Ratko Amanović i Nemanja Miković

Primena skrivenog Markovljevog modela u prepoznavanju govora na ograničenom rečniku

Analizirano je prepoznavanje govora nezavisno od govornika na ograničenom rečniku korišćenjem skrivenih Markovljevih modela (SMM). Korišćena karakteristična obeležja govornog signala su kepstralni koeficijenti Melove skale, po uzoru na rad Dejvisa i Mermelštajna (Davis S., Mermelstein P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28 (4): 357-66). Za potrebe istraživanja formirana je baza od 30 reči srpskog jezika, podeljenih u grupe od po 1, 2, 3 i 4 sloga, koje je izgovaralo 48 govornika. Za obeležavanje baze, računanje karakterističnih obeležja, treniranje i testiranje SMM-a korišćen je alat Hidden Markov model toolkit. Na snimljenoj bazi ostvarena je pozitivna prediktivna vrednost 95%, ukoliko je broj skrivenih stanja veći od 15.

Uvod

Prepoznavanje govora je proces koji omogućava mašini da identifikuje ljudski govor. Kao i kod svih govornih tehnologija, reč je o multidisciplinarnom problemu, za čije su rešavanje potrebna znanja iz raznih oblasti, počev od akustike, fonetike i lingvistike, pa do matematike, telekomunikacija, obrade signala i programiranja (Ghahramani 2001). Zadatak mašine jeste da se na osnovu ulaznih podataka u vidu zvuka neke reči ta reč prepozna. Prepoznavanje govora ima primenu u davanju komandi mašinama glasom, upotreba telefona bez ruku, zapisivanju teksta bez kucanja ili pisanja, automatskom prevođenju.

U radovima koji se bave problemima prepoznavanja govora najčešće se koriste algoritmi bazirani na skrivenim Markovljevim modelima – SMM (Rabiner 1989), dinamičkom savijanju vremena – DSV (Juang 1984) i neuronskim mrežama (Lippmann 1989).

U ovom radu je za prepoznavanje govora korišćen metod baziran na skrivenim Markovljevim modelima. Ideja rada jeste analiza sistema baziranog na SMM koji nezavisno od govornika prepoznaje izgovorenu reč iz unapred zadatog rečnika. Karakteristična obeležja govornog signala koja se koriste za treniranje SMM su bazirana na kepstralnim koeficijentima Melove skale (MFFC).

Prepoznavanje govora na srpskom jeziku je posebno zanimljivo za proučavanje zato što svako slovo označava jedan glas. Otuda naša pretpostavka da će broj skrivenih stanja SMM-a biti srazmeran broju slova i/ili slogova u reči. Kod prepoznavanja govora pomoću SMM-a nije poznato šta predstavljaju skrivena stanja.

Za potrebe istraživanja kreirana je baza od 30 različitih reči srpskog jezika, koje su razvrstane u grupe od po 1, 2, 3 i 4 sloga (tabela 1). Svaka grupa ima po šest reči, osim četvrte koja se sastoji od dvanaest reči. Unutar četvrte grupe postoje manje podgrupe od po tri reči koje su slične. Razlog za ovako formiranu četvrtu grupu je mogućnost poređenja rezultata sa referentnim radom (Davis i Mermelstein 1980), koji se bavi analizom problema prepoznavanja sličnih reči.

Ratko Amanović (1997), Smederevska Palanka, Đure Đakovića 15, učenik 3. razreda Palanačke gimnazije

Nemanja Miković (1997), Bečej, Sonje Marinković 53, učenik 3. razreda Tehničke škole

MENTORI:

Marko Bežulj, ISP / Microsoft development center Serbia

Miloš Stojanović, student Elektrotehničkog fakulteta Univerziteta u Beogradu

Natalija Todorčević, student Elektrotehničkog fakulteta Univerziteta u Beogradu

Prilikom sastavljanja baze učestvovalo je 48 različitih osoba, 24 muških i 24 ženskih, koji su po dva puta izgovorili svaku od 30 reči iz baze. Baza je podeljena na trening (40%), test (40%) i validaciju (20%).

Tabela 1. Spisak korišćenih reči

1 slog	2 sloga	3 sloga	4 sloga
hlad	šunka	kajgana	rastaviti
cvet	jaje	paprika	sastaviti
muž	tata	učenik	nastaviti
mit	peći	raketa	gramatika
nož	bašta	saditi	fanatika
konj	prozor	bandera	dramatika
			sušilica
			brusilica
			bušilica
			čeprkati
			pobrkati
			posrkati

Zarad potpunih rezultata idealno bi bilo ako bi istraživanje bilo vršeno na celom rečniku srpskog jezika. Međutim, zbog vremena potrebnog za snimanje i obeležavanje baze i zbog vremena potrebnog za procesiranje, to nije bilo moguće.

Za obeležavanje baze, računanje karakterističnih obeležja, treniranje i testiranje SMM-a korišćen je Hidden Markov model toolkit (HTK) (HTK 2002).

Na slici 1 prikazan je analizirani sistem za prepoznavanje govora. U ovom sistemu za svaku reč iz rečnika treniran je zaseban SMM. Klasa nepoznate reči se izračunava tako što se izračunata MFCC karakteristična obeležja nepoznate reči proslede na SMM za svaku reč iz baze, a rešavanjem problema evaluacije dolazimo do informacije kojoj klasi ta reč pripada.

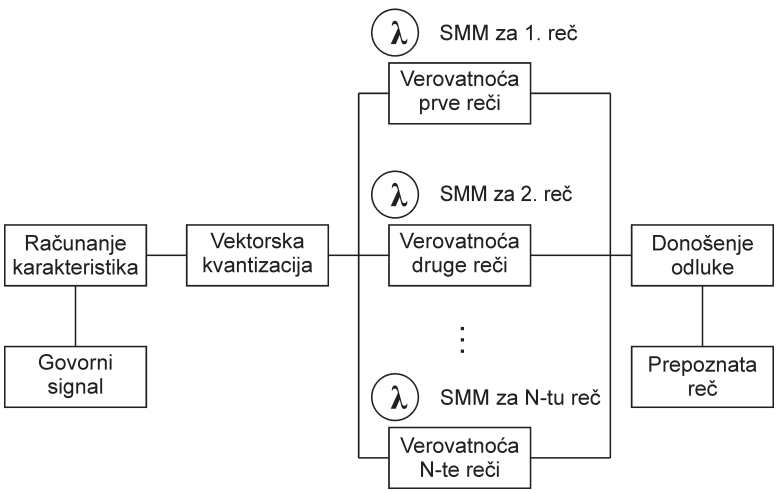
Kepstralni koeficijenti Melove skale

Prvi korak u prepoznavanju govora jeste određivanje karakteristika audio-signalâ koji nose informacije koje su potrebne za prepoznavanje određene reči. Jedna od najrasprostranjenijih karakteristika za prepoznavanje i obradu govora jesu kepstralni koeficijenti Melove skale (engl. Mel Frequency Cepstral Coefficients, MFCC). MFCC karakteristike su korišćene zato što su u referentnom radu dale dobre rezultate (Davis i Mermelstein 1980). Za računanje MFCC karakteristika korišćen je alat HCopy iz paketa HTK.

MFCC karakteristična obeležja računata su na sledeći način (Practicalcryptography.com; Young *et al*. 2009):

1. Primena prozorske funkcije: signal je podeljen na prozore širine 25 ms Hamming metodom, pri čemu svaki naredni prozor počinje 10 ms nakon početka prethodnog (Gales i Young 2007; Practicalcryptography.com)

2. Za svaki prozor izračunata je brza Furijeova transformacija (FFT) u 1024 tačke, pri



Slika 1. Blok šema analiziranog sistema za prepoznavanje govora

Figure 1. Block diagram of analyzed speech recognition system

čemu su se dobijene vrednosti kvadrirale kako bi se od amplitudskog spektrograma napona dobio spektrogram snage, jer je

$$P = \frac{U^2}{R}$$

(*P* je snaga, *U* je napon, a *R* je otpor).

3. Računanje filter banke:

3.1. Izabrane su dve tačke za najnižu i najvišu frekvenciju 0 i 8000 Hz (opseg ljudskog glasa). Frekvencije su preračunate u Melovu skalu pomoću formule

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right)$$

Jedan od razloga za dobre rezultate je to što je Melova skala logaritamska, što je čini približnijom ljudskom sluhu (Practicalcryptography.com).

3.2. Između ove dve izabrane tačke linearno je raspoređeno onoliko tačaka koliko ima i filtera, da bi svaki filter, osim prvog i poslednjeg, imao tri tačke koje bi obuhvatio. Zatim su dobijene vrednosti preračunate u herce formulom

$$M^{-1}(m) = 700 \exp\left(\frac{m}{1125}\right).$$

3.3. Sledeći korak je bio skaliranje dobijenih frekvencija od 0 do 512 (polovine broja tačaka Furijeove transformacije). Nad dobijenim tačkama formirana je filter banka. Svaki filter je obuhvatao tri uzastopne tačke: u prvoj je počinjao i ima vrednost 0, u drugoj je dostizao maksimum i imao vrednost 1, a u trećoj se vraćao u nulu. Prvi filter je počinjao u prvoj tački, drugi u drugoj, itd.

3.4. Proizvodi vrednosti filtera i vrednosti spektrograma snage su sabrani, a potom logaritmovani (jer je to bio korak ka dobijanju kepstrograma). Zatim je od tih vrednosti određena inverzna brza Furijeova transformacija i dobijen je kepstrogram. Iz kepstrograma je uzeto prvih 12 koeficijenata koji su potrebni za prepoznavanje govora. Koeficijenti gube brojnu vrednost sa porastom rednog broja *n* i zato su skalirani formulom

$$c_n' = \left(1 + \frac{L}{2} \cdot \sin\frac{n\pi}{L}\right) \cdot c_n$$

gde je *L* vrednost koja opisuje koliko puta povećavamo vrednost kepstralnih koeficijenata, c_n neskalirani kepstralni koeficijenti, a c_n' skalirani kepstralni koeficijenti.

4. Od tih skaliranih kepstralnih koeficijenata su izračunati delta kepstralni koeficijenti, koji daju vrednost promene MFCC karakteristika (Practicalcryptography.com), pomoću formule:

$$d_n = \frac{(c_{n+1}' - c_{n-1}') + 2(c_{n+2}' - c_{n-2}')}{10}$$

A od delta kepstralnih koeficijenata su izračunati Acceleration (Delta-Delta) koeficijenti, koji daju dodatne informacije o promeni MFCC karakteristika u vremenu (Practicalcryptography.com), pomoću formule:

$$a_n = \frac{(d_{n+1} - d_{n-1}) + 2(d_{n+2} - d_{n-2})}{10}$$

Kako jednačine za računanje koeficijenata d_n i a_n zavise od prošlih i budućih vrednosti, neophodna je modifikacija na početku i na kraju signala. Prvi ili poslednji koeficijent će se koristiti umesto onih koji nedostaju.

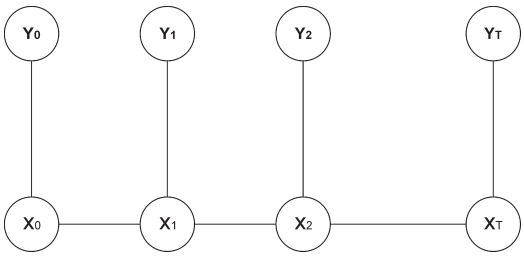
Parametri pri računanju MFCC karakterističnih obeležja su:

1. Broj koeficijenata iz kepstrograma: NUMCEPS = 12
2. Korišćena Hamming-ova prozorska funkcija: USEHAMMING = T
3. Koeficijent predobrade: PREEMCOEF = 1
4. Broj filter banaka: NUMCHANS = 26
5. Vrednost L u formuli za skaliranje: CEPLIFTER = 52

Skriveni Markovljevi modeli

Skriveni Markovljevi modeli (SMM) su metod za modeliranje vremenskih serija podataka. Koriste se u skoro svim sistemima za prepoznavanje govora, prepoznavanje oblika i drugim granama veštačke inteligencije (Ghahramani 2001). SMM određuje verovatnoću da su se neki skriveni procesi (X_0, X_1,..., X_T) dogodili na osnovu niza posmatranja (Y_0, Y_1,..., Y_T) (slika 2).

Kako bismo formirali odgovarajući model, neophodno je da znamo njegove karakteristike. Osnovne karakteristike SMM-a su broj stanja u modelu (N), broj diskretnih simbola u alfabetu posmatranja (M), matrica verovatnoćâ tranzicije



Slika 2. Skriveni Markovljev model: X – stanja, Y – opservacije.

Figure 2. Hidden Markov Model: X – states, Y – observations.

stanja (A), verovatnoća generisanja određenog posmatranja iz određenog stanja (B), inicijalna verovatnoća (π) (Rabiner 1989).

U problemima prepoznavanja govora nije poznata fizička reprezentacija skrivenih stanja SMM-a. Upravo zato ovaj rad pokušava da odgovori na pitanje šta je zapravo skriveno stanje SMM-a kod prepoznavanja govora i od čega zavisi. Na početku se matrica tranzicije stanja inicijalizuje nasumično, a zatim se treniranjem SMM-a dobija konačna matrica tranzicijâ. Rešavanjem ovog problema nije računata verovatnoća generisanja (B); međutim, možda bi upravo analiza generisanih posmatranja iz stanja pomogla pri razumevanju šta su to skrivena stanja.

Ne postoji metod za određivanje optimalne topologije SMM-a (Moreau 2002). Red modela određuje na koliko narednih stanja će trenutno stanje uticati. U ovom radu odabran je model drugog reda, kao na slici 3, zato što je ovakva topologija preporučena u referentnoj literaturi (Gales i Young 2007). Kod modela drugog reda će svako stanje uticati na dva sledeća stanja.

Prilikom analize SMM-a, rešavaju se tri osnovna problema (Rabiner 1989):

- problem ocenjivanja, koji rešavamo pomoću forward-backward algoritma, ovaj problem se pojavljuje kod testiranja sistema
- problem određivanja verovatnoće redosleda pojavljivanja skrivenih stanja, koji nismo rešavali
- problem optimizacije modela, koji rešavamo pomoću Baum-Welch algoritma prilikom treniranja modela.

Rezultati i diskusija

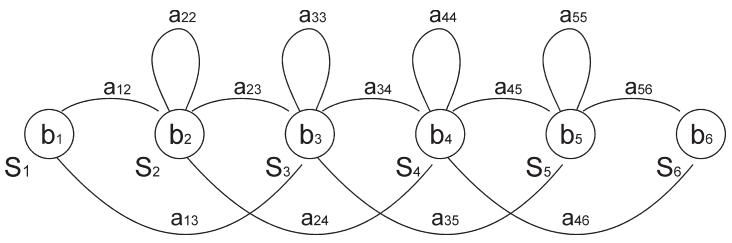
Performanse algoritma su okarakterisane pozitivnom prediktivnom vrednošću – PPV (engl. Positive predictive value, precision) i tačnom pozitivnom stopom – TPS (engl. True positive rate, recall) po formulama

$$\mathrm{PPV} = \frac{t_\mathrm{p}}{r_\mathrm{p}}$$

$$\mathrm{TPS} = \frac{t_\mathrm{p}}{p}$$

gde je t_p– broj tačno pozitivno detektovanih reči (reči koje su detektovane tačno), r_p – broj pozitivno detektovanih reči (reči koje su detektovane), a p – broj pozitivnih reči (reči koje treba da se detektuju).

Prepoznavanje se smatra uspešnim ako je PPV veća od 95%, zato što je u ovom slučaju TPS uvek 100%.



Slika 3. Izgled modela drugog reda. S_i – stanje i, a_{ij} – verovatnoća tranzicije iz stanja i u stanje j, b_i – verovatnoća generisanja posmatranja iz stanja i.

Figure 3. Example of the Hidden Markov Model of 2nd order used in this paper. S_i – state i, a_{ij} – probability of transition from state i to state j, b_i – observation probabilities for state i.

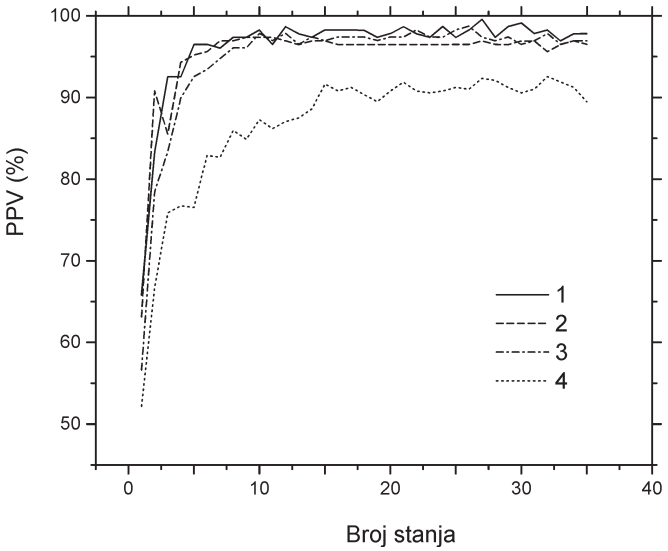
Na slici 4 je prikazan grafik PPV u zavisnosti od broja stanja SMM-a za reči od 1, 2, 3 i 4 sloga. Reči od 1, 2 i 3 sloga uspešno su prepoznate za 7 stanja i više, dok reči od 4 sloga nemaju uspešno prepoznavanje (maksimalna PPV 91%). Može se zaključiti da optimalni broj stanja nije srazmeran broju slogova.

Na slici 4 može se primetiti da uspešnost za grupu 4 (4 sloga, sa sličnim rečima) ne prelazi zadati prag od 95%, dok je u referentnom radu (Davis i Mermelstein 1980) postignuto uspešno prepoznavanje za 12 sličnih reči. Kako su u ovom radu slične reči duže nego u referentnom radu, pad uspešnosti prepoznavanja pripisan je dužini reči. Kako bi to bilo i potvrđeno, potrebno je napraviti bazu sa sličnim, ali kraćim rečima i ponoviti eksperiment.

Na slici 5 je prikazan grafik PPV u zavisnosti od broja stanja za reči od 3, 4, 5 i 6 slova.

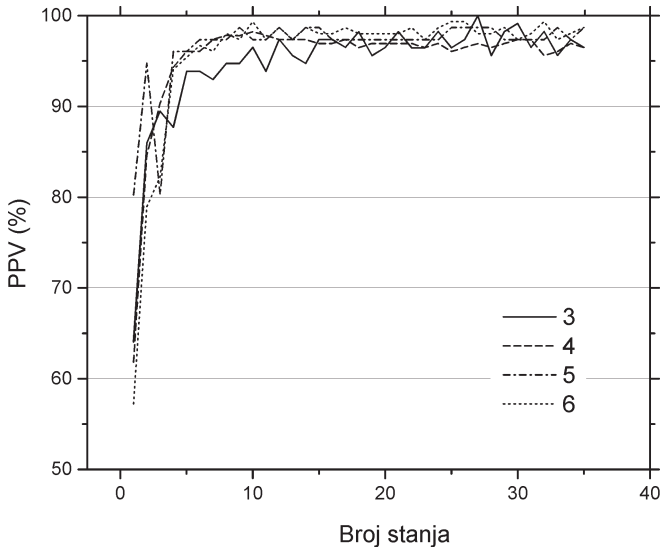
Na slici 6 je prikazan grafik PPV u zavisnosti od broja stanja za reči od 6, 7, 8 i 9 slova.

Sa slika 5 i 6 može se zaključiti da reči od 3, 4, 5 i 6 slova imaju uspešno prepoznavanje za 5 stanja i više, reči od 7 i 8 slova imaju uspešno



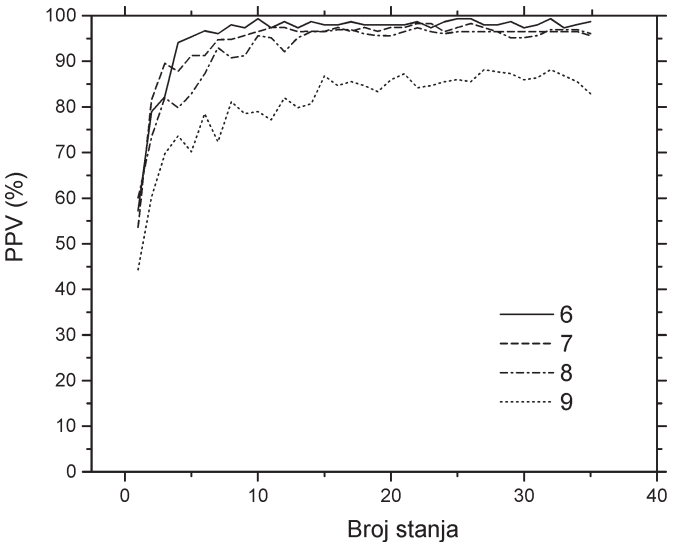
Slika 4.
Zavisnost PPV od broja stanja za 1, 2, 3 i 4 sloga

Figure 4.
Change of PPV as a function of number of hidden states for words with 1, 2, 3 and 4 syllables



Slika 5.
Zavisnost PPV od broja stanja za 3, 4, 5 i 6 slova

Figure 5.
Change of PPV as a function of number of hidden states for words with 3, 4, 5 and 6 letters



Slika 6.
Zavisnost PPV od broja stanja za reči od 6, 7, 8 i 9 slova

Figure 6.
Change of PPV as a function of number of hidden states for words with 6, 7, 8 and 9 letters

prepoznavanje za 7 stanja i više, a reči od 9 slova ne dostižu uspešno prepoznavanje, već dostižu maksimalnu PPV od 85%, i to tek za 15 stanja. Može se zaključiti da optimalni broj stanja nije srazmeran broju slova.

Zaključak

Prema rezultatima koji su dobijeni zaključeno je da optimalan broj stanja nije srazmeran broju slogova, kao ni broju slova. Ali, ne može se tvrditi da zaključak važi za sve reči u srpskom jeziku. Kako bi bilo moguće tvrditi da zaključak važi i za ostale reči u srpskom jeziku, potrebno je ponoviti metod koji je korišćen na celokupnom srpskom rečniku.

U prikazanom sistemu postignuto je uspešno prepoznavanje (PPV > 95%) za reči do 3 sloga, odnosno do 8 slova. Reči od 4 sloga odnosno 9 slova izabrane su tako da budu slične po uzoru na referentni rad (Davis i Mermelstein 1980); međutim, pri formiranju baze nije planirana kontrolna grupa koja u potpunosti reprodukuje karakteristike reči koje su autori koristili.

Literatura

Davis S. and Mermelstein P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28** (4): 357.

Gales M. and Young S. 2007. The Application of Hidden Markov Models in Speech Recognition. *FNT in Signal Processing*, **1** (3), 195.

Ghahramani Z. 2001. An Introduction To Hidden Markov Models And Bayesian Networks. *International Journal Of Pattern Recognition And Artificial Intelligence*, **15** (01): 9.

HTK 2002. Hidden Markov Model Toolkit. Cambridge University Engineering Department.

Juang B. 1984. On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition-A Unified View. *AT&T Bell Laboratories Technical Journal*, **63** (7): 1213.

Lippmann R. 1989. Review of Neural Networks for Speech Recognition. *Neural Computation*, **1** (1): 1.

Moreau N. 2002. HTK Basic Tutorial. Dostupno na: http://my.fit.edu/~vkepuska/HTK/HTK_basic_tutorial.pdf.

Practicalcryptography.com. Practical Cryptography. Dostupno na: http://practicalcryptography.com/miscellaneous

/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

Rabiner L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77** (2): 257.

Young S. Evermann G. Gales M. Hain T. Kershaw D. Liu X. Moore G. Odell J. Ollason D. Povey D. Valtchev V. and Woodland P. 2009. *The HTK Book*. Cambridge University Press.

Ratko Amanović and Nemanja Miković

Application of the Hidden Markov model in Speech Recognition on a Reduced Dictionary

This paper analyzes speaker independent speech recognition on a reduced dictionary using hidden Markov models (HMM) (Rabiner 1989). HMMs are used because they are better for speech recognition in relation to other algorithms (Juang 1984; Lippmann 1989). The features of the speech signals that are used are Mel frequency cepstral coefficients (MFCC) because of their good results in reference work (Davis and Mermelstein 1980). For the purpose of this research, a database of 30 words divided into groups of one, two, three and four syllables spoken by 48 people was made. The Hidden Markov Model Toolkit (HTK) was used to label the database, to calculate features and for training and testing HMMs. The number of hidden states of HMM in speech recognition is unknown. Because Serbian is a language where for one letter there is one phoneme, we expected that the number of hidden states of HMM would be proportional to the number of letters or the number of syllables. However, results show that the number of hidden states of HMM is not proportional to the number of syllables or letters. Further, it is shown that there is no optimal number of hidden states of HMM (figures 4, 5 and 6.). Accuracy achieved on the database was 95% if the number of hidden states was greater than 15.