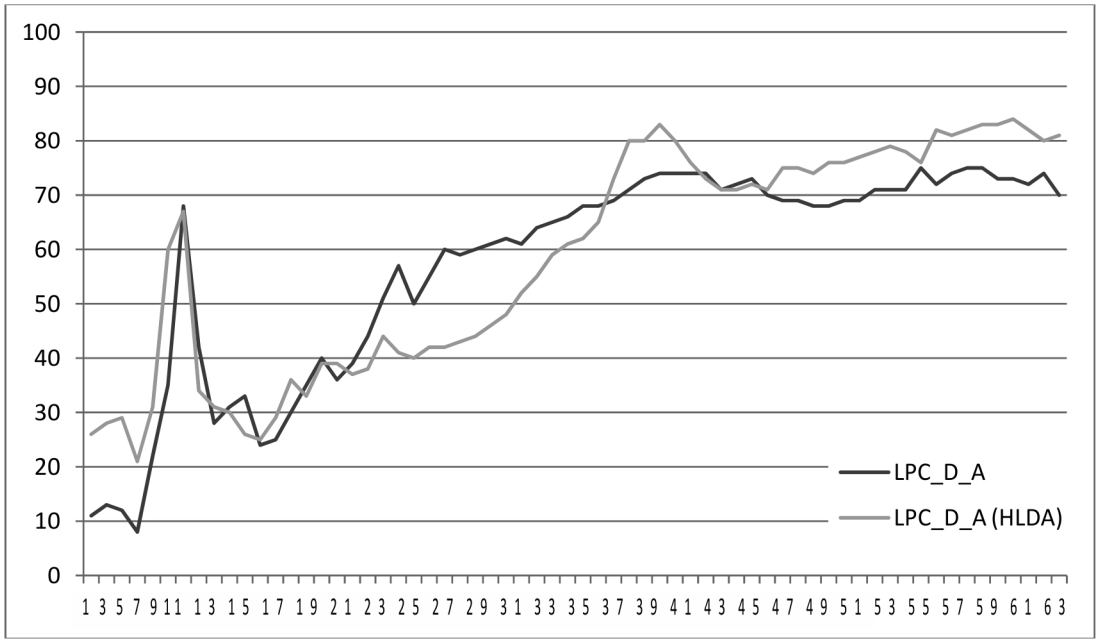


Slika 5: *Uspešnost prepoznavanja govornika [%] pre i nakon primene HLDA na PLP_D_A za analizatorski prozor od 25ms*

različite širine analizatorskog prozora. Slika 5 prikazuje uporedne rezultate uspešnosti prepoznavanja govornika za vektor obeležja PLP_D_A pre i nakon primene HLDA za analizatorski prozor širine 25*ms*. Eksperimenti pokazuju slične rezultate kao i prethodnim slučajevima; za mali broj Gausovih raspodela po stanju (1GMM i 2GMM) zabeleženo je poboljšanje performansi za oko 25% do 30%. Za veći broj Gausovih raspodela HLDA ne pruža nikakvo ili minimalno poboljšanje u prepoznavanju govornika.

Tabela 7: *Uspešnost prepoznavanja govornika [%] za vektor obeležja PLP_D_A nakon primene HLDA*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	62	65	67	59	59	61
2GMM	75	70	70	79	67	80
4GMM	38	28	27	29	30	62
8GMM	42	39	41	59	64	72
16GMM	84	78	85	75	70	68
32GMM	91	96	98	94	94	76
64GMM	98	95	94	90	94	79



Slika 6: *Uspešnost prepoznavanja govornika [%] pre i nakon primene HLDA na LPC_D_A za analizatorski prozor od 25ms*

U tabeli 8 prikazana je uspešnost prepoznavanja govornika za vektor obeležja LPC_D_A nakon HLDA transformacije za različite širine analizatorskog prozora. Slika 6 prikazuje uporedne rezultate uspešnosti prepoznavanja govornika za vektor obeležja LPC_D_A pre i nakon primene HLDA za analizatorski prozor širine 25*ms*. Rezultati primene HDLA na LPC obeležja su srazmerno lošiji u odnosu na rezultate MFCC i PLP obeležja. Uočen je porast uspešnosti prepoznavanja od oko 15% za mali broj Gausovih raspodela po stanju (1GMM, 2GMM i 4GMM), što je i dalje daleko od dovoljnog za bilo kakvu uspešnu primenu ovakvog sistema za prepoznavanje govornika. Za veći broj Gausovih raspodela HLDA ne pruža nikakvo ili minimalno poboljšanje u prepoznavanju govornika. Za slučaj 64GMM uspešnost prepoznavanja je porasla za 11%.

Tabela 8: *Uspešnost prepoznavanja govornika [%] za vektor obeležja LPC_D_A nakon primene HLDA*

	20ms	25ms	30ms	40ms	50ms	100ms
1GMM	24	26	27	27	28	26
2GMM	27	28	28	28	30	31
4GMM	20	21	18	23	18	23
8GMM	63	34	70	71	70	20
16GMM	35	39	37	30	44	34
32GMM	66	61	62	56	66	63
64GMM	74	81	69	85	80	33

5. ZAKLJUČAK

Postignuta tačnost sistema za prepoznavanje koji koriste PLP i MFCC obeležja uz odgovarajuću složenost modela obeležja je prilično visoka dostiže i 99%. LPC obeležja daju značajno lošije rezultate u odnosu na prethodno pomenuta iako su svi iskazi bili bez prisustva šuma, što je obično problem za LPC.

Ovo ispitivanje je takođe pokazalo da je optimalna širina prozora od 20 do 40*ms*. Iako se karakteristike govornika menjaju sporo, prozorske funkcije veće širine ne daju bolje rezultate. To ukazuje da je dinamika govora takođe značajna za prepoznavanje govornika. Rezultati pokazuju da za prozore širine 100*ms* i veći broj Gausovih komponenti po stanju dolazi do pada performansi sistema za prepoznavanje govornika bez obzira na to koji vektor obeležja se koristi.

Povećanje broja Gausovih raspodela (odnosno složenosti modela) ne mora obavezno da rezultuje boljim prepoznavanjem od strane mašine iako modeli nisu preobučeni (Tačnost prepoznavanja sistema sa 1 i 32 Gausove raspodele je viša nego za sistem sa 4 Gausove raspodele). Razlog tome može biti to da kada model ima samo jednu Gausovu raspodelu, onda su srednje vrednosti dovoljno udaljene u prostoru obeležja što pozitivno utiče na prepoznavanje govornika, iako je sam model veoma grubo aproksimiran u prostoru obeležja. U slučaju relativno malog broja GMM, model je (kao i u slučaju modela sa jednom GMM) i dalje suviše prost da bi opisao govornika u prostoru obeležja, ali deli taj prostor prema svakom govorniku na takav način da određeni regioni prostora obeležja budu predstavljeni bolje jednim govornikom u odnosu na nekog drugog. Kada broj GMM i dalje raste, aproksimacija raspodele obeležja je bolja, i time se smanjuje verovatnoća greške prepoznavanja.

Primena HLDA pružila je bolje rezultate tačnosti prepoznavanja govornika za jednostavne modele koji se sastoje od jedne ili dve Gausove raspodele po stanju, ali i dalje nedovoljno za neku praktičnu primenu. Povećavanjem složenosti modela govornika primena HLDA nije dala nikakva značajna poboljšanja, što dokazuje ispravnosti pretpostavke o dovoljno maloj korelisanosti koeficijenata MFCC, PLP i LPC vektora obeležja.

LITERATURA

[1] Speaker recognition, Joseph P. Campbell, Jr.Department of Defense Fort Meade, MD

[2] Zoran Ćirović, „Sistem za verifikaciju govornika zasnovan na elektroglotografskom signalu“, doktorska disertacija, ETF Beograd, 2011.

[3] “ An Exploration of Voice Biometrics” Lisa Myers, April, 2004

[4] S. Furui. Recent Advances in Speaker Recognition. *Pattern Recognition Letters*, 18:859–872, 1997.

[5] I. Jokić, S. Jokić, Z. Perić, M. Gnjatović, V. Delić, “Influence of the Number of Principal Components Used to the Automatic Speaker Recognition Accuracy”, scheduled for publication in the journal Electronics and Electrical Engineering – Kaunas: Technologija, ISSN 1392-1215, in No. 7(123), September of 2012.

[6] Campbell, J. Speaker recognition: a tutorial. Proceedings of the IEEE 85, 9 (September 1997), 1437–1462.

[7] Soong, F., A.E., A. R., Juang, B.-H., and Rabiner, L. A vector quantization approach to speaker recognition. AT & T Technical Journal 66 (1987), 14–26.

[8] Furui, S. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing 29, 2 (April 1981), 254–272.

[9] Joseph P. Campbell, Wade Shen, William M. Campbell, Reva Schwartz, Jean-François Bonastre, and Driss Matrouf “Forensic Speaker Recognition”, IEEE Signal Processing Magazine, March 2009

[10] Tree-based Gaussian Mixture Models for Speaker Verification by Francois Dirk Cilliers, Thesis presented at the University of Stellenbosch in partial fulfilment of the requirements for the degree of Master of Science in Electronic Engineering Department of Electrical and Electronic Engineering University of Stellenbosch Private Bag X1, 7602 Matieland, South Africa Study leader: Prof J.A. du Preez December 2005

[11] D. Reynolds, R. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE transactions on speech and audio processing, Vol. 3, No1, 1995, pp. 72-83

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19 – 41, 2000.

[13] Taufiq Hasan Student Member, IEEE and John H. L: Hansen A study on Universal Background Model training in Speaker Verification

[14] K.C. Sim and M.J.F. Gales, *Precision Matrix Modelling For Large Vocabulary Continuous Speech Recognition,* Cambridge University Engineering Department

[15] V. Delić, "Govorne baze na srpskom jeziku snimljene u okviru projekta AlfaNum," DOGS, 2000, str. 29-32.

[16] Cambridge University Engineering Department "The HTK Book", 2009.

[17] An Overview of Text-Independent Speaker Recognition: from Features to Supervectors, Tomi Kinnunen Haizhou Li, 2010.

Milan Dobrović, Telekom Srbija
Kontakt: milando@telekom.rs
Oblasti profesionalnog interesovanja: digitalna obrada slike, digitalna obrada govornog signala, prepoznavanje govornika, računarske mreže

Vlado Delić, Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Kontakt: vdelic@uns.ac.rs
Oblasti interesovanja: akustika, obrada audio signala, govorne tehnologije, komunikacija čovek-računar

Nikša Jakovljević, Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Kontakt: jakovnik@uns.ac.rs
Oblasti interesovanja: obrada signala, prepoznavanje govora, mašinsko učenje

Ivan Jokić, Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Kontakt: IBAHJOKIh@gmail.com
Oblasti interesovanja: obrada signala, prepoznavanje, govorne tehnologije,
automatsko prepoznavanje govornika