
UDK 801.41:681.3

Originalni znanstveni rad

Primljeno: 15. 5. 1990.

Milan Stamenković
VVTŠ KoV JNA, Zagreb

AUTOMATSKO PREPOZNAVANJE GOVORA

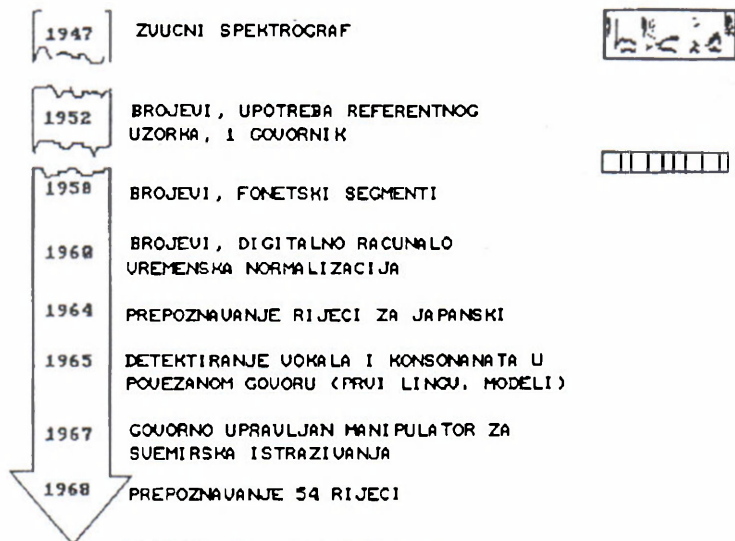
SAŽETAK

Osim što su definirana opća načela automatskog prepoznavanja govora, u ovom su radu razmatrani pristupi prepoznavanja govora temeljeni na tradicionalnoj teoriji prepoznavanja uzoraka i strojnom učenju u opreci s modelima koji uključuju fonetsko-lingvističke aspekte. Središnje mjesto rada jest opis sistema za prepoznavanje povezanoga govora, koji modelira procese govorne komunikacije.

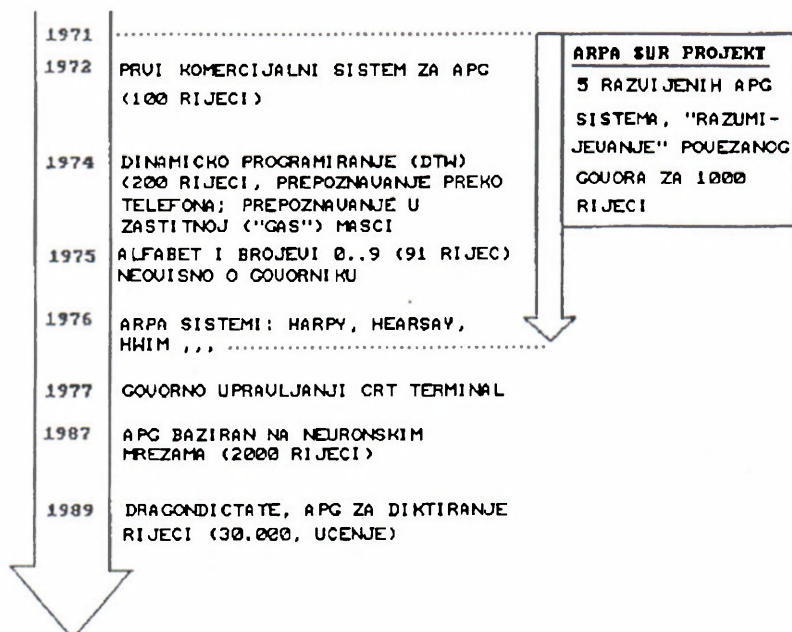
1. UVOD

Za razliku od pokušaja da se mehanički govor oponaša različitim vrstama uređaja, koji datiraju još iz drevnih vremena pa sve do danas (Witten, 1982) uspješna rješenja sistema za automatsku analizu i prepoznavanje govora (APG) znatno su novija. Naime, značajan napredak u analizi govora postignut je tek neposredno prije II svjetskog rata (1930-1940) kada je konstruiran uređaj za frekvencijsku analizu glasova koji je u naprednijoj verziji nazvan sonograf (Potter, 1947). Mjerenja spektra pojedinih glasova i njihovih međusobnih odnosa omogućila su da godine 1952. Davis, Biddulph i Balshek u Bell laboratoriju razviju prvi stroj za automatsko prepoznavanje brojeva (0. . 9). Njihov se uređaj sastojao od spektralnog analizatora za dva frekvencijska pojasa: ispod i iznad 900 Hz u kojem je bio ugrađen sklop za mjerenje broja prolazaka signala kroz referentnu razinu. Takav način mjerenja predstavljao je ekvivalente centralne frekvencije unutar svakog opsega. Dobivene vrijednosti bile su osi koordinatnog sustava (poslije poznat kao dualna formantna ravnina F_1 - F_2) koji je uspoređivan s etalonom za svaku pojedinu riječ. Za prepoznatu riječ proglašavana je ona čiji je referentni uzorak imao najveću mjeru korelacije s nepoznatim. Početni uspjesi bili su inspiracija za dalja istraživanja. Tako je ista grupa 1958. godine razvila sistem "Audrey", koji je za analizu upotrebljavao deset pojasnih filtera i izdvajao pojedine maksimume u spektru koji su bili vremenski usklađivani s referentnim uzorcima. Značajna novost u odnosu prema prvom radu bila je segmentacija govora na fonetske jedinice, koji su predstavljali pojedine tipove glasova. Postotak prepoznavanja za unaprijed određenog govornika i za isti vokabular praktično je bio 100% posto, pa ohrabrene takvim rezultatom niču mnoge istraživačke grupe, koje žele što prije konstruirati stroj za automatsko diktiranje. Međutim, tek kada se sistematski počeo rješavati taj problem, uočena je prava kompleksnost percepcije i razumijevanja govora koja nije mogla biti aproksimirana klasičnom teorijom prepoznavanja uzoraka. Tek nakon dvadesetak godina (Lca, 1980) napravljen je komercijalno kompletan sistem za prepoznavanje izolirano izgovorenih riječi. Sistem koji je tada ponudila tvrtka "Treshold Technology Incorporated" mogao je prepoznavati do 100 riječi pažljivo izgovorenih. Najviše su razvoju APG-a pridonijeli rezultati projekta SUR (Speech Understanding Research - istraživanja o razumijevanju govora) koji je voden u sklopu institucije (D)ARPA (Defence Advanced Research Project Agency) za potrebe američke vojske i NATO-pakta sredinom osamdesetih godina. U tom razdoblju (sl. 1) intenzivno je proučavan lingvistički model prepoznavanja govora koji uključuje više razina govorne komunikacije (fonetsku, leksičku, semantičku). Konkretni realizirani laboratorijski sistemi SDC, HWIM, HEARSAY-II i HARPY bili su namijenjeni za prepoznavanje povezanoga govora, neovisno o govorniku i mogli su voditi specijalizirani dijalog s operatorom.

RANA HISTORIJA SISTEMA ZA APG



NOVIJA HISTORIJA



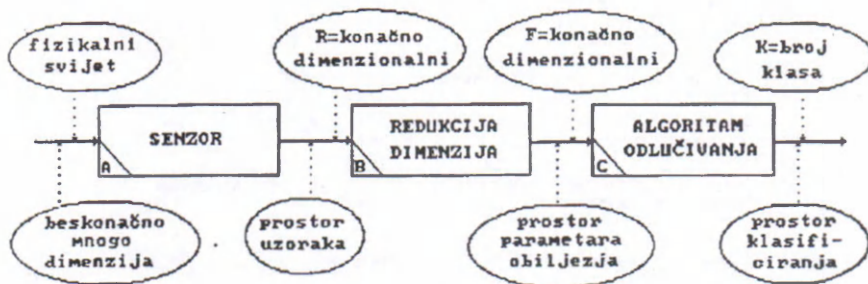
sl. 1 Povijesni pregled razvoja sistema za APG

Sistemi za prepoznavanje (segmentaciju) govora koji bi omogućavali slobodno diktiranje još nisu napravljeni. U posljednjem desetljeću mnogo se ulaže u koncepciju tzv. neuronskih mreža i strojnog učenja. Objavljeni rezultati već potvrđuju uspješnost prepoznavanja i nekoliko tisuća riječi za unaprijed pripremljenog govornika. Današnji "state of the art" u domeni APG je poopćeni HMM (Hidden Markov Model) model prepoznavanja, kojim se uspješno može prepoznavati i više desetaka tisuća (izoliranih) riječi za unaprijed pripremljenog govornika. Prvi takav komercijalni sistem predstavila je američka tvrtka "Apricot", na II Evropskoj konferenciji o govornoj komunikaciji i tehnologiji u Parisu rujna 1989.

Obrada govora na računalu u nas je tek u začetku. Prema prezentiranim radovima (ROJP III-1985, ROLP IV-1988) istraživanja su u domeni analize parametara obilježja signala i primjene tradicionalnih tehnika klasificiranja uzoraka. Opsežnija istraživanja segmentacije povezanoga govora s uključivanjem lingvističkog aspekta u obradu signala do sada nisu objavljena. Publicirani radovi uglavnom se bave prepoznavanjem izoliranih riječi za unaprijed pripremljenog govornika (najčešće brojeve 0 - 9). Svrha je ovog rada prikazati globalna načela prepoznavanja govora, klasifikaciju sistema za APG te da predloži model automatske segmentacije govora koja zahvaća lingvističke elemente govorne komunikacije.

2. PRISTUPI U RJEŠAVANJU PROBLEMA APG

Prvi sistemi za APG vjerno su slijedili opću shemu prepoznavanja uzoraka prikazanu na donjoj slici.



sl. 2-1 Konceptija strojnog prepoznavanja uzoraka

Pojave iz fizikalnog svijeta pobudjuju ulazne senzore koji okolni svijet svode na neku mjerljivu veličinu kodiranu na odgovarajući način. Nakon toga, iz mnoštva dobivenih kodova odabiru se samo oni koji prema definiranom kriteriju imaju najveću informativnost. Odabrane i transformirane prezentacijske kodove nazivamo parametri obilježja (engl. features). Redukcija promatranih parametara izuzetno je važna jer procesna moć postojećih računala uglavnom nije dovoljna za obradu svih izmjerenih veličina dobivenih s ulaznih senzora. Također, pažljivim odabirom parametara obilježja pojednostavljuje se klasifikacijski postupak i povećava njegova točnost. Iduća je faza klasifikacija (prepoznavanje) uzoraka prema definiranim pravilima. Primijenimo li opisanu shemu na segmentaciju (prepoznavanje) govora, tada je lako zaključiti da je ulazni senzor A zapravo mikrofon, dok odabir parametara čini vrsta primijenjene analize izvornog signala (analiza formanta, analiza F0, energija itd). Kao i kod općeg modela govorni segmenti mogu se klasificirati prema beskonačno mnogo kriterija, spomenimo neke: podjela po periodičnosti valnog oblika (vokalizirani - nevokalizirani), podjela prema iznosu energije, podjela prema visini F0, položaju formanta itd. Dalje ovisno o dužini glasovnog segmenta možemo govoriti o prepoznavanju izolirano izgovorenih riječi ili o prepoznavanju povezanog govora. Dakle, prepoznavanje uzoraka možemo shvatiti kao dvostruku transformaciju (Gonzalez, 1978) skupa uzoraka P na reducirani skup parametara obilježja F, nad kojim se primjenjuje neki od postupaka iz skupa postupaka klasificiranja C :

$$P \rightarrow F \rightarrow C \quad \dots (2-1)$$

Strukture uzoraka i parametara obilježja obično se predstavljaju kao linearne kombinacije nekih funkcija, dok je proces klasificiranja zapravo traženje najbolje aproksimacije referentne linearne kombinacije, koja predstavlja određenu klasu, s ulaznom linearnom kombinacijom, koja opisuje parametar obilježja. Da bi se preciznije odredila transformacija (2-1), nužno je definirati vektorski prostor i kriterij mjere sličnosti odnosno pripadnosti nekoj klasi, specifikacijom metričkog prostora (Kurepa, 1982), (Patrick, 1972).

DEF 2 - 1 Metrički prostor

Ureden par (X, d) nepraznog skupa X i funkcije $d: X \times X \rightarrow \mathbb{R}$ naziva se metrički prostor, a funkciju d zovemo razdaljinska funkcija ili metrika ako vrijedi:

1. $d(x, y) > 0$ za sve $x, y \in X$;
 2. $d(x, y) = 0 \Leftrightarrow x = y$
 3. $d(x, y) = d(y, x)$ za sve $x, y \in X$;
 4. $d(x, y) < d(x, z) + d(z, y)$, za sve $x, y, z \in X$;
- Vrijednost $d(x, y)$ nazivamo razdaljinom x i y.

Nažalost, klasična shema prepoznavanja uzoraka primjenjiva je samo na veoma ograničene sisteme za APG (vokabular do nekoliko desetaka riječi, pažljiv način izgovora itd) zbog više razloga npr:

a) Varijabilnost izgovora : Način izgovora (dinamika, tempo, naglasak, dezorganizacija na kraju izričaja itd) neposredno djeluje na snimljeni oblik signala. Na

primjer, riječ "zdravo" izgovorena kao pozdrav pri neformalnom susretu često je toliko izobličena da zadnja dva glasa praktično nisu niti izgovoreni (!), bez obzira na to što ih percipiramo (zahvaljujući pragmatičkoj redundanciji jezika). Podemo li od pretpostavke da se automatsko prepoznavanje izvodi isključivo na osnovi fizičkog signala, jasno je da nema niti teoretske mogućnosti da se detektiraju svi glasovi unutar riječi, jer niti ne postoje.

b) Problem koartikulacije: Zbog mehaničke inercije i drugih psiho-fizioloških ograničenja svaki se glas izgovara u odnosu prema prethodnim glasovima i onim koji slijede. To znači da je, uslijed različitih početnih uvjeta izgovora, svaki fonem na određeni način modificiran okolinom. Prijelazi od jednog glasa na drugi kontaminirani su, tako da se ne može točno odrediti granica među glasovima, nego se može govoriti samo o mjestu na kojem prevladava jedan od glasova.

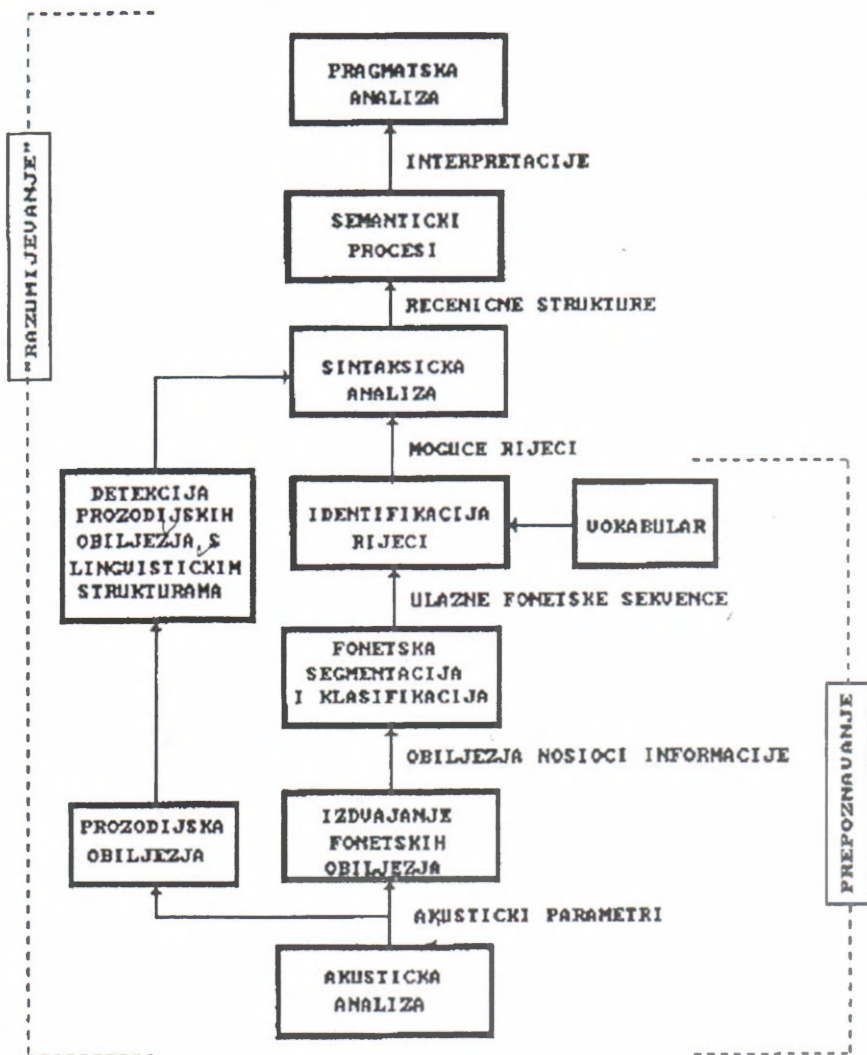
c) Varijabilnost govornika: Ovdje postoji problem da svaki govornik ima specifične govorne osobine koje možemo izraziti npr. kao srednji period F_0 , prosječni položaj formanata za pojedine glasove itd. Međutim, problem nastaje kada se parametri opisa realizacije pojedinih fonema jednoga govornika bolje podudaraju sa različitim glasovima drugoga govornika, čime je npr. fonem "m" jednog govornika uvijek sličniji fonemu "n" drugog itd.

a) Veličina vokabulara: Povećanjem broja riječi naglo se povećava broj sličnih riječi, što unosi nove zahtjeve za definiranje većeg broja elemenata koji se klasificiraju.

e) Tehnička ograničenja: Budući da se modeli automatskog prepoznavanja govora realiziraju realnim uređajima, uvijek je prisutan problem osjetljivosti pretpojačala, rezolucije digitalizacije signala itd. Zbog toga se npr. uvode dodatna ograničenja za uspješnost prepoznavanja govora: smanjena buka okoline, fiksno rastojanje mikrofona od usta itd. Jedno od ključnih ograničenja vezano za efikasnost modela jest i arhitektura i procesna snaga računala koja, prema procenama vodećih laboratorija (Kohonen, 1988), mora biti barem 10.000 MIPS-a (1 MIPS = milijun instrukcija u sekundi).

Da bi se riješili ti problemi, u sklopu ARPA SUR projekta razvijena je koncepcija za APG, koja efikasnost prepoznavanja (postotak grešaka, relativna neosjetljivost na način izgovora i govornika itd) postiže uključivanjem različitih razina govorne komunikacije. Sumarna konceptualna shema prikazana je na slici 2-2.

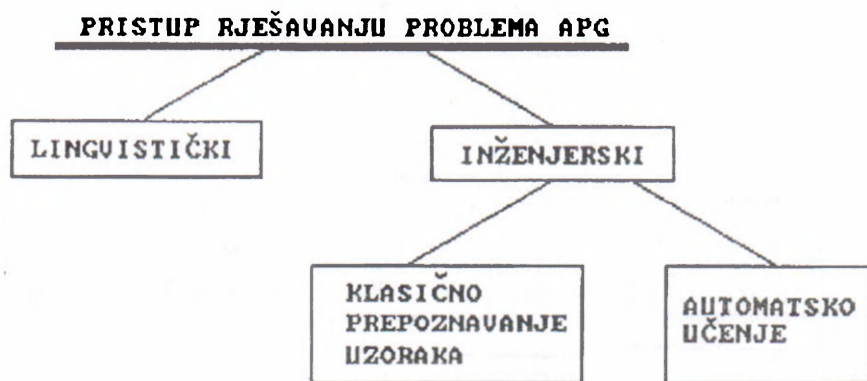
Prema predloženoj koncepciji, za prepoznavanje nije dominantna informacija o fizičkoj karakteristici govornog signala, nego naknadna simbolička obrada koja se realizira u procesu "razumijevanja". Nakon fizičko-akustičke analize formiraju se liste-kandidati, koji predstavljaju ulaz za sintaksičku razinu. Na toj se razini propuštaju one kombinacije koje zadovoljavaju sintaksna pravila, a potom služe kao parametri semantičke i pragmatske analize. Prema predloženom modelu udio klasičnog prepoznavanja (na osnovi fizikalne prirode signala i definiranih razdaljinskih funkcija) samo je 20-40%, dok se greške na toj razini korigiraju lingvističkom obradom. Međutim, ovdje nastaje problem modeliranja prirodnog jezika (sintakse, semantike i pragmatike) za koji i danas postoji tek gruba aproksimacija. Takva ograničenja (kompleksnost procesa) uvjetuju vezivanje sintakse za neko užo područje govornog diskursa, uglavnom za problemske situacije (kupovina, putovanje itd) o čemu će poslije biti više riječi.



sl. 2-2 Prosesi uključeni u prepoznavanju i "razumijevanju" govora

U posljednjih desetak godina sve je više radova koji APG pokušavaju riješiti modelom strojnog učenja (Kohonen, 1988) (Stamenković, 1988) (McClelland, 1988) Ideja od koje se ovdje polazi jest da nije potrebno eksplicitno znanje (unaprijed definirani algoritam) o načinu klasificiranja, nego da se ekvivalentno znanje može "naučiti" na osnovi primjera i odgovora. Znanje koje dobiveno na taj način može biti eksplicitno (induktivno učenje) ili implicitno (učenje s pomoću konektivnih modela). U slučaju eksplicitnog znanja moguće je neposredno ustanoviti koja su pravila uzeta kao distinktivno obilježje, dok kod implicitnog učenja to nije moguće.

Opisane globalne pristupe realizaciji sistema APG možemo prikazati na sljedeći način:



sl. 2-3 Pristupi u rješavanju APG

Dakle, razlikujemo sisteme za APG koji ne uključuju fonetsko-lingvističko znanje (inženjersko- tehnički pristup) i one koji ga uključuju.

Radi usporedbe performansi sistema za APG, definirano je nekoliko klasa prepoznavanja u ovisnosti o načinu izgovora:

a) **povezani govor** (continuous speech) jest normalan način izgovaranja riječi (bez umjetnih pauzi)

b) **izgovor po grupama riječi** (connected words) jest takav način izgovora gdje se pojedine grupe riječi izgovaraju bez razmaka, ali je razmak izmjeđu grupa cca 200-300 ms.

c) **izolirane riječi** (isolated words) jest način izgovora gdje je pauza između uzastopnih riječi cca 200- 300 ms.

Dodatni kriterij za sva tri načina izgovora uključuje govornika, tj. je li sistem neovisan (speaker independent) ili ovisan (speaker dependent) o njemu. Sistemi za APG koji su ovisni o govorniku zahtijevaju da se govornik unaprijed pripremi, tj. da svaki govornik prije procesa prepoznavanja snimi svoje referentne uzorke

(parametre) glasova, riječi ili rečenica. Najčešći parametri koji služe kao zajednički elementi distinktivnih obilježja opisani su u idućem poglavlju.

3. PARAMETRI OBILJEŽJA GOVORNOG SIGNALA

Nakon digitalizacije govora, sljedeći postupak prema segmentaciji jest definiranje pojedinih kvantitativnih i kvalitativnih svojstava signala na osnovi kojih se on analizira. Odabrana svojstva koja ulaze u metriku analize nazivamo parametri obilježja govornog signala. O vrsti analize, ovisit će i izbor parametara. Za većinu analiza iskustveno se uvode sljedeći parametri i njihove kombinacije:

- a) kratkovremenska energija
- b) broj prolaza kroz nulu
- c) trenutni spektar signala
- d) period osnovnog tona
- e) koeficijenti linearanc predikcije govora

Najjednostavnija prezentacija digitaliziranoga govornoga signala $x(n)$ dana je iznosom njegove energije:

$$E = \sum_n x(n)^2 \quad \dots (3-1)$$

Međutim, totalna energija izgovorene riječi nije informativna, jer različite riječi mogu imati približno istu energiju (npr. osamdeset i udaljenost) a ukupna energija iste riječi u mnogome ovisi o načinu izgovora (tempo, dinamika, akcent). Zbog toga se uvodi kratkovremenska energija koja opisuje energetska razinu segmenta unutar riječi za N uzoraka:

$$E(n) = \frac{1}{N} \sum_{m=0}^{N-1} [w(m) x(n-m)]^2 \quad \dots (3-2)$$

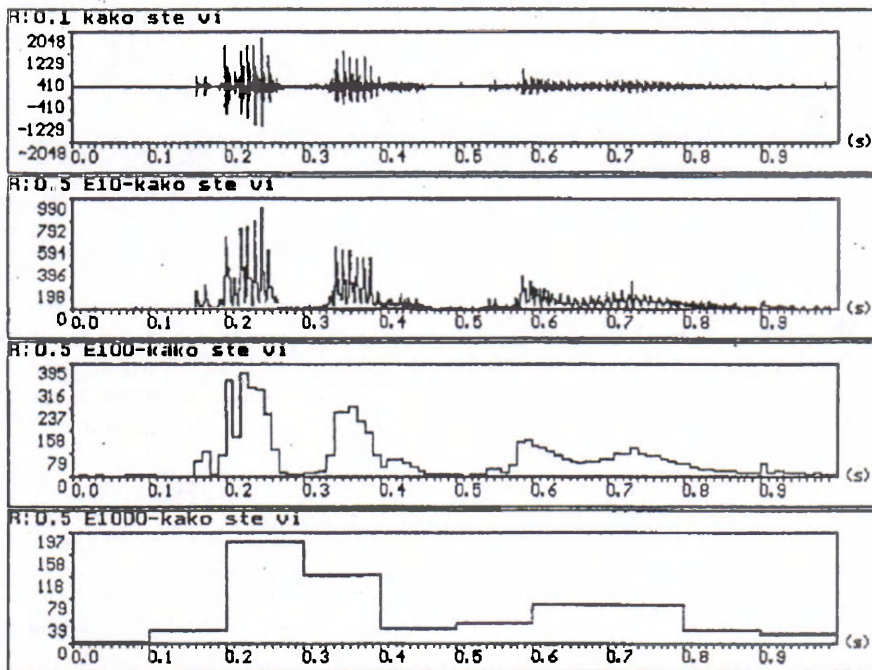
gdje je $w(m)$ funkcija kojom se ponderiraju pojedini uzorci i u literaturi se naziva prozor. Najčešće se upotrebljava tzv. pravokutni prozor:

$$w(m) = \begin{cases} 1 & \text{za } 0 \leq m \leq N-1 \\ 0 & \text{inače} \end{cases} \quad \dots (3-3)$$

Iznos $E(n)$ predstavlja srednju energiju signala po uzorku posljednjih N uzoraka do $x(n)$. Izraz 3-2 prenaplašava vokalizirane glasovne segmente u odnosu prema tišim govornim intervalima, pa se u praksi često koristi nešto izmijenjena definicija kratkovremenske energije, tj:

$$E(n) = \frac{1}{N} \sum_{m=0}^{N-1} |x(m) x(n-m)| \quad \dots (3-4)$$

Posebnu pozornost treba obratiti na širinu prozora $w(m)$ (tj. na njegovo trajanje). Ako je N veće, tada je zahvaćeno više uzoraka, pa se izraz 3-2 približava izrazu 3-1. Međutim, uzme li se premalo uzoraka, funkcija $E(n)$ postaje sve sličnija izvornom signalu $x(n)$ (po apsolutnim iznosima). Efekt širine vremenskog prozora na izgled funkcije kratkovremenske energije ilustriran je na slici 3-1 ("kako ste vi").



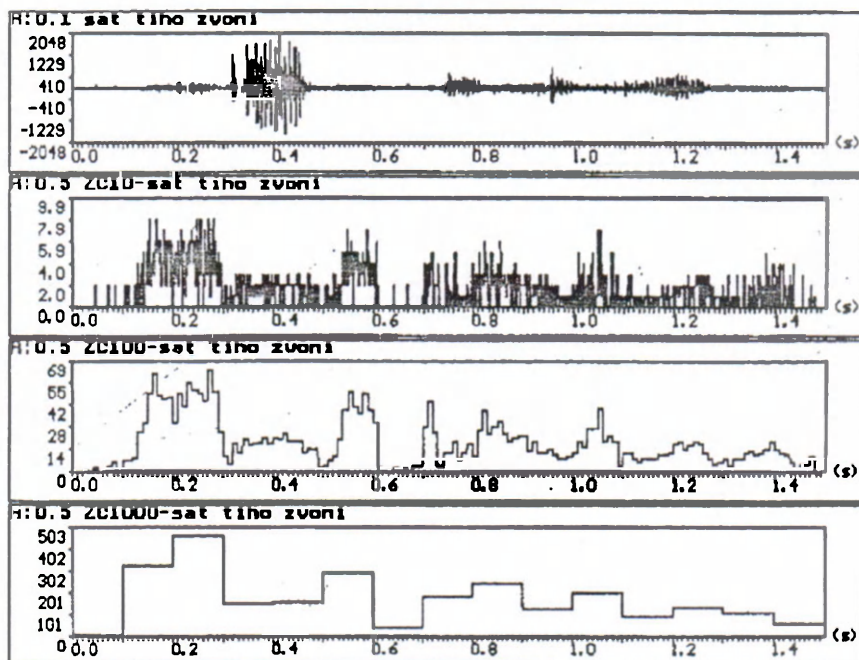
sl. 3-1 Utjecaj širine vremenskog prozora na oblik funkcije $E(n)$: a) PCM b) $N=10$ c) $N=100$ d) $N=1000$

Prema preporukama (Lea, 1980), (Schafcr, 1979) najprihvatljivija širina prozora je oko $\Delta t = 10$ ms, tj broj uzoraka za koje se izračunava srednja energija:

$$N = 0.01 \text{ [s]} \cdot F_u \text{ [Hz]} \quad \dots (3-5)$$

Broj prolaza signala kroz nulu predstavlja prvu ocjenu o općoj frekventijskoj prirodi signala. Naime, uzmemo li da je signal predstavljen nekom sinusnom (kosinusnom) funkcijom, tada je lako uspostaviti vezu između broja prolaza kroz nulu N i frekvencije :

$$F = \frac{1}{T} = \frac{1}{2} N_z \left[\frac{1}{S} \right] \quad \dots (3-6)$$



sl. 3-2 Veći iznos broja prolaza kroz nulu ukazuje na nevoalizirane glasove
b) $N=10$ c) $N=100$ d) $N=1000$

Funkciju jednog prolaza kroz nulu možemo definirati na sljedeći način:

$$N_z(n) = \begin{cases} 1 & \text{za } \text{sign}(x(n)) \neq \text{sign}(x(n-1)) \\ 0 & \text{inače} \end{cases} \quad \dots (3-7)$$

Analogno izrazima (3-2) i (3-4) promatramo određeni vremenski segment izražen brojem uzoraka N :

$$ZC(n) = \sum_{m=0}^{N-1} w(m) N_z(n-m) \quad \dots (3-8)$$

Izraz (3-8) primijenjen na govorni signal implicitno upozorava na glasove čija je energija koncentrirana u višim dijelovima spektra (iznad 2-3 KHz) npr. s, c, ć, š itd. Na slici 3-2 ("sat tiho zvoni") jasno se izdvajaju segmenti frikativa (h) i sibilanata (s, z).

Jedan od najvažnijih parametara govornog segmenta jest njegova frekvenzijska karakteristika (spektar). Prijelaz iz vremenske domene u frekvencijsku najčešće se ostvaruje s pomoću diskretne Fourierove transformacije :

$$X(e^{j\omega}) = \sum_{m=-\infty}^{m=+\infty} w(n-m) x(m) e^{-j\omega m} \quad \dots (3-9)$$

Indeks n odgovara posljednjoj digitaliziranoj vrijednosti signala. Funkcija vremenskog prozora $w(i)$ najčešće nije pravokutnog oblika, nego su u upotrebi tzv. Hamminog (3-10) ili Hanningov (3-11) prozor (Proakis, 1983):

$$w(m) = \begin{cases} 0.54 - 0.46 \cos(2\pi m/N) & \text{za } 0 \leq m \leq N-1 \\ 0 & \text{inače} \end{cases} \quad \dots (3-10)$$

$$w(m) = \begin{cases} 0.5 (1 - \cos(2\pi m/N)) & \text{za } 0 \leq m \leq N-1 \\ 0 & \text{inače} \end{cases} \quad \dots (3-11)$$

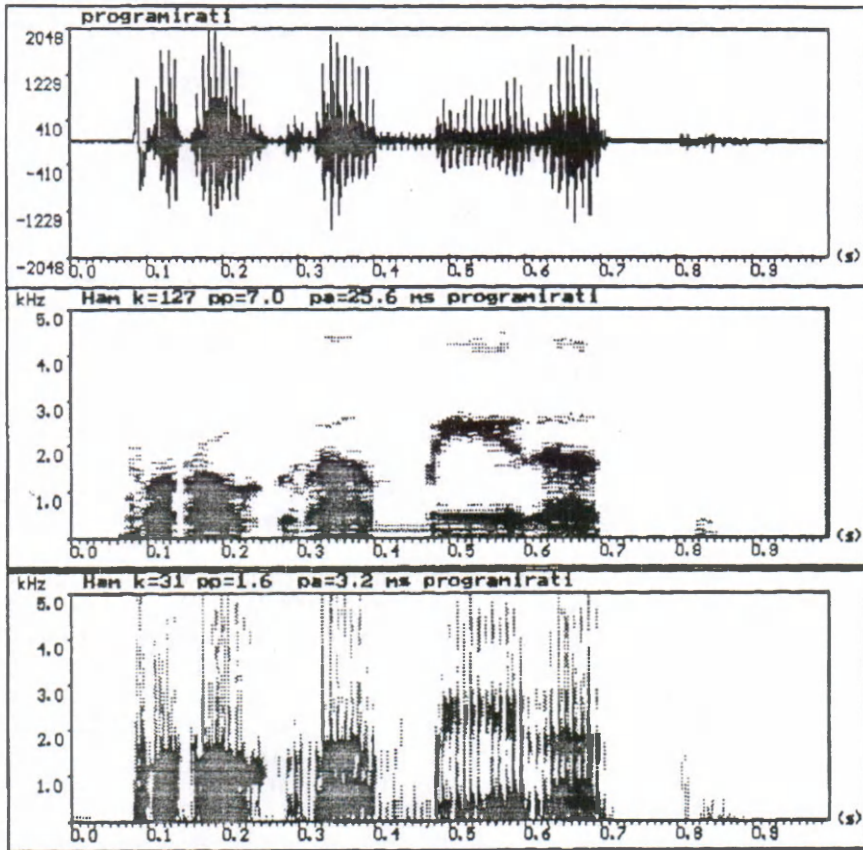
Pravokutni prozor ne upotrebljava se iz razloga jer uvodi dodatne lokalne maksimume u spektru zbog karakterističnog frekvencijskog odziva na pravokutni puls. Iz istih razloga kao i kod funkcije totalne energije signala (izraz 3-1), u praksi nije od interesa ukupni spektar signala, nego njegova raspodjela u pojedinim vremenskim intervalima. Zato se 3-9 zapisuje u izmijenjenom lokaliziranom obliku:

$$X(k) = \sum_{m=0}^{N-1} x(m) e^{-j2\pi km/N}, \quad k=0, 1, \dots, N-1 \quad \dots (3-12)$$

$$x(m) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi km/N}, \quad m=0, 1, \dots, N-1 \quad \dots (3-13)$$

Izraz 3-12 predstavlja Fourierovu transformaciju, gdje je $X(k)$ amplituda spektra na frekvenciji $f=2\pi km/N$. Izraz 3-13 jest inverzna Fourierova transformacija. Najčešće se uzima da je N potencija broja 2, čime se (3-12) i (3-13) mogu programski realizirati kao FFT algoritmi (Papamichalis, 1987). Klasična uska i

široka sonografska analiza se s pomoću FFT transformacije realizira izborom adekvatnog broja N (točaka) FFT analize i trajanja govornog segmenta (Δt).



sl. 3-3 FFT sonogoramski prikaz a) PCM signal b) $N=32$, $\Delta t=3, 2$ ms c) $N=128$, $\Delta t=25, 6$ ms.

Na slici 3-3 prikazana je FFT analiza riječi "programirati" pri b) $N=128$ i c) $N=32$ i upotrebom Hammingova prozora.

Osim FFT analize, za sonogramski prikaz često se upotrebljavaju i FIR (Finite Impulse Response) filtri, koji se mogu specificirati kao filtri propusnici opsega, niskopropusni ili visokopropusni filtri.

Period osnovnog tona (engl. Pitch period) predstavlja periodu harmonijskog, tona koji nastaje izgovaranjem vokaliziranih glasova. Često se obilježava i kao F_0 (u smislu 0-tog formanta) te predstavlja značajni prozodijski parametar govora. Nažalost, precizno ustanovljavanje F_0 u općem slučaju nije moguće zbog obezvučavanja na pojedinim prijelazima glasova ili pri kraju izgovora, pa su razvijeni mnogi algoritmi (Schafer, 1979), (Stamenković i Bakran, 1989) koji sa više ili

manje uspjeha detektiraju stvarni iznos F_0 . U ovom radu upotrijebljen je nešto izmijenjen AUTOC algoritam koji se temelji na centralnom ograničenju signala i funkciji autokorelacije (ACF - Autocorrelation Function). Centralno ograničenje signala jest linearna transformacija digitaliziranog signala i svrha joj je razbiti formantnu strukturu glasa te istodobno zadržati informaciju o periodičnosti. Matematička formulacija centralnog ograničenja dana je izrazom (3-14).

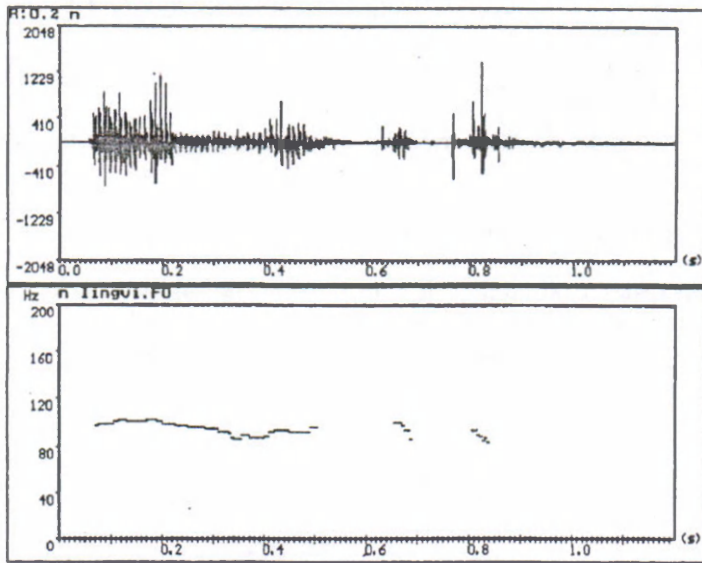
$$y((s(n))) = \begin{cases} s(n) - C & \text{za } s(n) < C \\ s(n) + C & \text{za } s(n) \geq C \\ 0 & \text{inače} \end{cases} \quad \dots \quad (3-14)$$

Digitaliziran uzorak $s(n)$ nakon centralnog ograničenja poprima vrijednost $y(s(n))$. Prag odsijecanja C iznosi 0.0-0.7 maksimalne amplitude signala unutar segmenta od $N/3$ uzoraka. Nakon primjene izraza (3 ← 14) izračunava se autokorelacija :

$$R(k) = \sum_{n=0}^{N-1} y(n) y(n+k), \quad k=0,1,\dots,K \quad \dots \quad (3-15)$$

Autokorelacija periodičnog (vokaliziranog) signala imat će izrazite maksimume, koji će se ponavljati periodom izvornog signala i upravo će predstavljati period osnovnog toga T , odnosno $F_0 = 1/T$. Nevokalizirani segmenti davat će autokorelacijsku funkciju bez naglašenih maksimuma, tj. amplituda će gotovo uvijek bit ispod vrijednosti $C R(0)$, gdje je C konstanta oko 0.3. Znači, period osnovnog tona određen je prvim indeksom k koji označava prvi lokalni maksimum takav da je $R(k)C > R(0)$. Broj uzoraka N predstavlja širinu autokorelacije i njegov izbor ovisi o donjoj frekvenciji F_0 , koja se želi analizirati. Radi univerzalnije primjene, često se uzima segment trajanja cca 30 ms ($F_0=33.3$ Hz), što znači da pri fekvenciji uzorkovanja od 10 KHz broj uzoraka N treba biti 300. Nakon izračunavanja autokorelacije svih uzoraka, pristupa se korekciji lažnih vrijednosti u toku višestrukih postupaka gladenja (engl. smoothing), provjere kontekstnih i drugih poznatih uvjeta vezanih za prirodu F_0 .

U posljednjih desetak godina sve se više upotrebljava tzv. LPC analiza govora (Linear Prediction Code), koja je izvorno motivirana rješavanjem problema komprimiranog telekomunikacijskog prijenosa digitalizirnoga govora. Daljnom matematičkom razradom LPC modela otkrivena je neposredna veza između tradicionalnih formantnih frekvencija te vrijednosti i broja tzv. parametara LPC analize čime LPC lagano potiskuje tradicionalnu sonografsku analizu. Suštinu LPC analize čini izračunavanje koeficijenata predikcije $a(j)$ (osnovnih LPC parametara):



sl. 3-4 a) PCM signal "lingvistika" b) F0

$$s(n) = \sum_{j=1}^k a(j) s(n-j) \quad \dots (3-16)$$

Na osnovi k posljednjih uzoraka i s pomoću koeficijenata predikcije izračunava se n -ti uzorak $s(n)$. Primjenom Z-transformacije (Rabiner et al., 1978) izraz (3-16) prijelazi u:

$$H(Z) = \frac{1}{1 - \sum_{j=1}^k a(j)Z^{-k}} \quad \dots (3-17)$$

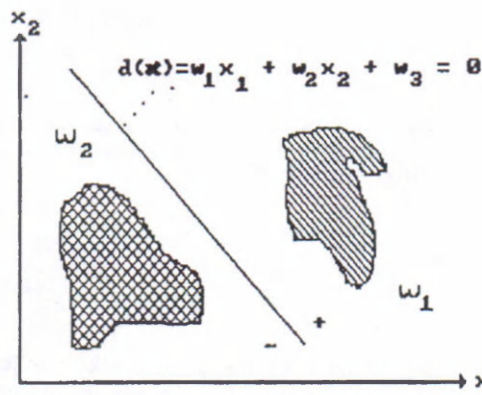
Uvrstimo li u brojnik konstantno pojačanja A , izraz (3-17) će predstavljati vremenski promjenljiv filter, čiji frekvencijski odziv možemo interpretirati kao prijenosnu funkciju vokalnog trakta. Algoritmom transformiranja koeficijenata predikcije (Witten, 1982) u tzv. koeficijente refleksije neposredno se izračunava formantna karakteristika signala. Minimalni broj potrebnih LPC parametara za specifikaciju n formantata iznosi $2n+1$.

4. MODELI PREPOZNAVANJA GOVORA

Različiti pristupi rješavanju problema APG uvjetovali su razradu mnogobrojnih modela analize i prepoznavanja govora. U ovom poglavlju prikazani su reprezentativni modeli već prihvaćene kategorijazcije pristupa prepoznavanju govora.

4. 1 Inženjersko-tehnički modeli

Inženjersko-tehnički modeli sistema za APG tretiraju uzorke govornog signala kao objekte u nekom prostoru klasifikacije bez fizikalne interpretacije njihova značenja. Općenito, problem se postavlja u definiranju funkcije (krivulje) diskriminacije među klasama objekata

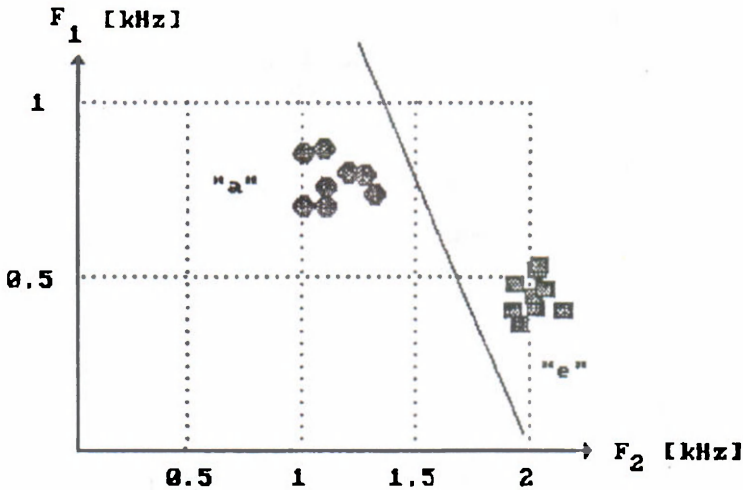


sl. 4. 1-1 Jednostavna diskriminacijska funkcija za dvije klase uzoraka

Pretpostavimo da postoje samo dvije klase objekata (ω_1 i ω_2) kao što je prikazano na slici 4. 1-1. Neka je dana jednačba linije koja razdvaja ω_1 i ω_2 :

$$d(x) = w_1 x_1 + w_2 x_2 + w_3 = 0 \quad \dots (4.1-1)$$

Oznake ω_1 , ω_2 i ω_3 jesu parametri, x_1 i x_2 koordinate varijabli dok $d(x)$ skraćeno predstavlja udaljenost $d(x_1, x_2)$. Sa slike 4. 1-1 vidi se da će svaki uzorak x (predstavljen koordinatama x_1 i x_2) iz klase ω_1 nakon uvrštavanja u 4. 1-1 rezultirati pozitivnu vrijednost $d(x)$. Isto tako, uzorci iz klase ω_2 leže na negativnoj strani od demarkacijske linije, tako da nakon uvrštavanja tih uzoraka $d(x)$ postaje negativno. Dakle, na osnovi predznaka od $d(x)$ dobivamo pravilo za pripadnost klasama, pa kažemo da uzorak x pripada klasi ω_1 ako $d(x) > 0$ odnosno pripada klasi ω_2 ako $d(x) < 0$. Za uzorke koji leže na demarkacijskoj liniji vrijednost $d(x) = 0$. Ovaj trivijalni primjer klasifikacije možemo ilustrirati usporedbom dvaju vokala, npr. "a" i "e" u ravni $F_1 - F_2$, što je prikazano na slici 4. 1-2.



sl. 4. 1-2 Razdvajanje vokala "A" i "E" (muški govornik).

Vrijednosti formanata dobiveni su FFT transformacijom (256 točaka, usrednjenih na širinu spektra od 340 Hz, Hammingov prozor) manualno segmentiranih odsječaka vokala digitaliziranih s 12 bita pri frekvenciji uzorkovanja 10 kHz.

Definiranje globalne diskriminacijske funkcije za M klasa (općenit slučaj) izvodi se s pomoću parcijalnih funkcija

$d_1(x), d_2(x), \dots, d(x)$ s osobinom da ako uzorak x pripada klasi ω_i tada:

$$d_i(x) > d_j(x), \quad j=1, 2, \dots, M, \quad j \neq i \quad \dots \quad (4.1-2)$$

Odluka o pripadnosti nekoj klasi donosi se na osnovi zadovoljenja parcijalnih uvjeta klasifikacije. Uzorak se tada dodjeljuje klasi čija diskriminacijska funkcija ima najveću numeričku vrijednost. Granica izmjeđu dviju klasa ω_i i ω_j , tada je dana sa:

$$d_i(x) - d_j(x) = 0 \quad \dots \quad (4.1-3)$$

Opisani način klasifikacije (prepoznavanja) u literaturi poznat je kao klasifikacija na osnovi odluke ("decision-theoretic approach"). Drugi model predstavlja tzv. sintaksičko prepoznavanje uzoraka ("syntactic approach"), gdje je funkcija udaljenosti nadomještena formalnom gramatikom, dok su objekti predstavljeni nizovima znakova koji čine jezik klase (Gonzalez i Thomson, 1978). Dakle, objekt pripada klasi ako je njegova nizovna interpretacija unutar jezika. Detaljnije o formalizmu generativne gramatike bit će objašnjeni u poglavlju 4. 2.

Doseg inženjersko-tehničkog pristupa jest prepoznavanje izoliranih riječi za unaprijed pripremljenog govornika. U iduća dva potpoglavlja bit će opisana dva najefikasnija modela za APG iz ove klase : dinamičko usklađivanje vremena (DTW - Dynamic Time Warping) i skriveni Markovljev model (HMM - Hidden Markov Model).

4. 1. 1 Dinamičko usklađivanje vremena (DTW)

Pretpostavimo da smo definirali rječnik od M riječi koje služe kao etalon. Svaka riječ Y neka je predstavljena s L obilježja:

$$Y_i = (y_1, y_2, \dots, y_L), \quad i=1..M. \quad \dots (4.1.1-1)$$

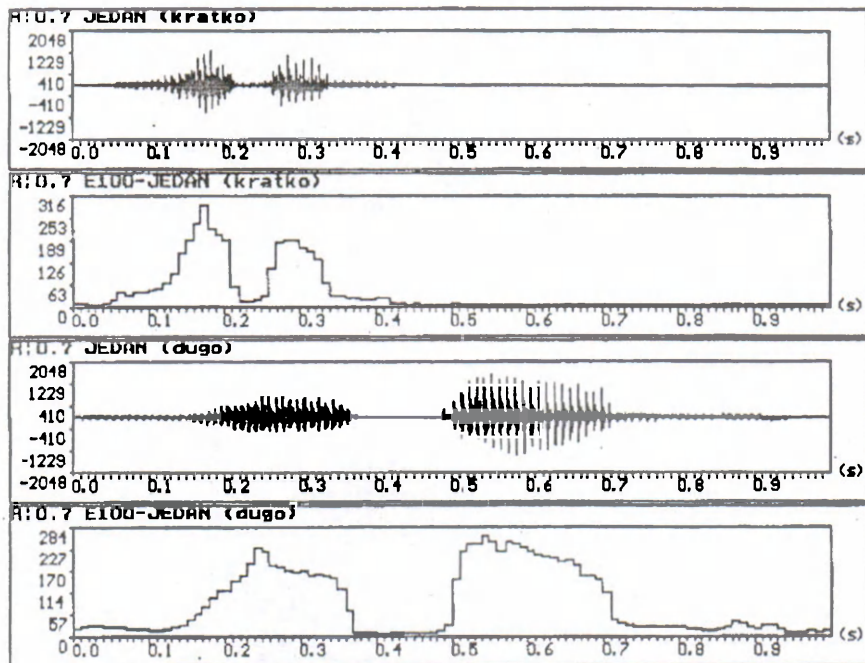
Vektor obilježja nepoznate riječi X jest:

$$X = (x_1, x_2, \dots, x_Z), \quad i=1..Z \quad \dots (4.1.1-2)$$

Proces prepoznavanja riječi X sada je određivanje indeksa riječi u rječniku s najmanjom udaljenošću :

$$j = \arg(\min_k | D(X, Y_k)) \quad \dots (4.1.1-3)$$

gdje je $D(X, Y_k)$ funkcija udaljenosti između k -te referentne i nepoznate riječi. Takav model prepoznavanja naziva se prepoznavanje prema etalonu (engl. template matching). Parametri y_1, \dots, y_k mogu biti kratkovremenska energija, FFT spektar, LPC koeficijenti, broj prolaza kroz nulu itd. Jedan od glavnih problema, koji se ovdje javlja, jest inherentna vremenska neusklađenost koja je posljedica različitog tempa izgovora riječi. To znači da dužina L vektora u etalonu i potrebna dužina Z izgovorene riječi mogu znatno varirati, a da je pritom izgovorena ista riječ. Da bismo ilustrirali taj efekt, promatrajmo sliku 4. 1. 1-1.

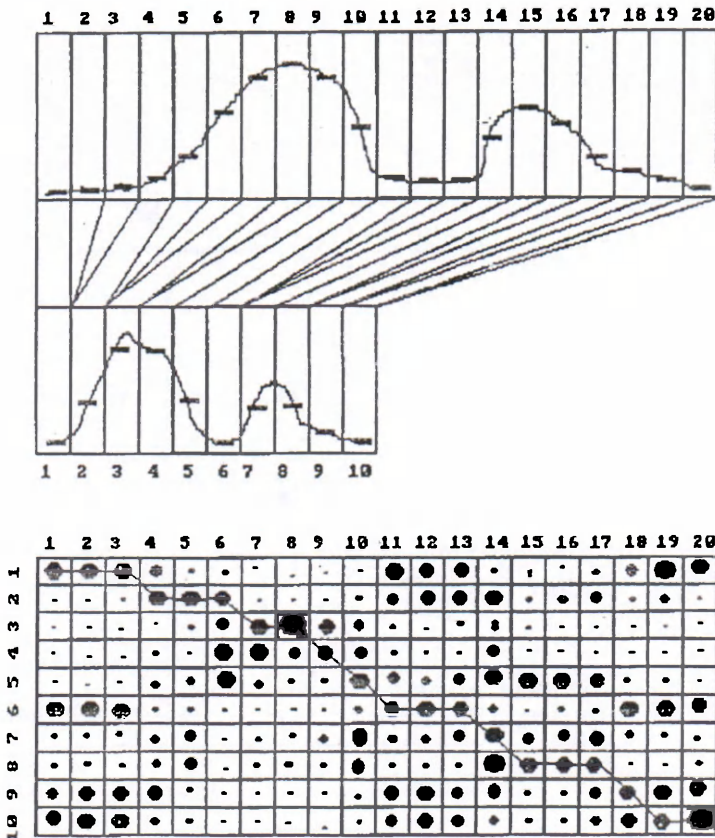


sl. 4. 1. 1-1 Riječ "jedan" izgovorena različitim tempom

Gornja slika prikazuje valni oblik i kratkovremensku energiju riječi "jedan", izgovorenu različitim tempom. Kratkovremenska energija izračunata je vremenskim prozorom od 10 ms. Označimo sa X i Y vektore kratkovremenske energije kratkog i dugog izričaja respektivno. Ako bismo definirali distancu između tih dviju riječi kao zbroj razlika energija u istim trenucima:

$$D(X, Y) = \sum_i |x_i - y_i| \quad \dots \quad (4.1.1-4)$$

odmah se vidi da mjera linearne distance nije adekvatna, jer se zbog različite vremenske pozicije ne uspoređuju ekvivalentni segmentni riječi (npr. cijela rječ "jedan" izgovorena kratko uspoređuje se s morfemom "je" u drugoj riječi, dok se morfem "dan" uspoređuje s tišinom). Izlaz iz ove situacije je, dakle, nelinearno usklađivanje indeksa vektora obilježja. Ideja vremenskog usklađivanja prikazana je na donjoj slici.

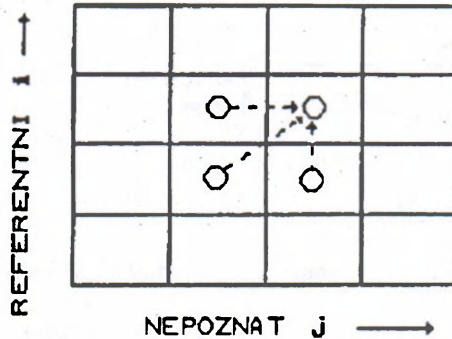


sl. 4. 1. 1-2 a) Nelinearna vremenska normalizacija b) matrica dinamičkog uspoređivanja.

Na slici 4. 1. 1-2 a) skicirane su krivulje kratkovremenske energije dugog i kratkog izričaja riječi "jedan" i segmentirane u ekvidistantnim razmacima. Debelim horizontalnim crticama označena je srednja razina energije u segmentu. Potupak određivanja distance vektora X i Y sada interpretiramo kao preslikavanje vremenskih osi nepoznatog uzorka na vremensku osi referentnog tako da odstupanja budu minimalna. Matrica preslikavanja ima dimenzije $L \times Z$ (za naš primjer $L=20$, $Z=10$). Za svaki element (i, j) u matrici preslikavanja definirana je tzv. lokalna distanca $d(i, j)$. Svrha nelinearnog usklađivanja vremena sada je pronalaženje puta kroz matricu tako da ukupna distanca bude minimalna. Kumulativna distanca najčešće se definira jednačbom dinamičkog programiranja (Nakanishi i Nakagava, 1987):

$$D(X, Y) = \min \sum_j d(u(j), j) \quad \dots \quad (4.1.1-5)$$

gdje je $u(j)$ funkcija usklađivanja. Za slučaj potpuno jednakih uzoraka optimalni put predstavljaće dijagonalu matrice dok u ostalim slučajevima oscilira oko nje (vidi sl. 4. 1. 1-2b). Kumulativna distanca je zbroj distanci između referentnog i testnog uzorka duž "najboljeg" puta. Prilikom kretanja kroz matricu definiraju se dopušteni pomaci npr.



sl. 4. 1. 1-3 Dozvoljeni pomaci duž dijagonale

što formalno zapisujemo :

$$g(i, j) = \min \left\{ \begin{array}{l} g(i-1, j) \\ g(i-1, j-1) \\ g(i, j-1) \end{array} \right\} + d(i, j) \quad \dots \quad (4.1.1-6)$$

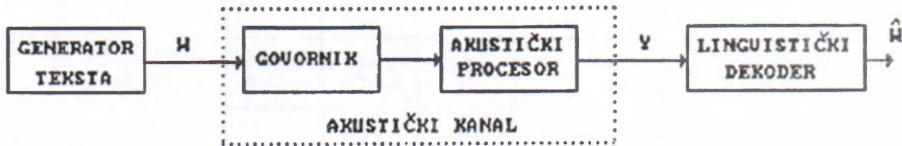
Nakon iterativnog postupka za $1 < I \text{ i } 1 < J$ ukupna distanca $D(X, Y)$ bit će jednaka $g(I, J)$. Takav način prepoznavanja govora poznat je pod nazivom dinamički usklađivanje vremena (DTW - Dynamic Time Warping), a elementi vektora obilježja najčešće predstavljaju FFT spektar.

Pretpostavka primjene DTW metode jest precizno određivanje početka i kraja izričaja, što znači da se on uspješno može primijenjivati samo za prepoznavanje izoliranih riječi. Drugi je nedostatak velika kompleksnost algoritam (broj računskih operacija), što otežava primjenu u realnom vremenu. U posljednjih nekoliko godina taj je problem riješen pojavom mikroprocссора specijalno namijenjenih za realizaciju DTW algoritama.

U poglavlju 4. 3. 2 o automatskoj segmentaciji govora opisan je način prepoznavanja vokala jednom varijantom DTW metode.

4. 1. 2 Skriveni Markovljev model (IIMM)

Proces prepoznavanja govora se u terminima teorije informacija može pojednostavljeno prikazati na sljedeći način:



sl. 4. 1. 2-1 Informacijsko-komunikacijski model APG

Govornik i akustički procesor predstavljaju kanal kojim se prenose simboli (tekst) prema lingvističkom dekoderu. Zadaća lingvističkog dekodera jest restaurirati originalnu poruku koja je podložna deformaciji zbog buke u kanalu. Akustički procesor (AP) funkcionira kao kompresor podataka koji transformira govorni valni oblik u niz parametar vektora, nakon kojih slijedi klasifikator uzoraka. Klasifikator uzoraka na svom izlazu daje povorku etiketa iz nekog konačnog alfabeta. Akustički procesor obično je vremenski sinkroniziran, tj. podaci na njegovom izlazu dostupni su u ekvidistantnim intervalima. Izlaz iz AP predstavlja, dakle, niz (vektor) \underline{Y} na osnovi kojeg lingvistički dekoder formulira očekivanu rječ w za originalnu \underline{W} . Da bi se minimizirala pogreška očekivane rječi, W treba biti odabran tako da:

$$P(\hat{w}|\underline{y}) = \max_{\underline{w}} P(w|\underline{y}) \quad \dots \quad (4.1.2-1)$$

Na osnovi Bayesovog pravila imamo:

$$P(w|\underline{y}) = \frac{P(w)P(\underline{y}|w)}{P(\underline{y})} \quad \dots \quad (4.1.2-2)$$

Budući da $P(\underline{y})$ ne ovisi o w , maksimiziranje $P(w|\underline{y})$ ekvivalentno je kao i maksimiziranje $P(w, \underline{y}) = P(w)P(\underline{y}|w)$. Ovdje je $P(w)$ apriorna vjerojatnost da će tekst generator producirati niz (rječ) w , a $P(\underline{y}|w)$ je vjerojatnost da će akustički kanal (zapravo AP) transformirati rječ w u izlazni niz \underline{y} . Da bi se na dekoderu odredilo $P(w)$, mora se definirati probabilistički model izvora koji generira simbol w (tekst-generator). Isto tako izračunavanje $P(\underline{y}|w)$ zahtijeva postojanje probabilističkog modela akustičkog kanala koji će uračunati varijete izgovora (dinamika, tempo, boja glasa itd). Pod pretpostavkom da postoje modeli za tekst-

generator (odnosno model jezika) i akustički kanal, biti će moguće neposredno odrediti simbol w s najvećom vjerojatnošću. Radi jednostavnosti, pretpostavimo da A emitira simbole u diskretnim i ekvidistantnim intervalima. Tada će se on zajedno s akustičkim kanalom moći predstaviti kao model nekog stohastičkog (Paušc, 1974) procesa. U praksi se kao zadovoljavajuća aproksimacija kompleksnosti procesa pokazao stohastički proces kod kojega distribucija slučajne varijable X u momentu $t=t_n$ ovisi samo o vrijednosti x_{n-1} procesa u trenutku $t_{n-1} < t$ koji je poznat kao Markovljev proces. Specijalan slučaj diskretnog Markovljeva procesa, kada se radi o diskretnim parametrima, nazivamo Markovljev lanac i on je osnova HMM modela.

Probabilistička funkcija Markovljeva lanca (Paušc, 1974) jest stohastički proces generiran s dva međuovisna mehanizma: Markovljevim lancem, koji ima konačan skup stanja, i skup slučajnih funkcija, od kojih je po jedna dodijeljena svakom stanju. U diskretnim trenucima vremena, proces se nalazi u nekom stanju i pritom je generiran neki simbol koji odgovora tekućem stanju. U sljedećem trenutku, Markovljev lanac mijenja stanje u skladu s matricom vjerojatnosti promjene stanja. Promatrač sa strane jedino "vidi" izlaz slučajnih funkcija združenih svakom stanju, ali ne može neposredno otkriti stanje Markovljeva lanca. Odatve potiče i naziv "Skriveni" Markovljev lanac (model). Danas je razvijeno više tipova HMM modela, a ovdje ćemo prikazati samo osnovni oblik $M=(Q, A, B)$, gdje su:

$Q = \{q_0, \dots, q_N\}$, gdje je q_0 početno a q_M završno stanje

$A = \{a_{ij}\}$ skup prijelaza gdje je a_{ij} vjerojatnost pijrelaza iz stanja i u stanje j .

$B = \{b_{ij}(k)\}$ Matrica izlaza: $a_{ij}(k)$ označava vjerojatnost emitiranja simbola kada se prelazi iz stanja i u stanje j .

Promatramo konačni uređeni niz (vektor) pojava:

$$O = O_1 O_2 O_3 \dots O_T \quad \dots \quad (4.1.2-3)$$

gdje je svaka pojava diskretnan simbol dobiven iz konačnog alfabeta, a njihova ukupnost predstavlja neku rječ w_i iz skupa riječi W , koje emituje izvor X . Pretpostavimo da postoji v definiranih HMM modela $M_1 \dots M_V$. Za svaki model može se izračunati vjerojatnost $\pi_i = P(O/M_i)$ za $1 \leq i \leq V$. Nakon toga rangiramo vjerojatnosti i klasificiramo nepoznatu riječ kao

$$w = w_i \Leftrightarrow \pi_i \geq \pi_j, \quad \text{za } 1 \leq j \leq V.$$

Naravno, prije nego se izračunaju ukupne vjerojatnosti, moraju biti određeni parametri za svaki od V HMM modela M , što se najčešće realizira algoritmom "naprijed-nazad" (Levinson, 1983). Iako je u osnovnoj verziji HMM model namijenjen za prepoznavanje izoliranih riječi, danas je predloženo više uspješnih modifikacija za prepoznavanje kontinuiranoga govora. S aspekta točnosti prepoznavanja DTW i HMM približno su isti, osim što HMM zauzima cca 10 puta

manje memorije računala i zahtijeva oko 20 puta manje računskih operacija (Rabiner et al., 1983).

4. 2 Modeli prepoznavanja govora bazirani na strojnom učenju

Izorna motivacija za istraživanje područja strojnog (automatskog) učenja bila je potreba za novim načinom programiranja računala ("kriza softverskog inženjerstva") kako bi se skratilo vrijeme obrade sve složenijih informacija te postavio nov odnos na relaciji komunikacije čovjek-računalo. Istraživanje strojnog učenja znatno je pridonijelo boljem razumijevanju prirode procesa učenja i predloženi su funkcionalni modeli zaključivanja i stjecanja novih znanja. Danas postoji mnoštvo klasifikacija i pristupa automatskom učenju (Michalski et al., 1984) npr. na osnovi strategije učenja: učenje memoriranjem (engl. Rote learning), učenje na osnovi rečenog (Learning by Being Told), učenje prema analogiji (Learning by Analogy), učenje na osnovi primjera (Learning from Examples), učenje samostalnim otkrivanjem (Learning from Observation and Discovery); prema formi naučenog znanja: strukturalno (znanja se neposredno mogu interpretirati) i statističko (znanja su kodirana na implicitan način) itd. Mehanizmi predstavljanja znanja u toku učenja također su raznovrsni: parametri algebarskih izraza, stabla odlučivanja, formalne gramatike, produkcijskih pravila, formalna logika, grafovi i mreže, semantički okviri (engl. frames), konfiguracije neuronskih mreža itd.

Detaljnije proučavanje automatskog učenja prelazi okvire ovog rada, a ovdje će biti prikazana dva modela učenja s aspekta moguće primjene u sistemima za prepoznavanje govora (prepoznavanje izoliranih vokala).

4. 2. 1 Eksplicitno učenje

Osnovna karakteristika mehanizama za eksplicitno (strukturalno) učenje jest da je naučeno znanje u obliku razumljivom čovjeku. Ovdje upotrijebljeni algoritam CORAL (Zagoruiko et al., 1985; Miškovic, 1989), spada u klasu induktivnih metoda učenja na osnovi primjera (konceptcija formiranja pojmova), a problem koji rješava može se opisati kao:

ZADANO: Skup $L = \{ l_1, \dots, l_M \}$ od M primjera opisanih skupom od N atributa. Svaki primjer pripada jednoj od K klasa.

NAĆI: Generalizirano distinktivno pravilo, za zadane primjere koje je upotrebljivo za klasificiranje novih primjera.

Objekt X prikazujemo skupom atributa $X = \{ x_1, \dots, x_N \}$ kojima je pridružen skup mogućih vrijednosti - $DOM(x_i)$. Ovisno o utjecaju poretka atributa na opis objekata razlikuju se:

- nominalni atributi (redoslijed atributa nije bitan)
- linearni atributi (vrijednosti predstavljaju ureden skup, odnosno prethodno definiran vektor obilježja)
- strukturirani atributi (atributi su parcijalno uređeni, najčešće hijerarhijski)

Algoritam CORAL predviđen je za rad s nominalnim i linearnim atributima. Globalna strategija algoritma je selekcija minimalnog broja informativnih

atributa koji obuhvaćaju što više objekata iste klase i istodobno što je moguće manje objekata iz drugih klasa. Opis klase predstavlja skup konjuktivnih formi od kojih svaka opisuje određen podskup primjera iste klase. Opći oblik distinktivnih pravila je dakle:

$$(P_1 \text{ and } \dots \text{ and } P_n) \text{ or } \dots \text{ or } (P_z \text{ and } \dots \text{ and } P_{z+k}) \quad \dots \quad (4.2.1-1)$$

gdje su P binarne konjunkcije oblika:

$$P_i = X_i \in [r_1, r_2] \quad \dots \quad (4.2.1-2)$$

Detalji o postupu formiranja pravila (4.2.1-1) i (4.2.1-2) dani su (Zagoruić et al., 1985), a ovdje će biti opisani samo ulazni objekti i bit će komentirani rezultati.

U eksperimentu prepoznavanja vokala sudjelovala su 3 muška govornika, koji su 14 puta izgovorili svaki vokal. Nakon digitalizacije (12-bitni A/D konvertor, frekvencija uzorkovanja 10 kHz) izračunata je FFT transformacija (Hammingov prozor) tako da je svaka klasa vokala bila predstavljena s 42 objekta. Radi usrednjavanja spektralne karakteristike, svaki objekt dobiven je kao srednja vrijednost FFT transformacije (32 spektralne vrijednosti) tri uzastopna segmenta trajanja 6.4 ms. Dakle, svaki vokal predstavljen je linearnim atributima:

$$X = [A_1, \dots, A_i, \dots, A_{32}] \quad \dots \quad (4.2.1-3)$$

čije su vrijednosti A iznosi energije na frekvenciji:

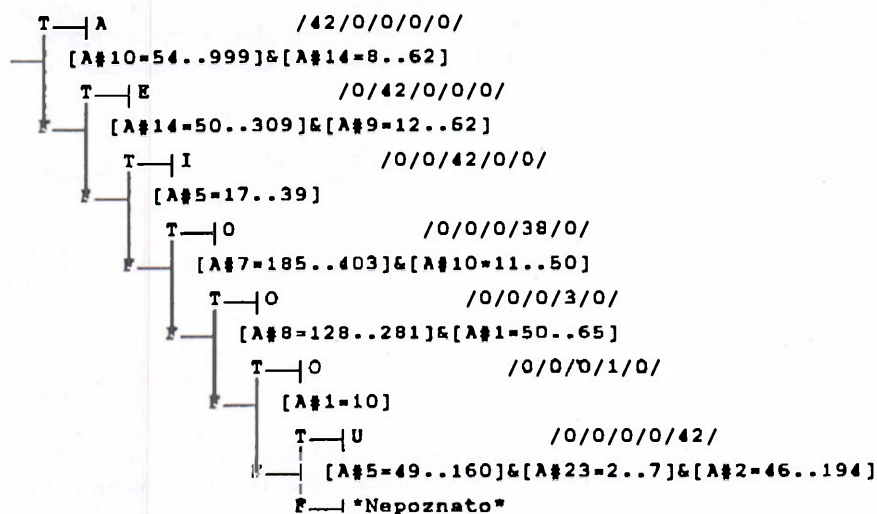
$$A_i = F_i = (i-1) \cdot 5/32 \text{ [kHz]} \quad \dots \quad (4.2.1-4)$$

Na slici 4.2.1.1-1 prikazano je klasifikacijsko stablo neposredno generirano programom CORAL (Mišković, 1989).

Uvjet za granjanje napisani su u razini čvora grananja. Desna kolona pokazuje broj uzoraka po klasama. Npr. dobiveno pravilo klasifikacije za vokal A glasi:

$$(A_{10} \in [54, 99]) \text{ and } (A_{14} \in [8..62]) \quad \dots \quad (4.2.1-5)$$

i dovoljno je da jednoznačno klasificira taj vokal što je zapisano u obliku: /42/0/0/0/0/



sl. 4. 2. 1-1 Klasifikacijsko stablo vokala

Osim konjunkcije koja stoji uz pojedini čvor stabla, redosljed grananja uključuje i sve dodatne uvjete tako da pravilo za vokal "E" ima sljedeći oblik:

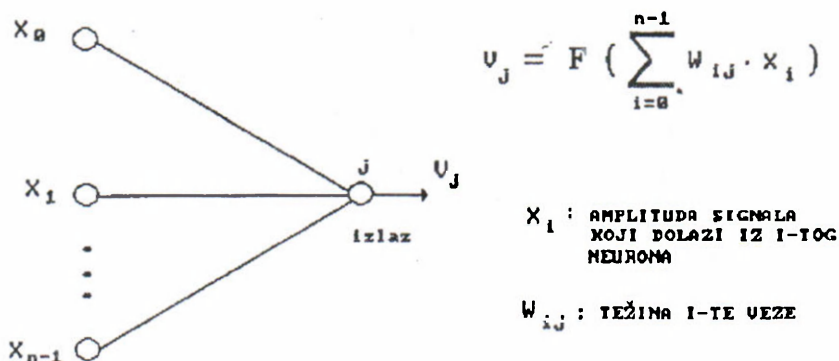
$$((\text{NOT} ((A_{10} \in [54, 99]) \text{ and } (A_{14} \in [8..62]))) \text{ and} \\ ((A_{14} \in [50, 309]) \text{ and } (A_9 \in [12..62])))$$

... (4.2.1-6)

Za navedene primjere objekti koji pripadaju klasi vokala "o" svrstani su u tri grupe. U prvu je svrstano 38 objekata, u drugoj 3, a u trećoj 1. Kao što se vidi iz stabla, klasifikacija je dijsunktna, tj. provedena je tako da se klase ne preklapaju (100 % točnost).

4. 2. 2 Implicitno učenje

Implicitno učenje obuhvaća one mehanizme gdje naučeno znanje nije lokalizirano na jednom mjestu, tj. ne može mu se dati fizikalni smisao (interpretacija). Algoritmi koji se koriste nekim od oblika te vrste učenja uglavnom se temelje na mnoštvu elementarnih međusobno povezanih procesora. Ovisno o prirodi modela mreže elementarni procesori u literaturi nazivaju se različito: ćelija, jedinica, neuron, perceptron itd. Topologija veza varira ovisno o upotrebljenom modelu učenja i definiranoj strukturi procesora. Motivacija za definiranje distribuirane obrade znanja jest imitacija rada ljudskog mozga, gdje je prirodni neuron najčešće formaliziran na sljedeći način (Yang et al., 1988):



sl. 4. 2. 2-1 Matematička prezentacija neurona

Neuroni imaju više ulaza (koji predstavljaju dendrite) i jedan izlaz (model za neurit) kojim su vezani za ulaze drugih neurona i šalju impulse aktivacije ili inhibicije. Neuroni koji dolaze u dodir s okolinom nazivaju se ulazno-izlazni (Input-Output) ili vidljivi (visible) dok su oni unutar mreže skriveni (Hidden). Organizacija skrivenih neurona može biti slojevita ili kompaktna ovisno o upotrijebljenom algoritmu učenja. Za slojevitou mrežu najčešće se koristi algoritam s povratnom propagacijom greške (Back Propagation Error) (McClelland i Rumelhart, 1988). Neki modeli računavaju i trenutni prag pobudnosti neurona koji se najčešće tretira kao veza s nekim jediničnim neuronom. U ovom radu bit će opisan eksperiment prepoznavanja izoliranih vokala konktivnim modelom poznatim u literaturi kao Boltzmannov stroj (BM-Boltzmann Machine) (Prager, 1988) (Stamenković, 1988.)

BM sastoji se od međusobno povezanih ćelija, koje mogu biti u stanju 0 ili 1 (model koji je ovdje upotrijebljen dopušta više od dva stanja, što će biti objašnjeno). Komunikacija s vanjskim svijetom ostvaruje se vidljivim ćelijama, koje imaju istu strukturu kao i skrivene. Globalni parametar mreže jest totalna energija:

$$E = - \sum_{i < j} w_{ij} s_i s_j + \sum_i b_i s_i \quad \dots \quad (4.2.2-1)$$

gdje su:

w_{ij} - veza između i-te i j-te ćelije

s_i - stanje i-te

s_j - stanje j-te ćelije

b_i - prag i-te ćelije

Ćelije se adaptiraju na okolinu tako što minimiziraju globalnu energiju. Razlika između k-te ćelije kada je ona u stanju 1 i 0 jest:

$$\Delta E_k = \sum_i (w_{ki} s_i) - b_k \quad \dots (4.2.2-2)$$

gdje su:

w_{ki} - veza između k-te i i-te ćelije

s_i - stanje i te ćelije

b_k - prag k-te ćelije

Promjena stanja k-te ćelije najčešće se definira stohastički:

$$P_k(s_k=1) = 1 / (1 + \exp(-\Delta E_k/T)) \quad \dots (4.2.2-3)$$

Varijabla T u gornjem izrazu ima fizikalnu interpretaciju "temperature" na kojoj se nalazi mreža. Mreža uči tako što se prilagođava okolini, koju predstavljaju stohastička funkcija gustoće stanja vidljivih ćelija. U prihvaćenom modelu, vidljive ćelije podijeljene su na dva skupa: I - ulazne ćelije i O - izlazne ćelije. U toku učenja, stanja ulaznih ćelija određena su ulaznim vektorom (koji parametarski predstavlja izgovoreni vokal), a na izlazne ćelije postavlja se vektor koji predstavlja kodirani odgovor. U toku prepoznavanja za nepoznati ulazni vektor očekuje se distribucija stanja izlaznih ćelija koja opisuje izgovorenu vrstu vokala. Jedan ciklus učenja (LC - Learning Cycle) sastoji se od tri koraka:

1. Određivanje p .
2. Određivanje p' .
3. Ažuriranje veza u funkciji od p i p' .

U prvom koraku a ulazne vidljive ćelije postavi se FFT vektor, dok su izlazne slobodne (bez forsiranja odgovora). BM se tada "zagrije" i "ohladi" do termičke ravnoteže te izračuna vjerojatnost p_{ij} da se i-ta i j-ta ćelija nalaze u stanju 1. Nakon toga, na izlazne ćelije priključuje se odgovor i ponavlja postupak. U trećem koraku ažuriraju se veze u funkciji od izračunate p i p' . Budući da je u realiziranom modelu (radi smanjenja broja ćelija) dopušteno da ćelija ima 16 stanja, vjerojatnosti p i p' izračunavaju se na sljedeći način:

$$p_{ij} = \sum_{\text{trening}} s_i s_j \quad \dots (4.2.2-4)$$

$$p'_{ij} = \sum_{\text{prepoznavanje}} s_i s_j \quad \dots (4.2.2-5)$$

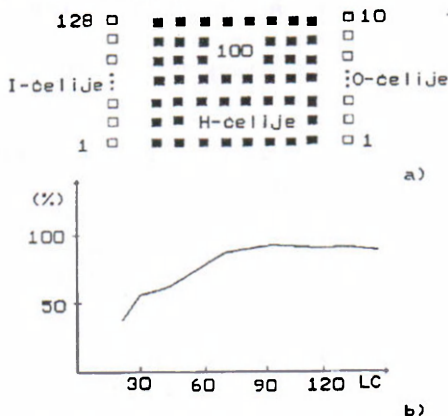
Detaljan opis algoritma učenja dan je u (Stamenković, 1988d) a ovdje ćemo još prikazati način obrade vokala i komentirati rezultate.

Bazu signala sačinjavalo je 20 uzoraka za svaki vokal (3 muška i 2 ženska govornika koji su izgovorili vokale po 4 puta). Frekvencija uzorkovanja bila je 10 KHz sa 8-bitnim A/D konvertorom. Centralni segment vokala u trajanju 25.6 ms automatski je ekstrahiran i podvrgnut FFT transformaciji. Amplitude spektra dobivene FFT transformacijom normalizirane su od 0 do 15 (16 razina) i predstavljale su element 128-dimenzionalnog ulaznog vektora za BM. Definirani vektori za kodiranje vokala bili su sljedeći:

vokal	vektor
I	0000110000
E	0001001000
A	0010000100
O	0100000010
U	1000000001

sl. 4. 2. 2-2 Kodirani izlazni vektori vokala

Vokali su kodirani prema kriteriju razdvojenosti formanata (prvih 5 bita) i prema poziciji momenta spektralne energije (drugih 5 bita). Struktura upotrijebljene BM i brzina učenja prikazani su na slici 4. 2. 2-3



sl. 4. 2. 2-3 Upotrijebljena BM a) Brzina učenja b)

BM u eksperimentu imala je 128 ulaznih ćelija čime je kodiran normaliziran (na 16 diskretnih razina koje predstavljaju stanja ulaznih ćelija) FFT spektar (128 točaka), 100 skrivenih i 10 izlaznih ćelija. Skrivenice ćelije bile su povezane međusobno, dok su vidljive (ulazne i izlazne) vezane samo za skrivenice.

Upotrijebljen konektivni model pokazao je zadovoljavajuće rezultate u prepoznavanju kvazistacionarnih segmenata govora. Osim zamjerke da se ne vidi na osnovi čega je donešena odluka o vrsti vokala, nedostatak BM (kao i svih konektivnih modela) predstavlja kompleksnost algoritma jer jedan ciklus učenja na PC AT računalu (8 MHz) trajao je oko 20 minuta.

4. 3 Fonetsko-lingvistički modeli

Za razliku od inženjersko- tehničkog pristupa prepoznavanju govora kod koga se govorni signal isključivo tretira kao realizacija nekog slučajnog procesa ili kao specifičan klasifikacijski postupak, fonetsko-lingvistički model implicitno uključuje obradu koja se proteže kroz više razina. Ako primijenimo tradicionalnu teoriju percepcije govora, tada možemo definirati 5 razina automatske obrade govora, koje su prisutne u fonetskom lancu (Lincard, 1977):

1. **fizička razina:** Govor se prenosi u obliku longitudinalnih valova i uzrokuje promjenu pritiska na ulaznom senzoru (mikrofonu), koji te promjene pretvara u naponske impulse. Impulsi se pojačavaju, filtriraju, digitaliziraju, a nakon toga izračunavaju se parametri obilježja.

2. **fonetska razina:** Na osnovi definirane metrike klasificiranja govorni segmenti svrstavaju se u pojedine fonetske grupe glasova, koje nužno ne moraju biti disjunktne.

3. **leksička razina:** Grupirani fonetski elementi prolaze kroz kontekstne provjere (npr. fonotaktička pravila) i provjere o mogućim kombinacijama nizova glasova (postojanje riječi u rječniku).

4. **sintaksička razina:** Nakon što se postavi lista mogućih riječi sintaksičkom se analizom provjerava korektnost ponuđenih sintaksičkih struktura.

5. **semantička razina:** Na osnovi prozodijskih elemenata (akcent, intonacija itd.) i semantičkog modela razumijevanja jezika određuje se globalna priroda analizirane govorne poruke.

Pojedini autori (Lea, 1980) definiraju i šestu razinu govorne komunikacije vezanu za pragmatiku prirodu govornog priopćavanja. Međutim, zbog izuzetne složenosti pragmatike informacije, koja je zapravo skupnost dotadašnjeg iskustva i broj percepcija svih čula, ne postoji adekvatna formalna interpretacija te razine. Zbog toga je nećemo razmatrati. Iz istih razloga semantička razina bit će analizirana samo djelomice.

Obrada govora na fizičkoj i (djelomice) fonetskoj razini preklapa se s već prikazanim načinima segmentiranja i izdvajanja vektora obilježja signala. Međutim, leksička i slijedeće razine ne odnose se više na fizikalne pojave, nego na manipulaciju simbolima. Zbog toga se matematički aparat, kojim se definiraju parametri modela na tim razinama u mnogome razlikuje od onog s prve dvije razine. Za opis simboličkih razina najčešće se upotrebljava teorija formalnih jezika (FJ) (Aho i Ullman, 1972) čije su osnovne definicije dane u sljedećem poglavlju. Nakon što su definirani uvodni pojmovi teorije FJ, u poglavlju 4. 3. 2 prikazana je hijerarhijska segmentacija govora do fonetske razine. Model sim-

boličke obrade na leksičkoj i sintaksičkoj razini prikazan je u poglavlju 4. 3. 3, gdje je opisan APG sistem za prepoznavanje povezanoga govora, koji se kao akustičko-fonetski processor koristi načinom segmentiranja iz poglavlja 4. 3. 2.

4. 3. 1 Osnovne postavke formalnih jezika

Polazni entitet u teoriji FJ jest znak. Međutim, za razliku od mjesta koji znak ima u semiologiji (tj. gdje se znak uvijek promatra kroz proces semjoze - odnosa između nosioca znaka, interpretanta i interpretatora (Moris, 1975), ovdje uzimamo samo intuitivnu predodžbu o njemu. To znači da se pojam znaka tretira jednako kao i npr. pojam točke u Euklidovoj geometriji. Najčešće, znaci predstavljaju grafeme nekog jezika, brojeve itd. Primjerice, znakovi jesu: 4, 7, (, a, d, , =, _ , ? , x, y, , . Proizvoljan konačan skup znakova x nazivamo alfabet i označavamo ga :

$$A = \{x_1, x_2, \dots, x_n\}.$$

Napišemo li znakove jedan do drugog tako da čine cjelinu, tada govorimo o simbolu (u literaturi se često susreću i nazivi string, riječ, rečenica koji će se i ovdje ravnopravno upotrebljavati). Broj znakova u simbolu predstavlja njegovu dužinu, tj.

$$|\alpha| = |x_1 x_2 x_3 x_4 \dots x_n| = n$$

Prazan simbol (string) jest simbol koji ne sadrži nijedan znak. Njegova je dužina 0 a označavat ćemo ga s λ tj. $|\lambda| = 0$.

Analogno potenciranju uvodi se skraćeni zapis repeticije znakova unutar stringa npr: $a^0 = \lambda$, $aaa = a^3$, $abbbbc = ab^4c$. Kombiniranje stringova dozvoljava nam relacija konkatencije.

DEF 4. 3. 1-1: konkatencija stringova

Neka su α i β stringovi nad alfabetom A tj.

$$\alpha = x_1 x_2 x_3 \dots x_n$$

$$\beta = y_1 y_2 y_3 \dots y_m$$

Konkatencija stringova α i β jest string γ dobiven na sljedeći način:

$$\gamma = \alpha \beta = x_1 x_2 \dots x_n y_1 y_2 \dots y_m = z_1 z_2 \dots z_{n+m}$$

Za razliku od alfabeta koji je konačan skup, skup svih stringova nad alfabetom, svakako je beskonačan. Zvezdica iznad oznake skupa npr. $L \leq A^*$ predstavljat će oznaku za tzv. zatvarač alfabeta, skup koji sadrži sve simbole (uključivši i λ) nad alfabetom A npr:

$$A = \{a\}, \quad A^* = \{\lambda, a, aa, aaa, aaaa, aaaa, \dots\} = \{a^n, n \geq 0\}$$

Formalni jezik L sačinjavaju rečenice iz skupa A^* , što znači da je $L \subseteq A^*$. String α , koji je sadržan unutar stringa β , nazivamo podstring stringa β .

U toku simboličke obrade jedan niz znakova (rečenica) preobličuje se u drugi. Od mnogih mehanizama preoblika spomenimo FST pretvarač (Finite State Transducer), koji će biti upotrijebljen na leksičkoj razini obrade govora (poglavlje 4. 3. 3):

DEF 4. 3. 1-2 FST pretvarač konačnog stanja

Deterministički pretvarač konačnog stanja jest 6-orka

$M = \{Q, \Sigma, \Delta, \delta, q_0, F\}$ gdje su:

$Q = \{q_0, q_1, q_2, \dots, q_n\}$	- konačni skup stanja
$\Sigma = \{a_1, a_2, a_3, \dots, a_m\}$	- ulazni alfabet
$\Delta = \{b_1, b_2, b_3, \dots, b_k\}$	- izlazni alfabet
$\delta : Q \times (\Sigma \cup \{\lambda\}) \rightarrow Q \times \Delta^*$	- funkcija prijelaza
$q_0 \in Q$	- početno stanje
$F \subseteq Q$	- skup završnih stanja

Na osnovi ulaznog simbola i internog stanja konačni pretvarač prelazi u slijedeće stanje i emitira novu etiketu (simbol). Obrada ulaznog niza etiketa obavlja se u toku smjene konfiguracija automata (q, x, y) , gdje su:

- 1) $q \in Q$, tekuće stanje
- 2) $x \in \Sigma^*$, preostali niz etiketa
- 3) $y \in \Delta^*$, emitirana etiketa

Definiramo relaciju \vdash (ili samo \vdash ako se pretvarač podrazumijeva) nad konfiguracijama pretvarača koju čitamo "prelazi u". Za sva stanja $q \in Q, x \in \Sigma^*$ takva da je $\delta(q, a) = (r, z)$ pišemo:

$$(q, ax, y) \vdash (r, x, yz)$$

Kažemo da je y izlaz za x ako je $(q_0, x, \lambda) \vdash^* (r, \lambda, y)$ pri $r \in F$. Zvezdica iznad oznake relacije označava ponavljanje proizvoljnog broja puta promjene konfiguracije pretvarača.

U teoriji FJ kažemo da je rečenica sintaksički ispravna ako pripada jeziku L i obrnuto. Budući da je jezik L obično beskonačan ili veoma velik skup, nužno je definirati neki konvergirajući mehanizam koji će provjeravati pripadnost neke rečenice jeziku. Jedno rješenje ovog problema jest formalizirana generativna gramatika Chomskog :

DEF 4. 3. 1-3: frazno-strukturirana gramatika $G=(N, P, S)$ je četvorka :

- 1) $N = \{ n_1, \dots, n_k \}$ - skup neterminalnih simbola
- 2) $\Sigma = \{ G_1, \dots, G_m \}$ - skup terminalnih simbola $N \cap \Sigma = \{ \}$
- 3) P - skup produkcionih pravila
 $P \subseteq (N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$
 $P = \{ (u,v) \mid u \in (N \cup \Sigma)^* N (N \cup \Sigma)^*, v \in (N \cup \Sigma)^* \}$
- 4) $S \in N$ - startni simbol

Elementi iz skupa N nazivaju se neterminalni simboli ili neterminali (ponekad se susreće naziv i lingvističke varijable), a elementi iz skupa Σ su terminalni simboli ili terminali (riječi). Treba naglasiti da su skupovi terminala i neterminala disjunktni. Skup P predstavlja skup produkcijskih pravila ili produkcija. Umjesto tradicionalne oznake za uređeni par, elementi (α, β) iz skupa P obično se prikazuju u formi produkcija :

$\alpha \rightarrow \beta$

Gornji prikaz čitamo : "alfa se producira u beta" . Produkciju oblika $\alpha \rightarrow \lambda$ nazivamo prazna produkcija ili λ - produkcija. Ovisno o obliku produkcijskih pravila gramatike su svrstane u 4 grupe (Aho, 1972) a najčešće su tzv. beskon-tekstne gramatike (CF, Context Free) koje imaju na lijevoj strani produkcije samo jedan neterminalni simbol.

Posljednji element gramatike S jest tzv. startni simbol od kojeg se polazi u procesu generiranja jezika. Startni element uvijek je neterminal. Za jezik L , koji je opisan gramatikom G , kažemo da je generiran gramatikom G i zapisujemo $L=L(G)$.

Dakle, rečenica α pripada jeziku L ako se može generirati gramatikom G . Generiranje jezika svodi se na smjenjivanje neterminalnih simbola drugim neterminalnim ili terminalnim simbolima. Sekvencu sukcesivne zamjene simbola nazivamo niz izvođenja dužine M , gdje je M broj zamjenjivanja simbola. Ako se prilikom svake zamjene zamjenjuje prvi s lijeva neterminalni simbol, za niz izvođenja kažemo da je derivacija s lijeva. Prijelaz niza simbola α u niz simbola α' označavamo $\alpha \Rightarrow \alpha'$. Primjerice, neka je dana gramatika $G = (\{ \sigma, \alpha, \beta, \delta, \gamma \} \cup \{ a, b, c, d \}, P, \sigma)$ sa skupom produkcija: $P = \{ \sigma \rightarrow a\alpha, \alpha \rightarrow \beta\alpha, \alpha \rightarrow \beta, \beta \rightarrow dc, \beta \rightarrow bb, \gamma, \gamma \rightarrow cc, \alpha \rightarrow \alpha \}$. Niz izvođenja $\alpha \Rightarrow a d c a$ je slijedeći

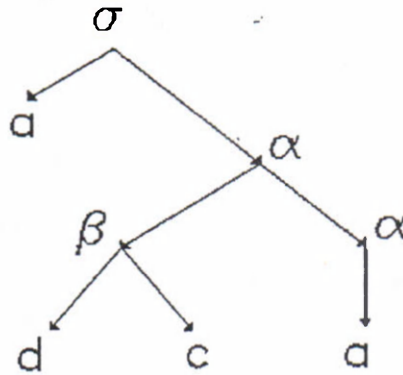
$\sigma \Rightarrow a\alpha \Rightarrow a\beta\alpha \Rightarrow a d c a \Rightarrow a d c a$

Derivacija s lijeva grafički se prikazuje stablom izvođenja:

DEF 4. 3. 1-4: stablo izvođenja. Neka je dana CF gramatika $G=(N, \Sigma, P, S)$. Stablo izvođenja nazivamo stablo (D, r) koje zadovoljava sljedeće uvjete:

- 1) Svaki vrh označen je simbolom iz $N \cup \Sigma$
- 2) Korijen stabla označen je sa S .
- 3) Ako je vrh n označen sa $A \in N$, a vrhovi $n_1 \dots n_k$ su direktni sljedbenici od n i označeni $X_1 \dots X_k$ respektivno, tada je $(A, X_1 \dots X_k) \in P$.
- 4) Ako je vrh n označen sa $A \in \Sigma$, tada je $Is(n)=0$.

Primjerice, stablo izvođenja za rečenicu iz prošlog primjera je:



sl. 4. 3. 1-1 Stablo izvođenja za rečenicu $\alpha = adca$

Budući da definicija 4. 3. 1-4 jednoznačno uspostavlja vezu između produkcija iz skupa P i strukture stabla izvođenja, to znači da za zadano stablo izvođenja možemo odrediti produkcije gramatike te niz izvođenja koji generira danu rečenicu.

4. 3. 2 Model automatske segmentacije govora na fizičko-fonetskoj razini

Automatska segmentacija govora koja bi uključivala lingvističke razine obrade govora mora biti vezana za određeni diskurs govora, koji se segmentira (definiranje leksike, sintakse, semantike). Budući da se na ovom mjestu daje samo koncepcija hijerarhijske obrade a ne specijaliziran sistem za segmentiranje konkretnog korpusa, hijerarhijska organizacija segmentiranja može se uspostaviti samo na fizičkoj i fonetskoj razini. Jedna moguća hijerarhija segmentiranja na tim razinama je sljedeća:

- A. segmentiranje prema kriteriju prisutnosti - - odsutnosti govora
- B. segmentiranje prema fonetskim kategorijama glasova
- C. segmentiranje unutar fonetskih kategorija

Segmentiranje se sastoji u dodjeljivanju nekog simbola (etikete) govornom segmentu s koji počinje u trenutku t i traje sekundi. Definirajmo sada alfabet etiketa kojima će se etiketirati govorni segmenti prema gornjoj hijerarhiji:

$$E = \{ ., \#, D, G, S, X, M, F, V, *, a, e, i, o, u \} \quad \dots \quad (4.3.2-1)$$

Fizikalno značenje definiranih etiketa dato je u tablici 4.3.2-1.

Etiketa	fizikalno značenje
.	tišina, odsustvo govornog signala
#	prisustvo signala
G	zvučni glasovi male energije (b,d,g)
S	silbilanti, frikativi male energije (s,c)
D	zvučni prekidni glasovi (dz,d)
X	bezvučni kompaktni glasovi (s,č,c)
F	bezvučni glasovi srednje energije (f,h)
M	nazali (m,n,nj)
V	vokali (a,e,i,o,u)
*	nedefiniran vokalizirani tranzijent
a e i o u	vokali (a,e,i,o,u)

tbl. 4.3.2-1 Etikete i njihova interpretacija

tbl. 4.3.2-1 Etikete i njihova interpretacija

Kod interpretacije etiketa treba biti obazriv, jer se ne smije zaboraviti da besprijekorna klasifikacija govornih segmenata nije moguća zbog kontinuirane prirode govora. Zato će predložene etikete pouzdano korelirati samo s općim karakteristikama govornog segmenta, koji se promatra, ali ne nužno i s pojedinim fonemima. Primjerice, zbog dezorganizacije glasova pri kraju izričaja treba očekivati etikete koje teže obezvučavanju i manjim iznosima energije.

Radi jednostavnijeg matematičkog modela te i zbog preporuka (Schafer i Markel, 1979) izabran je ekvidistantni segment govora kao jedinica segmentiranja s trajanjem $\tau = 10\text{ms}$. Znači, govorni signal s od N uzoraka predstavljamo kao zbroj disjunktivnih segmenata jednakog trajanja, tj. kao:

$$s = \sum_t s_{\langle t, t+\tau \rangle} \quad \dots \quad (4.3.2-2)$$

Niz etiketa segmenata možemo sada definirati kao rečenicu nekog jezika L , odnosno:

$$\alpha = \alpha_1 \alpha_2 \dots \alpha_j \dots \alpha_n \in E^* \quad \dots \quad (4.3.2-3)$$

Na prvoj razini segmentiranja (razina A) određujemo da li t-ti segment govora predstavlja govor ili tišinu. Segment s će se etiketirati etiketom ako su zadovoljene nejednadžbe :

$$E_t = \frac{1}{100} \sum_{i=1}^{100} |s(t+i)| < A_1 \cdot E_o \quad \dots (4.3.2-4)$$

$$ZC_t < A_2 \cdot ZC_o \quad \dots (4.3.2-5)$$

U izrazu 4.3.2-5 oznake E_o i ZC_o predstavljaju srednju energiju šuma i broj prolaza signala kroz nulu, respektivno, dok su E_t i ZC_t pripadne vrijednosti vezane za govorni segment koji počinje u trenutku t (i traje 10 ms, odnosno zahvata narednih 100 uzoraka). Konstante A_1 i A_2 ovise o načinu digitalizacije (broj bita A/D konvertora, karakteristike ulaznog filtra itd) i uvjetima okoline. Segmentima koji nisu etiketirani s ., dodjeljuje se # (oznaka prisustva govornog signala). Alfabet etiketiranja na prvoj razini jest:

$$E_1 = \{., \#\} \subset E \quad \dots (4.3.2-6)$$

Etiketiranje na višim razinama zahtijeva precizniji opis govornog segmenta. Zato svaki segment s opisujemo odabranim parametrima obilježja koje grupirano iskazujemo u formi vektora obilježja:

$$V_t = [x_1, x_2, x_3, x_4] \quad \dots (4.3.2-7)$$

Definirani elementi vektora obilježja su sljedeći :

- x_1 - kratkovremenska energija prema (3-4)
- x_2 - odnos snage višeg i niže opsega spektra
- x_3 - broj prolaza kroz nulu prema (3-8)
- x_4 - kvocijent kvadrata brojeva prolaza kroz nulu i kratkovremenske energije

Elementi vektora x i x izračunavaju se na sljedeći način :

$$x_2 = 1000 \frac{EH}{EL} = 1000 \frac{\sum_{j=2}^7 \text{FFT}(j)}{\sum_{j=2}^7 \text{FFT}(j)} \dots (4.3.2-8)$$

$$x_4 = 1000 \frac{ZC^2}{E^2} \dots (4.3.2-9)$$

gdje $\text{FFT}(k)$ predstavlja k -tu komponentu FFT spektralnog vektora od 8 točaka. Prosječne vrijednosti elemenata x_1, \dots, x_4 za sve glasove hrvatskosrpskog jezika, osim zabezvučne plozive (p, t, k) prikazane su na tabeli 4. 3. 2-2, gdje se pregledno vide globalne karakteristike pojedinih glasova. Primjerice, vjerojatno će zvučni glasovi b, d, g imati mali broj prolaza kroz nulu zbog izrazitog spektra u nižim frekvencijama i zbog dugog osnovog perioda. Iz istih razlog, ali s oprečnim vrijednostima, izdvajaju se šuškavi konsonanti č, š, ć itd. Kategorizacija tipa zvučni/bezvučni se za najveći broj glasova može ostvariti na osnovi kratkovremenske energije i odnosa energija u gornjem i donjem dijelu spektra. Budući da se vrlo malo glasova samo prema jednom kriteriju izrazito razlikuje od drugih glasova, pouzdanija i preciznija kategorizacija može se sprovesti samo na osnovi kombinacije elemenata ($x_1 \dots x_4$).

Formalizirajmo sada drugu razinu segmentacije, odnosno specificirajmo alfabet etiketa i uvjete etiketiranja. Dakle, na drugoj razini segmentacije (razina B) etikete iz skupa E transformiramo u etikete iz skupa. $E = \{ D, G, S, X, M, F, V, *, - \}$. Etiketne E segmenata razine A bile su dužine 1, jer je odluka o prirodni segmenta S_τ isključivala drugu (prisustvo/osdustvo govora). Međutim, na razini B etiketa S_τ bit će iz skupa E_2^* i predstavljati kompleksniji opis prirode govornog segmenta S_τ . Etiketa β_τ dobiva se konkatencijom mogućih elementarnih etiketa

$$\beta_\tau = b_1 b_2 \dots b_7 \quad b_i \in E_2 \quad \dots (4.3.2-10)$$

Uvjet $\psi_x(\tau)$ da bi se govornom segmentu koji počinje u trenutku τ i kojem je na prethodnoj razini etiketiranja (razina A) dodijeljena etiketa #, pridružila elementarna etiketa x , definiramo na sljedeći način :

$$\begin{aligned} \psi_x(\tau) : & (E_{\min} < E_\tau) \wedge (E_\tau < E_{\max}) \wedge \\ & (EHL_{\min} < EHL_\tau) \wedge (EHL_\tau < EHL_{\max}) \wedge \\ & (ZC_{\min} < ZC_\tau) \wedge (ZC_\tau < ZC_{\max}) \wedge \\ & (ZCE_{\min} < ZCE_\tau) \wedge (ZCE_\tau < ZCE_{\max}) \quad \dots (4.3.2-11) \end{aligned}$$

E	EH/EL	ZC	ZC^2/E^2	glas
684.40	7.35	18.57	1.24	a
53.98	10.80	5.50	11.43	b
24.63	388.31	24.67	981.60	c
97.16	3986.92	42.22	254.02	č
69.17	5780.00	48.00	625.78	ć
51.39	23.72	6.81	20.36	d
118.30	482.41	29.50	94.76	dž
74.67	840.52	17.33	145.61	d
402.22	39.61	11.47	0.99	e
86.32	42.74	14.80	49.50	f
56.22	20.19	8.50	27.62	g
155.42	14.01	25.61	68.52	h
243.11	346.72	15.36	5.26	i
115.88	20.20	7.02	5.54	j
190.48	2.59	9.50	2.99	l
93.90	53.23	9.75	14.62	lj
85.43	20.52	11.00	16.77	m
80.20	14.85	8.94	13.34	n
73.45	31.54	10.44	45.27	nj
641.51	6.03	13.59	0.57	o
349.95	5.22	14.67	3.26	r
63.82	296.74	20.20	169.06	s
126.09	1808.79	43.87	152.78	š
383.27	4.62	8.26	0.58	u
145.38	11.80	13.13	13.89	v
114.59	53.62	9.36	8.21	z
97.96	16.76	15.63	32.53	ž

gdje su : E_{τ} , EHL_{τ} , ZC_{τ} i ZCE_{τ} vrijednosti elemenata vektora obilježja prema 4. 3. 2-7 respektivno (govornog segmenta koji počinje u trenutku τ). Granične vrijednosti (E_{\min} , E_{\max} , EHL_{\min} , EHL_{\max} , ZC_{\min} , ZC_{\max} , ZCE_{\min} , ZCE_{\max}) dobivaju se višestrukim mjerenjem etalona koji su chstrahirani "ručnim" segmen-tiranjem (12 bitni AD/DA, frekvencija uzorkovanja 10 KHz).

Postupak etiketiranja na razini B možemo interpretirati kao prevodenje jezika $L_1 \in E_1 * u L_2 \in E_2 *$.

Posljednja razina etiketiranja (razina C) ovog modela uključuje klasificiranje unutar fonetskih kategorija. Zbog mnogih utjecaja na način izgovora i oblik govora (djelimice popisanih u poglavlju 2) na ovoj razini nij moguće pouzdano klasificirati sve glasove, pa je nužno odabrati određene podskupovne kategorije. U literaturi se uglavnom navode uspješna prepoznavanja izdvojenih vokala (Stamenković, 1981; Prager, 1988) pa su vokali i ovdje uzeti kao fonetska kategorija, ali će prepoznavanje biti unutar povezanoga govora. Dakle, segmen-tiranje na toj razini jest transformacija etiketa V u neku od etiketa iz skupa $E_3 = \{ a, e, i, o, u \}$ odnosno prevodenje iz jezika E_2 u jezik $\{ \{ E_2 \cup \{ a, e, i, o, u \} \} / \{ v \} \} *$. Za govorni segment etiketiran sa V izračunava se 32-kanalna FFT transformacija, tj. iz segmenta s izuzima se prvih 64 uzoraka. Prije izračunavanja FFT uzorci se množe Hammingovim prozorom (izraz.3-10). Odluka o tipu vokala (a, e, i, o, u) donosi se na osnovi minimalne udaljenosti definirane metrike prema položajuformanata i prema globalnoj spektralnoj razlici. S obzirom na in-herentnu frekvencijsku varijabilnost vokala, ne mogu se uspješno primijeniti linearne metrike oblika :

$$D(Y_r, Y_x) = \sum_i |Y_r(i) - Y_x(i)|^m \quad \dots \quad (4.3.2-12)$$

nego je nužno dopustiti određeni frekvencijski pomak dužine k između referentnog $Y_r(i)$ i nepoznatog elementa $Y_x(i+k)$ (vidi poglavlje 4. 1. 1). Ovdje primijenjeno klasificiranje vokala realizira se u dvije faze :

I - određivanje mogućih vokalskih kandidata

II - selekcioniranje na osnovu najmanje distance

Dakle, za segment $S_{<t, t+\tau>}$, koji je na razini B označen etiketom V, izračunava se FFT transformacija, koju prikazujemo vektorom :

$$Y = \langle y_1, \dots, y_{32} \rangle \quad \dots \quad (4.3.2-13)$$

Iz dobivenog spektralnog vektora izračunava se energija donjeg E_L i gornjeg E_H dijela spektra:

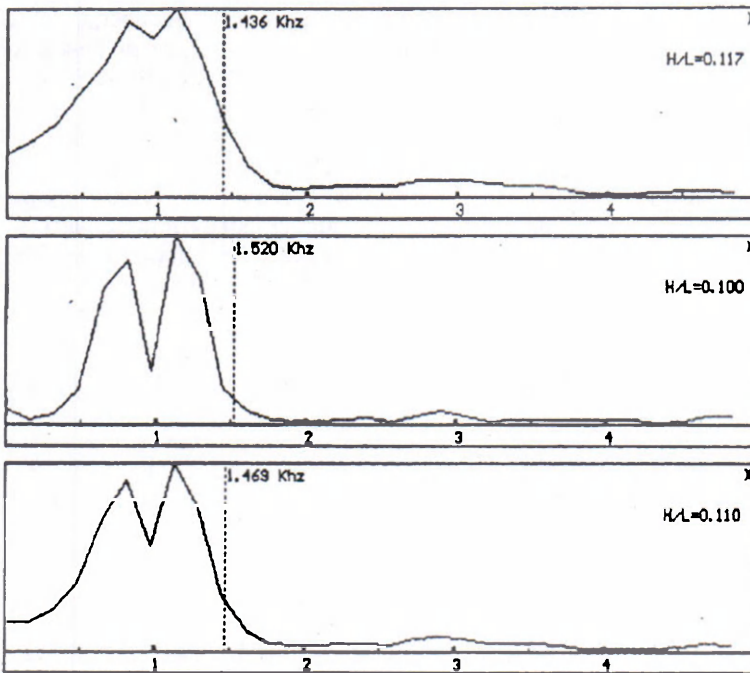
$$E_L = \sum_{j=1}^{16} Y_o(j) \quad \dots \quad (4.3.2-14)$$

$$E_H = \sum_{j=17}^{32} Y_o(j) \quad \dots \quad (4.3.2-15)$$

gdje je Y_0 normalizirana vrijednost spektralnog vektora :

$$Y_0(i) = Y(i) / \max_{n=1}^{32} |Y(n)| \quad \dots \quad (4.3.2-16)$$

S obzirom na to da je odabrana širina prozora segmentiranja 100 uzoraka (10 ms pri $F_u = 10$ KHz), radi preciznijeg izračunavanja, vektor Y zapravo predstavlja srednju vrijednost vektora Y_1 i Y_2 , koji se dobijaju nakon FFT analize 1. . 64 (Y_1) i 32. . 96 uzoraka (Y_2) etiketiranog segmenta. Primjeri izračunatih vektora Y_1 , Y_2 i Y_0 za vokal "a" prikazani su na slici 4. 3. 2-1.



sl. 4. 3. 2-1 Normalizirani spektralni vektori vokala "a"
a) Y_1 b) Y_2 c) Y_0

U fazi I određuju se vokali-kandidati na osnovi odnosa E_H i E_L . Da bi j-ti vokal bio kandidat, mora biti zadovoljen sljedeći uvjet :

$$E_{HLmin}(j) < \frac{E_H}{E_L} \leq E_{HLmax}(j) \quad \dots \quad (4.3.2-17)$$

Iznosi E_{HLmin} i E_{HLmax} dobiveni su na osnovi izmjerenih vrijednosti referentnih sgementa vokala. Da bi se smanjio utjecaj pogreške klasificiranja u

slučajevima jakog koartikulacijskog efekta, uvjet 4. 3. 2-17 oslabljuje se tako da se dopušta prekoračenje definiranih granica E_{HLmin} i E_{HLmax} s vjerojatnošću :

$$p(j) = \frac{1}{4} e^{-\left(\frac{E_{HLmin}}{E_H/E_L}\right)} \quad \dots (4.3.2-18)$$

$$p(j) = \frac{1}{4} e^{-\left(\frac{E_H/E_L}{E_{HLmax}}\right)} \quad \dots (4.3.2-19)$$

Nakon određivanja mogućih K ($K \leq 5$) kandidata vokala, izračunava se globalna minimalna distanca referentnih vokala i tekućeg segmenta (faza II) :

$$D(Y_j, Y_o) = \sum_{i=p}^q d(Y_j(i), Y_o(i)) \quad \dots (4.3.2-20)$$

Lokalna distanca $d(R(i), X(i))$ elementa referentnog vektora $R(i)$ i ispitivanog $X(i)$ definirana je na sljedeći način :

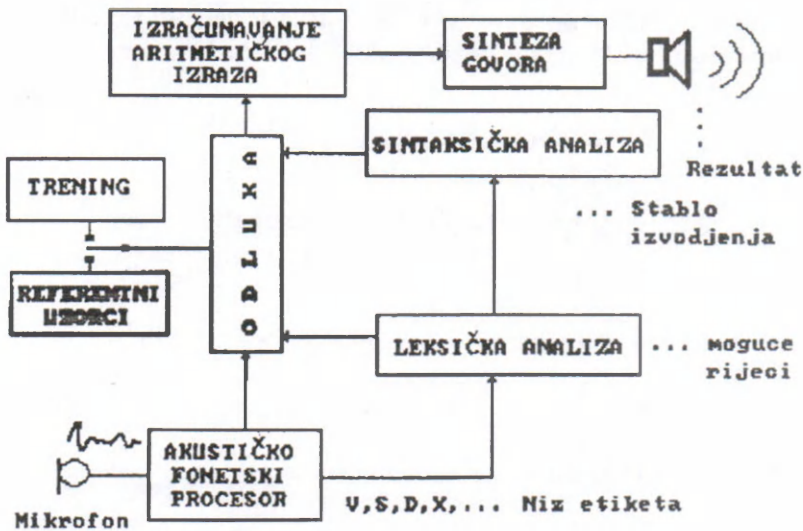
$$d(R(i), Y(i)) = \left[\min \{ |X(i)-R(i)|, |X(i)-R(i-1)|, |X(i)-R(i+1)| \} \right]^2 \quad \dots (4.3.2-21)$$

Da bi se ocjenila točnost načina klasificiranja, svaki je vokal bio izgovoren 100 puta (muški govornik). Dobiveni rezultati prikazani su na donjoj slici u obliku "matrice konfuzije" :

	A	E	I	O	U
A	98	1	1	0	0
E	1	97	2	0	0
I	0	2	98	0	0
O	1	0	0	97	2
U	0	0	0	2	98

tbl. 4. 3. 2-3 Uspješnost prepoznavanja vokala

Srednja točnost prepoznavanja s referentnim uzorcima istog govornika bolja je od 97%. Na slici 4. 3. 2-2 prikazan je proces segmentiranja govora na fizičkoj i fonetskoj razini prema opisanom modelu.



sl. 4. 3. 3-1 Opća slika sistema

Prva faza (akustičko-fonetska) prepoznavanja govora jest obrada signala iz mikrofona u akustičko-fonetskom procesoru, koji segmentira izgovorenu poruku na vremenske intervale od 10 ms i etiketira izdvojene segmente na način opisan u prethodnom poglavlju. Poslije etiketiranja slijedi faza leksičke analize. Na toj razini etikete se grupiraju i transformiraju u liste mogućih riječi (prema definiranom rječniku). Pripremljene alternative dalje se podvrgavaju sintaksičkoj analizi nakon čega se formira stablo izvođenja, koje predstavlja sintaksu izgovorene poruke (aritmetički izraz). Posljednja faza je izračunavanje izraza, a rezultat se iskazuje i sintetiziranim (digitaliziranim) govorom.

Leksička obrada je prva razina simboličke obrade koja se ne oslanja na fizikalni govorni signal nego na izdvojeni parametarski opis (u ovom slučaju - etikete). Predobradu leksičke analize sačinjava grupiranje etiketa dobivenih na fonetskoj razini u određene skupine koje predstavljaju pojedine glasovne kategorije ili konkretne foneme.

Prepoznati fonemi dalje se grupiraju u riječi i provjeravaju postoje li u vokabularu odgovarajući ekvivalenti. Definirani vokabular V jest skup koji sačinjavaju sljedeće riječi:

$$V = \{ \text{nula, jedan, dva, tri, četiri, pet, šest, sedam osam, devet, nula, plus, minus, podjeljeno sa, puta} \}$$

Predobrada leksičke analize jest, dakle, grupiranje i transformiranje postojećih etiketa koje neposredno predstavljaju pojedine glasove. Nažalost, budući da nije moguće postići točnost grupiranja od 100%, i ovdje moramo uvesti postupnu analizu. Na prvoj razini leksičke analize odabiremo neko svojstvo glasova, koje se može pouzdano i jednostavno opisati na osnovi vektora obilježja parametarskog zapisa govornog segmenta. Budući da je zvučnost/bezvučnost

jedno od takvih svojstava, možemo ga definirati kao kriterij grupiranja etiketa dobivenih na fizičkoj i fonetskoj razini. Na osnovi segmentiranih etiketa u fazi fonetske analize:

$$\begin{aligned}\alpha &= \alpha_1 \alpha_2 \alpha_3 \alpha_4 \dots \alpha_i \dots \alpha_k & (\text{razina: } A) \\ \beta &= \beta_1 \beta_2 \beta_3 \beta_4 \dots \beta_i \dots \beta_k & (\text{razina: } B) \\ \upsilon &= \upsilon_1 \upsilon_2 \upsilon_3 \upsilon_4 \dots \upsilon_i \dots \upsilon_k & (\text{razina: } C)\end{aligned}$$

definiramo globalni string zvučnosti nad alfabetom $E_Z = \{ B, Z, T \}$:

$$\mu = \mu_1 \mu_2 \mu_3 \mu_4 \dots \mu_i \dots \mu_k, \quad \mu_i \in E_Z \quad \dots \quad (4.3.3-1)$$

Interpretacija elemenata alfabeta E je sljedeća :

etiketa	značenje
B	bezzvučnost
T	tišina
Z	zvučnost

tbl. 4. 3. 3-1 Fizikalno značenje etiketa leksičke razine

Vrijednost podstringa μ_i bit će Z ako β_i sadrži barem jednu od etiketa D, G, M ili V (oznake zvučnih segmenata), dok će biti etiketiran s B ako β_i ne sadrži spomenute etikete nego neku od etiketa S, X ili F (bezzvučni segmenti). Odsustvo signala registrirano je etiketom $\alpha = \bullet$ pa u tom slučaju μ_i poprima vrijednost T. Na primjer, ako su za izričaj riječi četiri dobivene etikete :

```

α = #####. . . . #####
β = -----VVVVVV. . . . --V-V-----VVVVVV-
-----G-----
-S-S--S-----
XXXXXXXXXX-----
--F--F-----F-----
-----M-M-----
-----
υ = -----eieEiee. . . . --I*-----eIiiii-
```


tada je globalni string zvučnosti :

$$\mu = \text{BBBBBBBBBZZZZZZZTTTTZZZZZZZZZZZZZZZZZZ}$$

Sličan oblik stringa zvučnosti dobili bismo ako bismo segmentaciju provodili na osnovi algoritma analize perioda osnovnog tona i funkcije kratkovremenske energije. Primjerice ako je izračunati F0 za i- ti segment različit od 0, tada će poprimiti vrijednost Z, inače će mu biti pridružena etiketa T ili B ovisno o tome prelazi li energija segmenta zadani prag tišine. Međutim, zbog mnogobrojnih šumova (obezvučavanje i ozvučavanje glasova, utjecaj načina izgovora, fonološke okoline itd) koji su inherentno prisutni unutar govornog signala, dobiveni rezultati analize zvučnosti isključivo s pomoću vrijednosti F0 više će odstupati od "ručne" segmentacije.

Pridružimo sada svakoj riječi iz vokabulara pripadni string zvučnosti $\delta \in E_z^* = \{Z, T, B\}^*$, koji predstavlja opću karakteristiku zvučnosti izričaja pojedine riječi.

r ječ	δ	r ječ	δ	r ječ	δ
nula	Z	pet	TZT	plus	TZB
jedan	Z	šest	BZBT	minus	ZB
dva	Z	sedam	BZ	puta	TZTZ
tri	TZ	osam	ZBZ	podjeleno sa	TZBZ
četiri	BZTZ	devet	ZT		

tbl. 4. 3. 3-2 Stringovi zvučnosti

Na primjer, rječima nula, jedan i dva pridružen string zvučnosti je Z, što znači da svi govorni segmenti koji predstavljaju izričaje tih riječi imaju svojstvo zvučnosti. Iz istih razloga string zvučnosti $\delta = ZBZ$ predstavlja riječ "osam", jer ukazuje na alterniranje zvučnih i bezzvučnih segmenata. Budući da definirani string zvučnosti ne uključuje vremenski aspekt izričaja (tj. broj segmenata), nego samo promjene zvučnosti, nije moguće formirati listu riječi-kandidata za neki globalni string zvučnosti . Zato komprimirajmo string μ u string, μ_c s pomoću FST pretvarača:

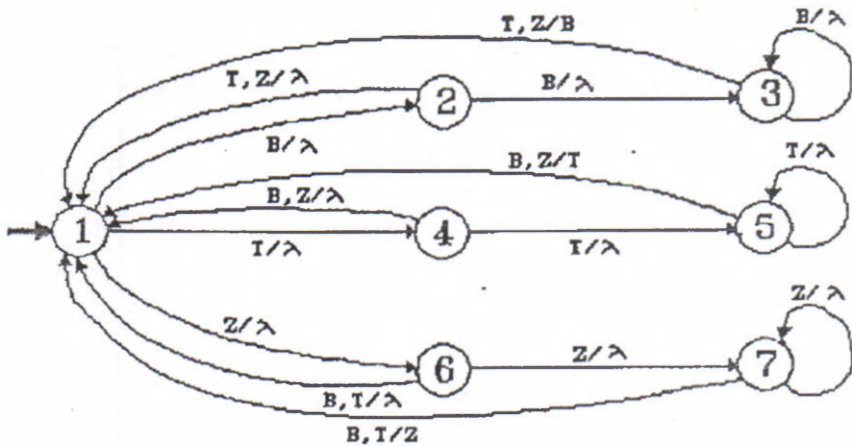
$$M = (\{1, 2, 3, 4, 5, 6, 7\}, \{B, T, Z\}, \{B T Z\}, \delta, 1, \{1, 2, 3, 4, 5, 6, 7\})$$

Funkcija ' α ' definirana je tabelarno :

q x	p y	q x	p y	q x	p y	q x	p y	q x	p y	q x	p y	q x	p y
1 B	2 λ	2 B	3 λ	3 B	3 λ	4 B	1 λ	5 B	1 T	6 B	1 λ	7 B	1 Z
1 T	4 λ	2 T	1 λ	3 T	1 B	4 T	5 λ	5 T	5 λ	6 T	1 λ	7 T	1 Z
1 Z	6 λ	2 Z	1 λ	3 Z	1 B	4 Z	1 λ	5 Z	1 T	6 Z	7 λ	7 Z	7 λ

Komprimirani globalni string zvučnosti uključuje samo promjene zvučnosti, pa je sada moguće izgraditi listu kandidata riječi. Na primjer, neka je $\mu_c = \text{BZTZZBZ}$. Moguće kombinacije (liste) riječi jesu:

1. četiri minus nula
2. četiri minus jedan
3. četiri minus dva
4. sedam tri osam
5. sedam tri nula sedam
6. sedam tri jedan sedam
8. sedam tri dva sedam



sl. 4. 3. 3-2 Grafički prikaz FST automata

Kad je određena lista mogućih riječi, pristupamo drugoj fazi leksičke analize - sužavanju izbora na osnovi dodatnih uvjeta o prirodi riječi iz vokabulara. Prilikom definiranja dodatnih uvjeta mora se paziti da oni budu dovoljno široki kako koartikulacijski efekti i način izgovora ne bi bitno utjecali na konačnu odluku o mogućim rječima (tj. isključivali prave kandidate), ali istodobno trebaju biti maksimalno uski kako bi se broj lista-kandidata što više reducirao. Imajući na umu ta dva suprotna zahtjeva i raspoložive etikete prema predloženom modelu segmentacije, definiramo dodatne uvjete na sljedeći način :

Svakoj riječi r iz vokabulara V pridružimo masku fonetske strukture $\emptyset \in \{ W, T, S, G, M, X \}^*$, koja će preciznije opisivati vremenski slijed segmenata unutar riječi prema donjoj tabeli.

rječ	δ	rječ	δ	rječ	δ
nula	MW	pet	TWT	plus	TWS
jedan	WGW	šest	YWST	minus	WS
dva	GW	sedam	SWGW	puta	TWTW
tri	TW	osam	VSW	podjeljeno sa	TWGSW
četiri	YWTW	devet	GWT		

tbl. 4. 3. 3-3 Fonetske maske

Interpretacija etiketa T, G, S je već objašnjena (tbl. 4. 3. 2-1). Nove etikete koje se pojavljuju su W i Y. Etiketa W označava zvučni segment velike energije, koji ne može biti etiketiran sa G ili M, dok Y predstavlja grupe bezzvučnih segmenata, koji su osim etikete S etiketirani i nekom drugom bezzvučnom etiketom.

Na primjer, fonetska maska za riječ "jedan" jest $\emptyset = WGW$, što znači da unutar izričaja mora postojati grupa segmenata koji su etiketirani sa G a omeđeni drugim zvučnim segmentima velike energije (npr. vokalima).

Fonetska maska izričaja dobiva se u fazi komprimiranja globalnog stringa zvučnosti, gde se svakom znaku $\mu_i \in \{ T, Z, B \}$ komprimiranog stringa zvučnosti $\mu_c = \mu_1 \mu_2 \dots \mu_n$ dodjeljuje indeks početnog (p) i završnog (z) segmenta koji je obuhvaćen i-tim znakom, odnosno:

$$\mu_c = \mu_1 \left| \begin{matrix} z_1 \\ p_1 \end{matrix} \right. \mu_2 \left| \begin{matrix} z_2 \\ p_2 \end{matrix} \right. \mu_3 \left| \begin{matrix} z_3 \\ p_3 \end{matrix} \right. \dots \mu_i \left| \begin{matrix} z_i \\ p_i \end{matrix} \right. \dots \mu_k \left| \begin{matrix} z_n \\ p_n \end{matrix} \right. \dots \quad (4.3.3-2)$$

Primjerice, komprimiranjem globalnog stringa zvučnosti koji se dobiva prema etiketama sa slike 4. 3. 2-2 imamo:

$$\mu_c = T \left| \begin{matrix} 5 \\ 0 \end{matrix} \right. Z \left| \begin{matrix} 16 \\ 6 \end{matrix} \right. B \left| \begin{matrix} 28 \\ 17 \end{matrix} \right. Z \left| \begin{matrix} 44 \\ 29 \end{matrix} \right. T \left| \begin{matrix} 49 \\ 45 \end{matrix} \right. Z \left| \begin{matrix} 68 \\ 50 \end{matrix} \right. B \left| \begin{matrix} 77 \\ 69 \end{matrix} \right. T \left| \begin{matrix} 86 \\ 78 \end{matrix} \right. Z \left| \begin{matrix} 112 \\ 87 \end{matrix} \right. T \left| \begin{matrix} 118 \\ 113 \end{matrix} \right.$$

Pokušajmo sada izgraditi liste riječi za nekoliko prvih etiketa zvučnosti za dobiveni μ_c . Prve dvije etikete TZ odgovaraju po zvučnosti i po obliku fonetske maske riječi tri, pa se ona stavlja na listu- kandidata. Budući da leksička analiza mora zahvatiti sve moguće kombinacije, pomičemo se za jedno mjesto i promatramo neovisno prvi zvučni segment Z. Prema tablici 4. 3. 3-2 za zvučni segment Z moguće riječi su: nula, jedan, dva. Međutim, izračunata fonetska maska od 6. do 16. segmenta (W) ne podudara se ni sa jednom maskom riječi-

kandidata (nula, jedan, dva), što znači da prva zvučna oblast sigurno ne predstavlja zasebnu riječ. Daljnom analizom dobivamo da riječi osam i podijeljeno sa odgovaraju po zvučnosti strukturi TZBZ, ali samo fonetska maska riječi osam zadovoljava uvjet. Postupak leksičke analize ponavlja se dok se ne iscrpe sve alternativne liste kandidata.

Kad je formirana lista kandidata mogućih nizova riječi, na sintaksičkoj razini se na osnovi sintaksnih pravila donosi konačna odluka o izgovorenoj poruci. Sintakсна pravila u ovom radu dana su u obliku CF gramatike:

$$G = (\{S, N, O\}, \{nula, \dots, devet, plus, \dots, podijeljeno\ sa\}, P, S)$$

$$S \rightarrow N O N \mid N O S$$

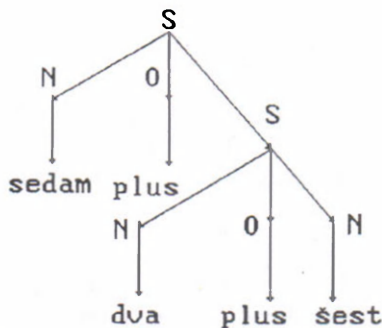
$$N \rightarrow nula \mid jedan \mid dva \mid tri \mid četiri \mid pet \mid šest \mid \\ sedam \mid osam \mid devet$$

$$O \rightarrow plus \mid minus \mid puta \mid podijeljeno\ sa$$

Pretpostavimo sada da smo kao izlaz leksičke analize dobili sljedeće liste-kandidate:

1. sedam plus dva minus minus
2. sedam plus dva plus plus
3. sedam plus dva plus šest

Primjenom algoritma sintaksičke analize (Ahoi Ullman, 1972), jedino će se dobiti stablo i niz izvođenja za rečenicu : "sedam plus dva plus šest", jer ona pripada definiranom jeziku aritmetičkih izraza:



sl. 4. 3. 3-3 Stablo izvođenja

Posljednji korak prema općoj slici sistema jest izračunavanje aritmetičkog izraza i govorna sinteza odgovora. Izračunavanje aritmetičkog izraza ovdje predstavlja semantiku prihvaćene sintaksičke strukture. Nakon izračunavanja aritmetičkog izraza, njegova se vrijednost prikazuje na monitoru i istodobno akustički realizira rudimentarnom sintezom govora (reprodukcija predznačenih digitaliziranih znamenki).

5. ZAKLJUČAK

Unatoč četrdesetogodišnjoj povijesti automatskog prepoznavanja govora i naglog razvoja digitalnih računala, načini da se riješi automatsko prepoznavanje povezanoga govora još nisu pronađeni. Može se reći da je jedino riješen problem prepoznavanja ograničenog broja izoliranih riječi za unaprijed pripremljenog govornika (Baker, 1989). Zaključak mnogih istraživanja jest da se uspješno prepoznavanje govora ne može ostvariti ako se ne ugradi inherentno simboličko znanje i određeni mehanizmi učenja (na simboličkoj razini). Iako se u posljednjem desetljeću naglo propagira pristup konektivnog modela prepoznavanja temeljenog na različitim topologijama neuronskih mreža koje bi same trebale (na osnovi primjera) "naučiti" fonetsko-lingvističko znanje, još nisu prevladani ARPA rezultati u domeni razumijevanja govora. Kompromisno rješenje koje se nametne, jest da se neuronske mreže iskoriste za učenje na percpcijskoj razini (u akustičko- fonetskom procesoru) dok bi se i dalje zadržala tradicionalna simbolička obrada.

REFERENCIJE

1. Aho, A. V. , Ullman, J. D. (1972) : The Theory of Parsing, Translation, and Compiling, Volume I : Parsing, Toronto: Prentice-Hall
2. Allerhand M. (1987): Knowledge- based Speech Pattern Recognition, London: Kogan Page
3. Baker, J. M. (1989): Dragondictate - 30K: Natural Language Speech Recognition with 30. 000 words, EUROSPEECH-89, Paris, (str. 161-163)
4. Charniak E. i McDermott (1985) : Introduction to AI, Massachusetts : Addison-Wesley
5. Gonzalez, R. C. i Thomson M. G. (1978): Syntactic Pattern Recognition, Massachusetts: Addison-Wesley
6. Hatzsaki K. et al. (1988) : Phoneme segmentation by an expert system based on spectrogram reading knowledge, SPEECH-88, (str. 927-931) Edinburgh, 22-26 aug.
7. Kohonen. T. (1988): Neural Phonetic Typewriter, Computer, Vol XXI, No 3, (str. 11-22)
8. Kurepa, S. (1982) : Matematička analiza 1 i 2, Zagreb: TK
9. Lea, W. A. , izdavač (1980) : Trends in Speech Recognition, New Jersey: Prentice-Hall
10. . Levinson, S. E. (1983) : An introduction to the Application of the Theory of Probabilistic Functions of a Markov Process in Automatic Speech Recognition, (str. 1035-1074), The Bell System Technical Journal, Vol. 62, No4, Part 1
11. Linard, J. S. (1977) : Les processus de la communication parlée, Paris: Masson
12. Mc Clelland, L. i Rumelhart, D. E (1988) : Explorations in parallel distributed processing, Massachusetts: MIT Press

13. Michalski, R. S. et al. (1984): "Machine learning: An artificial Intelligence Approach", Berlin : SpringerVerlag
14. Mišković, V. (1989): "Obrada Linearnih sributa u algoritmu za induktivno učenje CORAL", u priprermi za časopis "AUTOMATIKA"
15. Moris, Čarls. (1975) : Osnove teorije o znacima, Beograd: BIGZ
16. Nakanishi, H. i Nakagava, S. (1987): Speaker-Independent word recognition by less cost and stochastic dynamic time warping method, (str. 292-295) Vol2, J. Laver, izdavač, Zbornik "European Conference on Speech Technology ", Edinburgh
17. Papamichalis, P. E. (1987) : Practical Approaches to Speech Coding , New-Jersey: Prentice-Hall
18. Patrick, E. A. (1972): Fundamentals of Pattern Recognition, New-Jersey: Prentice-Hall
19. Paušc, Ž. (1974): "Vjerojatnost, informacija, stohastički procesi", Zagreb: ŠK
20. Plotnikov, V. N. (1988), Rečevoj dialog v sistemah upravljenija, Moskva : Mashinostroenic
21. Potter, R. K i Kopp G. A (1947) : Visible Speech, New York, : D. van Nostrand Co.
22. Proakis J. G. (1983) : Digital Communications, Singapore: McGraw-Hill, Carnegie-Mellon
23. Rumelhart, D. E. et al. (1986) : Learning representation by back-propagation errors, Nature, 323, (str. 533-536)
24. Prager, R. W et al. (1986): Boltzmann machines for speech recognition, Computer Speech and Language, Vol. I, (str. 3- 27)
25. Rabiner L. R. et al. (1983): On the Application of Vector Qunatization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition, (str. 1075- 1106, The Bell System Technical Journal, Vol. 62, No4, Part 1
26. Rabiner R. L. i Schafer W. Ronald (1978), Digital Processing of Speech Signals, Engkewood Cliffs: Prentice-Hall
27. Schafer R. W. , Markel, J. D., (1979), izdavači: Speech Analysis, New York: IEEE Press
28. Stamenković M. (1987): Primer govorno upravljalog procesnog sistema, MIPRO87, (str. 282-286), Opatija.
29. Stamenković, M. (1988.): Digitalno predstavljanje i analiza govora u vremenskoj domeni, Govor, br. 2. , (str. 109- 132)
30. Stamenković, M. i Bakran, J. (1988.): Fonetsko- lingvistički pristup prepoznavanja govora, MIPRO-88, Opatija
31. Stamenković, M. (1988.) : A learning speech recognition system, SPEECH-88, Edinburgh, (str. 773-780)
32. Stamenković, M. (1988) : Mašinsko učenje i prepoznavanjegovora, ETAN-88, Sarajevo
33. Stamenković M. i J. Bakran (1989) : An Intelligent PitchTracker Based on Formal Language Theory and Phonetic Knowledge, EUROSPEECH-89, Paris, (str. 470-473)
34. Zbornik ROJP III, Institut "Jožef Štefan", Ljubljana, 1985.

35. Zbornik ROLP IV, Institut "Jožef Stefan", Ljubljana, 1988.
36. Yang, F. et al. (1988): Utilisation d'un rescau de neurones pour la reconnaissance des mots isolés, SPEECH-88, Edinburgh, (str. 859-866)
37. Zagoruiko N. G. et al. (1985): Alogirtmy obnaruzenia empiriceskih zakonomiernosti, Novosibirsk: Nauka
38. Witten, I. H. (1982): Principles of Computer Speech, London: Academic Press

Milan Stamenković
VVTS KoV JNA, Zagreb

AUTOMATIC RECOGNITION OF SPEECH

SUMMARY

Besides the general principles of automatic speech recognition the paper deals with approaches to speech recognition based on the traditional theory of pattern recognition and machine learning opposed to models including phonetic and linguistic aspects. The central part of the paper is concerned with the description of the system for continuous speech recognition which shapes the speech communication processes.