Zavisnost performansi sistema za prepoznavanje govora od izbora obeležja, širine i preklapanja prozora

Lazar Berbakov, Nikša Jakovljević

***Sadržaj* — U radu je predstavljena zavisnost performansi sistema za automatsko prepoznavanje govora prilagođenog za srpski jezik od izbora obeležja, širine i preklapanja prozora. Za obuku sistema korišćena je govorna baza snimljena i labelirana na Katedri za telekomunikacije i obradu signala Fakulteta Tehničkih Nauka u Novom Sadu. Obuka i testiranje sistema su izvršeni upotrebom HTK alata razvijenih od tsrane Cambridge University Engineering Department. Variranjem različitih parametara dobijen je ukupan broj od 720 različitih sistema za prepoznavanje govora, čije su performanse poređene i u ovom radu grafički prikazane.**

***Ključne reči* —Sistem za prepoznavanje govora, HTK alati, PLP, MFCC, LPC.**

I. UVOD

Popularnost sistema za automatsko prepoznavanje govora je uslovila pojava dovoljno snažnih računara koji mogu da u praktično realnom vremenu obave veliki broj operacija potrebnih prilikom samog procesa prepoznavanja.

Prednost sistema za automatsko prepoznavanje govora ASR (Automatic *Speech Recognition*) nad ostalim interfejsima čovek – mašina su:

- Govor je najprirodniji način komunikacije kod ljudskih bića.
- Širok propusni opseg (prenosi se veća količina informacija od npr. unosa tastaturom).
- Nepotrebna obuka za korisnike sistema.
- Oči i ruke ostaju slobodne za druge zadatke.
- Mobilnost (mikrofon ima manje dimenzije od npr. tastature)
- Telefonski kanal je optimizovan za prenos govora jer prenosi samo učestanosti u opsegu od 300 - 3400 Hz, što je dovoljno za razumljivost govora.

Problemi koji su prisutni prilikom eksploatacije sistema za ASR su:

- Nemogućnost upotrebe u situacijama kada se zahteva tišina (biblioteke, pozorišta,...).
- Loše performanse u bučnom okruženju, koje su posledica male tačnosti prepoznavanja.
- Zavisnost od govornika (različiti dijalekti, ...).
- Privatnost korisnika biva narušena.

Pošto je razvoj sistema za automatsko prepoznavanje govora jako zavisan od jezika potrebno je poznavanje različitih naučnih disciplina kao što su: obrada signala, fonetika, akustika, statistika i lingvistika. Sistemi za ASR polako ulaze u upotrebu u mnogim delovima ljudskog života, a njihova najznačajnija primena je u sledećim oblastima:

- Pomoć slepim i slabovidim osobama predstavlja jednu od značajnijih primena jer im omogućuje da koristeći savremene tehnologije lakše dolaze do novih informacija i saznanja.
- Izdavanje komandi glasom gde se po pravilu koriste sistemi zavisni od govornika, čime se postiže i nešto viša tačnost prepoznavanja. Ova mogućnost često se koristi za iniciranje poziva u mobilnoj telefoniji, upravljanje uređajima u domaćinstvu kao i aplikacijama edukativnog i zabavnog karaktera.
- Govorni automati koji predstavljaju najpopularniju komercijalnu primenu govornih tehnologija i u sebi objedinjuju dve najznačajnije govorne tehnologije - prepoznavanje i sintezu govora.

Sistemi za prepoznavanje govora koji su razvijeni za potrebe ovog rada pripadaju grupi sistema nezavisnih od govornika namenjenih prepoznavanju kontiualnog govora.

U radu je predstavljena izrada fonetskog sistema za prepoznavanje govora prilagođenog za srpski jezik, primenom HTK alata za sintezu Markovljevih modela. Prednost fonetskog sistema nad sistemima koji se baziraju na rečima je njegova prilagodljivost, odnosno mogućnost za vrlo lako proširenje rečnika mogućih reči. U radu su korišćene tri vrste vektora obeležja:

- LPC (Linear Prediction Coefficients)
- MFCC (Mel Frequency Cepstral Coefficients)
- PLP (Perceptual Linear Prediction)

kao i različite širine i procenti preklapanja prozora.

Razvoj sistema za automatsko prepoznavanje govora obuhvata dve grupe koraka:

- obuka sistema
- testiranje sistema

II. TEORIJSKA OSNOVA

A. Govorna baza

Govorna baza korišćena prilikom obuke i testiranja sistema je izrađena na *Katedri za telekomunikacije i*

Lazar Berbakov Autor, Fakultet Tehničkih Nauka, Trg Dositeja Obradovića 6 Novi Sad, Srbija; (e-mail: lazar_yu@yahoo.com).

Nikša Jakovljević Autor, Fakultet Tehničkih Nauka, Trg Dositeja Obradovića 6 Novi Sad, Srbija; (e-mail: jakovnik@uns.ns.ac.yu).

obradu signala Fakulteta Tehničkih Nauka u Novom Sadu, a prilikom njenog labeliranja vodilo se računa o sledećem:

- Na plozivima kao što su *B*, *G*, *D*, *P*, *K* i *T*, se zbog njihove akustičke prirode, mogu uočiti dva potpuno različita dela sa stanovišta statistike signala. Oni su modelovani preko okluzije i ekplozije (npr. Bo i Be), te su tako i dobijena dva različita modela koji zajedno modeluju jedan ploziv.
- Obeležen je i glas *Ђ*, odnosno „*muklo A*“, koje se u savremenom srpskom jeziku ne zapisuje, ali se javlja u kontekstu sa slovom *R* ako mu sledi ili prethodi samoglasnik, i sasvim jasno se uočava na prikazu govornog signala.
- *sil* označava tišinu u govornom signalu.
- *int* su sve smetnje koje postoje u signalu kao što su tonsko i pulsno biranje, pucketanje, itd.
- *unk* označava glasove koje je generisao govornik, a nismo u stanju da ih klasifikujemo

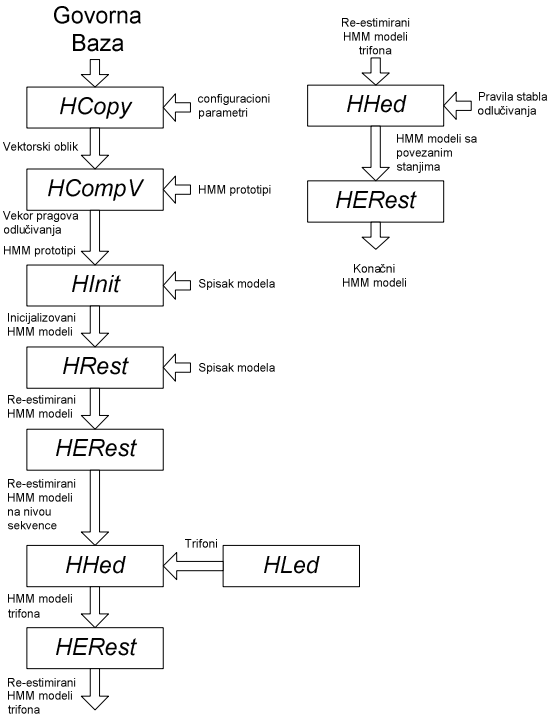
Baza za obuku se sastoji od 14496 govornih zapisa snimljenih od strane 340 govornika preko javne telefonske mreže, učestanošću odmeravanja od 8kHz, 8-bitnom A-law kvantizacijom. Baza za testiranje je snimljena u istim uslovima sa oko 40 govornika i disjunktna je u odnosu na bazu za obuku. Gramatika sadrži 149 reči, a omogućeni su prelazi iz svake reči u svaku reč.

B. *Obuka sistema*

Za obuku ASR sistema korišćeni su sledeći HTK alati:

- HCopy,
- HCompV,
- HInit,
- HRest,
- HERest,
- HLed,
- HHed,

a na Sl. 1. je algoritamski prikazana njihova upotreba:



Sl. 1 Obuka sistema za automatsko prepoznavanje govora

HCopy

Da bi se započela obuka sistema potrebno je iz govora izdvojiti samo one informacije koje su relevantne za prepoznavanje govora i to je upravo učinjeno korišćenjem alata *HCopy*. Ulazne podatke predstavljaju govorna baza za obuku u talasnom (*waveform*) obliku, i configuracioni fajl u kome su dati parametri na osnovu kojih se vrši konverzija govora u parametarski oblik, dok se kao izlaz dobija baza za obuku u parametarskom obliku. Konverzija iz talasnog u vektorski oblik je računarski veoma zahtevna, pogotovo zbog veličine baze za obuku, te se stoga ona vrši samo jednom, na početku obuke sistema.

HCompV

Za inicijalizaciju HMM modela globalnom srednjom vrednošću i varijansom govora koristi se alat *HCompV*. Podatak o globalnoj varijansi govora je veoma bitan u narednim koracima, i kao jedan od izlaznih fajlova ovog alata se javlja vektor podataka koji sadrži 1% (ili neki drugi izabrani procenat) prosečne vrednosti varijanse svih zvučnih fajlova. Ti podaci se u narednim koracima koriste kao prag, čime se sprečava da pojedini modeli usled malog broja vektora opservacija imaju male vrednosti varijansi.

HInit

Kada je pripremljen skup različitih prototipa HMM modela, može se pristupiti procesu inicijalizacije korišćenjem alata *HInit*. Osnovni princip *HInit* alata bazira se na konceptu HMM-a kao generatora vektora obeležja. Svaki uzorak za obuku sistema može se posmatrati kao izlazna sekvenca HMM čiji parametri treba da se izračunaju. Prema tome ako je poznato koja stanja su generisala svaki vektor obeležja u govornoj bazi za obuku, tada se nepoznate srednje vrednosti i varijanse posmatranog stanja HMM-a mogu izračunati na osnovu svih vektora obeležja povezanih sa datim stanjem.

Prvo se izvrši uniformna segmentacija govornih vektora obeležja i sukcesivni segmenti se pridružuju sukcesivnim stanjima modela, i izvrši se inicijalizacija parametara. Viterbijevim algoritmom se pronađe najverovatnija sekvenca stanja koja odgovara određenom govornom fajlu za obuku, a zatim se vrši izračunavanje parametara modela. Bočni efekat pronalaženja najverovatnije sekvence stanja predstavlja računanje log verodostojnosti. Ceo proces računanja parametara se ponavlja sve dok god se javlja povećanje log verodostojnosti.

HRest

Ovo je alat veoma sličan alatu *HInit*, s tim što on zahteva da su ulazne definicije HMM-ova već inicijalizovane a umesto Viterbijevog algoritma koristi tzv. *Baum-Welch* re-estimacioni postupak. On uključuje pronalaženje verovatnoće da se model nalazi u određenom stanju tokom trajanja nekog vremenskog intervala primenom *Forward-Backward* algoritma. Prema tome, dok Viterbijev algoritam donosi grubu odluku od strane kog stanja je generisan koji vektor obeležja. Baum-Welch algoritam omogućuje da jedan vektor bude pridružen nekolicini različitih stanja ali sa različitom verovatnoćom. Ovaj alat se koristi samo za obuku monofona (fonemi nezavisni od konteksta u kom se nalaze) jer kao i *HInit*

zateva da mu se kao ulazni podaci proslede prototip i ime modela.

HERest

Za razliku od prethodno opisanih alata HERest simultano ažurira sve HMM modele koristeći celu govornu bazu. Na početku HERest učitava sve HMM modele. Uz svaki govorni fajl mora postojati transkripcija tog govornog fajla na nivou fonema. HERest obrađuje svaki govorni fajl u jednom koraku. Nakon što ga učita u memoriju, koristi pridruženu transkripciju da kreira kompozitni model koji se dobija jednostavnim spajanjem pojedinih modela u skladu sa labelom u transkripciji. Nakon toga koristi se *Forward-Backward* algoritam kao i u alatu HRest, ali sada nad celim kompozitnim modelom.

HLed

Ovaj alat predstavlja jednostavan editor koji služi za manipulaciju label fajlovima. HLed nad label fajlovima vrši operacije koje su navedene u skript fajlu, a u radu je korišćen za generisanje liste svih trifona koji se javljaju u govornij bazi korišćenoj za obuku. Tome pribegavamo zbog toga što pojedini glasovi usled efekta koartikulacije imaju drugačije karakteristike u zavisnosti od konteksta u kom se nalaze, što je na dobar način modelovano upravo upotrebom modela trifona. Kao što se iz napred rečenog može videti, ovaj alat nije direktno vezan za računanje parametara HMM modela, već se koristi kao međualat u obuci ASR sistema.

HHed

HHed alat je alat namenjen manipulaciji nad strukturom HMM modela (povećanju broja mešavina u stanju, povezivanju stanja, prepovezivanju sa fonetskim modelima, itd.). Filozofija HTK alata je da se ASR sistem obučava u koracima. Počevši od monofona nezavisnih od konteksta u kom se nalaze, sistem se može poboljšavati kroz više faza. Svaka faza poboljšanja koristi HHed alat praćen sa re-estimacijom koristeći HERest. Prilikom razvoja ASR sistema predstavljenog u radu korišćen je u dva slučaja:

1. za kreiranje trifona od monofona
2. za spajanje parametara modela primenom stabla odlučivanja (Tree base clustering)

C. *Testiranje sistema*

Za testiranje i analizu performansi ASR sistema korišćeni su sledeći HTK alati:

- HParse,
- HVite,
- HResult,

HParse

Ovaj alat služi za kreiranje strukture (mreže), na osnovu gramatike u kojoj su specificirane sekvence reči koje su dozvoljene. Mreža na nivou reči reprezentuje gramatiku koja je napisana u formi Backus-Naur notacije gde se eksplicitno definišu dozvoljeni prelazi između pojedinih reči prilikom prepoznavanja.

HVite

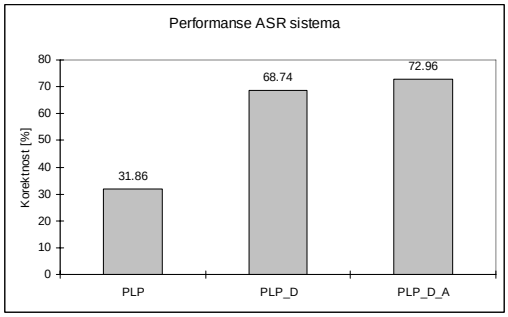
Nakon što je pripremljena mreža koja definiše reči koje se očekuju i kako se svaka od tih reči čita može se pristupiti prepoznavanju govora. Zadatak prepoznavača govora je da nađe najverodostojniju putanju kroz mrežu za dat nepoznat govorni fajl i HMM modele. Mreža za prepoznavanje se sastoji od skupa čvorova povezanih granama. Svaki čvor je ili HMM model ili kraj reči. Svaki HMM model za sebe predstavlja mrežu koja se sastoji od stanja povezanih granama, pa se samim tim mreža za prepoznavanje sastoji od stanja HMM modela povezanih granama prelaza.

HResult

Alat *HResult* poredeći transkripciju dobijenu prepoznavanjem sa originalnom transkripcijom generiše raznovrsne statistike vezane za performanse dobijenog ASR sistema.

III. REZULTATI

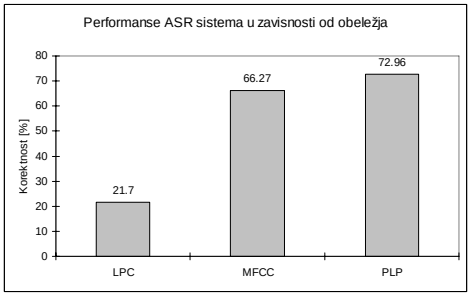
1. Ispitivanje uticaja dinamičkih koeficijenata na performanse ASR sistema. Prikazani su rezultati za ASR sisteme sa PLP obeležjima, 10 koeficijenata u vektoru obeležja, širina prozora 20ms i preklapanje između susednih prozora od 75%.



Sl. 2. Uticaj dinamičkih koeficijenata na performanse

Sa Sl. 2. se vidi da ukoliko se koriste samo statički koeficijenti (PLP) rezultujući ASR sistem ima izuzetno loše performanse, odnosno dobija se praktično neupotrebljiv sistem. Ukoliko se pored statičkih koeficijenata koristi i njihov prvi izvod (PLP_D) korektnost se praktično udvostručava. Razlog za ovakav skok korektnosti je vrlo verovatno posledica 2 faktora. Prvi faktor je da je informacija o tome šta je rečeno zapisana i u tranziciji formanata, i drugi da je na ovaj način smanjena korelisanost između susednih vektora obeležja što je bitna pretpostavka za funkcionisanje ASR sistema baziranog na HMM modelima. Dodavanje drugog izvoda (PLP_D_A) rezultuje poboljšanjem performansi ali ovaj put za daleko manji procenat.

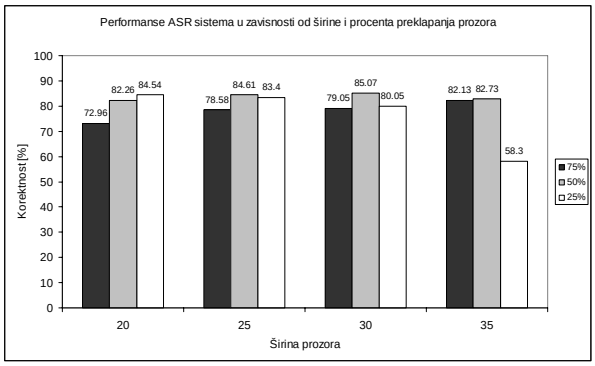
2. Zavisnost performansi ASR sistema od izbora obeležja (LPC MFCC PLP) Ovde su date vrednosti korektnosti za ASR sisteme sa 10 statičkih LPC, MFCC i PLP koeficijenata (statički parametri prvi i drugi izvod) širinom prozora 20 ms, i preklapanjem između susednih prozora od 75%.



Sl. 3. Uticaj izbora obeležja na performanse

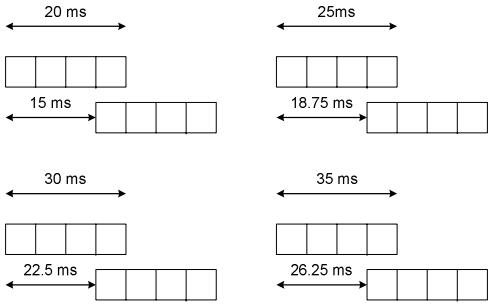
Sa Sl. 3. se vidi da se ASR sistem sa najlošijim performansama bazira na LPC obeležjima. ASR sistem koji koristi MFCC vektore obeležja očekivano daje bolje rezultate jer uzima u obzir percepciju promene učestanosti zvuka kod čoveka koja je opisana Mel skalom. Koristeći PLP vektore obeležja dobijaju se najbolji rezultati, jer oni pored Mel skale uzimaju u obzir način percepcije nivoa zvuka kod čoveka kao i efekat maskiranja. Najbolji LPC, MFCC i PLP sistemi imaju sledeće korektnosti prepoznavanja respektivno: 31.12 %, 84.13% i 86.01%.

3. Zavisnost performansi ASR sistema od širine i brzina pomeranja prozora. Kao ilustracija prikazane su performanse ASR sistema sa PLP obeležjima, statičkim parametrima, prvim i drugim izvodom, 10 statičkih koeficijenata i različitim širinama i procentima preklapanja prozora.



Sl. 4. Uticaj širine i preklapanja prozora na performanse

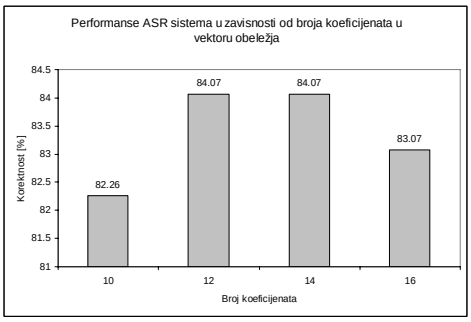
Na Sl. 4. je dat pregled performansi ASR sistema u zavisnosti od širine i procenta preklapanja prozora. Može se uočiti da pri svim dužinama prozora najbolje performanse ima onaj sistem koji koristi preklapanje između prozora od 50%. Pad performansi sistema sa povećanjem dužine prozora pri procentu preklapanja prozora od 25% je ilustrovano na Sl. 5.



Sl. 5. Ilustracija preklapanja prozora

Može se videti da se pri dužini prozora od 30 ms, prozoriranje vrši na svakih 26,25 ms, što je prilično dug vremenski interval. Ovim se gubi na vremenskoj rezoluciji pa samim tim opadaju i performanse ASR sistema.

4. Zavisnost performansi ASR sistema od broja koeficijenata. ASR sistemi sa LPC, MFCC i PLP obeležjima, statički parametri, prvi i drugi izvod, širina prozora 20ms, procenat preklapanja 50% i različit broj koeficijenata u vektoru obeležja.



Sl. 6. Uticaj broja koeficijenata na performanse

Na Sl. 6. vidimo da pri korišćenju PLP obeležja optimalne vrednosti broja koeficijenata predstavljaju 12 i 14. Pri korišćenju 16 koeficijenata broj kanala u filtar banci iznosi 34 čime se opisuju i vrlo fine promene u spektru kao što su pojedini harmonici, pa stoga opadaju i performanse sistema. Kod sistema baziranih na MFCC obeležjima trend opadanja performansi sistema sa porastom broja obeležja bi se primetio tek kod većeg broja obeležja. Ovo nije prikazano u radu jer je za najveći broj koeficijenata izabrano 16. Performanse sistema baziranih na LPC obeležjima pokazuju drugačiji trend u odnosu na prethodno navedene sisteme, jer se kod njih sa porastom broja obeležja smanjuju performanse.

5. Poredeći performanse sistema sa različitom organizacijom vektora obeležja po strimovima primećujemo da ne dolazi do poboljšanja istih ako koristimo veći broj strimova. Ovo je posledica topologije u kojoj se koristi samo jedna Gausova raspodela po stanju.

IV. ZAKLJUČAK

Kao što se i očekivalo najbolje performanse su postignute sa PLP obeležjima. Vrlo je bitno koristiti dinamička obeležja jer se vrši dekorelacija susednih frejmova, ali se opisuju i tranzicije formanata koje su i kod čoveka bitne za razumevanje. Ovo istraživanje je takođe pokazalo da je optimalna širina prozora između 25 i 35ms, kao i da rastojanje između susednih frejmova bude između 10 i 15ms.

LITERATURA

[1] Cambridge University Engineering Department "The HTK Book", December 2006.
[2] Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, New Jersey 1993.

ABSTRACT

The aim of this paper is to present ASR system performance dependances of used feature vector and windows sizes. For system training has been used speech corpus recorded and labeled at the Department of Communication and Signal Processing at the University of Novi Sad. The system has been realized using HTK tools, developed by Cambridge University Engineering Department.

ASR SYSTEM PERFORMANCES DEPENDANCE OF USED FEATURE VECTORS AND WINDOW SIZE

Lazar Berbakov, Nikša Jakovljević