University of Belgrade - School of Electrical Engineering
Signals and Systems Department

Bachelor Thesis

Speech Recognition using Hidden Markov Models

Student

Stefan Stojanović, 2016/0043

Mentor

Prof. Dr. Željko Đurović

Belgrade, *august* 2020.

Abstract

Hidden Markov models are statistical models applicable to many practical problems. Modeling phones or words using Hidden Markov models is one of the most common approaches in speech recognition. The advantages of such an approach are in its firm mathematical foundation, which enables simple implementation of the algorithms. The models are trained by re-estimation of parameters such that probability of observations is maximized under a given model.

In the first part of this thesis, we present the theoretical aspects of Hidden Markov models and derive re-estimation formulae for models in closed form. We give answers to three fundamental questions of Hidden Markov models - what is the probability of given observations, which states are the most probable for a given model, and how to reestimate parameters of the model. Then we consider aspects of Hidden Markov models which are important for speech recognition and practical implementation of the algorithms. We analyze models with continuous probability density functions of observations, training models with multiple observations, issues with scaling and initialization.

In the second part of the thesis, we apply Hidden Markov models to recognize isolated words independent of the speaker. The vocabulary consists of words zero to nine in the Serbian language. From the first set of experiments, we determined the optimal choice of the model parameters: number of states, number of components of Gaussian mixtures, choice of observation vector, the size of the training set, and the initialization method. In the second set of experiments, for the chosen parameters, we apply the model to recognize words independent of the speaker. For every subject, we form ten different models trained on the sequences of the other subjects. The average accuracy of this system is determined to be between 85.3% and 98.61% depending on the tested word.

Sažetak

Skriveni Markovljevi modeli su statistički modeli čija je primenljivost pokazana na mnoštvu praktičnih problema. Modeliranje glasova ili reči skrivenim Markovljevim modelima je jedan od najčešćih pristupa u prepoznavanju govora. Prednosti ovakvog modela ogledaju se u čvrstoj matematickoj osnovi, koja omogućava jednostavnu implementaciju algoritama. Obučavanje modela vrši se reestimacijom tako da se maksimizuje verovatnoća opservacija pod datim modelom.

U prvom delu ovog rada predstavljeni su teorijski aspekti skrivenih Markovljevih modela i pokazan je način dobijanja reestimacionih formula za modele u zatvorenoj formi. Dati su odgovori na tri osnovna pitanja skrivenih Markovljevih modela – koja je verovatnoća opservacija i koja su stanja najverovatnija za dati model, kao i kako reestimirati parametre modela. Potom su razmatrani aspekti skrivenih Markovljevih modela od važnosti za prepoznavanje govora i praktičnu implementaciju algoritama. Analizirani su modeli sa kontinualnom gustinom verovatnoće opservacija, obučavanje modela sa višestrukim opservacijama, problemi sa skaliranjem i inicijalizacijom.

U drugom delu rada primenjeni su skriveni Markovljevi modeli u prepoznavanju izolovanih reči nezavisno od govornika. Analizirani rečnik su činile cifre nula do devet izgovorene na srpskom jeziku. Prvim setom eksperimenata određen je optimalni izbor parametara modela: broj stanja, broj komponenti Gausove smeše, odabir opservacionog vektora, veličina skupa za obučavanje i način inicijalizacije. U drugom setu eksperimenta su ovako određeni parametri primenjeni u prepoznavanju reči nezavisno od govornika. Za svakog od ispitanika formirano je po deset modela koji su obučeni na sekvencama svih ostalih ispitanika. Procenjena je tačnost sistema testiranjem modela na ispitanicima i ostvarena je srednja tačnost od 85.3% do 98.61% u zavisnosti od testirane reči.

Zahvalnica

Zahvaljujem se redovnom profesoru Elektrotehničkog fakulteta Željku Đuroviću na upućenim primedbama i sugestijama, kao i na predloženoj temi diplomskog rada. Smatram da su skriveni Markovljevi modeli jedan od najlepših spojeva matematičkog formalizma sa nečim praktičnim i primenljivim u svakodnevnom životu. Stoga mi je ova tema, kao inženjeru, bila neizmerno zanimljiva te se nadam da sam uspeo da makar deo entuzijazma prenesem na sam diplomski rad i njegove čitaoce.

Zahvaljujem se i svojoj porodici, prijateljima i kolegama koji su mi značajno olakšali studiranje u prethodne četiri godine, a naročitu zahvalnost želim da iskažem svima koji su učestvovali kao ispitanici u ovom radu.

U Beogradu, avgusta 2020. godine *Autor*

Sadržaj

1 Uvod

> *"Smatrali smo da je pogrešno oponašati ljude. Na kraju krajeva, ako mašina treba da se pomera, pomera se točkovima - ne hodajući nogama. Ako mašina treba da leti, leti poput aviona - ne mašući krilima. Umesto opširnog istraživanja načina na koji ljudi slušaju i razumeju govor, želeli smo da nađemo prirodan način da mašine to same rade."*
>
> – Frederik Jelinek, 1987

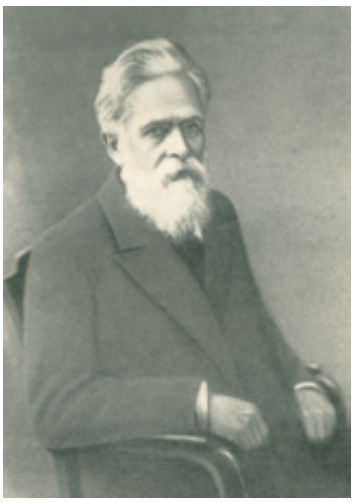
Govor je vokalni komunikacioni metod za izražavanje jezičkih konstrukcija jedinstven za ljudsku vrstu. Većina komunikacije među ljudima odvija se govorom - od iskazivanja osnovnih potreba u ranom detinjstvu, učenja, ostvarivanja emocionalnih i poslovnih odnosa - govor znatno olakšava naš svakodnevni život.

Sa druge strane, komunikacija između ljudi i mašina bila je donedavno zasnovana isključivo na negovornim metodama. Od nastanka prvih telekomunikacionih naprava postojala je međutim težnja da se neki vid komunikacije putem govora uspostavi. Iako su ti prvi napori doveli do nekih značajnih otkrića (koji će u nastavku biti predstavljeni), sam problem prepoznavanja i razumevanja govora nije doživeo jasan napredak. Prve naznake rešenja ovog problema pojavile su se pedesetih godina prošlog veka, a tridesetak godina nakon skromnog početka prepoznavanje govora doživljava nagli napredak zahvaljujući revitalizaciji rada ruskog matematičara Andreja Andrejeviča Markova.

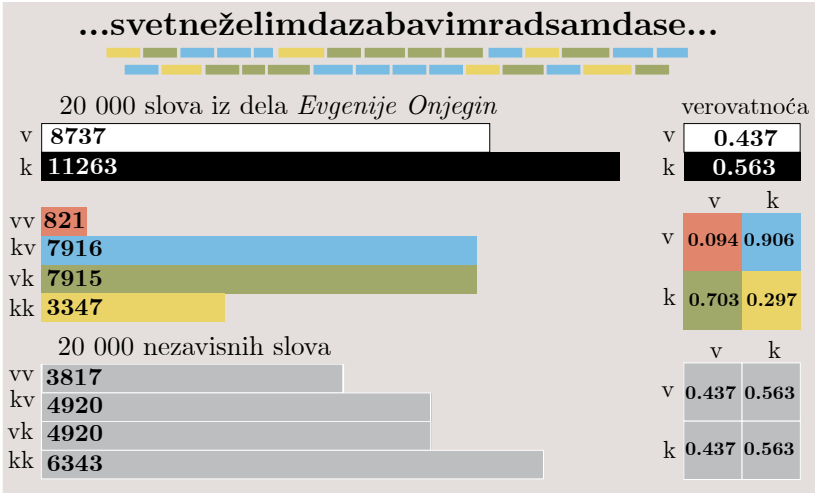
A. A. Markov (1856-1922) se bavio analizom stohastičkih procesa - događaja koji se odvijaju sa određenom verovatnoćom. Kao učenik poznatog petrogradskog matematičara P. L. Čebiševa, Markov nastavlja njegov rad na Zakonu velikih brojeva, i nakon dokaza J. Bernulija, mentora Čebiševa i sam daje dodatne dokaze za ovaj zakon [1]. Prema navodima datim u [1], Markovljeva ideološka i matematička neslaganja sa njegovim kolegom P. Nekrasovim, navode ga na izučavanje slučajnih procesa koji nisu u potpunosti nezavisni. On pokazuje da i za ovakve procese važi Zakon velikih brojeva i da nakon dovoljno dugog posmatranja procesa, njihovi parametri konvergiraju ka stacionarnim vrednostima. Ovakvi procesi, nazvani po autoru Markovljevi procesi, otkriveni 1906. godine, doveli su do najvećeg napretka u istoriji prepoznavanja govora tek sedam decenija kasnije.

Nažalost, sam Markov nije imao preveliku želju da se bavi primenama njegovog otkrića [1], ali 1913. godine u jedinstvenom radu te vrste Markov istražuje prvu primenu Markovljevih procesa u analizi dela *Evgenije Onjegin*, romana u stihovima ruskog pisca A. Puškina [2]. U ovom radu Markov ručno statistički analizira otprilike osminu romana, tj. deo romana u kome se nalazi dvadeset hiljada slova (preskačući razmake i druge prorede kao i ruske simbole za meke i tvrde glasove). U revolucionarnoj matematičkoj analizi jednog poetskog dela Markov pronalazi verovatnoće pojavljivanja suglasnika i samoglasnika kao i parova ova dva tipa glasova. Dodatno, zaključuje da raspored glasova u ovom delu nije nezavistan, tj. da pojavljivanje narednog slova u tekstu zavisi od prethodnog (videti sliku 2).

Otkrićem Markovljevih procesa Markov je stvorio teoriju neophodnu za opisivanje statističkih modela poznatih kao skriveni Markovljevi modeli. Skoro vek nakon Markovljeve smrti ovi modeli imaju primenu u vrlo širokom opsegu oblasti: u predikciji



Slika 1: A. A. Markov - tvorac Markovljevih procesa.



Slika 2: Analiza srpskog izdanja dela *Evgenije Onjegin* adaptirana iz [1]. Markov je ovakvu analizu (na originalnom tekstu) sproveo ručno 1913. godine. v - vokal, k - konsonant, vv, kv, vk i kk - odgovarajući parovi glasova.

ponašanja berze i finansijama [3, 4], analizi strukture genoma [5, 6], kriptoanalizi [7], opisivanju i predviđanju meteoroloških pojava [8, 9], prepoznavanju pokreta [10], sintezi govora [11], u određivanju službe reči u izgovorenoj rečenici (tzv. *part-of-speech tagging*) [12] i, jednu od najčešćih primena, u prepoznavanju govora [13, 14].

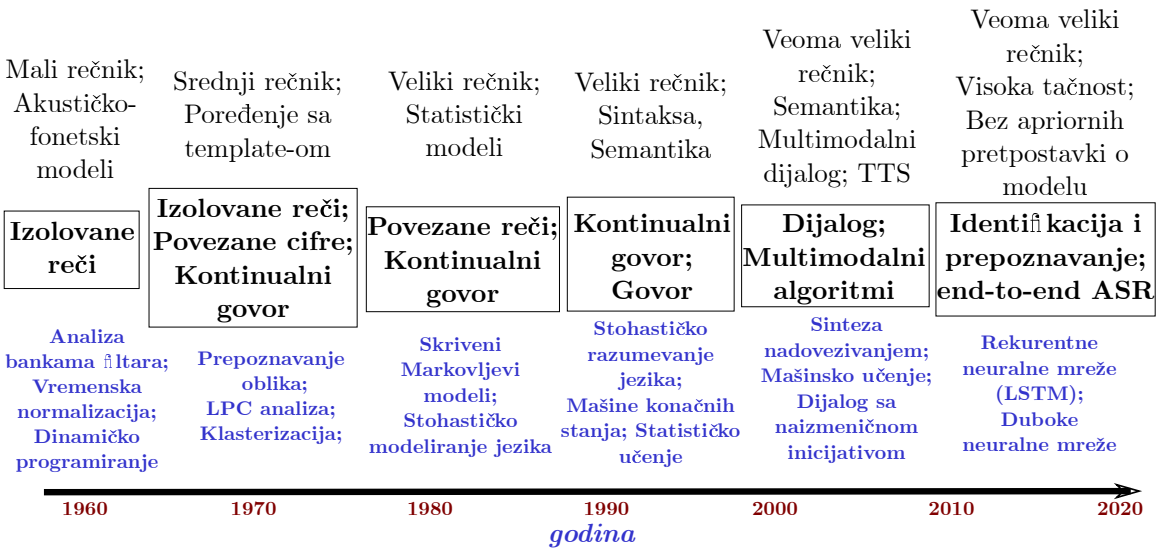
Pod prepoznavanjem govora (eng. (*Automatic*) *Speech Recognition* - (A)SR) podrazumeva se određivanje jezičke informacije sadržane u akustičkom talasu govornog signala. Primene prepoznavanja govora su mnogobrojne - unošenje teksta govorom; komunikacija sa različitim mašinama: automobilima, računarima, pametnim kućama; učenju jezika i asistenciji osoba sa nekim vidom invaliditeta (gluvoneme osobe, osobe koje nisu u mogućnosti da se služe rukama, oboleli od disleksije...); vojne; automatska komunikacija sa *call*-centrom... Skriveni Markovljevi modeli su imali primat u oblasti prepoznavanja govora do pre desetak godina. Da bismo bolje razumeli značaj i vremenski kontekst ovih modela u nastavku je prikazan kratak istorijski pregled otkića u oblasti prepoznavanja govora (za detaljniji pregled kao i reference za navedena otkrića videti [15]).

Prvi značajni rezultati iz oblasti obrade i prepoznavanja govora postignuti su u Belovim laboratorijama. Upravo tamo, tridesetih godina prošlog veka, H. Dadli konstruiše dva revolucionarna otkića: Vocoder, kojim se govorni signal filtrirao i modulirao što je omogućilo ostvarivanje više telefonskih komunikacija na istom medijumu istovremeno; i VODER, prvu električnu napravu za sintezu govora koja je koristila deset filtara za imitaciju rezonatorskih svojstva ljudskog govornog trakta. Dadlijev rad se oslanjao na otkrića H. Flečera, oca stereo zvuka, koji je primetio vezu između spektra govornog signala, njegovih karakteristika i ljudskog opažanja govora.

Pedesetih godina prošlog veka napravljeni su prvi koraci u prepoznavanju vokala (MIT), slogova (Radio Korporacija Amerike - RCA), određenih glasova (Univerzitetski Koledž u Engleskoj) i cifara (1952. - sistem Audrey u Belovim laboratorijama) koji su se pretežno zasnivali na analizi formanata glasova, odnosno njihovih spektralnih karakteristika. Tokom 1960-ih japanski istraživači dolaze do značajnih otkrića:

na Univerzitetu u Kjotou biva formiran prvi uređaj za prepoznavanje glasova koji ne prepoznaje izolovane reči, već koristi segmentaciju govora; na Nagoja Univerzitetu F. Itakura postavlja temelje linearnog prediktivnog kodiranja (LPC) za reprezentaciju govornog signala. Nešto kasnije, u SSSR-u T. Vincik postavlja temelje primene dinamičkog programiranja u prepoznavanju glasova. Iako iz ove istorijske perspektive deluje da su se istraživanja lagano nadovezivala, to nije bio slučaj. Naime, u članku [16] objavljenom 1969. godine jedan od direktora u Belovim laboratorijama J. Pirs daje vrlo negativne prognoze za oblast prepoznavanja govora što će dovesti do prestanka daljeg finansiranja istraživanja sve do njegovog povlačenja 1971. godine. Na sreću, bilo je i onih koji su više verovali u prepoznavanje govora. T. Martin (RCA) otkriva početkom sedamdesetih godina metode za vremensku normalizaciju govornih signala i osnuje prvu kompaniju koja se bavi prepoznavanjem govora. Podstaknute njegovim uspehom, agencija povezana sa američkom vojskom (ARPA) osnuje SUR (eng. *Speech Understanding Research* - program za istraživanje razumevanja govora) kojim su finansirani prvi sistemi za prepoznavanje reči iz, za to vreme, velikih rečnika. Neki od najznačajnijih sistema iz ovog programa su: Harpy (1976, Karnegi Melon Univerzitet - CMU), koji je sa prihvatljivom tačnošću prepoznavao 1011 različitih reči; i DRAGON (1975, Dž. Bejker iz CMU). Upravo sa Bejkerom započinje nova era primene skrivenih Markovljevih modela u prepoznavanju govora.

Bejker se u svom radu oslanja na rezultate koje je otkrio američki matematičar L. Baum sa svojim kolegama iz Instituta za odbrambenu analizu (IDA). Naime, Baum je pokazao kako je moguće, za određene vrste kriterijumske funkcije, reestimirati parametre tako da se kriterijumska funkcija uvek smanjuje. Ta analiza je obuhvatala i skrivene Markovljeve modele za slučaj opservacija sa diskretnim verovatnoćama pojavljivanja. Bejker upoznaje sa svojim radom i kolegu F. Jelineka (IBM), značajnu ličnost za razvoj SR tehnologije. Veliki problem tog vremena bilo je kako primeniti kontinualnu raspodelu verovatnoće za opisivanje opservacija. Početkom 1980-ih godina, istraživac iz IDA Dž. Ferguson održava niz predavanja iz oblasti skrivenih Markovljevih modela, koja će okupiti, kako će se kasnije pokazati, najznačajnije ljude



Slika 3: Istorijat istraživanja u oblasti prepoznavanja govora. Crnim slovima su obeležena najveća dostignuća do datog vremenskog okvira, a plavom metode koje su do dostignuća dovele. Slika je adaptirana iz [15].

za dalji razvoj i primenu ovih modela. Nakon ovih predavanja, u Belove laboratorije dolaze L. Rabiner i F. Juang koji nedugo zatim objavljuju niz radova o primeni skrivenih Markovljevih modela u prepoznavanju govora. Za razliku od Belovih laboratorija, kojima je cilj bio prepoznavanje govora nezavisno od govornika koje bi primenili u svojim telekomunikacionim servisima, IBM razvija sistem za prepoznavanje govora u drugom pravcu - sa ogromnim rečnikom (preko 20 hiljada reči) i obučavanjem (tj. zavisno od govornika) koje bi imalo ulogu u transkripciji - sistem Tangora.

Nakon toga započinje era skrivenih Markovljevih modela sa određenim unapređenjima. Istorijski pregled (sa više dostignuća nakon skrivenih Markovljevih modela) prikazan je na slici 3. Iako se krajem osamdesetih ponovo javlja interesovanje za veštačke neuralne mreže sa otkrićem višeslojnog perceptrona, ovi metodi nisu mogli značajno da pariraju Markovljevim modelima. Međutim, tridesetak godina kasnije, upravo će neuralne mreže dovesti do drugog najvećeg napretka u prepoznavanju govora. Kao nadogradnju rekurentnih neuralnih mreža S. Hohrajter i J. Šmidhuber otkrivaju LSTM (*Long short-term memory*) [17], neuralne mreže kojima je rešen problem iščezavajućeg gradijenta (*vanishing gradient*), koje će potom primeniti u SR. Neuralnim mrežama omogućena je fleksibilnost u rešavanju različitih zadataka iz oblasti procesiranja govora i smanjenje neophodnog apriornog znanja o modelu [18]. Četiri istraživacke grupe (Univerzitet u Torontu, Majkrosoft, Gugl, IBM) objavljuju 2012. godine rad o primeni dubokih neuralnih mreža u prepoznavanju govora [19] sa znatno poboljšanom tačnošću. Dodatno poboljšanje postignuto je primenom konvolucionih neuralnih mreža [20]. Pokazano je da je moguće obučiti duboke neuralne mreže *end-to-end*, što znači da je apriorno znanje o šumu, govornicima, rečniku nepotrebno [21,22]. Rezultati dobijeni primenom konvolucionih i LSTM neuralnih mreža pokazuju da je ovim metodama moguće dostići i ljudsku tačnost u određenim SR zadacima [23].

Iako nadmašeni od strane neuralnih mreža, skriveni Markovljevi modeli pružaju mogućnost za analizu govornog signala iz drugog ugla i samim tim mogu da doprinesu našem boljem razumevanju prepoznavanja govora. Problem prepoznavanja reči primenom skrivenih Markovljevih modela [24], pa i cifara [25] je istražen i u dobroj meri rešen. Valja napomenuti da je jedan od najčešće korišćenih softvera koji se bazira na Markovljevim modelima HTK [26], razvijen od strane laboratorije za mašinsku inteligenciju Univerziteta u Kembridžu.

Cilj ovog rada jeste formiranje sistema na bazi skrivenih Markovljevih modela koji prepoznaje cifre 0-9 izgovorene na srpskom jeziku. Razmatran je odabir parametara modela, kao i reprezentacije govornog signala, kako bi se ostvarile što bolje performanse. Ispitivana je robustnost sistema testiranjem modela na ispitanicima koji nisu učestovali u obučavanju. Dobijeni rezultati poređeni su sa rezultatima iz literature.

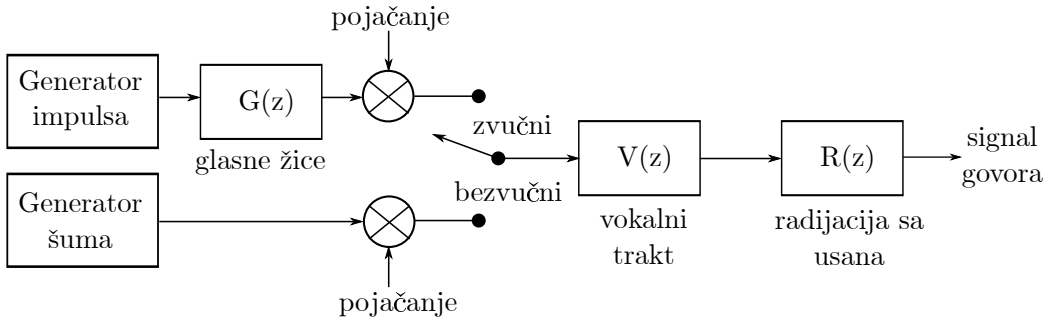
Rad je organizovan na sledeći način. U 2. poglavlju dat je teorijski okvir za prepoznavanje govora. U potpoglavlju 2.1 predstavljene su karakteristike govornog signala, u potpoglavlju 2.2 izdvajanje odgovarajućih obeležja koja se koriste kao opservacije u modelu. Potpoglavlje 2.3 sadrži detaljan opis skrivenih Markovljevih modela, *forward-backward* i Viterbijevog algoritma i reestimacije modela. U poslednjem potpoglavlju su prikazani praktični aspekti primene skrivenih Markovljevih modela u prepoznavanju govora. U 3. poglavlju prikazani su i diskutovani dobijeni rezultati - u 3.1 je analiziran uticaj parametara modela na ostvarenu tačnost, a u 3.2 je data analiza prepoznavanja govora na ispitanicima koji nisu deo obučavajućeg skupa.

2 Metodologija rada

2.1 Signal govora

U stvaranju ljudskog govora učestvuje više organa i celina: pluća, grkljan, glasne žice, ždrelo, usna i nosna šupljina. Sam proces produkcije glasa je predvođen nervnim sistemom koji upravlja mišicima vokalnog trakta.

Da bismo lakše analizirali glas sa stanovišta obrade signala, obično se govor posmatra kao signal na izlazu govornog sistema koji je pobuđen na odgovarajući način. Pobuda, na biološkom nivou, podrazumeva vazduh koji iz pluća nailazi na glasne žice i dalje nastavlja da se kreće vokalnim traktom. U zavisnosti od treperenja glasnih žica pobudu sistema možemo okarakterisati dvema grupama. Ukoliko se pri izgovaranju glasa glasne žice otvaraju i zatvaraju periodično, onda dolazi do stvaranja zvučnih glasova - kakvi su npr. samoglasnici. Sa druge strane, ukoliko glasne žice ne trepere prilikom izgovaranja, onda je taj glas bezvučan. Ovakva podela omogućava da pobudu za zvučne glasove predstavimo povorkom impulsa, a za bezvučne glasove širokopojasnim šumom. Dalje, sistemom $G(z)$ opisuje se formiranje periodične pobude za zvučne glasove na osnovu dobijenih impulsa.



Slika 4: Modeliranje ljudskog govornog aparata funkcijom prenosa $H(z) = G(z)V(z)R(z)$. Gornja grana odgovara produkciji zvučnih glasova, a donja bezvučnih. Slika je formirana na osnovu [27].

Prilikom prolaska kroz vokalni trakt, ovakva pobuda se filtrira filtrom čije karakteristike zavise od oblika vokalnog trakta, tj. aktiviranih mišića. Pretpostavićemo da je takav filtar dat funkcijom prenosa V(z). Dodatno, na osnovu akustičkih modela, stiže se do veze između pritiska i protoka vazduha na usnama koji je diferencijalnog oblika: $R(z) = R_0(1 - z^{-1})$ [27]. Za dalju analizu pogodno je predstaviti ceo sistem funkcijom prenosa od vazdušnog protoka u plućima do pritiska na usnama:

$$H(z) = G(z)V(z)R(z) \tag{1}$$

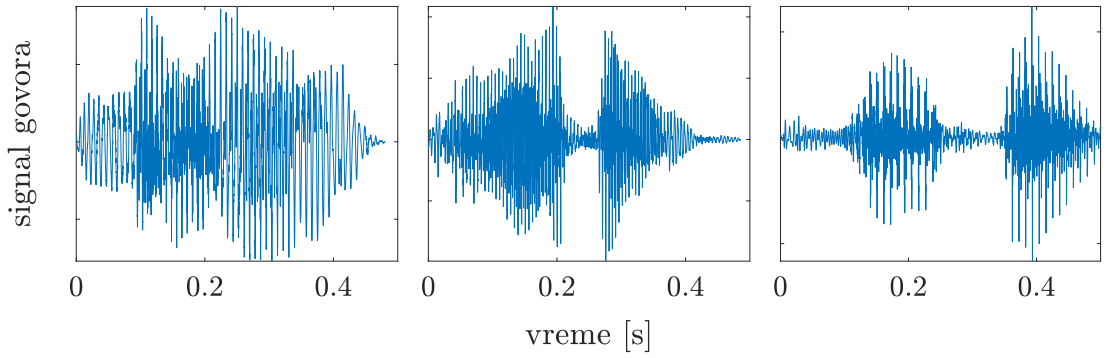
Na slici 4 je prikazan celokupni sistem kojim je modelirano formiranje govora.
Važno je napomenuti da se ovakav sistem za većinu glasova (sem za nazale) može opisati samo polovima sistema, tj. nule sistema nisu neophodne, pa se sistem $H(z)$ može predstaviti kao:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_k a_k z^{-k}} \tag{2}$$

gde su a_k odgovarajući polovi sistema.

2.2 Obrada signala i izdvajanje obeležja

Većina standardnih jezika sadrži više od tridesetak glasova - srpski jezik tačno trideset. U nekim jezicima, u zavisnosti od toga koji glasovi im prethode i slede, isti glasovi ne moraju biti izgovoreni na isti način. To znači da je raznolikost glasova koje jedan čovek može da proizvede dosta veliki. Pored raznolikosti glasova, način na koji jedan čovek izgovara isti glas je najčešće promenljiv - menjamo brzinu izgovaranja, jačinu govora, akcenat... Sa druge strane, različiti ljudi imaju različitu građu vokalnog trakta, samim tim i drugačije izgovaraju iste glasove. Na slici 5 su prikazani signali govora tri ispitanika pri izgovaranju reči "jedan" - primećuje se da se ovi signali dosta razlikuju. Dakle, ljudski govor je izrazito raznovrstan sa izraženom intra- i intervarijabilnošću. To čini problem prepoznavanja govora naročito teškim. Većina metoda (pre neuralnih



Slika 5: Signal govora za reč "jedan" izgovorenu od strane troje različitih ispitanika.

mreža) zasnivala se na pronalaženju i izdvajanju odgovarajućih obeležja govora koji bi analizu učinili jednostavnijom. U prethodnom poglavlju je uveden sistem $H(z)$ kojim je modelirano formiranje govora. Jedna od prvih metoda za izdvajanje obeležja nazvana linearno prediktivno kodiranje (LPC) zasniva se upravo na estimaciji polova modela $H(z)$.

2.2.1 LPC reprezentacija govora

Neka je dat diskretizovani govorni signal $x[n]$ i neka je sa $\hat{x}[n]$ data aproksimacija tog signala definisana kao:

$$\hat{x}[n] = \sum_{k=1}^{p} \hat{a}_k x[n-k] \tag{3}$$

tj. kao linearna kombinacija p zakašnjenih odbiraka signala x. Primetimo da je ovakva predikcija posledica pretpostavljenog autoregresivnog modela $H(z)$ reda p - AR(p):

$$H(z) = \frac{X(z)}{U(z)} = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \iff x[n] = u[n] + \sum_{k=1}^{p} a_k x[n-k] \tag{4}$$

gde smo pretpostavili da znamo tačne koeficijente a i da je broj polova tačno p. Jasno je da predikcija (3) odgovara tačnom obliku signala x kada nema pobude ($u = 0$). U nastavku će biti prikazan način dobijanja optimalnih (u smislu srednje kvadratne greške) koeficijenata $\{\hat{a}_k\}_{k=\overline{1,p}}$.

Greška napravljena predikcijom oblika (3) je:

$$e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{p} \hat{a}_k x[n-k] \tag{5}$$

Srednja kvadratna greška predikcije jednaka je:

$$E = \sum_n e^2[n] = \sum_n \Big(x[n] - \sum_{k=1}^{p} \hat{a}_k x[n-k] \Big)^2 \tag{6}$$

Minimizacijom izraza po nepoznatim koeficijentima $\{\hat{a}_m\}_{m=\overline{1,p}}$. tj. postavljanjem $\partial E / \partial \hat{a}_m = 0$ za svako $m = \overline{1,p}$ dobija se sistem jednačina oblika:

$$\sum_n x[n]x[n-m] = \sum_n x[n-m] \sum_{k=1}^{p} \hat{a}_k x[n-k] = \sum_n x[n] \sum_{k=1}^{p} \hat{a}_k x[n-k+m] \tag{7}$$

gde je u poslednjoj jednakosti izvršena smena indeksa. Označimo autokorelacionu funkciju stacionarnog signala x kao: $r_{xx}[m] = \sum_n x[n]x[n+m]$ (primetite da je u izrazu za E i r_{xx} izbačena ista konstanta radi preglednosti). Sada se izraz (7) može zapisati kao:

$$r_{xx}[-m] = \sum_{k=1}^{p} \hat{a}_k r_{xx}[m-k] \tag{8}$$

za svako $m = \overline{1,p}$. Kako je autokorelaciona funkcija simetrična ($r_{xx}[m] = r_{xx}[-m]$), u matričnom zapisu ovaj sistem je dat sa:

$$\begin{bmatrix} r_{xx}[1] \\ r_{xx}[2] \\ \vdots \\ r_{xx}[p] \end{bmatrix} = \begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \dots & r_{xx}[p-1] \\ r_{xx}[1] & r_{xx}[0] & \dots & r_{xx}[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[p-1] & r_{xx}[p-2] & \dots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} \tag{9}$$

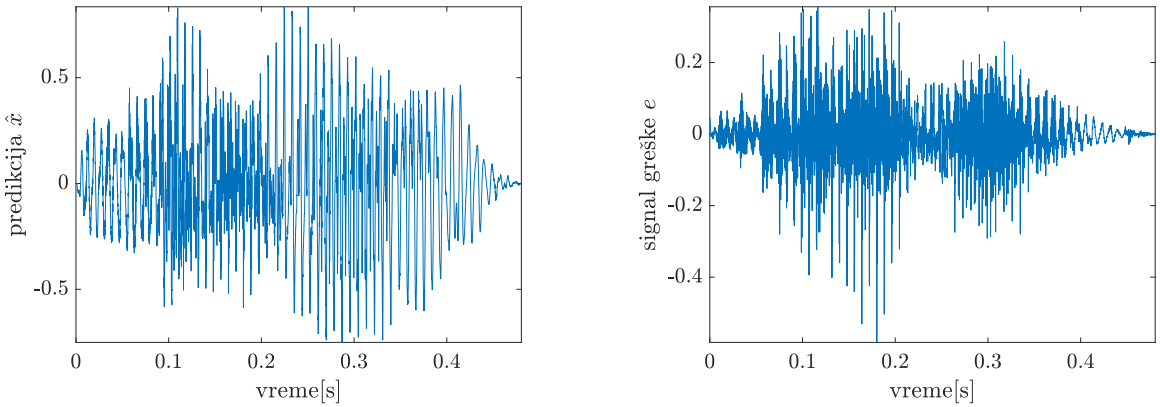
Navedena kvadratna matrica je takođe Toeplitz-ova matrica, što znači da je kompleksnost traženja koeficijenata $\hat{a}_k$ (kompleksnost $\mathcal{O}(p^2)$) manja nego kod običnih matrica.

U prethodnom izvođenju pretpostavljeno je da je signal x stacionaran signal. Iako signal govora nije stacionaran, posmatranjem dovoljno kratkih segmenata ovog signala pretpostavka o stacionarnosti je primenljiva. Treba napomenuti da postoje i druge metode koje se mogu primeniti za estimaciju parametara $\{\hat{a}_k\}_{k=\overline{1,p}}$ poput kovarijacione metode koja je optimalna (ukoliko je signal zaista izlaz AR(p) modela) ali su rešenja manje stabilna.

Na slici 6 (levo) prikazana je predikcija govornog signala sa slike 5 (levo) za red modela $p = 10$. Preslušavanjem ovog signala jasno se prepoznaje reč "jedan". Na slici 6 (desno) prikazana je greška estimacija, koja je veća u bezvučnim delovima reči kada se glasovi ne mogu u potpunosti reprezentovati AR(p) modelom.

2.2.2 Kepstralna reprezentacija govora

Pored LPC koeficijenata postoji grupa drugih koeficijenata koji se mogu koristiti za opisivanje govornog signala i uglavnom se oni mogu sračunati na osnovu LPC koeficijenata. Pokazano je da kepstralni koeficijenti ostvaruju najbolje rezultate iz ove



Slika 6: Predikcija (levo) i greška predikcije (desno) govornog signala primenom $p = 10$ LPC koeficijenata.

grupe koeficijenata u primeni prepoznavanja govora [25].

Neka je sa $\mathcal{F}$ označen operator diskretne Furijeove transformacije (DFT). Tada se kepstar signala x definiše sa:

$$c[n] = \mathcal{F}^{-1}\Big\{\log|\mathcal{F}\{x\}|\Big\} = \sum_{k=0}^{N-1} \log\Big|\sum_{m=0}^{N-1} x[m]e^{-j\frac{2\pi}{N}mk}\Big|e^{j\frac{2\pi}{N}kn} \tag{10}$$

Razlika kepstra u odnosu na inverznu DFT ogleda se u računanju logaritma modula diskretne Furijeove transformacije. Glavna prednost ovakvog pristupa jeste princip separacije, koji je predstavljen u nastavku. Spojimo najpre prenosnu funkciju vokalnog trakta i radijacije na usnama u jedinstvenu funkciju $V'(z) = V(z)R(z)$. Tada na osnovu jednačine (1) važi:

$$H(z) = G(z)V'(z) \iff h[n] = g[n] * v'[n] \tag{11}$$

gde je sa $*$ označen operator konvolucije. Traženjem kepstra signala h dobija se:

$$c_h[n] = \mathcal{F}^{-1}\Big\{\log|\mathcal{F}\{g * v'\}|\Big\} = \mathcal{F}^{-1}\Big\{\log|G(e^{j\omega})V'(e^{j\omega})|\Big\} = \tag{12}$$

$$= \mathcal{F}^{-1}\Big\{\log|G(e^{j\omega})| + \log|V'(e^{j\omega})|\Big\} = c_g[n] + c_{v'}[n] \tag{13}$$

Dakle, moguće je linearno razdvojiti kepstralne koeficijente koje potiču od glasnih žica (ekscitacije) i vokalnog trakta. Dodatno, pokazuje se da su koeficijenti koji odgovaraju ekscitaciji $c_g[n]$ pomereni ka višim kvefrencijama (većim vrednostima n), dok vokalnom traktu odgovaraju niži koeficijenti. Samim tim, odabirom prvih nekoliko kepstralnih koeficijenata c_h uzima se samo onaj deo signala koji je određen vokalnim traktom. Upravo to čini kepstralne koeficijente vrlo efikasnim u reprezentaciji govora.

Pokažimo sada odnos koji važi između kepstralnih i LPC koeficijenata. Radi jednostavnosti izvođenja odredićemo vezu sa kompleksnim kepstrom $\tilde{h}$ (čiji realni deo odgovara kepstru definisanom u (10)).

$$\log H(z) = -\log\left(1 - \sum_{k=1}^{p} \hat{a}_k z^{-k}\right) = \sum_{m=-\infty}^{\infty} \tilde{h}[m]z^{-m} \tag{14}$$

Diferenciranjem izraza po z dobija se:

$$\frac{-\sum_{n=1}^{p}(-n)\hat{a}_n z^{-n-1}}{1-\sum_{k=1}^{p}\hat{a}_k z^{-k}} = \sum_{m=-\infty}^{\infty}(-m)\tilde{h}[m]z^{-m-1} \tag{15}$$

Potom se množenjem odgovarajućim faktorom dobija:

$$-\sum_{n=1}^{p}(-n)\hat{a}_n z^{-n} = \sum_{m=-\infty}^{\infty}(-m)\tilde{h}[m]z^{-m} - \sum_{k=1}^{p}\sum_{m=-\infty}^{\infty}\hat{a}_k m\tilde{h}[m]z^{-m-k} \tag{16}$$

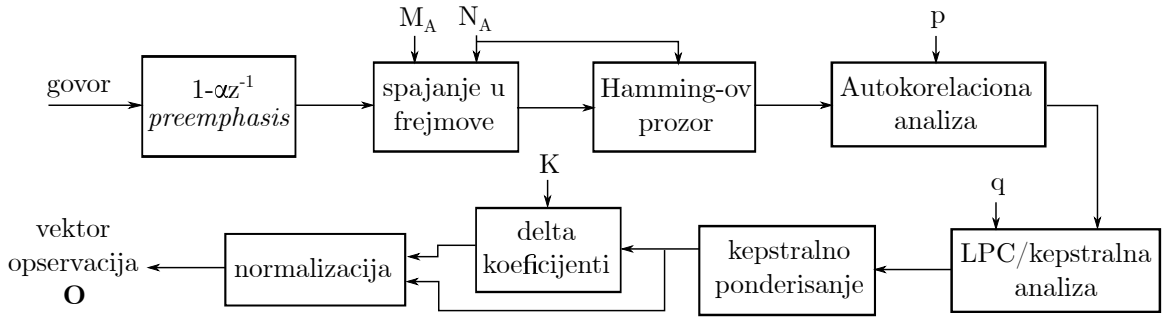
Pažljivim uvođenjem smene i izjednačavanjem članova po z dobijaju se sledeći izrazi:

$$\tilde{h}[n] = \begin{cases} \hat{a}_n + \sum_{k=1}^{n-1}\frac{k}{n}\tilde{h}[k]\hat{a}_{n-k}, & 0 < n \leq p \\ \sum_{k=n-p}^{n-1}\frac{k}{n}\tilde{h}[k]\hat{a}_{n-k}, & n > p \\ 0, & \text{inače} \end{cases} \tag{17}$$

Često se umesto standardnih kepstralnih ili LPC koeficijenata koriste koeficijenti koji u obzir uzimaju samu prirodu govora i ljudskog auditornog sistema. Takvi su npr. koeficijenti Perceptualne Linearne Predikcije (PLP) ili Mel-frekvencijski kepstralni koeficijenti (MFCC). MFCC se dobijaju nakon primene diskretne kosinusne transformacije (DCT) pa su sami koeficijenti izrazito nekorelisani. Pokazano je da ovi koeficijenti postižu najbolje rezultate u problemima prepoznavanja govora [28]. U rezultatima će biti upoređeni dobijeni rezultati za prepoznavanje govora kepstralnim koeficijentima sa onima dobijenim primenom MFCC generisanih pomoću ugrađene Matlab funkcije. Za više teorijskih informacija o MFCC pogledati [27].

2.2.3 Blok šema sistema za izdvajanje obeležja

Nakon definisanja koeficijenata kojima će govor biti predstavljen, u nastavku je data osnovna blok šema sistema (slika 7) za izdvajanje obeležja govornog signala (postupak je preuzet iz [13]).



Slika 7: Blok šema sistema za izdvajanje obeležja govornog signala. Šema je adaptirana iz [27].

Nakon snimanja audio signala sa odgovarajućom frekvencijom odabiranja, signal se

najpre "naglašava" (eng. *preemphasis*) tako što se pojačavaju više učestanosti signala. To se postiže sistemom sa diferencijalnom transfer funkcijom oblika $1 - \alpha z^{-1}$ gde je $\alpha \approx 0.95$. Potom, kako bismo osigurali stacionarnost signala, posmatra se signal na prozorima dužine $N_A \approx 45\ ms$, a preklapanje ovih prozora obično iznosi $M_A \approx 15\ ms$. Kako bi granični efekti bili potisnuti, dobijeni frejmovi se množe Hamingovom prozorskom funkcijom (odbirci na krajevima su oslabljeni, a u sredini pojačani).

Potom se pristupa računanju p odbiraka autokorelacione funkcije, koji se koriste za računanje LPC koeficijenata prema formuli (9). Od p LPC koeficijenata se formira novih q kepstralnih koeficijenata prema formuli (17). Kako svi kepstralni koeficijenti ne sadrže istu količinu korisnih informacija, oni se ponderišu množenjem prozorskom funkcijom oblika:

$$W_c(m) = 1 + \frac{q}{2}\sin\left(\frac{\pi m}{q}\right), \qquad m = \overline{1, q} \tag{18}$$

Ponekad je pogodno da se, sem samih kepstralnih koeficijenata, u opservacionom vektoru nađu i koeficijenti koji odgovaraju brzini promene kepstralnih koeficijenata - delta koeficijenti. Označimo najpre sa $c_l[m]$ m-ti kepstralni koeficijent u l-tom prozoru. Tada se ovaj izvod po vremenu aproksimira polinomom koji se računa na osnovu $2K + 1$ prozora centriranih oko trenutnog:

$$\Delta c_l(m) = \sum_{k=-K}^{K} k c_{l-k}(m), \qquad m = \overline{1, q} \tag{19}$$

Na kraju je pogodno normalizovati dobijene koeficijente kako bi imali iste opsege vrednosti. Normalizacija znatno olakšava dalje numeričke postupke sa vektorima opservacija. U postupcima koji slede svi koeficijenti iz opservacija su normalizovani tako da imaju nultu srednju vrednost i jediničnu varijansu.

2.3 Skriveni Markovljevi modeli

U ovom poglavlju biće predstavljeni skriveni Markovljevi modeli (eng. *Hidden Markov Models* - HMM), dati odgovori na tri osnovna pitanja u vezi sa ovim modelima i biće razmatrani problemi koji nastaju pri praktičnoj implementaciji ovih modela i njima odgovarajućih algoritama.

2.3.1 Markovljevi procesi

Uvedimo za početak koncept Markovljevih procesa, stohastičkih procesa koji dobro opisuju sekvencijalne događaje kakav je i sam govor. Zbog jednostavnosti i praktične primenljivosti, biće razmatran najjednostavniji tip Markovljevih procesa - Markovljev lanac, kod kojeg je vreme diskretno, a prostor stanja konačan.

Svaki stohastički proces X može se posmatrati kao funkcija dveju promenljivih t i ω:

$$X(t,\omega): \quad t \in T,\ \omega \in \Omega$$

gde je T kontinualni ili diskretni skup, a Ω skup svih mogućih slučajnih događaja. Fiksiranjem vrednosti promenljive t, $X(t,\omega) = X_t$ postaje slučajna promenljiva, dok se fiksiranjem jedne realizacije slučajnog događaja ω dobija deterministička funkcija $X_\omega(t)$.

Definicija 2.1. *Neka je $X(t,\omega)$, $t \in T$ stohastički proces, pri čemu su X_t slučajne promenljive koje uzimaju vrednosti iz diskretnog skupa stanja D_X i neka su date tačke $t_i \in T$ $(i = \overline{1,N})$ takve da važi poredak $t_1 < t_2 < \cdots < t_{N-1} < t_N$. Tada je $X(t,\omega)$ Markovljev lanac ako važi:*

$$P(X_{t_N} = x_{k_N} | X_{t_1} = x_{k_1}, X_{t_2} = x_{k_2}, \ldots, X_{t_{N-1}} = x_{k_{N-1}}) = P(X_{t_N} = x_{k_N} | X_{t_{N-1}} = x_{k_{N-1}}) \tag{20}$$

gde je $x_{k_i} \in D_X$ za svako i.

Iz prethodne definicije može se zaključiti da ukoliko je neki proces Markovljev njegova vrednost u budućnosti zavisi od prošlih trenutaka samo preko sadašnje vrednosti - tj. sva predistorija procesa sadržana je u prethodnom trenutku. U nastavku će vrednost slučajnog procesa u trenutku t_i biti obeležena sa X_i, a Markovljev lanac $X_1, X_2, \ldots, X_N$ kao $\{X_k\}$. Jednačina (20) se može skraćeno zapisati kao:

$$P(X_N = x | X_1, X_2, \ldots, X_{N-1}) = P(X_N = x | X_{N-1}), \qquad x \in D_X \tag{21}$$

Teorema 2.1. *Neka je $\{X_k\}$ Markovljev lanac. Tada važi:*

$$P(X_1, X_2, \ldots, X_N) = P(X_1) \prod_{i=1}^{N-1} P(X_{i+1} | X_i) \tag{22}$$

Dokaz. Primenom definicije uslovne verovatnoće i matematičke indukcije lako je pokazati da važi:

$$P(X_1, X_2, \ldots, X_N) = P(X_1) P(X_2|X_1) P(X_3|X_1, X_2) \cdots P(X_N|X_1, X_2, \ldots, X_N) \tag{23}$$

Dalje, kako je $\{X_k\}$ Markovljev lanac sledi da svaka slučajna promenljiva X_i uslovno zavisi samo od X_{i-1}, odakle sledi tražena formula. □

Po ovoj teoremi, za opisivanje Markovljevog lanca dovoljno je znati početnu raspodelu $P(X_1)$ i verovatnoće prelaza iz stanja i u stanje $(i+1)$: $P(X_{i+1}|X_i)$. Posebno od interesa su homogeni Markovljevi lanci kod kojih $P(X_{i+1}|X_i)$ zavisi isključivo od razlike t_{i+1} i t_i a ne i od samih vremenskih trenutaka. Ukoliko usvojimo pretpostavku da su ovi vremenski trenuci ekvidistantni (sa jediničnim rastojanjem), analiza Markovljevih lanaca se znatno olakšava.

Definicija 2.2. *Neka je $\{X_k\}$ Markovljev lanac i neka skup mogućih stanja D_X sadrži tačno N elemenata. Matrica $\mathbf{A} = [a_{ij}]$ takva da je:*

$$(\forall k) \quad a_{ij} = P(X_{k+1} = j | X_k = i), \qquad i, j = \overline{1, N} \tag{24}$$

naziva se matricom prelaza Markovljevog lanca $\{X_k\}$.

Kako je verovatnoća nenegativna veličina i proces iz svakog i-tog stanja mora da pređe u jedno od drugih N stanja, za elemente matrice prelaza važe sledeće jednačine:

$$(\forall i, j) \quad a_{ij} \geq 0 \tag{25a}$$

$$(\forall i) \quad \sum_{j=1}^{N} a_{ij} = 1 \tag{25b}$$

Ovakva matrica $\mathbf{A}$, koja ispunjava uslove (25a) i (25b), naziva se stohastička matrica. Ranije je napomenuto da je pored verovatnoća prelaska iz jednog stanja u drugo neophodno poznavati i apriorne verovatnoće stanja. Stoga, uvedimo vektor apriornih verovatnoća stanja $\mathbf{\Pi} = [\pi_i]$ kao:

$$\pi_i = P(X_1 = x_i), \qquad i = \overline{1, N} \tag{26}$$

Primetite da je za dato $\mathbf{A}$ i $\mathbf{\Pi}$ moguće odrediti verovatnoću da se proces nalazi u bilo kojem stanju u bilo kojem zadatom trenutku. Opisivanje ovakvim modelom može biti prigodno u slučaju kada opservacije ne postoje ili kada je moguće direktno opservirati stanja. Međutim, sa stanovišta prepoznavanja govora, stanja nikad nisu direktno opservabilna. Tada je neophodno koristiti posebno formirane modele procesa koji se nazivaju skriveni Markovljevi modeli.

Skriveni Markovljevi modeli

Ukoliko zamislimo stanja Markovljevih procesa kao glasove koje izgovaramo, tada opservacije predstavljaju akustički signal koji je snimljen pri izgovaranju tih glasova. Dakle, sama stanja nisu direktno opservabilna. Obeležimo date opservacije kao vektor opservacija dužine T: $\mathbf{O} = [\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_T]$. Na osnovu teorijskih razmatranja iz podsekcije 2.2.3 sledi da je svaki vektor opservacija $\mathbf{O}_t$ sačinjen od odgovarajućih koeficijenata formiranih u t-tom frejmu sekvence. Pretpostavimo za sada da je skup vrednosti koje opservacije mogu uzeti konačan i to dužine M, i označimo ga sa $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M]$.

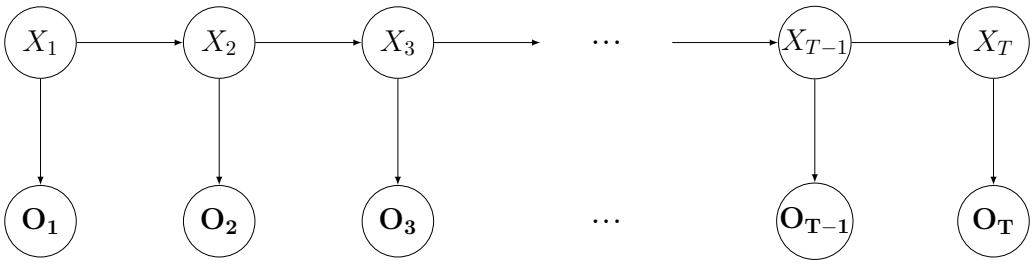
Neka su sa $\mathbf{X} = [X_1, X_2, \ldots, X_T]$ označena stanja Markovljevog lanca koja mogu uzeti jednu od N vrednosti: $x_1,\ x_2, \ldots, x_N$. Skriveni Markovljevi modeli, pored pretpostavke Markovljevih lanaca (20), pretpostavljaju sledeću uslovnu nezavisnost:

$$P(\mathbf{O}_t = \mathbf{v}_k | X_1 = x_{k_1}, X_2 = x_{k_2}, \ldots, X_t = x_{k_t}) = P(\mathbf{O}_t = \mathbf{v}_k | X_t = x_{k_t}) \tag{27}$$

Dakle, verovatnoća opservacije simbola $\mathbf{v}_k$ u trenutku t, pri datom trenutnom stanju X_t, ne zavisi od ostalih stanja. Definišimo sada matricu $\mathbf{B} = [b_{ij}]$, dimenzija $M \times N$, koja sadrži verovatnoće opservacija simbola $\mathbf{v}_i$ $(i = \overline{1, M})$ u stanju x_j $(j = \overline{1, N})$:

$$b_j(i) = b_{ij} = P(\mathbf{O}_t = \mathbf{v}_i | X_t = x_j) \tag{28}$$

Kao i kod matrice prelaza pretpostavićemo da su vrednosti elemenata b_{ij} matrice $\mathbf{B}$ nezavisne od vremena - tj. da zavise samo od stanja i opservacije. Na osnovu pretpostavki (20) i (27) može se formirati graf skrivenih Markovljevih modela u kojoj su prikazane uslovne zavisnosti između odgovarajućih promenljivih - stanja i opservacija (slika 8).



Slika 8: Uslovne zavisnosti u mreži skrivenih Markovljevih modela.

Kako je proces prelaska iz stanja u stanje stohastičke prirode, a i opservacija određenog simbola u datom stanju je takođe stohastički proces, kaže se da se skrivenim Markovljevim modelima opisuju dvostruki stohastički procesi.

Svaki HMM je okarakterisan brojem stanja N, kardinalnošću skupa vrednosti opservacija M, matricom prelaza $\mathbf{A}$, matricom verovatnoće opservacije simbola $\mathbf{B}$ i matricom apriornih verovatnoća $\mathbf{\Pi}$. Obično se HMM definisan ovim parametrima skraćeno označava kao $\lambda = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$.

Pokazalo se da je za praktičnu upotrebu skrivenih Markovljevih modela potrebno rešiti tri osnovna problema:

1. Problem nalaženja verovatnoće: Za dati model $\lambda = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$, koja je verovatnoća da opservirana sekvenca bude $\mathbf{O} = [\mathbf{O}_1, \ldots, \mathbf{O}_T]$?

2. Problem estimacije optimalnih stanja: Za dati model $\lambda = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ i opservacije $\mathbf{O}$ koja je sekvenca stanja $\mathbf{X} = [X_1, \ldots, X_T]$ optimalna?

3. Problem estimacije parametara: Ako je opservirana sekvenca $\mathbf{O}$ kako naći parametre HMM $\lambda = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ tako da je opservirana sekvenca najverovatnija?

U narednim sekcijama biće prikazana rešenja sva tri problema.

2.3.2 Prvi problem HMM

Pretpostavimo za početak da je poznata sekvenca stanja $\mathbf{X} = [X_1, \ldots, X_T]$. Na osnovu pretpostavke (27) o nezavisnosti trenutne opservacije od drugih stanja pri datom trenutnom stanju dobija se sledeći izraz:

$$P(\mathbf{O}|\mathbf{X}, \lambda) = \prod_{t=1}^{T} P(\mathbf{O}_t | X_t = x_t, \lambda) = \prod_{t=1}^{T} b_{x_t}(\mathbf{O}_t) \tag{29}$$

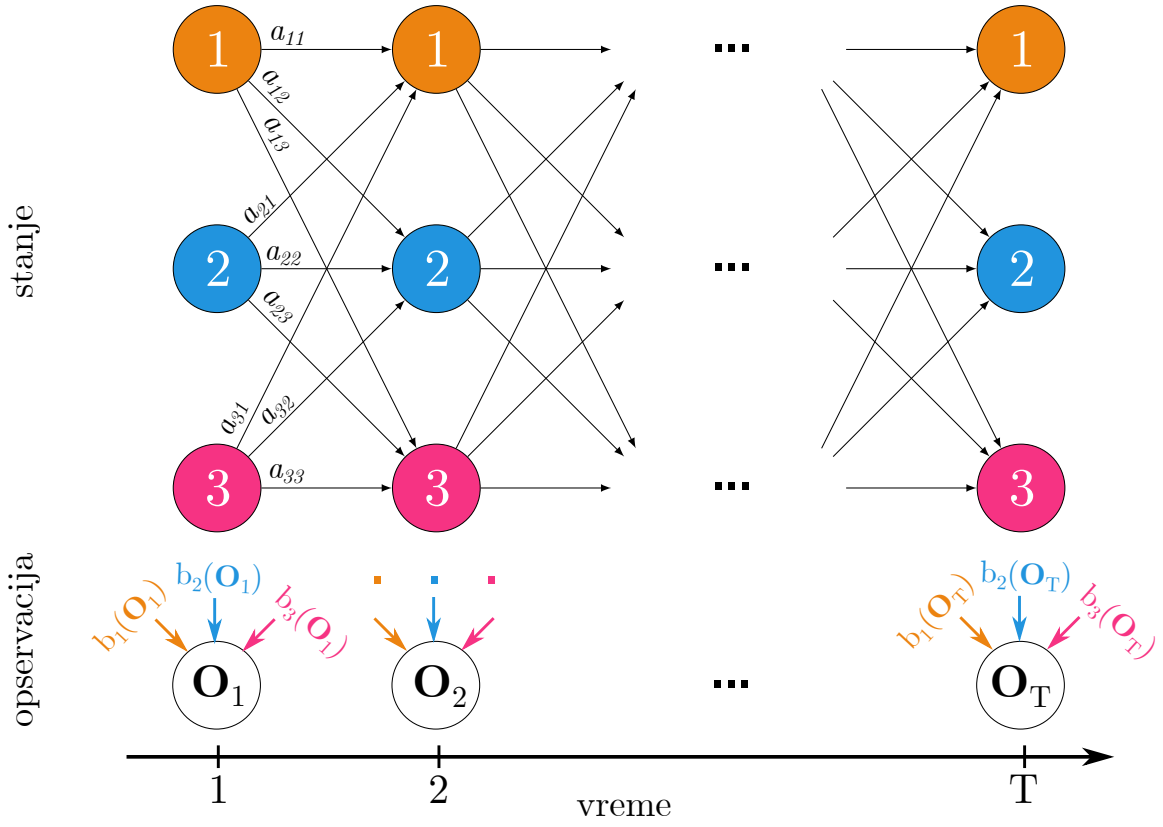
Potom, na osnovu definicije diskretnih Markovljevih procesa prvog reda (tj. pretpostavke (20)) sledi:

$$P(\mathbf{X}|\lambda) = P(X_1 = x_1, \ldots, X_T = x_T|\lambda) = P(x_1)\prod_{t=2}^{T} P(x_t|x_{t-1}) = \pi_{x_1} a_{x_1 x_2} \ldots a_{x_{T-1} x_T}$$

gde je sa $P(x_t)$ skraćeno označeno $P(X_t = x_t)$. Primenom marginalizacije promenljivih $\mathbf{X}$ i definicije uslovne verovatnoće dobijamo:

$$\begin{aligned} P(\mathbf{O}|\lambda) &= \sum_{\mathbf{X}} P(\mathbf{O}, \mathbf{X}|\lambda) = \sum_{\mathbf{X}} P(\mathbf{O}|\mathbf{X}, \lambda) P(\mathbf{X}|\lambda) \\ &= \sum_{x_1, \ldots, x_T} \pi_{x_1} b_{x_1}(\mathbf{O}_1) a_{x_1 x_2} b_{x_2}(\mathbf{O}_2) \ldots a_{x_{T-1} x_T} b_{x_T}(\mathbf{O}_T) \end{aligned} \tag{30}$$

gde je poslednja suma po $x_1, \ldots, x_T$ suma po svim mogućim kombinacijama dodele vrednosti $x_1, \ldots, x_T$ slučajnim promenljivama $X_1, \ldots, X_T$.



Slika 9: *Trellis* struktura problema za HMM sa tri stanja.

Pokazuje se da je kompleksnost računanja $P(\mathbf{O}|\lambda)$ primenom ove jednačine $\sim \mathcal{O}(2TN^T)$ tj. raste eksponencijalno sa dužinom sekvence. Sa druge strane, *forward-backward* algoritam, opisan u nastavku, ima složenost $\sim \mathcal{O}(N^2 T)$, tj. linearnu u odnosu na T. *Forward-backward* algoritam ima *trellis* (rešetkastu) strukturu (slika 9) zbog koje je broj računanja znatno manji nego kod direktnog sumiranja iz izraza (30).

Najpre se definiše *forward* promenljiva α:

$$\alpha_t(i) = P(\mathbf{O}_1, \ldots, \mathbf{O}_t, X_t = x_i|\lambda), \quad t = \overline{1, T}, \ i = \overline{1, N} \tag{31}$$

Iako nije neophodna za rešavanje prvog problema, *backward* promenljiva β definiše se analogno *forward* promenljivi i biće korišćena u rešavanje drugog problema. *Backward* promenljiva data je sa:

$$\beta_t(i) = P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T | X_t = x_i, \lambda) \tag{32}$$

Značaj ovih promenljivih ogleda se u lakom induktivnom računanju α_{t+1} i β_{t-1} na osnovu α_t i β_t, redom. Naime, primetimo da važi:

$$\begin{aligned}
\alpha_{t+1}(j) =& P(\mathbf{O}_1, \ldots, \mathbf{O}_{t+1}, X_{t+1} = x_j | \lambda) = \\
=& P(\mathbf{O}_{t+1} | \mathbf{O}_1, \ldots, \mathbf{O}_t, X_{t+1} = x_j, \lambda) P(\mathbf{O}_1, \ldots, \mathbf{O}_t, X_{t+1} = x_j | \lambda) = \\
=& P(\mathbf{O}_{t+1} | X_{t+1} = x_j, \lambda) \sum_{i=1}^{N} P(\mathbf{O}_1, \ldots, \mathbf{O}_t, X_t = x_i, X_{t+1} = x_j | \lambda) = \\
=& P(\mathbf{O}_{t+1} | X_{t+1} = x_j, \lambda) \sum_{i=1}^{N} P(X_{t+1} = x_j | X_t = x_i, \lambda) P(\mathbf{O}_1, \ldots, \mathbf{O}_t, X_t = x_i | \lambda) = \\
=& b_j(\mathbf{O}_{t+1}) \sum_{i=1}^{N} a_{ij} \alpha_t(i)
\end{aligned} \tag{33}$$

Analogno se može pokazati da važi: $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{O}_{t+1}) \beta_{t+1}(j)$. Odredimo sada vrednost *forward* promenljive u početnom koraku indukcije:

$$\alpha_1(i) = P(\mathbf{O}_1, X_1 = x_i | \lambda) = P(X_1 = x_i | \lambda) P(\mathbf{O}_1 | X_1 = x_i, \lambda) = \pi_i b_i(\mathbf{O}_1) \tag{34}$$

Na kraju, primetimo da važi:

$$\sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} P(\mathbf{O}_1, \ldots, \mathbf{O}_T, X_T = x_i | \lambda) = P(\mathbf{O}_1, \ldots, \mathbf{O}_T | \lambda) = P(\mathbf{O} | \lambda) \tag{35}$$

Dakle, rešenje prvog problema, tj. verovatnoća sekvence opservacija za dati model λ data je sumom varijabli $\alpha_T(i)$. Na osnovu svega navedenog, *forward-backward* algoritam sadrži sledeća tri koraka:

1. inicijalizacija:

$$\alpha_1(i) = \pi_i b_i(\mathbf{O}_1) \tag{36a}$$

$$\beta_T(i) = 1 \tag{36b}$$

2. indukcija ($t = 1, \ldots, T-1$):

$$\alpha_{t+1}(j) = b_j(\mathbf{O}_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij} \tag{37a}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{O}_{t+1}) \beta_{t+1}(j) \tag{37b}$$

3. terminacija:

$$P(\mathbf{O} | \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{38}$$

Napominjemo da je za rešavanje prvog problema HMM potrebno koristiti samo jednačine sa *forward* promenljivom α.

2.3.3 Drugi problem HMM

Traženje optimalne sekvence stanja nema jedinstveno rešenje jer se optimalnost rešenja može definisati na više načina. U nastavku razmatramo dva tipa optimalnosti - individualno najverovatnija stanja i najverovatniju sekvencu stanja.

Primetimo najpre da je verovatnoća da se proces sa opservacijama $\mathbf{O}$ nađe u stanju x_i u trenutku t srazmerna proizvodu *forward* i *backward* promenljive:

$$\begin{aligned} P(X_t = x_i, \mathbf{O}|\lambda) &= P(\mathbf{O}_1, \ldots, \mathbf{O}_t, X_t = x_i, \mathbf{O}_{t+1}, \ldots, \mathbf{O}_T|\lambda) = \\ &= P(\mathbf{O}_1, \ldots, \mathbf{O}_t, X_t = x_i|\lambda) P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T|\mathbf{O}_1, \ldots, \mathbf{O}_t, X_t = x_i, \lambda) = \\ &= P(\mathbf{O}_1, \ldots, \mathbf{O}_t, X_t = x_i|\lambda) P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T|X_t = x_i, \lambda) = \alpha_t(i)\beta_t(i) \end{aligned} \tag{39}$$

Potom uvedimo promenljivu γ kojom je opisana verovatnoća da se proces nađe u određenom stanju ukoliko su date opservacije $\mathbf{O}$:

$$\gamma_t(i) = P(X_t = x_i|\mathbf{O}, \lambda) = \frac{P(X_t = x_i, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{40}$$

Problem nalaženja individualno najverovatnijih stanja se sada može zapisati kao:

$$X_t^* = \arg\max_{i=\overline{1,N}} \gamma_t(i) \tag{41}$$

gde je X_t^* optimalno stanje u trenutku t. Iako je svako stanje najverovatnije, ovakav odabir sekvence stanja $[X_1^*, X_2^*, \ldots, X_T^*]$ može biti nepogodan ukoliko su neki od prelaza stanja nemogući (ili malo verovatni). Ilustrujmo to na primeru prepoznavanja glasova. Neka je najverovatniji glas u trenutku t na osnovu opservacija glas š, a u trenutku $t + 1$ glas dž. Iako su ova stanja individualno najverovatnija, verovatnoća da glas dž prati glas š u srpskom jeziku je dosta mala, pa ova sekvenca ne bi bila verovatna. Stoga se češće primenjuje kriterijum optimalnosti čitave sekvence stanja.

Problem nalaženja optimalne sekvence stanja je komplikovaniji, ali kao i u slučaju *forward-backward* algoritma postoji jednostavan algoritam sa induktivnim korakom. Takav algoritam je nazvan po američkom elektroinženjeru Viterbijev algoritam. Cilj ovog algoritma je određivanje

$$\arg\max_{\mathbf{X}} P(\mathbf{X}|\mathbf{O}, \lambda) = \arg\max_{\mathbf{X}} \frac{P(\mathbf{X}, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} = \arg\max_{\mathbf{X}} P(\mathbf{X}, \mathbf{O}|\lambda) \tag{42}$$

gde je iskorišćena činjenica da imenilac $P(\mathbf{O}|\lambda)$ ne zavisi od promenljivih $\mathbf{X}$.
Uvedimo sledeću promenljivu:

$$\delta_t(i) = \max_{x_1, \ldots, x_{t-1}} P(X_1, \ldots X_{t-1}, X_t = x_i, \mathbf{O}_1, \ldots, \mathbf{O}_t|\lambda) \tag{43}$$

i uočimo da važi rekurzija:

$$\delta_{t+1}(j) = \Big[\max_i \delta_t(i) a_{ij}\Big] b_j(\mathbf{O}_{t+1}) \tag{44}$$

Jednačina (44) se dokazuje analogno jednačini (37a) uz zamenu operacije sume operacijom maksimuma. I sam Viterbijev algoritam ima vrlo sličnu implementaciju kao

forward-backward algoritam.

Pre nego što Viterbijev algoritam bude prikazan, uvedimo pokazivače Ψ_t koji će pokazivati na najverovatnije stanje u $t-1$ trenutku ukoliko je dato stanje u t-tom trenutku. Pronalaženjem najverovatnijeg poslednjeg stanja X_T^* i praćenjem odgovarajućih pokazivača dolazi se do optimalne sekvence. Dakle, Viterbijev algoritam za nalaženje optimalne putanje $[X_1^*, X_2^*, \ldots, X_T^*]$ sa maksimalnom verovatnoćom $P^* = \max\ P(\mathbf{X}, \mathbf{O}|\lambda)$ ima sledeće korake:

1. inicijalizacija:

$$\delta_1(i) = \pi_i b_i(\mathbf{O}_1) \tag{45a}$$

$$\Psi_1(i) = 0 \tag{45b}$$

2. indukcija $(t = 1, \ldots, T-1)$:

$$\delta_{t+1}(j) = \Big[\max_i \delta_t(i) a_{ij}\Big] b_j(\mathbf{O}_{t+1}) \tag{46a}$$

$$\Psi_{t+1}(j) = \arg\max_i \delta_t(i) a_{ij} \tag{46b}$$

3. terminacija:

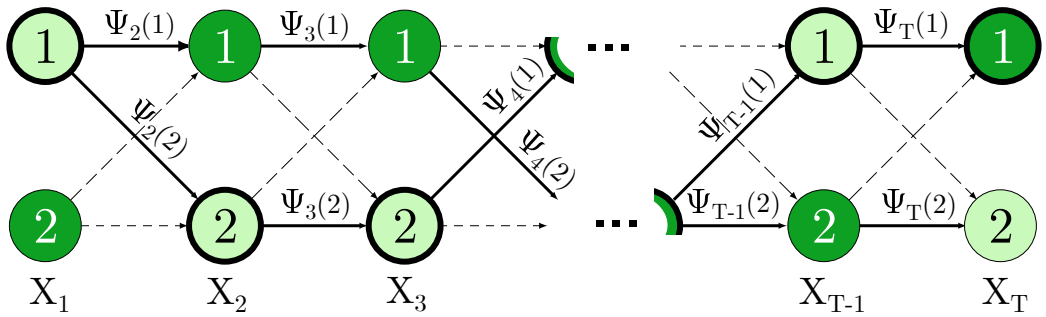
$$P^* = \max_i \delta_T(i) \tag{47a}$$

$$X_T^* = \arg\max_i \delta_T(i) \tag{47b}$$

4. *backtracking* $(t = T-1, \ldots, 1)$:

$$X_t^* = \Psi_{t+1}(X_{t+1}^*) \tag{48}$$

Na slici 10 je prikazana ilustracija Viterbijevog algoritma na modelu sa dva stanja. Podebljane strelice predstavljaju pokazivače na prethodno najbolje stanje, a pode-



Slika 10: Ilustracija Viterbijevog algoritma u modelu sa dva stanja. Optimalna sekvenca jeste: $[1, 2, 2, \ldots, 1, 1]$.

bljana stanja predstavljaju najverovatniju sekvencu stanja. U prva tri koraka algoritam propagira vrednosti δ i Ψ nadesno do poslednjeg stanja X_T. Potom se, na osnovu formiranog niza pokazivača, vraća do početnog stanja. Primetiti da stanja koja su najverovatnije do trenutnog trenutka (tamno zelena boja, imaju najveće δ_t u trenutku t) ne moraju biti u najverovatnijoj putanji (sem poslednjeg stanja u sekvenci koje je uvek verovatnije od svih drugih terminalnih stanja).

2.3.4 Treći problem HMM

Cilj trećeg problema jeste traženje parametara λ tako da funkcija $P(\mathbf{O}|\lambda)$ bude maksimalna. Problem estimacije modela je srž Markovljevih modela, jer se upravo rešavanjem ovog problema dolazi do načina prevođenja obučavajućih podataka u sam model. Traženje maksimuma funkcije može se rešiti kao optimizacioni problem raznim gradijentnim metodama, ali se češće koristi Baum-Velč algoritam, ekvivalentan EM (*Expectation-Maximization*) metodu u statistici.

Ovaj algoritam je nazvan po matematičaru Leonardu Baumu koji je prvi predložio reestimacionu proceduru za Markovljeve lance [29], a potom i reestimacione algoritme za generalniji oblik Markovljevih modela [30, 31]. Po njegovoj metodi za reestimaciju parametara svaki reestimirani model $\overline{\lambda}$ ima veću verovatnoću $P(\mathbf{O}|\overline{\lambda})$, tj. važi: $P(\mathbf{O}|\overline{\lambda}) \geq P(\mathbf{O}|\lambda)$ uvek kada se model $\overline{\lambda}$ reestimira na osnovu modela λ.

Rešavanje problema reestimacije parametara započećemo teoremom koja je primenljiva za HMM, ali važi i za druge oblike modela. Sve verovatnoće u ovom izvođenju podrazumevaju da su opservacije $\mathbf{O}$ date, tj. predstavljaju uslovne verovatnoće za dato $\mathbf{O}$. Neka je su sa λ i $\overline{\lambda}$ označeni modeli i neka se razmatra proces sa stanjima $x \in D_X$, gde D_X nije neophodno diskretni skup.
Definišimo pomoćnu funkcija Q kao:

$$Q(\lambda, \overline{\lambda}) = \int_{x \in D_X} p(x, \lambda) \log p(x, \overline{\lambda}) \, d\mu(x) \tag{49}$$

Teorema 2.2. *[31] Ako je $Q(\lambda, \overline{\lambda}) \geq Q(\lambda, \lambda)$ tada je $P(\overline{\lambda}) \geq P(\lambda)$. Jednakost važi samo ako je $p(x, \lambda) = p(x, \overline{\lambda})$ skoro svuda u odnosu na meru $d\mu(x)$.*

Dokaz. Primetimo najpre da je $\log(x)$ konkavna funkcija na intervalu $(0, \infty)$. Pretpostavimo da važi da je $Q(\lambda, \overline{\lambda}) \geq Q(\lambda, \lambda)$. Dokazaćemo da je tada $\log \frac{P(\overline{\lambda})}{P(\lambda)} \geq 0$ što je ekvivalentno traženom iskazu. Preuređivanjem izraza dobija se:

$$\log \frac{P(\overline{\lambda})}{P(\lambda)} = \log \left(\frac{1}{P(\lambda)} \int_{x \in D_X} p(x, \overline{\lambda}) d\mu(x) \right) = \log \int_{x \in D_X} \frac{p(x, \lambda) d\mu(x)}{P(\lambda)} \cdot \frac{p(x, \overline{\lambda})}{p(x, \lambda)} \tag{50}$$

Kako je po definiciji $\displaystyle\int_{x \in D_X} \frac{p(x, \lambda) d\mu(x)}{P(\lambda)} = 1$, a funkcija $\log(x)$ je konkavna funkcija, primenom Jensenove nejednakosti dobija se:

$$\log \frac{P(\overline{\lambda})}{P(\lambda)} \geq \int_{x \in D_X} \frac{p(x, \lambda) d\mu(x)}{P(\lambda)} \log \frac{P(x, \overline{\lambda})}{P(x, \lambda)} = \frac{1}{P(\lambda)} \left(Q(\lambda, \overline{\lambda}) - Q(\lambda, \lambda) \right) \geq 0 \tag{51}$$

gde poslednja nejednakost sledi na osnovu početne pretpostavke. Jednakost važi kada je $\frac{P(x, \overline{\lambda})}{P(x, \lambda)} = const.$ skoro svuda za meru $\frac{P(x, \lambda) d\mu}{P(\lambda)}$, odnosno za $P(x, \overline{\lambda}) = P(x, \lambda)$ skoro svuda za meru $d\mu$. □

Dakle, ono što ova teorema omogućuje jeste maksimizacija funkcije Q umesto direktne maksimizacije verovatnoće modela $P(\lambda|\mathbf{O})$. Pokazaće se da je maksimizacija ove funkcije dosta jednostavnija, naročito u slučaju skrivenih Markovljevih modela.

Neka je $\lambda = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ trenutni model, $\overline{\lambda} = (\overline{\mathbf{\Pi}}, \overline{\mathbf{A}}, \overline{\mathbf{B}})$ model nakon reestimacije parametara, a D_X konačan skup mogućih stanja. Tada važi:

$$P(\lambda) = P(\mathbf{\Pi}, \mathbf{A}, \mathbf{B}) = \sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) \tag{52}$$

gde su $P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ verovatnoće oblika (30). Pomoćna funkcija Q tada ima sledeći oblik:

$$\begin{aligned} Q(\lambda, \overline{\lambda}) =& Q(\mathbf{\Pi}, \mathbf{A}, \mathbf{B}; \overline{\mathbf{\Pi}}, \overline{\mathbf{A}}, \overline{\mathbf{B}}) \\ =& \sum_{x_1, \ldots, x_T \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) \left(\log(\overline{\pi}_{x_1}) + \sum_{t=1}^{T-1} \log(\overline{a}_{x_t x_{t+1}}) + \sum_{t=1}^{T} \log(\overline{b}_{x_t}(\mathbf{O}_t)) \right) \end{aligned} \tag{53}$$

Iz dobijenog izraza se vidi da se reestimacija parametara $\overline{\mathbf{\Pi}}$, $\overline{\mathbf{A}}$ i $\overline{\mathbf{B}}$ može vršiti nezavisno, što znatno olakšava reestimaciju. Pokažimo kako se dobijaju formule za reestimaciju parametara $\overline{\mathbf{A}}$ maksimizacijom funkcije $Q(\lambda, \overline{\lambda})$. Drugi član iz zagrade u izrazu (53) se može zapisati kao:

$$\sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) \sum_{t=1}^{T-1} \log(\overline{a}_{x_t x_{t+1}}) = \sum_{i=1}^{N} \sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) \sum_{t: x_t = i} \log(\overline{a}_{i x_{t+1}}) \tag{54}$$

Na ovaj način možemo da maksimizujemo navedeni izraz nezavisno po svakom od N sabiraka ($i = \overline{1, N}$). Primetite da u svakom takvom sabirku učestvuju samo parametri oblika $\overline{a}_{ij}$ gde je i fiksirano - određuju se parametri u jednoj vrsti matrice $\overline{\mathbf{A}}$ nezavisno od drugih.

Neka je sa $N_{ij}(x)$ označen broj prelaza iz i-tog u j-to stanje za sve prelaze stanja (iz svakog stanja t u stanje $t+1$). Primetimo da broj prelaza naravno zavisi od sekvence stanja. Posmatranjem svake od $i = \overline{1, N}$ suma iz jednačine (54) dobija se:

$$\begin{aligned} \sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) \sum_{t: x_t = i} \log(\overline{a}_{i x_{t+1}}) =& \sum_{j=1}^{N} \sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) N_{ij}(x) \log(\overline{a}_{ij}) = \\ =& \sum_{j=1}^{N} \log(\overline{a}_{ij}) \left[\sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) N_{ij}(x) \right] \end{aligned} \tag{55}$$

Maksimizacijom svake od $i = \overline{1, N}$ funkcija oblika (55) sa ograničenjima: $\sum_{j=1}^{N} \overline{a}_{ij} = 1$ i $\overline{a}_{ij} \geq 0$ primenom Lagranževog metoda množitelja dobija se:

$$\overline{a}_{ij} = \frac{\sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) N_{ij}(x)}{\sum_{k=1}^{N} \sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) N_{ik}(x)} \tag{56}$$

Nađimo sada jednostavan način da sračunamo gornji izraz na osnovu prethodnih parametara modela λ. Podsetimo se definicije promenljive γ_t (izraz (40)) u kojoj smo $\gamma_t(i)$ predstavili kao verovatnoću da se proces nađe u stanju i u trenutku t za zadate opservacije i model. Analogno tome, uvedimo promenljivu $\xi_t(i, j)$ koja predstavlja

verovatnoću da se proces nađe u i-tom stanju u trenutku t i u narednom trenutku $t+1$ pređe u stanje j:

$$\xi_t(i,j) = P(X_t = x_i, X_{t+1} = x_j | \mathbf{O}, \lambda) = \frac{P(X_t = x_i, X_{t+1} = x_j, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} \tag{57}$$

Pokažimo da je ovaj izraz jednostavno izračunati pomoću *forward* i *backward* varijable:

$$\begin{aligned} P(X_t = x_i, X_{t+1} = x_j, \mathbf{O}|\lambda) = P(\mathbf{O}_1, \dots \mathbf{O}_t, X_t|\lambda) P(X_{t+1}|X_t, \lambda) P(\mathbf{O}_{t+1}|X_{t+1} = x_j, \lambda) \\ \cdot P(\mathbf{O}_{t+2}, \dots, \mathbf{O}_T | X_{t+1} = x_j, \lambda) = \\ = \alpha_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) \beta_{t+1}(j) \end{aligned} \tag{58}$$

U prethodnom izvođenju su iskorišćene uslovne nezavisnosti date u izrazima (20) i (27). Dodatno, verovatnoću $P(\mathbf{O}|\lambda)$ iz imenioca razlomka u izrazu (57) možemo sračunati marginalizacijom izraza (58) po promenljivama X_t i X_{t+1}. Konačno, dobija se izraz:

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) \beta_{t+1}(j)}{\sum_k \sum_l \alpha_t(k) a_{kl} b_l(\mathbf{O}_{t+1}) \beta_{t+1}(l)} \tag{59}$$

Bez strogog dokazivanja možemo da primetimo da promenljiva $\xi_t(i,j)$ i brojilac izraza (56) opisuju verovatnoću prelaska iz i-tog u j-to stanje na osnovu datih opservacija i modela. Kako su sa N_{ij} označeni svi prelazi stanja nezavisno od vremena, potrebno je sabrati doprinose oblika $\xi_t(i,j)$ za svaki trenutak t. Stoga se izraz za reestimaciju parametara $\overline{\mathbf{A}}$ može zapisati kao:

$$\overline{a}_{ij} = \frac{\sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) N_{ij}(x)}{\sum_{k=1}^{N} \sum_{x \in D_X} P(x, \mathbf{\Pi}, \mathbf{A}, \mathbf{B}) N_{ik}(x)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{k=1}^{N} \sum_{t=1}^{T-1} \xi_t(i,k)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{60}$$

Ovim je određena formula za reestimaciju parametara $\overline{\mathbf{A}}$ na osnovu poznatih promenljivih.

Na analogan način moguće je doći do izraza za reestimaciju parametara $\overline{\mathbf{\Pi}}$ i $\overline{\mathbf{B}}$. Dobijaju se sledeći izrazi:

$$\overline{\pi}_i = \gamma_1(i), \quad i = \overline{1, N} \tag{61}$$

$$\overline{b}_j(k) = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, \quad j = \overline{1, N}, \ k = \overline{1, M} \tag{62}$$

$\text{za } \boldsymbol{O}_t = \boldsymbol{v}_k$

Tumačenje sve tri reestimacione formule je očito: parametar $\overline{\pi}(i)$ je jednak verovatnoći da se proces nađe u stanju i u prvom trenutku; parametar $\overline{a}_{ij}$ verovatnoći prelaska iz i-tog u j-to stanje; parametar $\overline{b}_j(k)$ verovatnoći da se proces nađe u stanju j i opservira simbol $\mathbf{v}_k$. Naravno, sve ove verovatnoće se određuju na osnovu tekućeg modela.

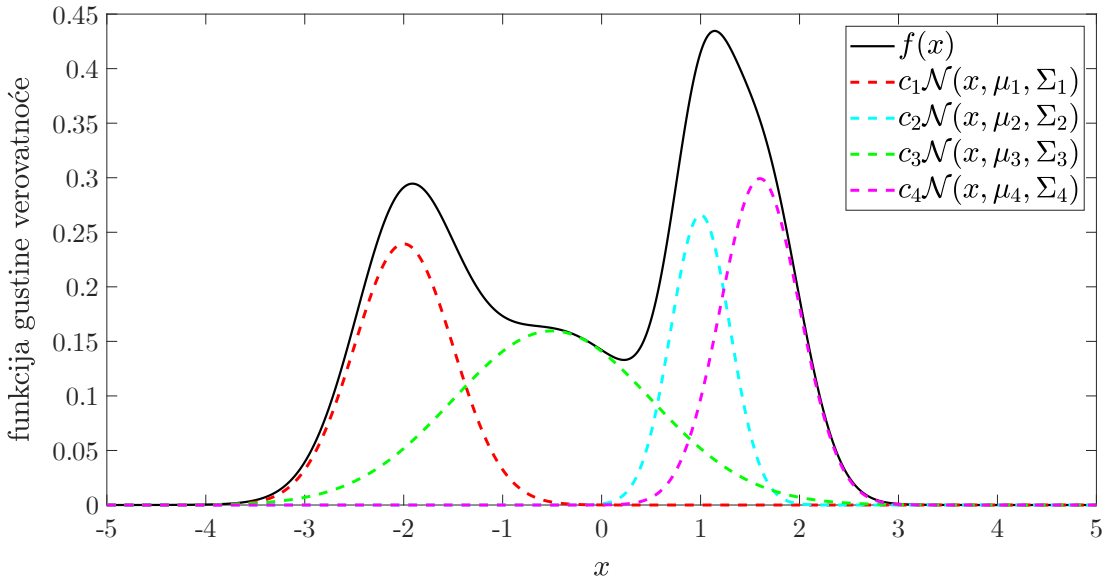
Ovim je data teorija neophodna za formiranje skrivenih Markovljevih modela. U narednom poglavlju biće analizirani posebni aspekti skrivenih Markovljevih modela koji su bitni za modeliranje govora.

2.4 Skriveni Markovljevi modeli za prepoznavanje govora

2.4.1 Kontinualne funkcije gustine verovatnoće opservacija

Analiza skrivenih Markovljevih modela prikazana u prethodnom potpoglavlju uzima u razmatranje samo modele sa diskretnim i konačnim skupom mogućih simbola koje opservacije mogu uzeti. Međutim, kao što je prikazano u poglavlju o izdvajanju obeležja, opservacije u prepoznavanju govora su uglavnom odgovarajući LPC/kepstralni koeficijenti. Stoga je za korišćenje prethodno opisanih modela sa konačnim skupom simbola neophodno ove koeficijente kvantizovati. Iako je ovaj pristup prihvatljiv [13], pristup koji mnogo bolje generalizuje dato znanje zasniva se na parametrizaciji opservacija.

Ukoliko se pretpostavi raspodela za opservacije i na odgovarajući način odrede parametri te raspodele, dobija se pogodan metod za opisivanje opservacija. Metode za primenu takvih raspodela u skrivenim Markovljevim modelima i naročito u njihovom obučavanju postepeno su bile unapređivane. Tako je početkom sedamdesetih bio poznat postupak za reestimaciju parametara log konkavnih raspodela [30,31] (normalna raspodela, Laplasova raspodela,...), da bi potom bio proširen i na eliptički simetrične raspodele [32] (Košijeva raspodela,...) i na kraju na konačne smeše (eng. *mixtures*) log konkavnih ili eliptički simetričnih raspodela [33].



Slika 11: Konačna smeša sa $M = 4$ normalno raspodeljenih komponenti.

Definišimo konačnu smešu M normalnih raspodela kao:

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mathbf{O}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}), \quad j = \overline{1, N} \tag{63}$$

gde su c_{jm} koeficijenti koji određuju doprinos m-te komponente smeše raspodeli za j-to stanje, $\mathcal{N}(\mathbf{O}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$ je normalna raspodela sa matematičkim očekivanjem $\boldsymbol{\mu}_{jm}$ i kovarijacionom matricom $\boldsymbol{\Sigma}_{jm}$. Kako bi raspodela b_j bila normirana, tj. kako bi važilo $\int_{-\infty}^{\infty} b_j(\boldsymbol{x})d\boldsymbol{x} = 1$ za svako stanje $j = \overline{1, N}$, koeficijenti c_{jm} moraju da zadovoljavaju

sledeće uslove:

$$\sum_{m=1}^{M} c_{jm} = 1, \quad j = \overline{1, N} \tag{64a}$$

$$c_{jm} \geq 0, \quad j = \overline{1, N},\ m = \overline{1, M} \tag{64b}$$

Raspodelama oblika (63) moguće je aproksimirati do željene tačnosti svaku konačnu, neprekidnu funkciju gustine verovatnoće [13]. Povećanjem broja M, odnosno elemenata smeše, može se proizvoljno smanjivati greška aproksimacije. Međutim, prevelikim odabirom vrednosti M gubi se moć generalizacije - setimo se da ovom raspodelom želimo da opišemo verovatnoću pojavljivanja LPC/kepstralnih koeficijenata pri izgovaranju reči i to nezavisno od govornika. Na slici 11 je prikazano formiranje složene funkcije gustine verovatnoće pomoću četiri normalne raspodele sa odgovarajućim parametrima.

Na osnovu analiza datih u [32] i [33] i postupka sličnom onom prikazanom u sekciji 2.3.4 dolazi se do sledećih reestimacionih formula za parametre c_{jm}, $\boldsymbol{\mu}_{jm}$ i $\boldsymbol{\Sigma}_{jm}$ [13]:

$$\overline{c}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j, m)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j, k)} \tag{65}$$

$$\overline{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j, m) \mathbf{O}_t}{\sum_{t=1}^{T} \gamma_t(j, m)} \tag{66}$$

$$\overline{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j, m)(\mathbf{O}_t - \boldsymbol{\mu}_{jm})(\mathbf{O}_t - \boldsymbol{\mu}_{jm})^T}{\sum_{t=1}^{T} \gamma_t(j, k)} \tag{67}$$

gde je $\gamma_t(j, m)$ proširenje promenljive $\gamma_t(j)$ koje u obzir uzima doprinos svake komponente $m = \overline{1, M}$ smeše:

$$\gamma_t(j, m) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \cdot \frac{c_{jm}\mathcal{N}(\mathbf{O}_t, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})}{\sum_{k=1}^{M} c_{jk}\mathcal{N}(\mathbf{O}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})} \tag{68}$$

Naglasimo još da je radi dalje reestimacije pogodno promenljive $\overline{c}_{jm}$ ograničiti nekom donjom granicom $\epsilon > 0$ tako da primenjujemo pravilo:

$$\overline{c}_{jm} < \epsilon \implies \overline{c}_{jm} = \epsilon \tag{69}$$

Korišćenje pune kovarijacione matrice $\overline{\boldsymbol{\Sigma}}_{jm}$ (sa svim elementima) može dovesti do numeričke nestabilnosti (prilikom inverzije). Empirijski je pokazano da je korišćenje

punih matrica ekvivalentno korišćenju dijagonalnih matrica sa većim brojem komponenti smeše [25]. Stoga će se u analizi koja sledi koristiti isključivo dijagonalne kovarijacione matrice. Kao i kod parametara $\overline{c}_{jm}$ pogodno je dijagonalne elemente (varijanse) ograničiti nekom konstantom $\epsilon > 0$:

$$\overline{\mathbf{\Sigma}}_{jm}(k,k) < \epsilon \implies \overline{\mathbf{\Sigma}}_{jm}(k,k) = \epsilon, \quad k = \overline{1,M} \tag{70}$$

2.4.2 Skaliranje

Primena reestimacionih formula (60), (61), (65), (66) i (67) nije primenljiva u praktičnim situacijama kada je dužina sekvence opservacija iole veće dužine. Naime, *forward* i *backward* promenljive (α i β) se smanjuju sa dužinom sekvence i vrlo brzo imaju vrednosti koje su na ili ispod granice numeričke preciznosti (dolazi do *undeflow*-a). Stoga je neophodno ove promenljive na adekvatan način skalirati i izmeniti reestimacione formule kako ova promena ne bi uticala na reestimirane parametre.

Pretpostavimo da postoje skalirane promenljive $\widehat{\alpha}$ i $\widehat{\beta}$ takve da se izraz (60) može zapisati kao:

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \widehat{\alpha}_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) \widehat{\beta}_{t+1}(j)}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} \widehat{\alpha}_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) \widehat{\beta}_{t+1}(j)} \tag{71}$$

Primena induktivnog koraka (37a) na skaliranu promenljivu $\widehat{\alpha}$ daje:

$$\alpha'_t(j) = \sum_{i=1}^{N} \widehat{\alpha}_{t-1}(i) a_{ij} b_j(\mathbf{O}_t) \tag{72}$$

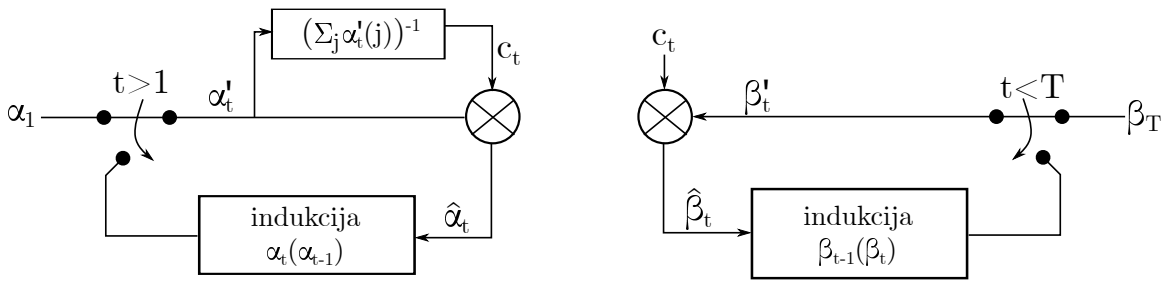
Odredimo zatim konstantu za skaliranje kao:

$$c_t = \left(\sum_{j=1}^{N} \alpha'_t(j) \right)^{-1} \tag{73}$$

i skalirajmo α'_t tako da dobijemo $\widehat{\alpha}_t$:

$$\widehat{\alpha}_t(j) = c_t \alpha'_t(j) = \prod_{\tau=1}^{t} c_\tau \alpha_t(j) = C_t \alpha_t(j) \tag{74}$$

Primetimo da je u slučaju malih vrednosti α'_t konstanta c_t jako velika, pa se množenjem sa tom konstantom promenljiva $\widehat{\alpha}_t$ povećava. Potrebno je postaviti i početni korak $\alpha'_1(j) = \alpha_1(j)$. Ceo opisani algoritam za skaliranje je prikazan na slici 12 (levo).



Slika 12: Blok šema za skaliranje *forward* (levo) i *backward* (desno) varijable.

Kako je red veličine promenljive α približno jednak redu veličine *backward* varijable β, istim konstantama skalira se i $\widehat{\beta}$:

$$\widehat{\beta}_{t+1}(j) = c_{t+1}\beta'_{t+1}(j) = \prod_{\tau=t+1}^{T} c_\tau \beta_{t+1}(j) = D_{t+1}\beta_{t+1}(j) \tag{75}$$

gde su parametri β' dobijeni primenom induktivnog koraka (37b):

$$\beta'_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{O}_{t+1}) \widehat{\beta}_{t+1}(j) \tag{76}$$

a početan korak indukcije je $\beta'_T(i) = 1$. Blok šema za skaliranje *backward* varijable data je na slici 12 (desno).

Zamenom izraza (74) i (75) u izraz (71) dobija se:

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} C_t \alpha_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) D_{t+1} \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} C_t \alpha_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) D_{t+1} \beta_{t+1}(j)} \tag{77}$$

Kako je $C_t D_{t+1} = \prod_{\tau=1}^{T} c_\tau = C_T$ za svako $t = \overline{1, T}$, ovi članovi u izrazu (77) se poništavaju tako da se dobija identičan izraz za parametre $\overline{a}_{ij}$ kao pre skaliranja.

Nakon skaliranja ne važi da je $P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \widehat{\alpha}_T(i)$. Međutim, kako važi:

$$\prod_{t=1}^{T} c_t \sum_{i=1}^{N} \alpha_T(i) = C_T \sum_{i=1}^{N} \alpha_T(i) = 1 \tag{78}$$

sledi da je:

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \frac{1}{\prod_{t=1}^{T} c_t} = \frac{1}{C_T} \tag{79}$$

Kako bi se i ovde izbegao *underflow*, često se računa log verovatnoća kao:

$$\log[P(\mathbf{O}|\lambda)] = -\sum_{t=1}^{T} \log(c_t) \tag{80}$$

Slične postupke skaliranja moguće je (često i neophodno) primeniti i u drugim delovima modela - npr. računanje log vrednosti promenljivih δ_t u Viterbijevom algoritmu. Implementacija nije zahtevna tako da će na ovom mestu biti preskočeno ponavljanje izraza za Viterbijev algoritam (za više detalja videti [13]).

2.4.3 Višestruke sekvence opservacija

Pretpostavimo da želimo da obučimo skriveni Markovljev model tako da prepoznaje odgovarajuću reč. Prema dosadašnjoj analizi potrebno je snimiti signal govora prilikom izgovaranja te reči, obraditi signal na odgovarajući način i formirati vektor

opservacija. Potom se ta sekvenca opservacija koristi za reestimaciju parametara. Međutim, vrlo često je jedna sekvenca opservacija nedovoljna kako bi dobijeni model bio upotrebljiv. Neki od razloga za to jesu veliki broj parametara modela u odnosu na dužinu i sadržaj opservacija ili nemogućnost generalizacije modela na osnovu jedne sekvence opservacija. Stoga se obično za obučavanje modela koriste višestruke sekvence opservacija. Do sada smo koristili notaciju $\mathbf{O}$ da označimo jednu sekvencu opservacija tj. $\mathbf{O} = [\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_T]$. Međutim, nadalje ćemo jednu sekvencu označavati sa $\mathbf{O}^{(k)}$, a višedimenzionalni vektor koji sadrži sve sekvence opservacija kao $\mathbf{O} = \left[\mathbf{O}^{(1)}, \mathbf{O}^{(2)}, \ldots, \mathbf{O}^{(K)}\right]$, gde je svaka sekvenca opservacija $\mathbf{O}^{(k)} = \left[\mathbf{O}_1^{(k)}, \mathbf{O}_2^{(k)}, \ldots, \mathbf{O}_{T_k}^{(k)}\right]$.

Ukoliko pretpostavimo da su sekvence opservacija nezavisne, tada je verovatnoća opservacije svih sekvenci data izrazom:

$$P(\mathbf{O}|\lambda) = \prod_{k=1}^{K} P(\mathbf{O}^{(k)}|\lambda) = \prod_{k=1}^{K} P_k \tag{81}$$

gde su P_k pojedinačne verovatnoće za svaku od sekvenci opservacija. Tada se reestimaciona formula za parametre $\overline{\mathbf{A}}$ može zapisati kao ponderisana suma članova dobijenih iz pojedinačnih sekvenci:

$$\overline{a}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) a_{ij} b_j(\mathbf{O}_{t+1}^{(k)}) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_{t+1}^k(j)} \tag{82}$$

Pokazuje se da je računanje izraza (82) jednostavnije u domenu skaliranih varijabli $\widehat{\alpha}$ i $\widehat{\beta}$. Na osnovu ranije navedenog izraza $C_t^{(k)} D_{t+1}^{(k)} = C_T^{(k)}$, koji važi za svaku od $k = \overline{1, K}$ sekvenci opservacija, dobija se:

$$\overline{a}_{ij} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \widehat{\alpha}_t^k(i) a_{ij} b_j(\mathbf{O}_{t+1}^{(k)}) \widehat{\beta}_{t+1}^k(j)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \widehat{\alpha}_t^k(i) \widehat{\beta}_{t+1}^k(j)} = \frac{\sum_{k=1}^{K} C_T^{(k)} \sum_{t=1}^{T_k-1} \alpha_t^k(i) a_{ij} b_j(\mathbf{O}_{t+1}^{(k)}) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} C_T^{(k)} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_{t+1}^k(j)} \tag{83}$$

Prisetimo se iz izraza (79) da je $C_T^{(k)} = [P(\mathbf{O}^{(k)}|\lambda)]^{-1} = P_k^{-1}$ i primetimo da je izraz (83) jednak izrazu (82).

2.4.4 Inicijalna procena parametara HMM

Obučavanje skrivenih Markovljevih modela bazirano je na reestimacionim procedurama. Stoga je potrebno formirati inicijalni model λ_0 koji će biti korišćen u prvoj reestimaciji. Ukoliko se u modelu koriste kontinualne funkcije gustine verovatnoće za opis opservacija u nekoj stanju, jako je važna dobra inicijalna procena parametara $\boldsymbol{\mu}_{jk}$, $\boldsymbol{\Sigma}_{jk}$ i c_{jk}. Razmotrimo prvo inicijalizaciju ostalih parametara - $\mathbf{A}$ i $\boldsymbol{\Pi}$.

Model koji je implementiran u nastavku je poznat kao *left-right* model. To znači

da je prelazak iz stanja i u bilo koje stanje j gde je $j < i$ zabranjen, odnosno da je $P(X_t = x_i | X_{t-1} = x_j) = 0$, što nas dovodi do zaključka da je $\mathbf{A}$ gornje trougaona matrica. Ovakvim odabirom modela može se naglasiti vremenska priroda signala - kao i vreme i sama stanja mogu da se pomeraju samo unapred.

Reestimacijom parametara koji su u prethodnom modelu imali nultu vrednosti, oni zadržavaju nultu vrednost. Stoga je dovoljno inicijalno postaviti odgovarajuće elemente tranzicione matrice na nultu vrednost. Ostale vrednosti matrice moguće je izabrati proizvoljno. Matricu apriornih verovatnoća ćemo izabrati tako da ona bude konstantna tokom čitavog obučavanja modela, kako bismo osigurali da je prvo stanje uvek početno stanje opservacije. Jedan od mogućih izbora matrice $\mathbf{A}$ i jedini preporučeni izbor matrice $\mathbf{\Pi}$ je:

$$\mathbf{A} = \begin{bmatrix} 0.5 & 0.5 & 0 & \dots & 0 & 0 \\ 0 & 0.5 & 0.5 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0.5 & 0.5 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}, \qquad \mathbf{\Pi} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{84}$$

Iako nasumična inicijalizacija parametara $\boldsymbol{\mu}_{jk}$, $\mathbf{\Sigma}_{jk}$ i c_{jk} ne daje uvek dobre rezultate, ona je prvi korak inicijalizacije našeg modela. Nakon ove proizvoljne inicijalizacije parametara vrši se klasterizacija opservacija *K-means* algoritmom.

Proces klasterizacije obavlja se zasebno za svako od $n = \overline{1, N}$ stanja. Potrebno je pre procesa klasterizacije dodeliti svakoj opservaciji stanje kojem pripada. Ovo se može uraditi nasumično, međutim kako očekujemo da prvih nekoliko opservacija odgovara prvom stanju, potom sledećih par drugom itd. naša inicijalna pretpostavka jeste da ukoliko postoji N stanja, opservacije $\mathbf{O}_1^{(k)}$ do $\mathbf{O}_{T_k/N}^{(k)}$ odgovaraju prvom stanju, $\mathbf{O}_{T_k/N+1}^{(k)}$ do $\mathbf{O}_{2T_k/N}^{(k)}$ drugom itd. (postupak ponoviti za svaku sekvencu opservacija).

Fiksirajmo stanje n i neka su sa $\mathbf{O}_1, ..., \mathbf{O}_L$ označeni svi vektori opservacija (iz svih sekvenci) koji pripadaju n-tom stanju. Opservacije koje odgovaraju stanju n dele se nasumično u jednu od M grupa (klastera), a potom se za svaki od klastera računa srednja vrednost opservacija koje pripadaju tom klasteru: $\boldsymbol{\mu}_{nm}$ za $m = \overline{1, M}$. Sledeći postupak se iterativno ponavlja ($t = \overline{1, T}$):

$$\text{ind}_t = \arg\min_m ||\mathbf{O}_t - \boldsymbol{\mu}_{nm}||_2^2 \tag{85}$$

odnosno, pronalazi se klaster čija je srednja vrednost najbliža trenutnoj opservaciji i u narednoj iteraciji preraspoređujemo tu opservaciju u taj klaster. Kada se odrede svi indeksi za $t = \overline{1, T}$ računaju se nove srednje vrednosti klastera kao:

$$\boldsymbol{\mu}_{nm} = \frac{1}{N_m} \sum_{t:\text{ind}_t = m} \mathbf{O}_t \tag{86}$$

gde je $N_m = \sum_{t:\text{ind}_t=m}$ broj opservacija koje pripadaju m-tom klasteru. Algoritam se iterativno ponavlja sve dok ne dođe do konvergencije, tj. dokle god postoje opservacije koje su preraspodeljene. Na osnovu finalne raspodele određuju se sledeći parametri:

$$\mathbf{\Sigma}_{nm} = \frac{1}{N_m} \sum_{t:\text{ind}_t = m} (\mathbf{O}_t - \boldsymbol{\mu}_{nm})(\mathbf{O}_t - \boldsymbol{\mu}_{nm})^T \tag{87}$$

$$c_{nm} = \frac{N_m}{\sum_{m=1}^{M} N_m} \tag{88}$$

Naglasimo još jednom da se postupak ponavlja nezavisno za svako od $n = \overline{1, N}$ stanja. Podsetimo da ovakvi inicijalni parametri $\mathbf{\Sigma}$ i c treba da ispunjavaju uslove (69) i (70), kao i uslov da je matrica $\mathbf{\Sigma}$ dijagonalna matrica. Stoga je neophodno odgovarajuće parametre postaviti na vrednost ϵ ili nulu na kraju inicijalizacije.

Nažalost, kako su opservacije uvek višedimenzionalne (obično imaju i više od 10 dimenzija) direktna vizuelizacija klasterizacije nije moguća. Međutim, pokazuje se u praksi da je ovaj jednostavan tip klasterizacije sasvim prigodan za inicijalizaciju skrivenih Markovljevih modela.

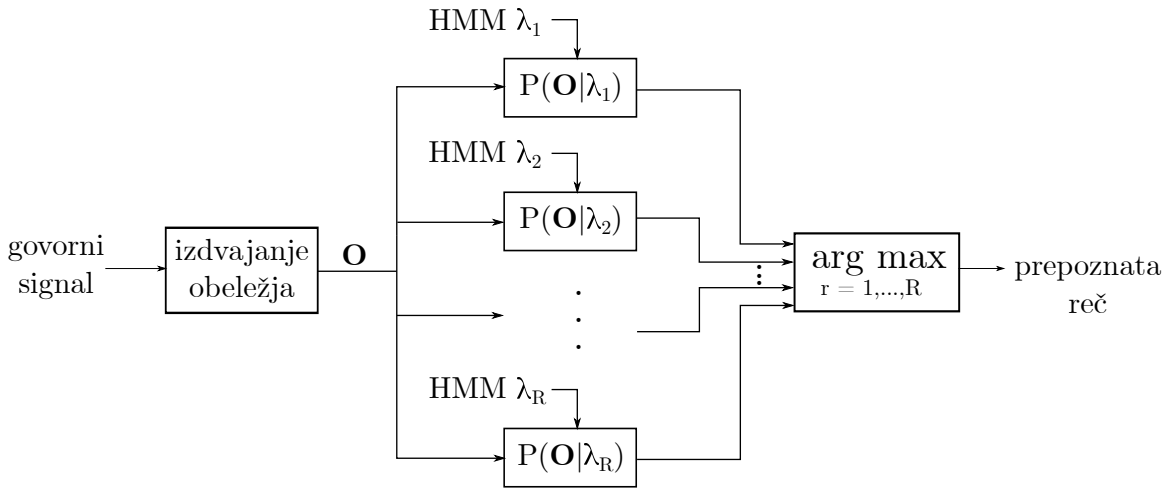
2.4.5 Blok šema sistema za prepoznavanje reči

Reestimacijom modela dobijaju se modeli koji imaju veću vrednost $P(\mathbf{O}|\lambda)$. Međutim, u praksi je moguće reestimirati model konačan, i po mogućstvu što manji, broj puta. Stoga je neophodno uvesti određenu metriku za određivanje konvergencije ovog algoritma. Kako je $P(\mathbf{O}|\lambda)$ vrednost koju treba maksimizovati, obično se rastojanje između dva modela za opservaciju $\mathbf{O}$ dužine T usvaja u obliku [34]:

$$D(\lambda, \overline{\lambda}) = \frac{1}{T}\Big(\log P(\mathbf{O}|\overline{\lambda}) - \log P(\mathbf{O}|\lambda)\Big) \tag{89}$$

U slučaju višestrukih opservacija potrebno je izraze oblika $D^{(k)}(\lambda, \overline{\lambda})$ sabrati kako bi se dobila konačna procena. U eksperimentima koji su predstavljeni u nastavku je podešena vrednost praga D_0 tako da algoritam završava reestimaciju nakon par koraka. Pokazuje se da već nakon par koraka reestimacije vrednost $D(\lambda, \overline{\lambda})$ neznatno mala, te da je model konvergirao.

Na slici 13 je prikazana blok šema sistema za prepoznavanje izolovanih reči iz rečnika dužine R. Nakon obučavanja zasebnog modela λ_r za svaku od $r = \overline{1, R}$ reči, za formirani vektor opservacija se računaju verovatnoće oblika $P(\mathbf{O}|\lambda_r)$. Argument r za koji je data verovatnoća najveća odgovara rednom broju prepoznate reči u rečniku. Na taj način može se prepoznati bilo koja od R reči iz rečnika.



Slika 13: Blok šema sistema za prepoznavanje izolovanih reči iz rečnika dužine R.

3 Rezultati i diskusija

Opisani skriveni Markovljevi modeli su upotrebljeni u prepoznavanju izolovanih reči iz rečnika dužine $R = 10$. Rečnik čine cifre u srpskom jeziku: "nula", "jedan", "dva", "tri", "četiri", "pet", "šest", "sedam", "osam" i "devet". Radi lakšeg prikazivanja rezultata u tabelama ove reči će biti predstavljene ciframa. Audio sekvence su snimljene tako da su reči izolovane - svaka sekvenca sadrži isključivo jednu reč.

U radu je učestvovalo 15 ispitanika. Svaki od ispitanika je snimljen mikrofonom kako izgovara svaku od reči po nekoliko puta. Ispitanici čine heterogenu grupu - učestovalo je 9 muških i 6 ženskih ispitanika, starosti od 13 do 50 godina. Za snimanje sekvenci nisu korišćeni isti tehnički uslovi (mikrofon) i sekvence su snimane u različitim okruženjima.

U prvom setu eksperimenata (podsekcija 3.1) je određivan optimalan izbor parametara za skrivene Markovljeve modele. U drugom delu (podsekcija 3.2) su ovako optimalno određeni modeli primenjeni za prepoznavanje cifara nezavisno od govornika.

3.1 Izbor parametara modela

Za eksperimente u ovoj podsekciji izvršena je podela podataka na sledeći način. Odabrano je po 7 ponavljanja svake od reči za svakog od ispitanika i formiran skup za obučavanja od 105 sekvenci za svaku reč. Potom su formirani skupovi za testiranje:

- skup za testiranje 1: sastoji se od ostalih sekvenci koje nisu ušle u izbor sekvenci za obučavanje, a potiču iz iste sesije snimanja,
- skup za testiranje 2: sekvence snimljene sa izraženo varijabilnim izgovaranjem reči od strane jednog od ispitanika; sekvence iz ove sesije ne učestvuju u setu za obučavanje.

Kako različiti ispitanici imaju različit broj sekvenci koje učestvuju u skupu za testiranje 1, vršeno je ponderisanje tačnosti za svaku reč na sledeći način. Neka ispitanik k ima S_k sekvenci koje učestvuju u skupu za testiranje 1, i neka je T_k broj ispravno prepoznatih reči za tog ispitanika. Tada se ukupna tačnost T računa kao:

$$T = 100\% \cdot \frac{1}{15} \sum_{k=1}^{15} \frac{T_k}{S_k}$$

Na ovaj način svi ispitanici imaju jednak doprinos ukupnoj tačnosti prepoznatih reči.

3.1.1 Broj stanja modela

U prvom eksperimentu je analiziran uticaj reda modela, tj. broja stanja u modelu N. Formirano je po deset modela koji odgovaraju svakoj od cifara za svaki odabir N. Testirane su vrednosti $N \in \{2, 3, 4, 5, 6, 7, 8\}$. Dodatno, testiran je slučaj kada svakoj reči odgovara model sa brojem stanja jednakim broju glasova u toj reči (označeno sa "N = # glasova").

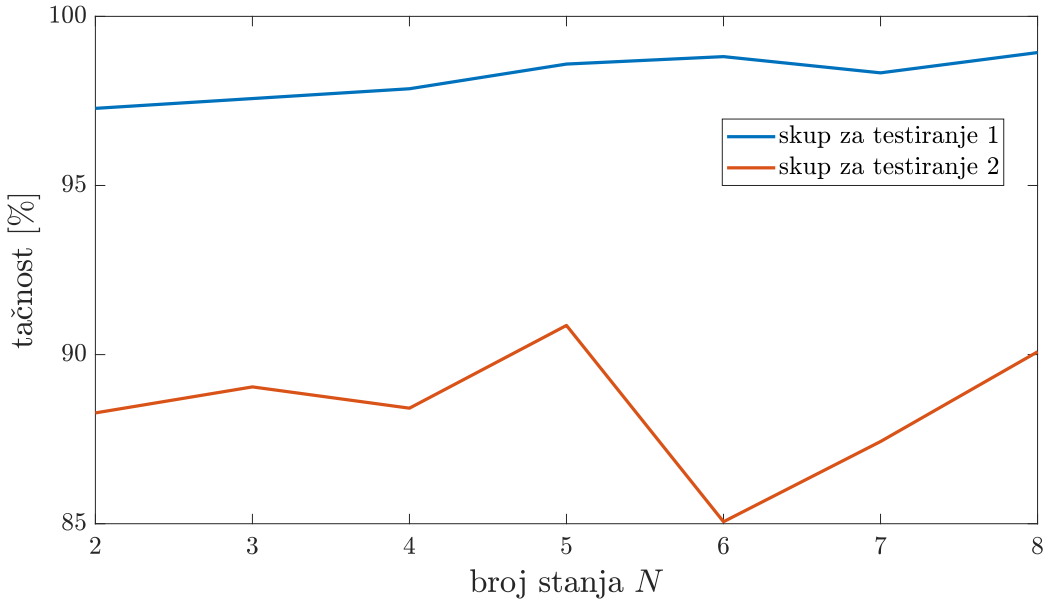
Korišćeno je 12 MFCC koeficijenata kao opservacije i $M = 6$ komponenti u Gausovoj smeši. U tabeli 1 prikazani su dobijeni rezultati. U gornjoj tabeli su rezultati

dobijeni testiranjem za skupu za testiranje 1, a u donjoj tabeli na skupu za testiranje 2. U poslednjoj koloni sa desne strane navedene su vrednosti usrednjene po svim ciframa za dati broj stanja.

N \ reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
2	99.26	95.00	97.67	96.50	99.39	91.22	95.97	98.71	100.00	99.05	97.28
3	100.00	95.67	96.71	97.33	100.00	91.89	97.44	98.57	100.00	98.06	97.57
4	99.26	97.33	96.46	97.00	100.00	94.11	97.44	98.38	100.00	98.57	97.86
5	100.00	97.33	100.00	97.33	100.00	95.22	99.49	97.94	100.00	98.57	98.59
6	100.00	98.00	100.00	97.33	100.00	95.22	98.97	100.00	100.00	98.57	98.81
7	100.00	98.00	97.96	97.00	100.00	95.33	98.97	97.43	100.00	98.57	98.33
8	100.00	98.00	100.00	97.00	100.00	96.56	99.49	99.67	100.00	98.57	98.93
# glasova	99.26	96.33	96.94	97.83	100.00	91.75	99.49	99.05	100.00	98.57	97.92

N \ reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
2	100.00	100.00	100.00	100.00	75.00	61.54	54.55	100.00	100.00	91.67	88.28
3	100.00	100.00	100.00	100.00	75.00	69.23	63.64	90.91	100.00	91.67	89.05
4	100.00	100.00	100.00	100.00	75.00	53.85	63.64	100.00	100.00	91.67	88.42
5	100.00	100.00	100.00	100.00	66.67	69.23	81.82	90.91	100.00	100.00	90.86
6	100.00	100.00	100.00	76.92	75.00	61.54	54.55	90.91	100.00	91.67	85.06
7	100.00	100.00	100.00	92.31	75.00	61.54	45.45	100.00	100.00	100.00	87.43
8	100.00	100.00	100.00	92.31	91.67	61.54	63.64	100.00	100.00	91.67	90.08
# glasova	100.00	100.00	92.31	100.00	83.33	76.92	54.55	90.91	100.00	100.00	89.80

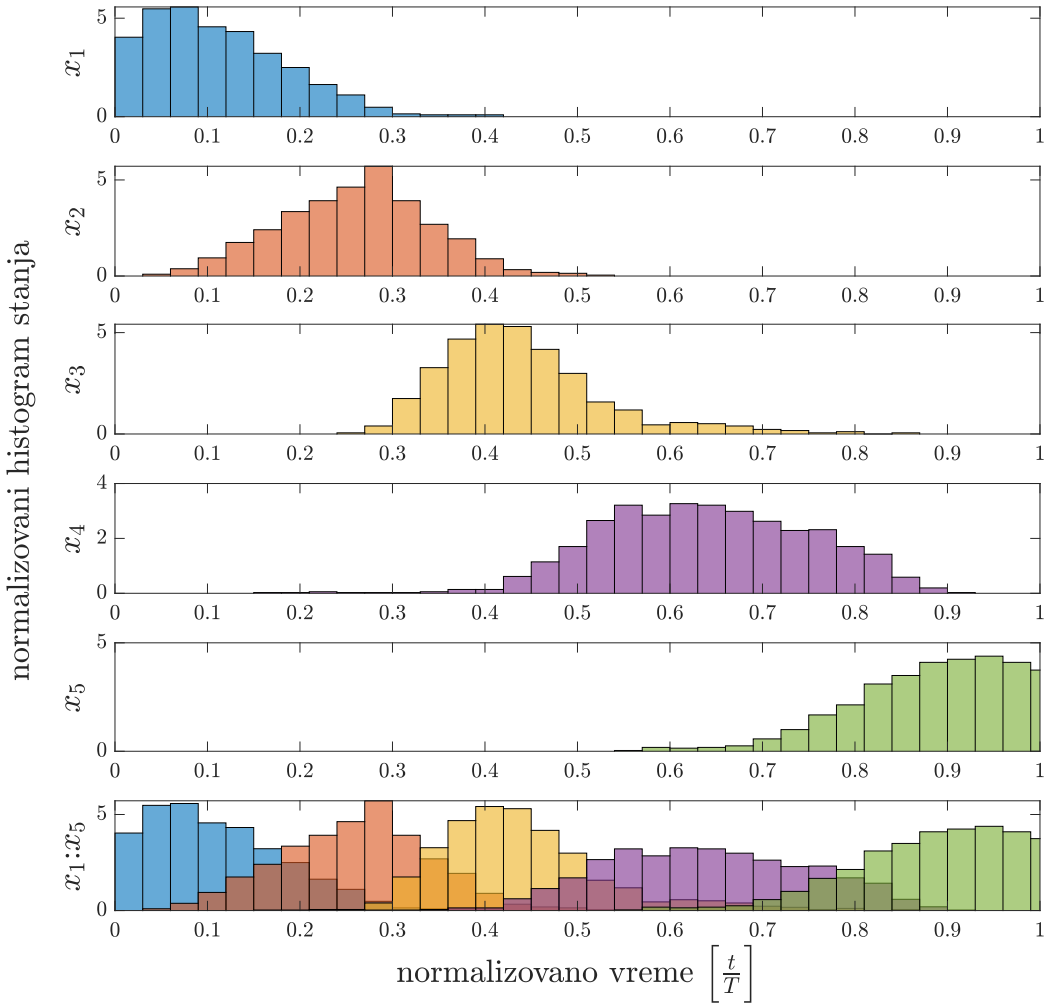
Tabela 1: Srednja tačnost [%] prepoznatih cifri (0-9) na skupu za testiranje 1 (gore) i skupu za testiranje 2 (dole) u zavisnosti od broja stanja u modelu N.



Slika 14: Zavisnost tačnosti [%] od broja stanja N u modelu.

Na slici 14 je prikazana zavisnost srednje vrednosti tačnosti od broja stanja u modelu za oba testirajuća skupa. Kao što je i očekivano, dobijena je bolja tačnost u prepoznavanju reči iz sesija snimanja koja su deo obučavajućeg skupa. Međutim, obe tabele pokazuju visoku tačnost za sve reči. U gornjoj tabeli primećujemo da se sa povećanjem broja stanja, tačnost konstantno povećava. Ovo može biti posledica sličnosti obučavajućeg i testirajućeg skupa 1, jer se sa većim brojem stanja pruža više mogućnosti za modeliranje sekvenci iz obučavajućeg skupa. Samim tim, sekvenca iz testirajućeg skupa 1 će biti lakše prepoznata.

Kada se ova analiza primeni na testirajućem skupu 2, vidi se da postoji pad nakon $N = 5$. Pošto su ove sekvence manje korelisane sa obučavajućim skupom, nije pogodno da model ima veliki broj stanja jer se tom prilikom gubi na generalnosti modela. Stoga će u nastavku biti usvojena vrednost $N = 5$.



Slika 15: Normalizovani histogram stanja u zavisnosti od normalizovanog vremena. Prva pet podgrafika odgovaraju histogramima prvog do petog stanja; šesti podgrafik predstavlja uporedni prikaz sva pet stanja. Reč je "četiri", $N = 5$, $M = 6$, opservacije čine 12 MFCC koeficijenata.

Na grafiku 15 su prikazani histogrami stanja u zavisnosti od normalizovanog vremena. Histogram je formiran na osnovu opservacija za reč "četiri". Za svaku od sekvenci je Viterbijevim algoritmom određena optimalna sekvenca stanja i nakon normalizacije dužine sekvence, stanja su prikazana u vidu histograma. Vidimo da sva stanja traju

približno jednako dugo i možemo približno da procenimo u kojem trenutku je dato stanje najverovatnije. Moguće je dodatno ispitivati u kojoj meri se pozicije ovako estimiranih stanja poklapaju sa položajem glasova u reči.

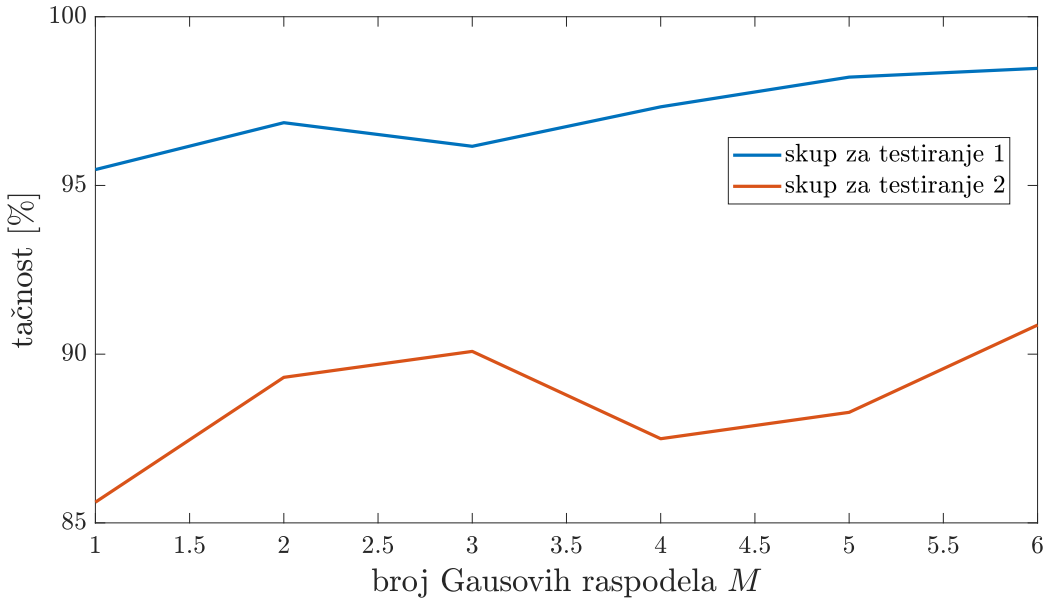
3.1.2 Broj Gausovih komponenti za opisivanje opservacija

U narednom eksperimentu analiziran je uticaj parametra M na tačnost modela. U tabeli 2 prikazani su dobijeni rezultati za set za testiranje 1 i 2, a na grafiku 16 je predstavljena zavisnost srednje vrednosti tačnosti od parametra M.

M \ reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
1	98.52	93.43	98.33	97.33	89.19	95.32	94.95	95.83	99.52	92.29	95.47
2	100.00	95.43	98.50	96.67	99.39	91.98	94.87	96.48	100.00	95.24	96.86
3	100.00	96.10	97.41	97.33	93.33	92.75	96.92	92.48	100.00	95.24	96.16
4	99.26	96.33	98.21	97.33	100.00	93.41	98.46	95.03	100.00	95.24	97.33
5	100.00	97.00	97.29	97.00	100.00	95.22	98.46	99.05	100.00	98.06	98.21
6	100.00	98.00	99.17	97.33	100.00	94.56	98.46	98.56	100.00	98.57	98.47

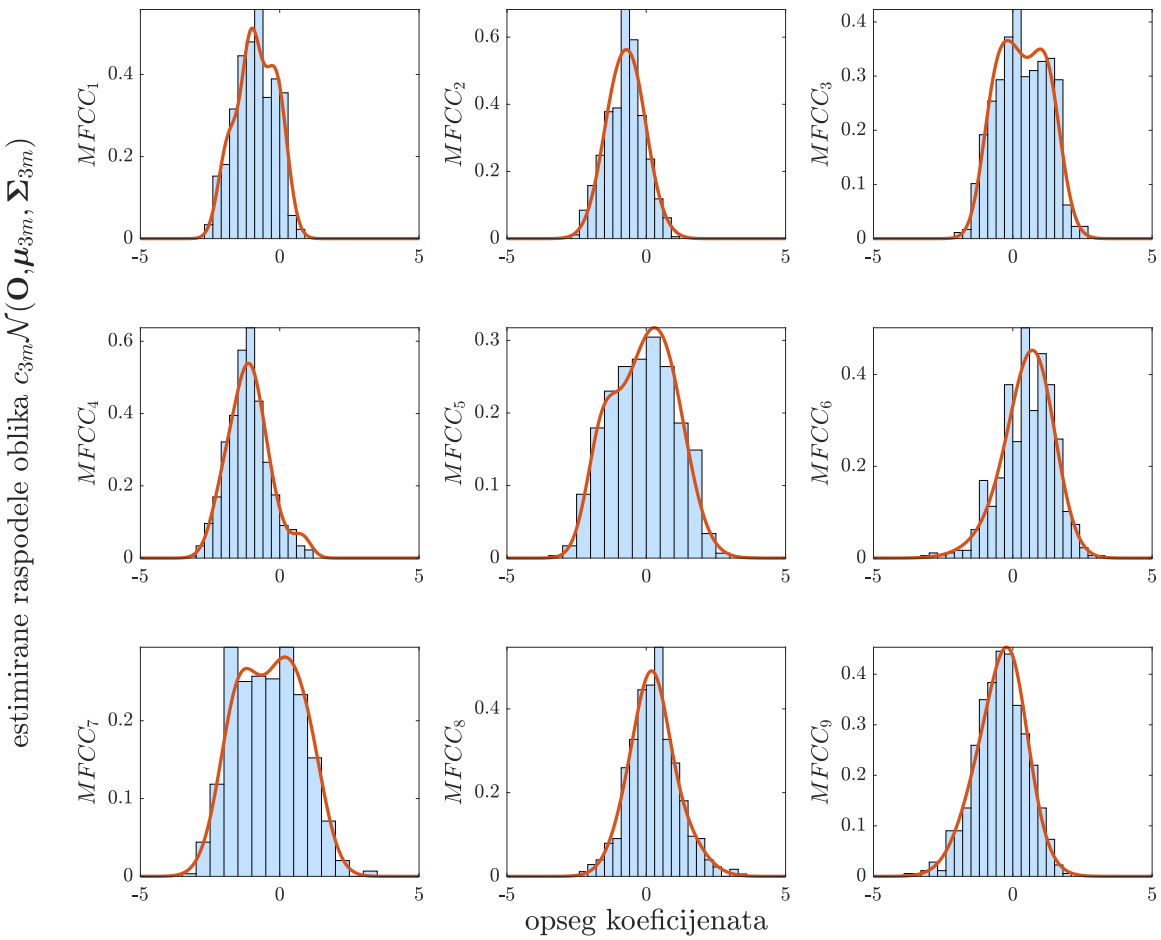
M \ reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
1	100.00	100.00	92.31	92.31	75.00	69.23	54.55	72.73	100.00	100.00	85.61
2	100.00	100.00	92.31	100.00	83.33	53.85	63.64	100.00	100.00	100.00	89.31
3	100.00	100.00	100.00	100.00	83.33	53.85	63.64	100.00	100.00	100.00	90.08
4	100.00	100.00	100.00	84.62	83.33	61.54	45.45	100.00	100.00	100.00	87.49
5	100.00	100.00	100.00	84.62	66.67	76.92	54.55	100.00	100.00	100.00	88.28
6	100.00	100.00	100.00	100.00	66.67	69.23	81.82	90.91	100.00	100.00	90.86

Tabela 2: Srednja tačnost [%] prepoznatih cifri (0-9) na skupu za testiranje 1 (gore) i 2 (dole) u zavisnosti od broja M Gausovih komponenti za opisivanje opservacija.



Slika 16: Zavisnost tačnosti [%] od broja smeša u modelu M.

Iz tabela primećujemo da odabir parametra M nema preveliki uticaj na tačnost. Za slučaj $M = 1$ dostiže se tačnost od preko 95% za skup 1 i preko 85% za skup 2. To znači da je verovatnoća opservacije $b_j(\mathbf{O})$ dobro aproksimirana jednom normalnom raspodelom. Na slici 17 su prikazani histogrami za prvih 9 od 12 MFCC koeficijenata korišćenih kao opservacije. Vidimo da većina koeficijenata ima približno normalnu raspodelu, ali da postoje i neki koje je potrebno predstaviti sa više normalnih komponenti.



Slika 17: Histogrami dvanaestodimenzionalnih opservacija (12 MFCC koeficijenata) za reč "četiri¨ - prikazane raspodele za prva devet koeficijenata u stanju $n = 3$ i za Gausovu smešu od $M = 6$ elemenata. Crvenom linijom je prikazana estimirana funkcija gustine verovatnoće.

Treba napomenuti da sa povećanjem parametara, kako M tako i N, raste vreme neophodno za obučavanje modela. Stoga je preporučivo postaviti ove parametre na male, ali dovoljne vrednosti. Mi ćemo se ipak ovde odlučiti za vrednost $M = 6$ koja daje maksimalnu srednju vrednost tačnosti na drugom skupu za testiranje.

3.1.3 Izbor koeficijenata za opservacije

U ovom eksperimentu je analiziran izbor koeficijenata kojima će biti reprezentovane opservacije. Ispitivane su tri vrste koeficijenata: LPC, kepstralni i MFCC. U sva tri slučaja odabrano je po 12 koeficijenata da predstavljaju opservacije. Dobijeni rezultati su prikazani u tabeli 3.

U skladu sa dosadašnjim istraživanjima, korišćenjem LPC je ostvarena najmanja tačnost, a znatno veća kepstralnim i MFCC koeficijentima. Razlika između ove dve grupe koeficijenata je nešto manja od 1%, što upućuje da je moguće koristiti oba tipa koeficijenata u prepoznavanju ovih reči. Odlučujemo se za MFCC koeficijente koji su ostvarili malo bolju tačnost.

reč / koeficijenti	0	1	2	3	4	5	6	7	8	9	srednja vrednost
LPC	76.86	86.76	71.36	88.23	90.37	78.69	95.63	67.75	91.11	87.64	83.44
kepstralni	100.00	97.33	98.56	97.33	98.28	93.78	99.33	91.70	99.52	99.05	97.49
MFCC	100.00	98.00	99.17	97.33	100.00	94.56	98.46	98.56	100.00	98.57	98.47

Tabela 3: Srednja tačnost [%] prepoznatih cifri (0-9) na skupu za testiranje 1 u zavisnosti od primenjenih koeficijenata za reprezentaciju opservacija.

Potom je analiziran broj MFCC koeficijenata koji čine opservacioni vektor. Obučeni su modeli za vrednost parametra $q \in \{4, 6, 8, 10, 12, 14, 16\}$. Rezultati su prikazani u tabeli 4 i grafiku 18.

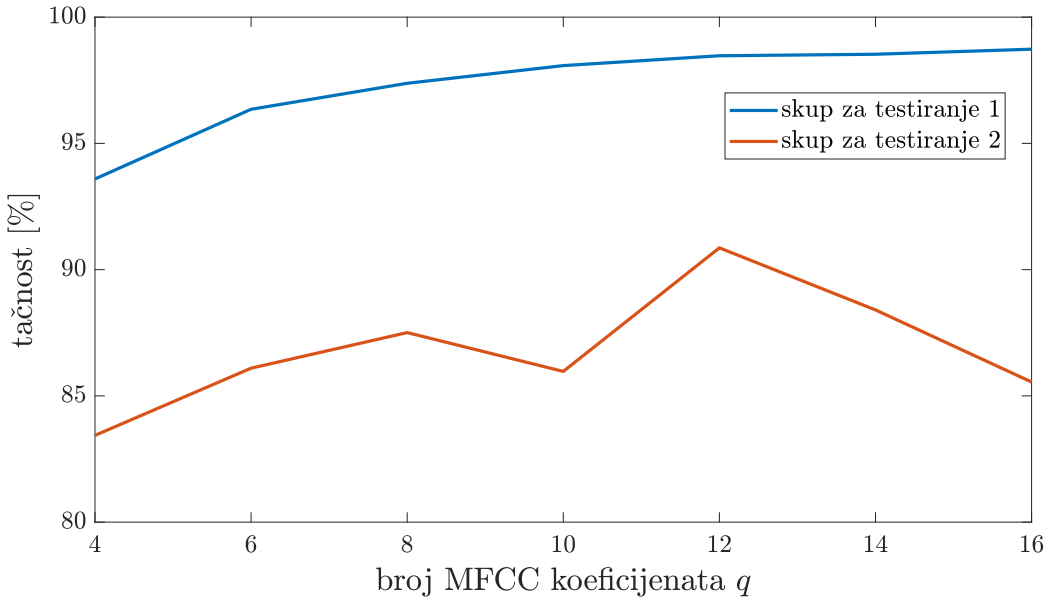
q_{MFCC} / reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
4	88.86	98.06	92.51	92.09	100.00	91.51	90.56	89.11	99.17	93.98	93.59
6	98.67	98.67	97.03	98.33	98.28	90.51	96.12	90.87	100.00	94.97	96.35
8	98.06	95.59	99.05	97.59	99.39	93.52	98.46	94.37	99.52	98.24	97.38
10	100.00	97.33	98.38	97.33	99.39	94.75	98.46	97.10	100.00	98.06	98.08
12	100.00	98.00	99.17	97.33	100.00	94.56	98.46	98.56	100.00	98.57	98.47
14	100.00	98.00	99.17	98.67	100.00	95.22	98.97	100.00	100.00	95.24	98.53
16	100.00	98.67	98.91	97.33	100.00	95.22	98.97	99.67	100.00	98.57	98.73

q_{MFCC} / reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
4	100.00	92.86	76.92	84.62	75.00	76.92	63.64	72.73	100.00	91.67	83.44
6	100.00	100.00	76.92	84.62	91.67	61.54	72.73	81.82	100.00	91.67	86.10
8	100.00	100.00	92.31	100.00	75.00	61.54	54.55	100.00	100.00	91.67	87.51
10	100.00	100.00	100.00	84.62	75.00	53.85	54.55	100.00	100.00	91.67	85.97
12	100.00	100.00	100.00	100.00	66.67	69.23	81.82	90.91	100.00	100.00	90.86
14	100.00	100.00	100.00	84.62	83.33	61.54	54.55	100.00	100.00	100.00	88.40
16	100.00	100.00	100.00	100.00	66.67	61.54	54.55	72.73	100.00	100.00	85.55

Tabela 4: Srednja tačnost [%] prepoznatih cifri (0-9) na skupu za testiranje 1 (gore) i skupu za testiranje 2 (dole) u zavisnosti od broja MFCC koeficijenata q koji se koriste kao vektor opservacija.

Kao i kod izbora reda modela, mogu se izvući dva generalna zaključka. Prvi je da se povećanjem parametara q povećava tačnost ostvarena na prvom setu podataka. Na neki način dolazi do preobučavanja modela. Drugi zaključak jeste da na drugom skupu podataka, tačnost počinje da pada nakon vrednosti $q = 12$. Kako je rečnik u ovim eksperimentima jako mali (svega deset reči), nije neophodan veliki broj koeficijenata

za dobro razlikovanje reči. Takođe, povećanje broja koeficijenata može dovesti do lošijeg obučavanja modela (npr. usled nedovoljne količine podataka za obučavanje). Dakle, opredeljujemo se za vrednost $q = 12$ kao optimalnu vrednost.



Slika 18: Zavisnost tačnosti [%] od broja MFCC koeficijenata q za predstavljanje opservacija.

Poslednji eksperiment iz ove grupe eksperimenata analizirao je uticaj dodavanja delta MFCC koeficijenata opservacionom vektoru. Pored 12 prvih MFCC koeficijenata, dodato je 12 delta koeficijenata sa parametrom $K = 2$. Rezultati su prikazani u tabeli 5.

reč / koeficijenti	0	1	2	3	4	5	6	7	8	9	srednja vrednost
MFCC + ΔMFCC	100.00	98.67	99.74	96.67	100.00	95.89	100.00	98.52	100.00	99.05	98.85

Tabela 5: Srednja tačnost [%] prepoznatih cifri (0-9) na skupu za testiranje 1 nakon dodavanja delta MFCC koeficijenata u opservacione vektore.

Ovakav izbor parametara svakako dovodi do povećanja tačnosti: 98.85% u odnosu na prethodnih 98.47%. Međutim, formiranje 24-dimenzionalnog opservacionog vektora dovodi do dugog obučavanja modela. Takođe, nije jasno da li se ovo povećanje tačnosti zasniva na primeni delta koeficijenata ili primeni većeg broja koeficijenata. Kako je u eksperimentu u sekciji 3.2 neophodno obučavanje velikog broja modela, dodavanje delta koeficijenata nije praktično.

3.1.4 Veličina obučavajućeg skupa podataka

Sasvim je očekivano da sa povećanjem dostupnih sekvenci za obučavanje raste i tačnost modela. Međutim, potrebno je odrediti koji je minimalan broj sekvenci neophodan za dostizanje prihvatljive tačnosti. Stoga su u ovom eksperimentu testirani modeli koji su obučavani na $N_{\text{sekvenci}} \in \{15, 30, 45, 60\}$ sekvenci. Rezultati su dati u tabeli 6.

Pored očekivanog trenda porasta tačnosti sa povećanjem obučavajućeg skupa, primećujemo da je tačnost kod modela obučenih sa samo 15 sekvenci skoro 95%. Naravno, ne treba zaboraviti da je set za testiranje 1 dosta korelisan sa obučavajućim skupom. Ali, jasno je da je za prepoznavanje reči iz ovakvog rečnika dovoljna po jedna sekvenca za svakog od ispitanika kako bi model bio dobro obučen. Kako je potrebno da obučeni model bude što generalniji, u eksperimentu u sekciji 3.2 biće korišćeno svih 105 sekvenci iz skupa za obučavanje.

$N_{sekvenci}$ \ reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
15	100.00	93.10	98.33	85.64	98.89	90.56	91.72	95.81	99.52	95.24	94.88
30	100.00	94.71	97.67	94.85	100.00	94.47	98.46	96.27	99.52	98.06	97.40
45	100.00	94.43	96.58	95.83	97.78	90.36	98.97	94.37	99.52	95.24	96.31
60	100.00	95.67	98.08	96.17	100.00	94.08	97.62	98.24	99.52	98.06	97.74

Tabela 6: Srednja tačnost [%] prepoznatih cifri (0-9) na skupu za testiranje 1 u zavisnosti od broja sekvenci za treniranje modela.

3.1.5 Inicijalizacija matrice prelaza

Izrazom (84) dat je jedan od mogućih načina inicijalizacije matrice prelaza $\mathbf{A}$. Međutim, ovakvom inicijalizacijom sve vrednosti matrice koje imaju nultu vrednost ostaće jednake nuli i nakon obučavanja. Samim tim, svi direktni prelazi iz i-tog u $i+k$-to stanje, gde je $k \geq 2$ su nemogući. Predložen je alternativni način inicijalizacije, gde se nulte vrednosti iznad glavne dijagonale matrice $\mathbf{A}$ postavljaju na neke male nenulte vrednosti. U tabeli 7 prikazani su dobijeni rezultati: u gornjem redu za matricu inicijalizovanu kao u (84), a u donjem redu na predloženi način. Kako bi se dobila što drastičnija razlika u tačnosti ova dva modela, modeli su obučeni sa $N = 8$ stanja, tako da je matrica prelaza $\mathbf{A}$ dimenzija 8×8.

$\mathbf{A}$ inicijalno \ reč	0	1	2	3	4	5	6	7	8	9	srednja vrednost
$a_{ij}=0$ $(j=i,i+1)$	100.00	98.00	100.00	97.00	100.00	96.56	98.97	99.00	100.00	98.57	98.81
$a_{ij}=0$ $(j<i)$	100.00	98.00	100.00	97.00	100.00	96.56	99.49	99.67	100.00	98.57	98.93

Tabela 7: Srednja tačnost [%] prepoznatih cifri (0-9) na skupu za testiranje 1 u zavisnosti od načina inicijalizacije matrice prelaza $\mathbf{A}$. Modeli sadrže $N = 8$ stanja.

Iz tabele primećujemo da je čak i za ovako velike matrice $\mathbf{A}$, razlika u dobijenoj tačnosti neznatna. Dodatno, nakon analize obučenih modela primećeno je da većina matrica $\mathbf{A}$ inicijalizovanih na drugi način ima neznatne vrednosti na pozicijama koje su udaljene više od jednog stanja. Dakle, u ovom problemu način inicijalizacije matrice $\mathbf{A}$ nema značajan uticaj na krajnju tačnost sistema.

U ovoj sekciji su određeni optimalni parametri za treniranje modela sa datim opservacijama i rečnikom. Utvrđeno je da je optimalan izbor parametara: broj stanja $N = 5$, broj komponenti Gausove smeše $M = 6$, opservacije u vidu $q = 12$ MFCC koeficijenata.

3.2 Prepoznavanje govora nezavisno od govornika

U ovoj podsekciji je određivana tačnost prepoznavanja reči na ispitanicima koji nisu deo obučavajućeg skupa, tj. prepoznavanje govora nezavisno od govornika.

Kako je broj ispitanika u eksperimentu relativno mali, za validaciju modela korišćen je *K-fold cross-validation* metod. Po ovom metodu za svakog od K ispitanika se formiraju modeli, koji se potom testiraju samo na tom ispitaniku. Na taj način određeno je K procena tačnosti, koje se potom usrednjuju kako bi se dobila konačna tačnost. Pri obučavanju modela za testiranje k-tog ispitanika, koriste se sve sekvence za obučavanje drugih ispitanika. Time je osigurano da je formirani model nezavistan od ispitanika na kome se testira.

U tabeli 8 prikazani su dobijeni rezultati: ispitanici su predstavljeni po vrstama, a reči po kolonama. U poslednjoj vrsti data je srednja tačnost za datu reč.

reč / ispitanik	0	1	2	3	4	5	6	7	8	9
1	100.00	100.00	88.24	94.12	100.00	58.82	100.00	100.00	86.67	100.00
2	100.00	100.00	93.33	100.00	100.00	100.00	100.00	100.00	100.00	100.00
3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
4	100.00	100.00	100.00	96.30	100.00	96.77	92.59	77.78	100.00	100.00
5	100.00	100.00	93.33	100.00	100.00	100.00	7.14	100.00	80.00	100.00
6	100.00	90.91	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
7	100.00	100.00	100.00	100.00	100.00	42.11	20.00	85.00	100.00	100.00
8	56.25	100.00	100.00	100.00	91.67	93.75	100.00	50.00	100.00	100.00
9	88.89	95.24	100.00	100.00	100.00	71.43	80.95	90.48	100.00	80.95
10	100.00	76.47	100.00	100.00	100.00	94.12	100.00	100.00	100.00	100.00
11	100.00	91.67	100.00	100.00	87.50	80.00	100.00	100.00	90.91	100.00
12	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00	100.00	44.44
13	91.67	90.91	92.86	66.67	100.00	100.00	100.00	100.00	100.00	93.75
14	100.00	92.86	92.86	100.00	100.00	92.31	100.00	84.62	100.00	100.00
15	100.00	100.00	100.00	92.31	100.00	100.00	100.00	91.67	100.00	100.00
srednja vrednost	95.79	95.87	97.37	96.63	98.61	88.62	86.71	85.30	97.17	94.61

Tabela 8: Srednja tačnost [%] prepoznatih cifri (0-9) za svakog od ispitanika nakon treniranja modela na svim drugim ispitanicima.

Ostvarena je najmanja srednja tačnost za cifre 7, 6 i 5, dok je za ostale cifre srednja tačnost preko 90%. Ostvarena tačnost je u skladu sa onom ostvarenom u literaturi [25]. Lošija tačnost ostvarena na ciframa 7, 6 i 5 može biti posledica slabih ploziva (t u pet i šest) kao i nazala (m u sedam), glasova koji imaju malu energiju. Zanimljivo je primetiti i da su cifre 1, 8 i 9, koje takođe imaju glasove iz ove dve grupe na kraju reči, znatno bolje prepoznate.

U tabeli 9 je data matrica konfuzije usrednjena po svim ispitanicima. Dobijene vrednosti tačnosti su usrednjene na način opisan u sekciji 3.1.

prepoznata reč

tačna reč	0	1	2	3	4	5	6	7	8	9
0	95.79	0.97	0.74	0.42	0.00	0.00	2.08	0.00	0.00	0.00
1	1.10	95.87	0.78	0.00	0.00	0.00	0.56	1.21	0.48	0.00
2	0.92	0.00	97.37	0.00	0.00	0.00	0.00	0.48	0.00	1.23
3	0.56	0.00	0.00	96.63	1.87	0.00	0.56	0.00	0.00	0.39
4	0.00	0.00	0.00	0.00	98.61	0.00	0.00	0.83	0.00	0.56
5	0.00	0.42	0.00	0.67	0.00	88.62	0.00	1.02	0.00	9.28
6	1.67	1.91	0.00	0.25	6.35	0.00	86.71	3.11	0.00	0.00
7	0.74	5.54	0.00	0.00	1.69	0.00	0.63	85.30	3.70	2.38
8	1.33	0.00	0.00	0.00	0.00	0.00	0.00	1.49	97.17	0.00
9	0.74	1.05	0.63	0.00	0.00	2.96	0.00	0.00	0.00	94.61

Tabela 9: Matrica konfuzije za prepoznavanje cifara 0-9. Vrednosti tačnosti su date u procentima i usrednjene po svim ispitanicima.

Iz matrice konfuzije možemo očitati kada najčešće dolazi do zamene cifara iz rečnika. Izgovorena reč "pet" će u 9.28% slučajeva biti prepoznata kao reč "devet". Moguće objašnjenje za ovaj fenomen jeste način izgovaranja ploziva t. Posmatranjem vremenskih sekvenci signala koji odgovaraju reči "pet" može se primetiti da, u zavisnosti od ispitanika, signal ima samo jedan, glavni deo (ukoliko ispitanik ne izgovara glasno glas t) ili još jedan dodatni, odvojeni deo (ukoliko ispitanik naglašava glas t). Stoga je moguće da kod ispitanika koji naglašavaju glas t u reči "pet" zbog tog drugog dela signala, reč "pet" bude prepoznata kao "devet".

Drugi par reči kod kojeg najčešće dolazi do permutacije jeste "šest" i "četiri". Glasovi š i č su prednjonepčani i bezvučni, a oba su praćena vokalom e u datim rečima. Samim tim, ove dve reči imaju sličan početak. Takođe, moguće je da varijacije u izgovaranju reči "četiri" kao "četri" doprinose povećanju sličnosti sa rečju "šest".

Poslednji par sa relativno velikom greškom klasifikacije jeste "sedam" i "jedan". Jasno je da su ove dve reči fonetski vrlo slične i da je za ispravnu klasifikaciju neophodno dobro prepoznavanje glasa s odnosno j. Kako su to prvi glasovi u datim rečima, od suštinske je važnosti da je segmentacija reči izvršena na adekvatan način tako da je prvi glas reči očuvan.

U ostalim slučajevima je greška klasifikacije manja od 4%, tako da ostali parovi reči neće biti posebno analizirani. Potrebno je izvršiti dodatna istraživanja da li se u kontrolisanim uslovima i sa kvalitetnijim mikrofonima ostvaruju slični rezultati, odnosno da li zaista postoji korelisanost između prethodno navedenih parova reči.

4 Zaključak

Predloženi sistem sačinjen od skrivenih Markovljevih modela je sa visokom tačnošću prepoznao reči izgovorene na srpskom jeziku čak i u slučaju modela formiranih nezavisno od ispitanika na kojima je model testiran. Pokazan je eksperimentalni način za dobijanje optimalnih parametara modela, koji imaju vrednosti približno jednake onima predloženim u literaturi. Napravljena je razlika između testiranja na opservacijama koje su snimljene u istim sesijama snimanja kao i opservacije iz skupa za obučavanje, i opservacijama koje su naknadno snimljene. Kao što je i očekivano, dobijena tačnost upućuje na veću korelisanost prvog seta za testiranje sa setom za obučavanje. Stoga je preporučivo da optimalni parametri budu određeni na osnovu drugog skupa podataka za testiranje. Da bi parametri bili bolje procenjeni, poželjno je formirati veću bazu opservacija i potom testirati modele na većem broju ispitanika koji nisu učestvovali u obučavanju modela.

Veličina rečnika u ovom radu je svega deset reči, tako da nije jasno u kojoj meri je ovakav sistem primenljiv u prepoznavanju reči iz rečnika veće dužine. Stoga je potrebno izvršiti dalje analize i eksperimentalno utvrditi koja je maksimalna veličina rečnika za koju je tačnost prepoznavanja reči i dalje zadovoljavajuća. Takođe, kako je opisani sistem vrlo fleksibilan, moguće je izmeniti sistem tako da prepoznaje pojedine glasove, a ne cele reči. Modeliranjem glasova praktično se postiže prepoznavanje govora za neograničeni broj reči (pod uslovom da postoji model za svaki glas). Naravno, ovakav pristup je složeniji od ovog predloženog u radu i potrebno je analizirati u kojoj meri je on ostvariv. Kao što je pokazano i u poslednjem eksperimentu, naročitu pažnju je potrebno posvetiti modeliranju glasova sa malom energijom - kakvi su plozivi i nazali.

Dodatno unapređenje predloženog sistema jeste prepoznavanje govora, odnosno čitavih rečenica. Za takav sistem neophodno je uzeti u razmatranje i sintaksička i semantička pravila srpskog jezika. Takođe, neophodno je rešiti problem segmentacije govora na reči (ili glasove) na odgovarajući način. Kao prvi korak u formiranju takvog sistema predlažemo prepoznavanje govora sačinjenog isključivo od cifara $0-9$ i primene ovog sistema sa manjim izmenama u takvoj analizi.

Literatura

[1] Brian Hayes et al. First links in the markov chain. *American Scientist*, 101(2):252, 2013.

[2] A. A. Markov. An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19(4):591–600, 2006.

[3] Md Rafiul Hassan and Baikunth Nath. Stock market forecasting using hidden markov model: a new approach. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pages 192–196. IEEE, 2005.

[4] Rogemar S Mamon and Robert James Elliott. *Hidden Markov models in finance*, volume 4. Springer, 2007.

[5] Gary A Churchill. Hidden markov chains and the analysis of genome structure. *Computers & chemistry*, 16(2):107–115, 1992.

[6] Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.

[7] Chris Karlof and David Wagner. Hidden markov model cryptanalysis. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 17–34. Springer, 2003.

[8] Mosaad Khadr. Forecasting of meteorological drought using hidden markov model (case study: The upper blue nile river basin, ethiopia). *Ain Shams Engineering Journal*, 7(1):47–56, 2016.

[9] Walter Zucchini and Peter Guttorp. A hidden markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923, 1991.

[10] Yanghee Nam and KwangYun Wohn. Recognition of space-time hand-gestures using hidden markov model. In *Proceedings of the ACM symposium on virtual reality software and technology*, pages 51–58, 1996.

[11] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013.

[12] Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer speech & language*, 6(3):225–242, 1992.

[13] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[14] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[15] Biing-Hwang Juang and Lawrence R Rabiner. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1:67, 2005.

[16] John R Pierce. Whither speech recognition? *The journal of the acoustical society of america*, 46(4B):1049–1051, 1969.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[19] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[20] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. Deep convolutional neural networks for lvcsr. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8614–8618. IEEE, 2013.

[21] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[22] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[23] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

[24] K-F Lee and H-W Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

[25] Lawrence R Rabiner, B-H Juang, Stephen E Levinson, and M Mohan Sondhi. Recognition of isolated digits using hidden markov models with continuous mixture densities. *AT&T technical journal*, 64(6):1211–1234, 1985.

[26] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3(175):12, 2002.

[27] Lawrence R Rabiner and Ronald W Schafer. *Introduction to digital speech processing*. Now Publishers Inc, 2007.

[28] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[29] Leonard E Baum and John Alonzo Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

[30] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[31] Leonard E Baum et al. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3(1):1–8, 1972.

[32] L Liporace. Maximum likelihood estimation for multivariate observations of markov sources. *IEEE Transactions on Information Theory*, 28(5):729–734, 1982.

[33] B-H Juang. Maximum-likelihood estimation for mixture multivariate stochastic observations of markov chains. *AT&T technical journal*, 64(6):1235–1249, 1985.

[34] B-H Juang and Lawrence R Rabiner. A probabilistic distance measure for hidden markov models. *AT&T technical journal*, 64(2):391–408, 1985.