JEDAN PRISTUP KA STATISTIČKOM PREPOZNAVANJU REČI IZ OGRANIČENOG REČNIKA

Miloš Blagojević, Željko Đurović, *Elektrotehnički fakultet u Beogradu*

Nagrađeni rad mladog istraživača – komisija AU

Sadržaj – *U radu je predstavljen algoritam za prepoznavanje pojedinačnih reči, iz ograničenog skupa reči. Opisani algoritam se zasniva na primeni metoda statističkog prepoznavanju oblika. Ukratko su predstavljene teorijske osnove na kojima su zasnovane pojedine faze algoritma, počevši od osnovne obrade signala i detekcije početka i kraja reči, pa do projektovanja linearnih klasifikatora i redukcije dimenzija.*

1. UVOD

U ovom radu je prikazan jedan način na koji se može pristupiti rešavanju problema prepoznavanja govora. Predloženi algoritam za prepoznavanje govora se zasniva na rezultatima teorije statističkog prepoznavanja oblika.

Na početku rada je dat opis formirane baze govornih signala. Baza se sastoji od reči iz 10 različitih klasa.

U glavi 3. opisana je metoda detekcije granica reči unutar snimljnog signala koji pored reči sadrži i termički šum. Signali su predstavljeni preko svoje *Teager* energije. Ovakva predstava signala omogućila je dobru detekciju kraja reči i u slučajevima kada kraj reči čini visokofrekventni signal male snage (suglasnici m,n,t).

Potom je opisan metod izbora vektora karakterističnih obeležja snimljenih signla. Vektor koji je odabran za predstavu signala sastoji od 48 elemenata..

Glave 5. i 6. su posvećene opisu korišćenih metoda statističkog prepoznavanja oblika. Dat je kretak pregled najvažnijih formula i njihove primene u klasifikaciji signala. Redukcijom sa 48 na dve dimenzije vektora karakterističnih obeležja dobija se predstava signala u dvodimenzionalnom prostoru. Pokazalo se da su tako dobijene klase separabilne i da je moguća upotreba jednostavnih linearnih klasifikatora.

Na kraju rada je opisan postupak klasifikacije i dati su rezultati testiranja sistema.

2. FORMIRANJE BAZE

Problem prepoznavanja govora posmatran je na ograničenom skupu reči. Za elemente rečnika su odabrane cifre (nula,...,devet), što znači da postoji 10 različitih klasa. Formirana je baza sa glasovima različitih govornika.. U snimanju baze učestvovalo je 10 govornika (6 muških i 4 ženska), pri čemu je svaki govornik više puta ponavljao istu reč. Signal je sniman sa frekvencijom odabiranja 8kHz, u trajanju od 2s. Kako se radi o prepoznavanju pojedinačnih reči, trajanje snimanja od 2s je bilo savim dovoljno da omogući kvalitetno formiranje baze.

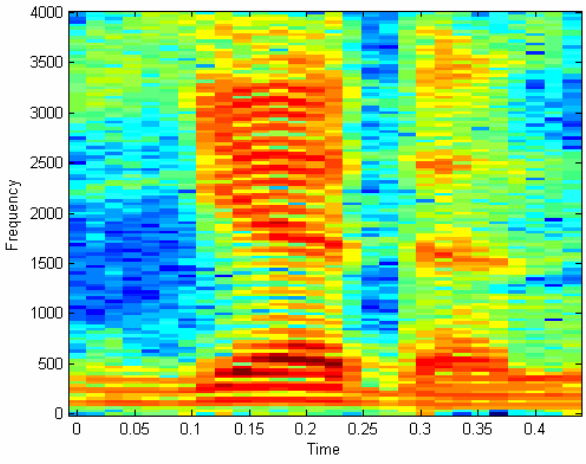
Prilikom snimanja signala primećeno je da u zavisnosti od opreme koja se koristi za snimanje (pre svega od mikrofona i podešavanja računara), zavisi kvalitet snimka i nivo šuma u signalu. Stoga su svi elementi baze snimani sa dva slična mikrofona na jednom računaru sa istim podešavanjima.

Kasnijom analizom utvrđeno je da problem za prepoznavanje predstavljaju glasovi sa takozvanim *clipp*-ovanjem. *Clipp*-ovanje se javlja u slučajevima kada pojedini odbirci signala koji se snima, prelaze maksimalnu vrednost koju oprema za snimanje dozvoljava, pa se za vrednost ovih odbiraka dodeljuje maksimalna definisana vrednost. Ovim se menja priroda izvornog signala, pa su svi svi signali koji su imali više *clipp*-ovanih odbiraka izbačeni iz baze.

3. DETEKCIJA POČETKA I KRAJA REČI

Kako je već rečeno svi snimljeni signali su imali isti trajanje (2s). Sledeći korak u našem radu predstavlja izdvajanje dela signala koji nosi informaciju o reči, od termičkog šuma iz te 2s. Da bi se to izvelo potrebno je pronaći algoritam za preciznu detekciju početka i kraja reči.

Početna ideja je bila da se pokuša segmentacija signala na osnovu njegove energije Međutim ovaj pristup nije dao dovoljno dobre rezultate, pri čemu je najveći problem predstavljala detekcija kraja reči koje sadrže slovo T (pet, šest, devet). Isprobani neki drugi pristupi rešavanju ovog problema, a kao najbolje rešenje se pokazalo korišćenje *Teager* energije za predstavljanje signala.



Slika 1. *Spektrogram reči 'jedan'*

Osnovna ideja *Teager* algoritma je da se istaknu komponente signala na višim frekvencijama. To se postiže množenjem odgovarajućih spektralnih komponenti kvadratom frekvencije. Prvo se računa spektrogram celog signala. Za računanje spektrograma je korišćena MATLAB funkcija *specgram*. Signal se prvo podeli na preklapajuće segmente. Odabrano je da širina svakog segmenta bude 256 odbiraka, a preklapanje među susednim segmentima 128. Potom je za svaki segment izračunata *Fourier*-ove transformacije u NFFT tačaka. Postavlja se pitanje izbora pravih vrednosi za navedene parametare (širina prozora, veličina preklapanja i NFFT). Odabrane su one vrednosti za koje su dobijeni zadovoljavajući rezultati, ali neka detaljnija istraživanja nisu izvršena.

Nakon formiranja spektrograma, vrednosti dobijenih spektralnih komponenti su pomnožene sa kvadratom učestanosti. Na taj način dobijena je *Teager* predstava signala, a zatim se *Teager* energija određenog segmenta dobija kao koren sume *Teager* spektralnih komponenti koje pripadaju tom segmentu.

Posmatrajmo jedan segment signala čija je veličina jednaka definisanoj veličini prozora. Označimo sa $X(\omega_k)$ spektralnu komponentu unutar posmatranog segmenta koja odgovara učestanosti ω_k. Odgovarajuću *Teager* spektralnu komponentu dobijamo kao:

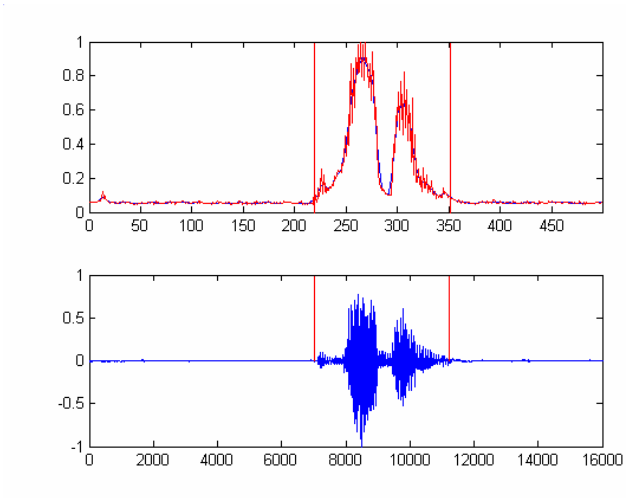
$$f_k = \omega_k^2 X(\omega_k) \tag{1}$$

Odavde *Teager* energiju koja odgovara *i*-tom segmenta signala (T_i) računamo kao:

$$T_i = \left(\sum_{k=1}^{K} f_k \right)^2, \tag{2}$$

gde je K broj tačaka u kojima se računaju spektralne komponente.

Dobijena predstava signala je prilično dinamična kriva sa velikim skokovima, pa je potrebno dodatno je isfiltrirati da bi se smanjila verovatnoća pogrešne detekcije granica reči. To je urađeno usrednjavanjem krive na prozoru dužine N (u ovom slučaju odabrano je N=3).



Slika 2. *Teager energija i talasni oblik reči 'jedan'*

Granice reči su određivane poređenjem nivoa *Teager* energije snimljenog signala sa odabranim pragovima.. Vrednost pragova detekcije je određena na osnovu procenjenog nivoa šuma i jednaka je *Teager* energiji šuma uvećanoj za 2.5% maksimalne vrednosti energije. Vrednost praga je utvrđena eksperimentalno i predstavlja neku vrstu kompromisa. Smanjenjem vrednosti praga svakako bi se smanjio procenat nedetektovanih delova problematičnih reči, ali na drugoj strani sistem bi bio manje robustan i jako osetljiv na smetnje koje se mogu pojaviti na početnom delu snimljenog signala.

Procenjeni nivo šuma je karakterisan *Teager* energijom signala na početku i kraju snimljene sekvence. Podrazumevano je da se korisni signal nalazi u centralnom delu snimljenog niza, odnosno da se na krajevima niza nalazi samo termički šum.

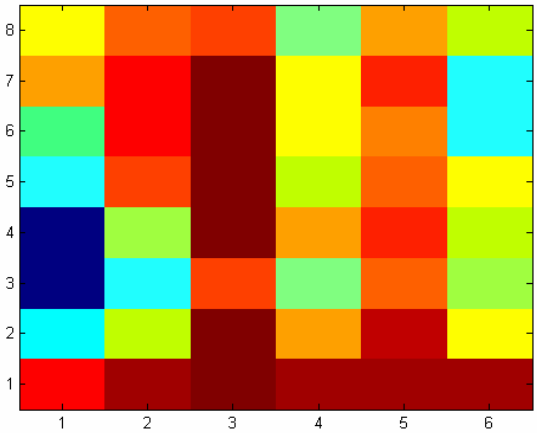
Na ovaj način dobijena je bazu skraćenih signala i nastavljeno je sa daljom obradom signala iz te baze.

4. IZBOR KARAKTERISTIČNIH OBELEŽJA

Vektor karakterističnih obeležja signala je formiran na osnovu spektrograma signala. Spektrogram je dobijen na već opisan način, a posmatrana je njegova decibelska razmera. Zatim je formirana mreža kojom je spektrogram podeljen na *N*x*M* pravougaonih delova. Izbor vremenske i frekvencijske rezolucije mreže, odnosno širine pravougaonih segmenata po vremenskoj i frekvencijskoj osi, bitno utiče na tačnost sistema za prepoznavanje. Suviše velika vremenska rezolucija utiče negativno na tačnost sistema za prepoznavanje. Najpogodnije vrednosti rezolucija su određene eksperimentalno. Za rezoluciju po vremenu je odabrano N=6, a za rezoluciju po frekvenciji M=8. Dobri rezultati se dobijaju i za grupu vrednosti bliskih odabranim.

Za svaki od pravougaonih segmenata izračunata je srednja vrednost logaritama spektralnih komponenti unutar njega. Na ovaj način dobijen je 48-dimenzionalni vektor koji predstavlja vektor karakterističnih obeležja za dati signal.

Ovim postupkom svi signali iz baze skraćenih signala su predstavljeni odgovarajućim vektorom karakterističnih obeležja.



Slika 3. *Predstava signala sa 48 elemenata*

5. REDUKCIJA DIMENZIJA

Sledeći korak predstavlja redukcija 48-dimenzionalnog vektora karakterističnih obeležja na manji broj dimenzija kako bi se smanjila matematička kompleksnost dalje procedure i omogućila lakša klasifikacija. U tu svrhu treba naći odgovarajuću transformacionu matricu A kojom se slučajni n-dimenzionalni vektor X preslikava (redukuje) u m-dimenzionalni vektor Y, pri čemu je:

$$Y = A^T X \tag{3}$$

Redukcijom dimenzija se gubi deo informacije sadržane u originalnom vektoru, ali ideja je da se pogodnim izborom elemanata transformacione matrice gubitak informacija o razlikama između klasa učini što manjim. Redukcija dimenzija je vršena pomoću *Karhunen-Loeve* metode, a korišćen je kriterijum na bazi mere rasipanja. U našem slučaju vršena je redukcija na svega dve dimenzije, što se pokazalo kao dovoljno.

Korišćeni kriterijum redukcije dimenzija je za cilj imao da se od 48 elemenata vektora odaberu samo 2, vodeći računa da o tome da posmatrane klase budu što je moguće više razdvojene (u novoformiranom prostoru). Na ovaj način su sve snimljene reči predstavljene u dvodimenzionalnom

prostoru, čime je omogućeno vizuelno razlikovanje različitih klasa, a takođe i primena jednostavnih linearnih klasifikatora.

Pretpostavimo da postoji L različitih klasa, označimo ih sa ω_i, $i = 1,...,L$. Svakoj od tih klasa pridružimo vektor matematičkog očekivanja M_i i kovarijacionu matricu Σ_i; $i = 1,...,L$.

Ako sa P_i označimo apriornu verovatnoću pojave i-te klase, tada za združeni vektor matematičkog očekivanja M_0 možemo pisati:

$$M_0 = E\{X\} = \sum_{i=1}^{L} P_i M_i \tag{4}$$

Definišimo matricu unutar-klasnog rasejanja (*within class scatter matrix*) S_W kao:

$$S_W = \sum_{i=1}^{L} P_i E\left\{(X - M_i)(X - M_i)^T / \omega_i\right\} = \sum_{i=1}^{L} P_i \Sigma_i \tag{5}$$

a matricu međuklasnog rasejanja (*between class scatter matrix*) S_B kao:

$$S_B = \sum_{i=1}^{L} P_i (M_i - M_0)(M_i - M_0)^T \tag{6}$$

Ostalo je još da se pronađe transformaciona matrica A, takva da slučajni vektor $Y = A^T X$ minimizira odabrani kriterijum:

$$J_1 = tr(S_W^{-1} S_B). \tag{7}$$

Da bi ovaj kriterijum imao minimalnu vrednost potrebno je da matrica S_W ima što veću vrednost, a matrica S_B što manju, odnosno da rastojanje između klasa bude što veće, a rastojanje elemenata unutar klase što manje. Ovo je u saglasnosti sa postavljenim ciljem da klase međusobno budu što više razmaknute i separabilne, a elementi unutar pojedinačnih klasa grupisani.

Rešavanjem ovog problema se dobija da je:

$$A = \left[\Psi_1 \Psi_2 ... \Psi_m\right], \tag{8}$$

gde su Ψ_i sopstveni vektori matrice ($S_W^{-1} S_B$) kojima odgovaraju najveće sopstvene vrednosti iste matrice ($S_W^{-1} S_B$).

6. KLASIFIKACIJA

Ideja ovog algoritma je da se prepoznavanje reči vrši samo pomoću linearnih klasifikatora koji razdvajaju dve klase. Ukupan broj klasa je jednak broju različitih reči, odnosno 10. Za svake dve klase projektovan je po jedan linearni klasifikator, pa je ukupan broj potrebnih klasifikatora 45.

Projektovanje linearnog klasifikatora ima smisla u slučaju kada su klase separabilne i kada ta separabilnost potiče od rastojanja u vektorima srednjih vrednosti. Da bi to bilo ostvareno potrebno je za svaki par klasa pronaći transformacione matrice i izvršiti redukciju dimenzija, na način opisan u prethodnoj glavi. Pokazalo se da su za bazu u kojoj su reči dobro skraćene, klase dovoljno separabilne i da je linearni klasifikator dobro rešenje.

Sledeći korak je projektovanje linearnih klasifikatora za svaki par klasa, u obliku:

$$\begin{aligned} h(X) = V^T X + v_0 < 0 \Rightarrow X \in \omega_1 \\ h(X) = V^T X + v_0 > 0 \Rightarrow X \in \omega_2 \end{aligned}, \tag{9}$$

gde su ω_1 i ω_2 klase čije se razdvajanje vrši, a $h(X)$ je linerana diskriminaciona funkcija.

Definišimo sledeće parametre:

$$\begin{aligned} \eta_i &= E\{h(X)/\omega_i\} = E\left\{V^T X + v_0 / \omega_i\right\} \quad i = 1,2 \\ \eta_i &= V^T M_i + v_0 \end{aligned} \tag{10}$$

$$\begin{aligned} \sigma_i^2 &= \mathrm{var}\{h(X)/\omega_i\} = \mathrm{var}\left\{V^T X + v_0 / \omega_i\right\} \quad i = 1,2 \\ \sigma_i^2 &= V^T \Sigma_i V \end{aligned} \tag{11}$$

Projektovanje linearnog klasifikatora podrazumeva pronalaženje parametara V i v_0 tako da se minimizira neki unapred određeni kriterijum, $f\left(\eta_1, \eta_2, \sigma_1^2, \sigma_{21}^2\right)$.

Rešenje za V se može napisati u obliku:

$$V = \left[s\Sigma_1 + (1-s)\Sigma_2\right]^{-1} \left(M_2 - M_1\right), \tag{12}$$

gde je:

$$s = \frac{\dfrac{\partial f}{\partial \sigma_1^2}}{\dfrac{\partial f}{\partial \sigma_1^2} + \dfrac{\partial f}{\partial \sigma_2^2}}. \tag{13}$$

Optimalna vrednost za v_0 se dobija iz:

$$\frac{\partial f}{\partial h_1} + \frac{\partial f}{\partial h_2} = 0. \tag{14}$$

Za kriterijum koji se minimizira odabran je:

$$f = \frac{P_1 \eta_1^2 + P_2 \eta_2^2}{P_1 \sigma_1^2 + P_2 \sigma_2^2} \tag{15}$$

Odavde se dobija:

$$V = \left[P_1 \Sigma_1 + P_2 \Sigma_2\right]^{-1} (M_2 - M_1) \tag{16}$$

$$v_0 = -V^T \left(P_1 M_1 + P_2 M_2\right) \tag{17}$$

7. DOBIJENI REZULTATI

Kada se jednom formiraju sve potrebne transformacione matrice i klasifikatori sistem je spreman za proces prepoznavanja reči i više se ne vraćamo na te korake. Razdvajanje procesa "učenja" od procesa prepoznavanja omogućava da se sam proces prepoznavanja odvija dosta brže.

Ulaz u sistem predstavlja signal govora snimljen sa frekvencijom odabiranja 8kHz i trajanjem 2s. Zatim taj signal prolazi kroz obradu sličnu onoj kroz koju su prošli svi signali originalne baze. Prvo se vrši segmentacija signala na način

opisan u glavi 2. a potom i formiranje vektora karakterističnih obeležja o čemu je bilo reči u glavi 3. Nakon toga se izvršava proces klasifikacije signala.

Ideja je da se za svake dve klase utvrdi kojoj klasi je verovatnije (po odabranom kriterijumu) da izgovorena reč pripada U tu svrhu se koriste unapred pripremljni parameti lineranih klasifikatora (V i v_0) i transformacione matrice (A**).** Broj testova jednak je broju formiranih linearnih klasifikatora i svaka klasa se javlja 9 puta u testovima.

Princip odlučivanja je zasnovan na majoritetnoj logici. Izgovorenu reč, čije se prepoznavanje vrši, svrstavamo u onu klasu koja je nakon svih testova najviše puta bila izabrana (poželjno svih 9 puta).

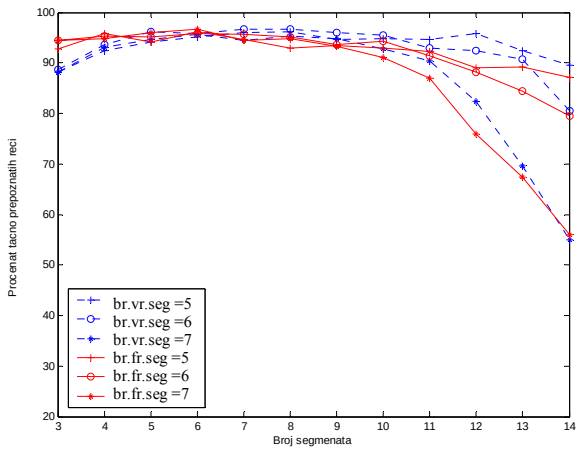
Kao što je i očekivano tačnost rada opisanog algoritama je veća za govornike čiji se glasovi nalaze u bazi, nego za one čiji glasovi nisu u bazi. Problem predstavljaju oni govornici čiji glas nije sličan ni sa jednim glasom iz baze. Ovo se može popraviti povećavanjem baze i ponovnim obučavanjem sistema.

Tačnost rada sistema opada i u slučaju da se testiranja vrše na opremi čije se karakteristike znatno razlikuju od karakteristika opreme na kojoj je baza snimana.

Proces testiranja sistema je izveden na signalima iz baze. Za testiranje je korišćena metoda leave-one-out. Testiranje je sprovedeno tako što se iz baze skraćenih reči ukloni jedna (probna) reč i nakon toga obuči sistem na preostalom delu baze. Nakon toga se probna reč dovodi na ulaz sistema i vrši se njena klasifikacija a rezultat se beleži. Postupak ponavlja, a iz baze skraćenih reči se svaki put uklanja druga reč.

Na osnovu opisanog testiranja, za sistem formiran sa parametrima korišćenim u ovom radu, dobijena je tačnost prepoznavanja od 97%.

Na slici 4. su prikazani rezultati testiranja sistema obučavanog sa različitim rezolucijama po vremenu i frekvenciji. Punom linijom prikazani su rezultati dobijeni u slučajevima kada je broj segmenata po frekvenciji konstantan, a menja se broj segmenta po vremenu, dok su isprekidanom linijom prikazani rezultati koji odgovaraju obrnutoj situaciji (broj segmenata po vremenu je konstantan, a menja se broj segmenta po frekvenciji).



Slika 4. *Rezultati testiranja*

8. ZAKLJUČAK

Sistem opisan u ovom radu demonstrira mogućnosti primene rezultata iz oblasti statističkog prepoznavanja oblika, za prepoznavanje reči (signala govora).

Karakteristika predstavljenog sistema je što se sve operacije neophodne za proces obučavanja sistema izvršavaju samo jednom i nezavisno od procesa prepoznavanja reči. Ovo ima i dobru i lošu stranu. Dobro je što se na ovaj način proces prepoznavanja reči odvija brzo sa relativno malom računskom kompleksnošću. Problem može predstavljati činjenica da sistem nije lako dodatno obučiti, pa ako želimo da dodamo samo jedan novi uzorak signala u bazu moramo ponovo sprovesti celu proceduru obučavanja sistema i iznova izračunati sve transformacione matrice i parametre linearnih klasifikatora .

Nedostatak opisanog procesa odlučivanja je to što se uvek donosi odluka da signal pripada nekoj od klasa , pa čak i kada je izgovorena neka reč koja nije slična ni sa jednom od definisanih klasa. Ovo je posledica toga što je broj formiranih klasa jednak broju reči za čije se prepoznavanje sistem obučava Stoga se ovaj nedostatak može relativno lako otkloniti formiranjem dodatne klase koja bi obuhvatila sve slučajeve u kojima je izgovorena neka druga reč i gde je odstupanje od postojećih reči isuviše veliko.

Ovaj algoritam predstavlja samo jednu ideju za pristup rešavanju problema prepoznavanja glasa i ima puno prostora za njegovo dalje unapređenje i modifikaciju.

LITERATURA.

[1] Lingyun Gu, Stephen Zahorian, "A new robust algorithm for isolated word endpoint detection",

[2] Li Deng, Douglas O'Shaughnessy, "Speech Processing", Marcel Dekker 2003

[3] K. Fukunaga, "Statistical Pattern Recognition", Academic Press, 1990

[4] Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments"

[5] L.Lamel, L.Rabiner, A.Rosenberg, J.Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", *IEEE Trans.Acoust., Speech, Signal Processing, vol. ASSP-29, No.4, August 1981*

Abstract: The paper presents the algorithm for recognition of spoken words belonging to limited dictionary. The algorithm is based on statistical pattern recognition. The paper presents the short overview of theoretical background of different phases of the algorithm including basic signals filtering, startpoint and endpoint detection, dimension reduction and linear classifiers design.

ONE APPROACH FOR STATISTICAL RECOGNITION OF SPOKEN WORDS BELONGING TO LIMITED DICTIONARY

M.Blagojević, Ž. Đurović