KAZALO

1. UVOD

Govor je najbolj naravna in neposredna oblika komunikacije med ljudmi. Enaka sposobnost se vse bolj nestrpno pričakuje tudi od informacijsko komunikacijske tehnologije[1], ki nas obdaja. Njen razvoj namreč omogoča vedno večjo avtomatizacijo najrazličnejših opravil, katerih velik del bi lahko bil nedvomno enostavneje opravljen s pomočjo govorne komunikacije. Nesporno potreba po "strojnem" obvladovanju govornih dialogov strmo narašča in predstavlja eno od osnovnih gibal nadaljnjega razvoja.

Dandanes se govorni signal prenaša poleg vse večjih razdalj še po vse večjem številu različnih prenosnih poti. Tipičen primer je spletna telefonija, ki je trenutno v velikem razmahu. Zaradi omenjene raznolikosti in dolžine prenosnih poti se v novejših komunikacijskih sistemih pojavlja kar nekaj težav, ki jih je potrebno rešiti. Med njimi so trenutno najbolj pereče časovna zakasnitev in motnje, ki nastajajo pri zajemu, prenosu in reprodukciji govornih signalov. V teh pogojih so tudi sistemi za razpoznavanje govora[2] precej manj uspešni oziroma učinkoviti; poleg vnaprej predvidenih se namreč srečajo še s celo vrsto neznanih aditivnih, konvolutivnih ter nelinearnih popačitev govornih signalov.

Kar nekaj časa je razvoj na področju avtomatskega razpoznavanja govora[3] uspešno potekal z metodami, ki so temeljile bolj na praktičnem ("inženirskem") delu s konkretnimi SRG kot pa na reševanju teoretičnih vprašanj v njihovi zasnovi. Ta so bila ves čas bolj v ozadju. Deloma tudi zaradi tega, ker je bilo poznavanje tovrstnih procesov pri človeku še precej omejeno. V zadnjem času pa vse bolj prihajajo na plan slabosti takega pristopa. Zato sta ponovno v ospredju človek in njegova govorna percepcija kot vzorčni primer učinkovitosti in uspešnosti tudi v najtežjih akustičnih pogojih.

Danes poskušamo intenzivneje spoznavati človekove slušne sposobnosti in jih posnemati. Že v preteklosti so te sicer služile kot izvor nekaterih preprostejših prijemov, ki so dandanes v SRG zelo razširjeni; takratni uspešni začetek pa ni imel pravega logičnega nadaljevanja. Danes je jasno, da je potrebno lastnosti človekove govorne percepcije precej podrobneje spoznati, da bi se njegovi uspešnosti lahko zares približali. Trenutno poznavanje pa še ni na zadovoljivi ravni. Dosti boljše je v primeru procesa poslušanja, ki se odvija v človekovem

---

[1] *V nadaljevanju IKT.*

[2] *V nadaljevanju SRG.*

[3] *V nadaljevanju ARG.*

ušesu, bistveno slabše pa v primeru procesa razpoznavanja informacijske vsebine govora, ki poteka na višjem nivoju - v možganih.

Na področju ARG pogosto prihaja do t.i. razvojnega paradoksa. Na eni strani so natančno umerjeni in zato uspešni SRG, ki zaradi tega celo omejujejo nadaljnji razvoj. Na drugi strani so bolj "napredno" zasnovani sistemi, ki pa v neposrednih primerjavah zaradi manjšega števila preizkusov in opravljenih raziskav nimajo enakovrednega izhodišča. Na celotnem področju govornih komunikacij je opazen še vse bolj izrazit razkorak med potrebami in zmožnostmi obstoječih sistemov. V tej "stiski" so zato največkrat v prednosti kvantitativno boljše rešitve, četudi so alternativni pristopi dolgoročno morda obetavnejši.

Obvladovanje govorne komunikacije je že za človeka samega kar zahtevna naloga. Vsak otrok namreč v svojem razvoju porabi kar nekaj let, da razvije svoje govorne sposobnosti do osnovne ravni, medtem ko nadaljnji proces izpopolnjevanja traja še občutno dlje. Iluzorno je torej pričakovati hitro rešitev problema avtomatskega razpoznavanja govora brez postopnega in dolgotrajnega razvojnega procesa, v katerem bo potrebno človekovo govorno percepcijo še precej podrobneje spoznati.

SRG je v splošnem sestavljen iz dveh procesov: parametrizacije in razpoznavanja. Po analogiji ustrezata dogajanju v človekovem ušesu in možganih. Povezavi in prenosu znanj med področjema proučevanja človekovega sluha in ARG je bila posvečena večina mojega dosedanjega dela [77, 78-82, 120]. V disertaciji pa je v središču pozornosti bolj specifičen del procesa parametrizacije – natančneje – uporaba okenskih funkcij[1] pri kratkočasovni Fourierovi transformaciji[2]. Najbolj pogosto uporabljana okna imajo namreč kar nekaj potencialnih slabosti in večinoma niso oblikovana v skladu z lastnostmi in značilnostmi človekovega sluha. Tudi interakcija med obema glavnima procesoma v SRG skupaj z vplivom okna ni dovolj raziskana. Zato si tovrstno področje zasluži večjo pozornost, predvsem z vidika reševanja dveh najbolj aktualnih praktičnih problemov: zagotovitve krajše časovne zakasnitve in večje robustnosti razpoznavanja, ki zagotavlja uspešnost kljub prisotnosti motenj. Pri robustnih SRG je torej razlika v uspešnosti ob prisotnosti motenj in brez njih majhna. Ta lastnost je vedno večjega pomena, saj število različnih pogojev delovanja SRG skokovito narašča.

---

[1] *Okenska funkcija definira okno v zveznem prostoru. Okensko zaporedje (v nadaljevanju okno) je njena diskretna realizacija.*

[2] *V nadaljevanju KČFT; angl. STFT – "Short Time Fourier Transform".*

1.1 ISKANJE OPTIMALNIH OKEN

Okna se lahko širše obravnavajo tudi kot linearni, časovno invariantni diskretni sistemi s končnim enotinim odzivom[1]. Ti so na področju digitalnega procesiranja signalov zelo popularni. Še posebej pri tem izstopa podskupina z linearnim faznim odzivom[2], ki slovi po enostavnosti načrtovanja in realizacije. Ima pa tudi nekatere slabosti - večja časovna kompleksnost, daljša zakasnitev[3] in slabši amplitudni odziv. Zahteva po linearnosti faze ima namreč za posledico liho oziroma sodo simetrijo odziva na enotin impulz. Ta sicer poenostavi postopek načrtovanja, hkrati pa pomeni omejitev razsežnosti prostora možnih rešitev pri izpolnjevanju zadanih zahtev. Na nekaterih področjih je linearnost faze še vedno nepogrešljiva, pri razpoznavanju govora pa ni razlogov za njeno koristnost [60, 68, 70, 72, 77].

Znanje s področja KEO filtrov se v veliki meri lahko uporabi tudi za načrtovanje oken v kratkočasovni frekvenčni analizi. Spekter govornega signala se v času namreč dokaj hitro spreminja, zato se analizira v končnih časovnih odsekih - okvirjih. Te se formalno dobi z množenjem neskončno dolgega signala s končno dolgim utežnostnim zaporedjem – oknom. Amplitudni odziv posameznega okvirja je posledično konvolucijski integral med Fourierovima transformoma okna in neskončno dolgega signala. To pomeni, da se dejanski frekvenčni odziv ne da povsem natančno izračunati; lahko pa se dobi boljši ali slabši približek. Uspešnost je pri tem odvisna predvsem od lastnosti okna, ki je lahko določeno eksplicitno ali načrtovano v skladu z izbranim kriterijem optimalnosti.

V medsebojnih primerjavah okenskih funkcij sta najpogosteje uporabljeni vrednosti širine glavnega vala in maksimalne višine stranskih valov v amplitudnem odzivu [8]. Vendar sta omenjeni lastnosti tesneje povezani s splošno frekvenčno analizo signalov in z načrtovanjem KEO filtrov [117, 118, 120]. Pomen obeh lastnosti za ARG pa je zaradi posebnosti področja potrebno še podrobneje raziskati.

Lastnosti oken so običajno med seboj povezane. To pomeni, da ni mogoče poljubno izboljšati enih, ne da bi se pri tem poslabšale druge. Torej se lahko iščejo le kompromisne rešitve, ki izboljšujejo pomembnejše lastnosti na račun drugih, manj pomembnih. Za uspešno uporabo oken na specifičnih področjih (nedvomno to velja tudi za ARG) je zato potrebno ločiti

---

[1] *V nadaljevanju KEO.*

[2] *Linearnost faznega odziva (krajše faze) pomeni enako časovno zakasnitev frekvenčnih komponent signala pri prehodu skozi sistem.*

[3] *Predstavlja dodatno oviro pri delu v realnem času – npr. dvosmerni govorni komunikaciji.*

pomembne lastnosti od manj pomembnih ter ugotovljeno upoštevati pri načrtovanju bolj primernih oken.

Na področju ARG trenutno obstaja malo nam znanih virov ali raziskav o nesimetričnih oknih [85, 86]. Med njimi se le naše raziskave ukvarjajo s problemom njihove optimalnosti [78-82]. Zato je za oceno pomembnosti posameznih lastnosti oken potrebno uporabiti znanje z drugih področij, predvsem s splošne frekvenčne analize signalov in ARG. Povsem odprto vprašanje pa je, ali to za oblikovanje sodobnejših SRG sploh zadostuje. Nekatere lastne raziskave so namreč dokaj nedvoumno pokazale, da je vpliv izbrane okenske funkcije na uspešnost in robustnost SRG večji od pričakovanj [77, 78-82]. Zato je eden od glavnih ciljev disertacije prispevati k boljši raziskanosti omenjenega vpliva, ki je bil v preteklosti deležen premajhne pozornosti.

Kljub trenutnemu razmahu govornih tehnologij se obstoječe stanje le počasi spreminja. Večina današnjih SRG še vedno uporablja Hammingovo simetrično okno, ki je zaradi svoje enostavnosti in solidnih lastnosti zelo popularno tudi na drugih področjih. V manjšem številu se v SRG pojavijo še druge "klasične", eksplicitno določene okenske funkcije[1]. Te večinoma predstavljajo suboptimalen kompromis med različnimi lastnostmi. Med njimi je malo takih, ki so po katerem od znanih kriterijev optimalne [92, 93]. Po mojem mnenju je to zadosten razlog za dvom, da je uporaba teh oken sprejemljiva tudi s stališča nadaljnjega razvoja na področju ARG.

Okna se lahko načrtujejo tudi z metodami s področja KEO filtrov. Vendar so le te večinoma prilagojene uporabi dveh najpogostejših kriterijev: maksimalne absolutne napake[2] in srednje kvadratne napake. Prav tako je pri filtrih največkrat želeni amplitudni odziv odsekoma konstanten. Primernost teh izhodišč za uporabo na specifičnem področju ARG je vsekakor potrebno podrobneje raziskati. Pri tem ne kaže zanemariti še morebitne alternativne možnosti. Obojemu je v disertaciji posvečeno veliko pozornosti.

1.2 POSEBNOSTI PODROČJA PROCESIRANJA GOVORA

Na širšem področju procesiranja[3] govornih signalov je zelo pomembna zahteva po čim krajši splošni časovni zakasnitvi. Pri tem je uporaba simetričnih oken precejšnja omejitev; dolžina okna je namreč edini vzvod za skrajšanje časovne zakasnitve, vendar se s tem poslabša tudi

---

[1] *Npr. Hannovo, Blackmanovo, trikotno okno…*

[2] *Imenuje se tudi Čebišev oziroma "minimaks" kriterij.*

[3] *Področje poleg razpoznave in sinteze govornih signalov vsebuje še zajem ter prenos signalov na daljavo – večinoma v digitalni obliki.*

frekvenčna ločljivost analize. Uporaba nesimetričnih okenskih funkcij na tem področju omogoča relativno velike časovne prihranke, kar je za delovanje v realnem času zelo pomembno.

Sistemi za razpoznavanje govora so dokaj specifična aplikacija frekvenčne analize. Za razpoznavanje daljših glasov (npr. samoglasniki) so pomembnejše stacionarne lastnosti, medtem ko so za določene glasove (npr. zaporniki) odločilni kratkotrajni zvočni dogodki. Za dobro reprodukcijo govornih signalov so pomembne lastnosti vsakega individualnega govorca, ki so pri od govorca neodvisnih[1] SRG povsem odveč; slednji namreč poskušajo izvesti abstrakcijo, ki iz govora ohrani le informacije, povezane z vsebino in ne z individualnimi posebnostmi glasu oziroma govorca. Zato je proces računanja značilk, ki sledi sami frekvenčni analizi, dokaj specifičen in zapleten. Prav tako v tem primeru optimalnost okna ni zanimiva samo v smislu kriterijev s področja splošne frekvenčne analize, ampak predvsem z vidika uspešnosti oziroma robustnosti razpoznavanja govora. V tem kontekstu je izbira optimalnih lastnosti okna še precej težja.

Kar nekaj lastnosti človekovega slušnega zaznavanja je v precejšnjem razkoraku v primerjavi z zasnovo in načinom delovanja večine današnjih SRG [67-74, 77]. Med njimi so za oblikovanje oken najpomembnejše lastnosti same frekvenčne analize v ušesu in neobčutljivost človekovega sluha na fazne popačitve govornega signala. Do sedaj opravljene raziskave tako na področju frekvenčne analize človekovega sluha kot tudi ARG zmanjšujejo pomen frekvenčne ločljivosti [68-70]. Prav tako je kar nekaj razprav o tem, da je monotonost[2] amplitudnega odziva pomembnejša, kot se zdi na prvi pogled [66]. Na drugi strani pa že omenjena človekova neobčutljivost na fazne popačitve govornih signalov navaja na sprostitev zahteve po linearnosti faze pri načrtovanju oken. S stališča načrtovanja je problem v tem primeru kompleksnejši, a hkrati prinaša večji prostor potencialnih rešitev in s tem možnost izboljšav. Te so lahko še bolj izrazite, če se po vzoru lastnosti frekvenčne analize človekovega sluha poveča širina glavnega vala v amplitudnem odzivu okna [77]. Nenazadnje pa obstaja še povsem praktičen razlog za opustitev zahteve po linearnosti faze. V vseh SRG se namreč kot osnova za nadaljnji proces razpoznavanja uporabi le amplitudni odziv, fazna informacija pa se običajno zanemari.

---

[1] *Tovrstni sistemi so danes v veliki večini.*

[2] *Monotonost amplitudnega odziva pomeni, da je njegova vrednost v vsaki naslednji točki manjša ali enaka prejšnji.*

1.3 OSNOVNA IZHODIŠČA IN SMERNICE

Problem določitve vpliva okenskih funkcij na uspešnost SRG je dokaj zapleten. V prvi vrsti zaradi tega, ker sta oba glavna procesa, parametrizacija in razpoznavanje, še vedno obravnavana kot dve zaključeni celoti, največkrat celo z dokaj različnih izhodišč. Prvi je bolj pod vplivom splošne frekvenčne analize, medtem ko je drugi bližje področju razpoznavanja vzorcev. Zelo malo pa je znanega o njunem medsebojnem vplivu, čeprav je jasno, da je končna uspešnost SRG v največji meri odvisna prav od sinergijskega učinka obeh.

V opisanem kontekstu dobi vrednotenje vpliva okna na uspešnost SRG dvojno razsežnost. Nekoliko lažje bo mogoče opredeliti vpliv okna na dobljene predstavitve govornega signala v procesu parametrizacije. Vrednotenje bo namreč z vidika splošne frekvenčne analize bolj objektivno, saj bodo ugotovitve neodvisne od procesa razpoznavanja. Dejanski vpliv na končno uspešnost SRG pa bo mogoče ovrednotiti le kvantitativno na osnovi rezultatov praktičnih preizkusov. Zelo aktualno je namreč vprašanje, v kakšni meri lahko obstoječi proces razpoznavanja sploh izkoristi morebitne boljše opise signalov iz procesa parametrizacije. Lahko se namreč zgodi, da bodo v prihodnosti konceptualno naprednejše zasnove SRG te izboljšave še bolje izkoristile. Mogoče je tudi pričakovati, da se bosta oba procesa zlila v celoto [80] in bo problem učinkovite interakcije med njima implicitno lažje rešljiv.

V opisanih okoliščinah bo tudi interpretacija rezultatov praktičnih preizkusov dokaj specifična. Dokazana prednost uporabe nesimetričnih oken z vidika časovne zakasnitve namreč ne bo vplivala na kvantitativno uspešnost razpoznavanja SRG; zaznaven bo le vpliv amplitudnega odziva oken. Seveda tudi v primeru zanemarljivega ali morebiti celo negativnega vpliva na uspešnost še vedno nesporno ostaja prednost krajše časovne zakasnitve. Prav tako se lahko uspešnost sodobnejše zasnovanih sistemov še precej spremeni. Tako je že vnaprej mogoče trditi, da bodo praktični rezultati le trenuten odgovor na zastavljeno vprašanje vpliva oken na uspešnost procesa razpoznavanja. Oba glavna procesa v SRG se bosta v nadaljnjem razvoju nedvomno vedno tesneje povezovala. Če se ob tem upošteva še trend približevanja lastnostim človekovega sluha, je mogoče pričakovati še boljše rezultate.

V disertaciji bo merjena t.i. inherentna robustnost SRG, pri kateri motnje niso prisotne v procesu učenja. Pogosto se namreč pri podobnih preizkusih motnje dodajo že v učno množico. Vendar se v tem primeru bolj meri sposobnost obvladovanja vnaprej znanih motenj v procesu učenja kot pa uspešnost SRG v bolj realni situaciji, kjer se pojavijo tudi nepredvidene motnje.

Glede na povedano so bila pri praktičnih preizkusih in raziskovalnem delu širše upoštevana naslednja načela:

- *v čim večji meri ohraniti obstoječe podatke, postopke, strukturo SRG in vse ostale parametre, ki jih ni nujno potrebno spreminjati*
- *ohraniti računsko kompleksnost in uspešnost SRG*
- *ohraniti nesprotni*[1] *način delovanja zaradi večje primerljivosti rezultatov*
- *uspešnost SRG ovrednotiti predvsem v pogojih, ki niso prisotni v procesu učenja – "inherentna robustnost"*
- *izogniti se posegom, ki bi v duhu "grobe sile" prinašali večjo uspešnost SRG (podvajanje statističnih modelov, natančno uglaševanje sistema s pomočjo velikega števila preizkusov, razširjanje učne množice z motnjami, itd)*
- *glede na omejitve razpoložljive aparaturne opreme in časa selektivno izvajati praktične preizkuse v različnih obsegih sorazmerno glede na pričakovano pomembnost rezultatov*
- *poskrbeti za čim večjo raznolikost zasnov, govornih zbirk in ostalih pogojev delovanja*
- *medsebojno primerjavo uspešnosti obeh referenčnih SRG izvesti le na osnovi enakih testnih množic; podrobnejše primerjave zaradi velikih razlik v zasnovah in implementacijah presegajo okvir tega dela*
- *uspešnost SRG izraziti v odstotkih uspešno razpoznanih besed*[2]

1.4 KRATEK OPIS VSEBINE

Disertacija vsebuje šest večjih vsebinskih sklopov. Po uvodu se v 2. poglavju nahaja kratek oris področja ARG. Opisani so atributi, ki določajo kompleksnost SRG. Posebna pozornost je posvečena procesu parametrizacije. V njem se namreč kot prvi korak izvede kratkočasovna frekvenčna analiza, ki je za vrednotenje vpliva oken še posebej zanimiva. V tej fazi se namreč pod vplivom okna oblikuje časovno frekvenčni opis signala, ki vstopi v nadaljnje procese pri razpoznavanju govora. V nadaljevanju so še podrobnejši opisi izračunov standardnih značilk, ki so bili uporabljeni v praktičnih preizkusih. Poglavje zaključi opis drugega glavnega procesa v SRG – razpoznavanja.

Naslednje (3.) poglavje podrobneje opisuje referenčno okolje za vrednotenje inherentne robustnosti SRG. Vpliv izbranih oken je namreč mogoče proučiti le s pomočjo praktičnih

---

[1] *Angl. "off-line".*

[2] *Angl. "WSR" – Word Success Rate.*

preizkusov z veliko mero raznolikosti vseh parametrov, ki vplivajo na uspešnost razpoznavanja. Zato referenčno okolje vsebuje govorni zbirki z dveh jezikovnih področij in dva različno zasnovana referenčna SRG. Vsi so podrobneje opisani skupaj s postopkom za simulacijo motenj oziroma težjih akustičnih pogojev delovanja referenčnih SRG.

V 4. poglavju se nahaja obširnejša analiza možnosti načrtovanja nesimetričnih oken. Najprej je pojasnjen pomen oken v frekvenčni analizi govornih signalov in potencialne prednosti nesimetričnih oken. Predstavljeni so nekateri možni kriteriji za njihovo načrtovanje. Med metodami so najprej opisane tiste s področja KEO filtrov, za njimi pa še nekatere enostavnejše metode z uporabo NEO[1] parametričnih modelov in kombiniranjem obstoječih simetričnih oken. Skozi celotno poglavje je pozornost usmerjena še v možnosti uporabe splošnejših optimizacijskih metod na tem področju.

Rezultati praktičnih preizkusov so predstavljeni v 5. poglavju. Spremljajo jih zanimivi zaporedni preizkusi uspešnosti že naučenih SRG in poglobljena analiza predstavljenih rezultatov. Poglavje se zaključi s končnimi ugotovitvami.

V 6. poglavju so povzete končne ugotovitve analize praktičnih rezultatov in raziskav, opisanih v disertaciji. Podana so splošna priporočila za uporabo nesimetričnih oken v SRG. Zatem je opredeljen še pomen opravljenih raziskav in možnosti nadaljnjega dela na tem področju.

[1] *Sistem z neskončnim odzivom na enotin impulz – "NEO - Neskončni Enotin Odziv".*

2. AVTOMATSKO RAZPOZNAVANJE GOVORA

Razpoznavanje govora je proces preslikave signala v tekstovni zapis zaporedja razpoznanih enot (fonemov, besed, besednih zvez, stavkov, povedi). Večinoma je to že končni rezultat, v nekaterih primerih pa le osnova nadaljnjih postopkov za razumevanje vsebine (npr. simultano prevajanje, vodenje govornih dialogov, …). V prvem primeru gre za ožji in v drugem za širši pomen razpoznavanja govora. V skladu z osnovnimi izhodišči je v tej disertaciji poudarek na razpoznavi osnovnih govornih enot – besed – ter uporabi enostavnih in standardnih postopkov.

Splošna zasnova in pomembnejše značilnosti SRG so opisane v podpoglavju 2.1. Sledi predstavitev procesa parametrizacije (podpoglavje 2.2), ki vključuje postopek kratkočasovne frekvenčne analize; v njem imajo pomembno vlogo okna, ki so v središču pozornosti. Sledijo še opisi standardnih postopkov za računanje kompaktnih reprezentativnih predstavitev zajetega govornega signala. Ti vstopajo v proces razpoznavanja, ki je opisan v podpoglavju 2.3. V njem se izvede primerjava z referenčnimi opisi posameznih enot v slovarju SRG in določi končni izid razpoznavanja.

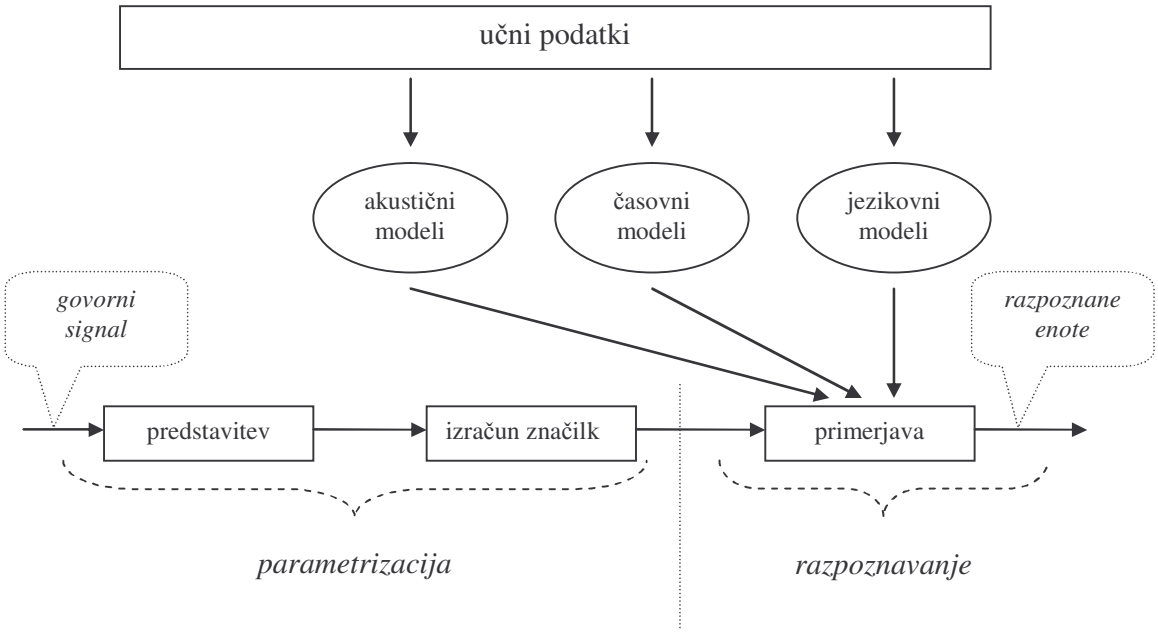
2.1 SPLOŠNA ZASNOVA SISTEMOV ZA RAZPOZNAVANJE GOVORA - SRG

Tovrstni sistemi so v splošnem označeni z več atributi. Glede na svoj namen so lahko dokaj enostavni (npr. sistemi za razpoznavanje manjšega števila ločenih besed) ali zelo kompleksni (npr. razpoznavanje spontanega govora v šumnem okolju). V splošnem velja, da je problem razpoznavanja težji pri večjem številu besed oziroma enot v slovarju ali izrazitejši prisotnosti motenj. K težavnosti lahko prispevajo tudi nekatere na prvi pogled manj pomembne stvari (npr. podobnost izgovorjav posameznih besed v slovarju). Nekaj najbolj značilnih atributov za razvrstitev SRG po zahtevnosti obvladanega problema prikazuje Tabela 2.1. Vrednosti atributov so v posameznih vrsticah navedene po naraščajoči zahtevnosti.

Atributi	*Zaloga vrednosti po naraščajoči zahtevnosti*
časovna povezanost govora	posamezne besede, fraze, vezan govor, tekoč govor
število govorcev	znan govorec, več znanih govorcev, neznan govorec[1]
velikost slovarja	majhen slovar (nekaj besed), srednji (1000 besed), velik (več kot 20000 besed)
Akustični vpliv okolice (aditivni šum)	laboratorij, tiho okolje, hrupno okolje (ulica, tovarna, avtomobil, itd)
prenosni medij	kvalitetna avdio oprema, digitalna telefonija, analogna in spletna telefonija, mobilna telefonija
način govora	narekovanje, prebiranje, spontan govor

Tabela 2.1: Pomembnejši atributi SRG.

Zgradbo splošnega sistema za razpoznavanje govora prikazuje Slika 2.1. Razpoznavanje govora v ožjem smislu lahko razdelimo na dva glavna koraka. V prvem se pripravi kompakten, reprezentativen opis vhodnega govornega signala – t.j. proces parametrizacije. V njem se najpogosteje uporablja kratkočasovna Fourierova transformacija skupaj z ostalimi tehnikami s področja digitalnega procesiranja signalov. Iz vzorcev vhodnega govornega signala se v postopku frekvenčne ali (in) časovne analize najprej oblikuje začetna predstavitev signala. Ta predstavlja osnovo za nadaljnje postopke računanja končnega zaporedja vektorjev



Slika 2.1: Zgradba splošnega SRG.

[1] *Oznaka ponazarja neomejeno število govorcev.*

značilk[1], ki vstopa v proces razpoznavanja oziroma primerjave.

Dobljen opis se v drugem koraku primerja z že prej pripravljenimi referenčnimi opisi[2] oziroma modeli – t.j. proces razpoznavanja. Najprej se vhodni opis na osnovi različnih kriterijev (statistična podobnost, spektralna razdalja, …) primerja z akustičnimi referenčnimi modeli, ki opisujejo tovrstne značilnosti posameznih govornih enot v slovarju sistema. Zaporedje posameznih klasifikacij se nato primerja s časovnimi modeli, ki opisujejo dinamiko spreminjanja akustičnih značilnosti govora. Dobljena ocena časovnega ujemanja se na koncu ovrednoti še v jezikovnih modelih, ki opisujejo relacije elementarnih govornih enot v daljših zaporedjih (npr. besede, fraze, stavki, povedi). Za končni rezultat razpoznavanja se izberejo najboljša (najverjetnejša) hipoteza in njej ustrezno zaporedje govornih enot.

V procesu primerjave se uporabijo tehnike in metode, ki temeljijo na različnih pristopih. Trenutno sta daleč najpogostejša le dva formalna koncepta, in sicer :

- *statistični modeli; opisujejo globalne lastnosti signalov in podobnosti med trenutnimi in referenčnimi opisi signalov*
- *mrežni modeli; sestavljeni iz večjega števila enostavnih, med seboj povezanih vozlišč, ki razvrščajo (klasificirajo) vhodne vzorce v predhodno ustrezno naučene razrede*

Statistični pristop je, predvsem po zaslugi prikritih Markovih modelov[3], na tem področju prisoten že zelo dolgo. Kljub nekaterim pomanjkljivostim je še vedno najbolj razširjen. Pozornost pa zasluži tudi mrežni koncept (najbolj popularen predstavnik so nevronske mreže), ki je na področju razpoznavanja govora precej mlajši in zato tudi manj dozorel. Glede na nekatere potencialne prednosti je neupravičeno zapostavljen; deloma tudi zaradi že v uvodu omenjenega razvojnega paradoksa.

V novejšem času se vse bolj pogosto pojavljajo SRG, ki uporabljajo kombinacijo obeh najbolj popularnih pristopov. V teh t.i. hibridnih sistemih se skuša izkoristiti zgolj njune prednosti. Daleč najpogostejša je zasnova z nevronsko mrežo kot akustičnim klasifikatorjem in statističnimi (HMM) modeli za časovno poravnavo. Poenostavljen primer takega sistema je drugi referenčni SRG (podpoglavje 3.3), medtem ko je prvi (podpoglavje 3.2) tipičen predstavnik realizacije statističnega pristopa na akustičnem in časovnem nivoju.

---

[1] *Končno množico parametrov, s katerimi opišemo končni izsek nekega signala, imenujemo tudi vektor značilk, posamezne komponente pa značilke.*

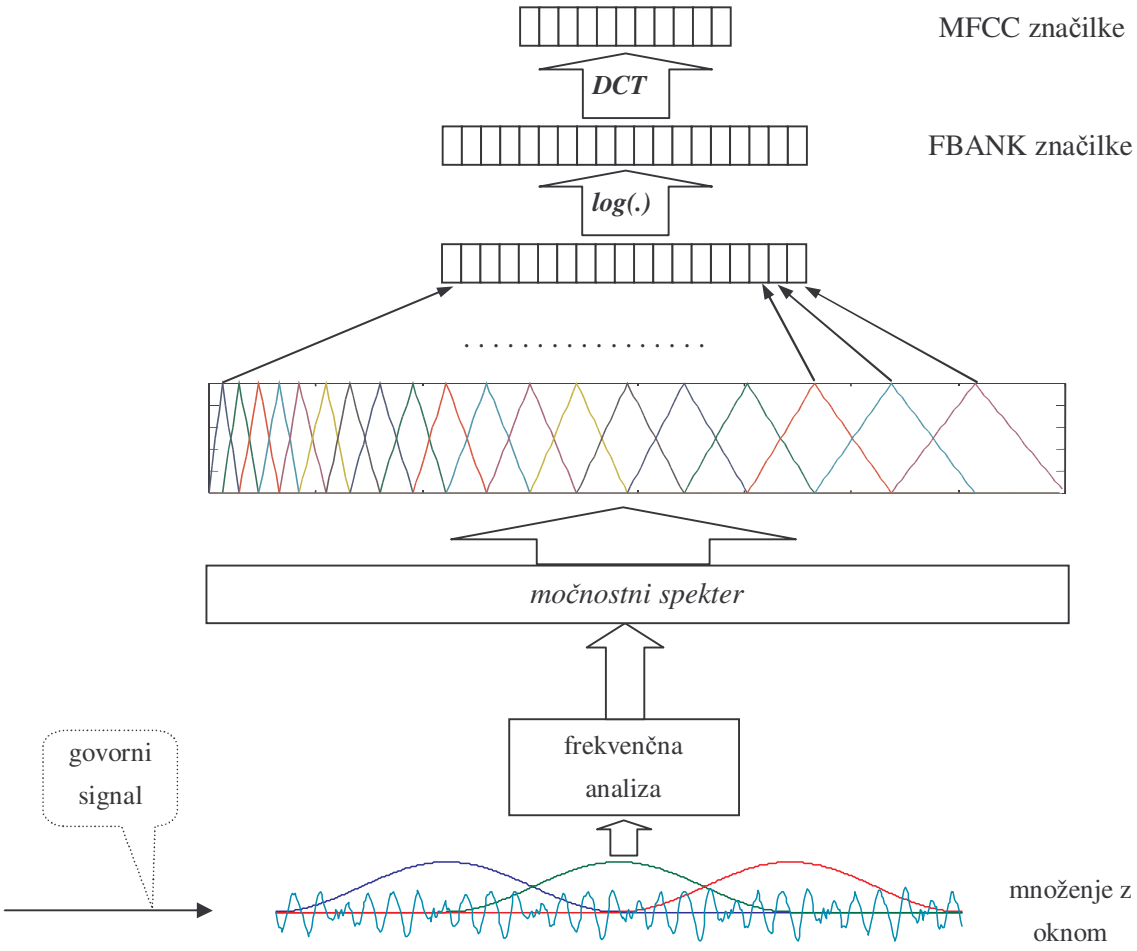
[2] *Opisi se pripravijo v postopku učenja.*

[3] *V nadaljevanju HMM ali angl. "Hidden Markov Models".*

2.2 PARAMETRIZACIJA

Razpoznavanje govora je zahteven problem, predvsem zaradi izrazite variabilnosti govornega signala. Enolična preslikava med tekstovnim zapisom in njegovo akustično realizacijo v obliki govora namreč ne obstaja. Če želijo SRG uspešno opraviti svojo nalogo, morajo signale predstaviti oziroma obravnavati tako, da ločijo informacijsko vsebino od variabilnosti.

Današnji SRG omenjen problem rešujejo na več načinov. V fazi parametrizacije poskušajo generirati opise signalov, ki vsebujejo čim več za razpoznavanje pomembnih informacij in čim manj variabilnosti ter individualnih značilnosti vsakega govorca. To še posebej velja za sisteme, ki razpoznavajo govor "neznanih"[1] govorcev. Postopek najbolj razširjene parametrizacije z uporabo MFCC[2] oziroma FBANK[3] značilk prikazuje Slika 2.2.



Slika 2.2: Parametrizacija govornega signala.

[1] *Govorci, ki niso zajeti v učni množici.*

[2] *Standardna oznaka kepstralnih značilk; angl. "Mel Frequency Cepstral Coefficients".*

[3] *Standardna oznaka značilk pasovnega spektra, angl. "FBANK" oziroma "FilterBank".*

Govor se vzorči in v obliki diskretnega zaporedja vstopa v sistem. Tukaj se najprej razdeli na med seboj prekrivajoče se izseke oziroma okvirje[1]. Vzorci v vsakem od njih se pomnožijo z izbranim oknom (najpogostejše je Hammingovo simetrično okno). Po frekvenčni analizi se izračuna še amplitudni oziroma močnostni spekter. Postopek je podrobneje opisan v podpoglavju 2.2.1.

Dobljeni močnostni spekter se nato s pomočjo prekrivajočih trikotnih utežnostnih funkcij združi v neenakomerno razporejene širše frekvenčne pasove, v katerih se oblikujejo za razpoznavanje pomembne informacije. Amplitude v posameznih pasovih se še logaritmirajo in rezultat je standardna predstavitev signala – FBANK značilke. Te se pogosto še dekorelirajo s pomočjo diskretne kosinusne transformacije (krajše DCT[2]). Ob običajnem zmanjšanju števila značilk nastane končni rezultat - vektor MFCC značilk. Postopek je podrobneje opisan v podpoglavju 2.2.2.

2.2.1 Kratkočasovna Fourierova analiza govornega signala

Vhodni signal se najprej razdeli na končne (lahko tudi prekrivajoče) krajše izseke, v katerih se kasneje izračuna Fourierov transform (Slika 2.3). Formalno je operacija delitve definirana kot produkt vzorcev neskončnega signala in okenskega zaporedja, ki ima neničelne vrednosti le znotraj aktualnega okvirja.

S pomikanjem okna po vhodnem signalu nastane zaporedje posameznih okvirjev

$$s_w(n,m) = w(n-m)\, s'(m), \qquad -\infty<m<\infty \;\; in \;\; n=n_t,\; n_t=n_0+tN,\; t=0,1,2,\ldots, \tag{2.1}$$

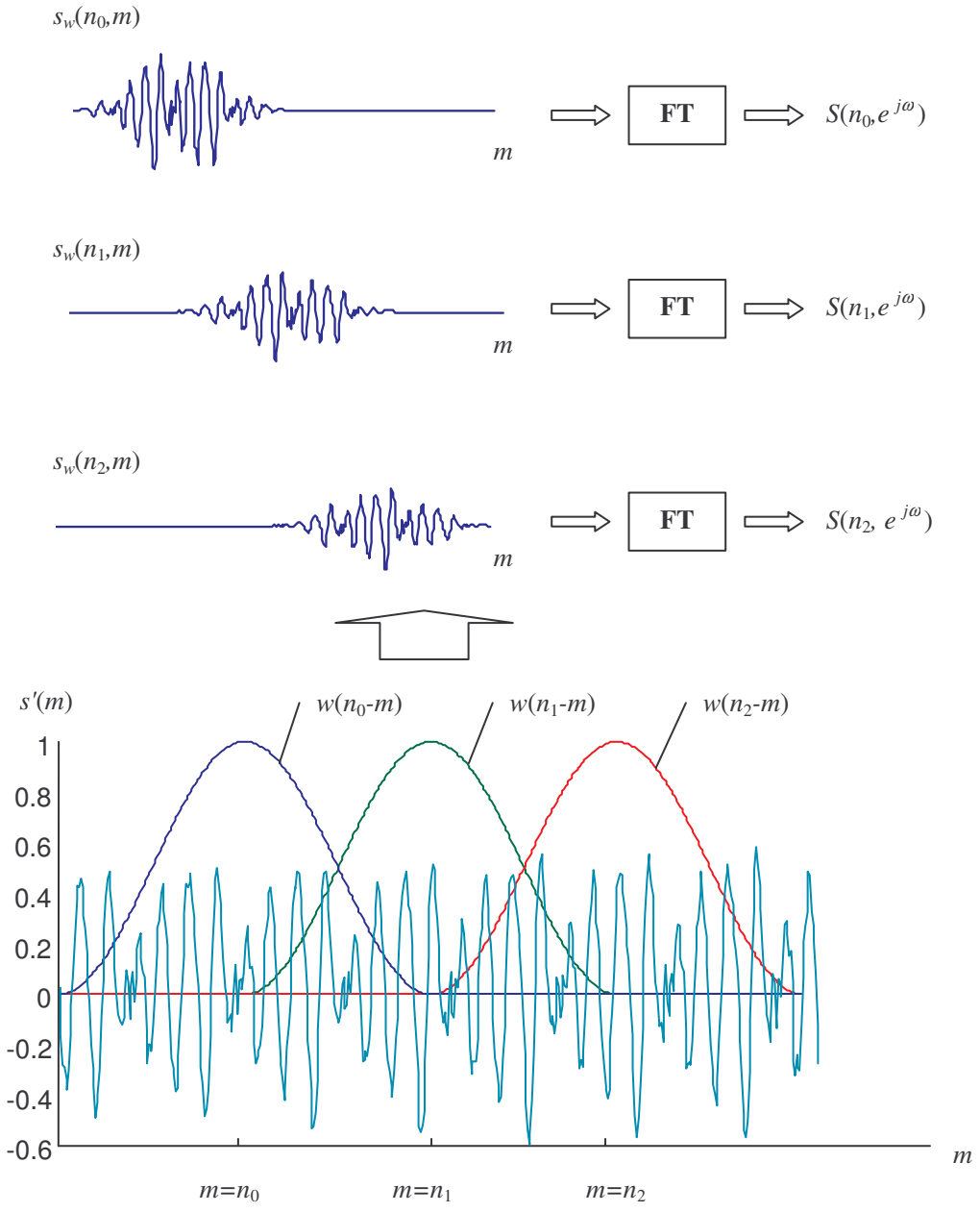
kjer je $s'(m)$ zaporedje vzorcev neskončnega signala, $w(n-m)$ okensko zaporedje, n indeks središčne točke posameznega okvirja, N razmik med dvema okvirjema in t številka okvirja. Dolžina vseh okvirjev je enaka dolžini okna. V vsakem okvirju se izračuna frekvenčni odziv, ki se doda dvodimenzionalnemu, časovno frekvenčnemu opisu signala $S(n,e^{j\omega})$

$$S(n,e^{j\omega}) = \sum_{m=-\infty}^{\infty} s_w(n,m)\; e^{-j\omega m}. \tag{2.2}$$

---

[1] *Pri vseh praktičnih preizkusih v disertaciji je bila dolžina okvirja 32 ms, razmik med dvema sosednjima okvirjema pa 10 ms.*

[2] *Angl. "Discrete Cosine Transform".*



Slika 2.3: Računanje kratkočasovnega spektra signala.

Zaradi uporabe okna dobljeni frekvenčni spekter ni več povsem enak dejanskemu. Določen je z izrazom, ki opisuje konvolucijski integral

$$S(n,e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S'(e^{j(\omega+\theta)})\ W(e^{j\theta})e^{j\theta n}\,d\theta, \tag{2.3}$$

kjer je $S'(e^{j\omega})$ frekvenčni odziv neskončno dolgega signala in $W(e^{j\omega})$ frekvenčni odziv okna. Izraz (2.3) pomeni, da se vedno izračuna le boljši ali slabši približek dejanskega frekvenčnega odziva. Izbira okna torej pomembno vpliva na dobljen frekvenčni odziv. Lastnosti okna oziroma s tem povezano popačenje frekvenčnega odziva se določijo pri samem načrtovanju, kjer je poleg zahtevanih lastnosti pomembna še izbira ustreznega kriterija.

Običajno je zaradi želene lastnosti linearnosti faznega odziva okensko zaporedje simetrično. Ker pa je znano, da je uspešnost človekovega razpoznavanja govora praktično neodvisna od morebitnih faznih popačenj signala [68], ni prav nobenega razloga, da ne bi v frekvenčni analizi uporabili nesimetričnega okna. Že iz načrtovanja digitalnih KEO filtrov je jasno, da se lahko z nesimetrijo precej pridobi. Ob zadosti širokih prepustnem in prehodnem pasu se na ta račun dobi filtre z boljšim amplitudnim odzivom [77]. Podrobnejša obravnava uporabe nesimetričnih oken pri razpoznavanju govora se nahaja v 4. poglavju.

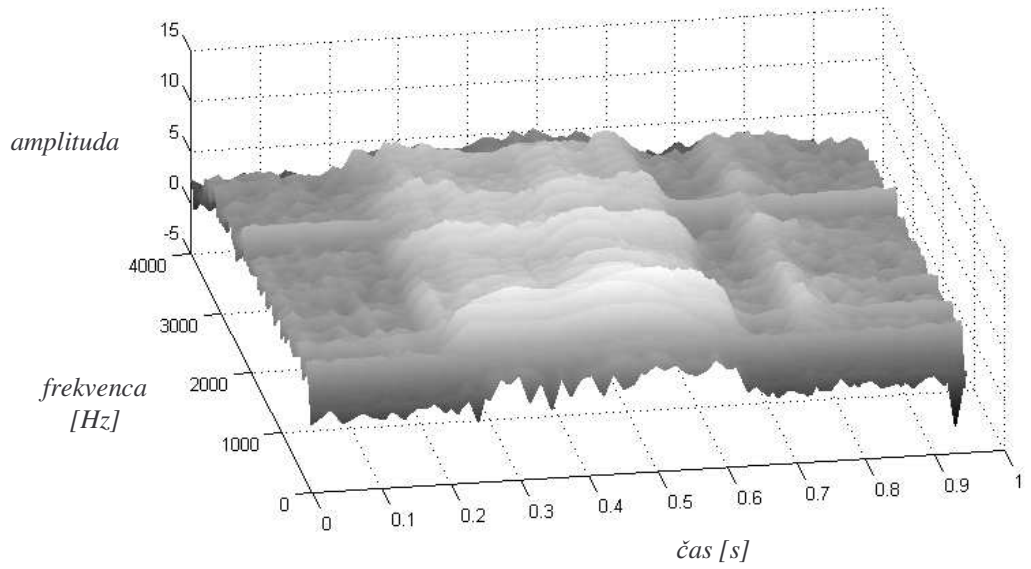
Končna oblika časovno-frekvenčnega opisa signala - t.i. amplitudni kratkočasovni spekter – se dobi še z izračunom absolutne vrednosti zaporedja frekvenčnih odzivov. Spekter se pogosto še logaritmira[1] v skladu s človekovo percepcijo amplitude signala. Dobljena osnovna časovno-frekvenčna predstavitev se imenuje logaritemski kratkočasovni spekter, njegova grafična predstavitev pa tudi spektrogram (Slika 2.4).

Posamezne komponente kratkočasovnega spektra se že lahko imenujejo tudi značilke; v nekaterih primerih je namreč omenjen spekter že končni opis vhodnega signala. Če se upoštevajo vrednosti posamezne značilke v odvisnosti od časa[2], se dobi t.i. signal značilke, ki ponazarja časovno diskretni potek njene vrednosti – v tem primeru je to logaritem amplitude posamezne frekvenčne komponente v točki ω_0:

$$M(t,e^{j\omega}) = log(\,|S(n_t,e^{j\omega})\,|), \qquad \omega = \omega_0,\ \ n_t = n_0 + tN,\ \ t = 0, 1, 2, \ldots \tag{2.4}$$

[1] *Pogosto se pred logaritmiranjem amplituda kvadrira – uporabi se t.i. močnostni spekter.*

[2] *Čas je ponazorjen s številko okvirja t.*



Slika 2.4: Primer spektrograma besede "devet".

2.2.2 Računanje kompaktnejših predstavitev signala

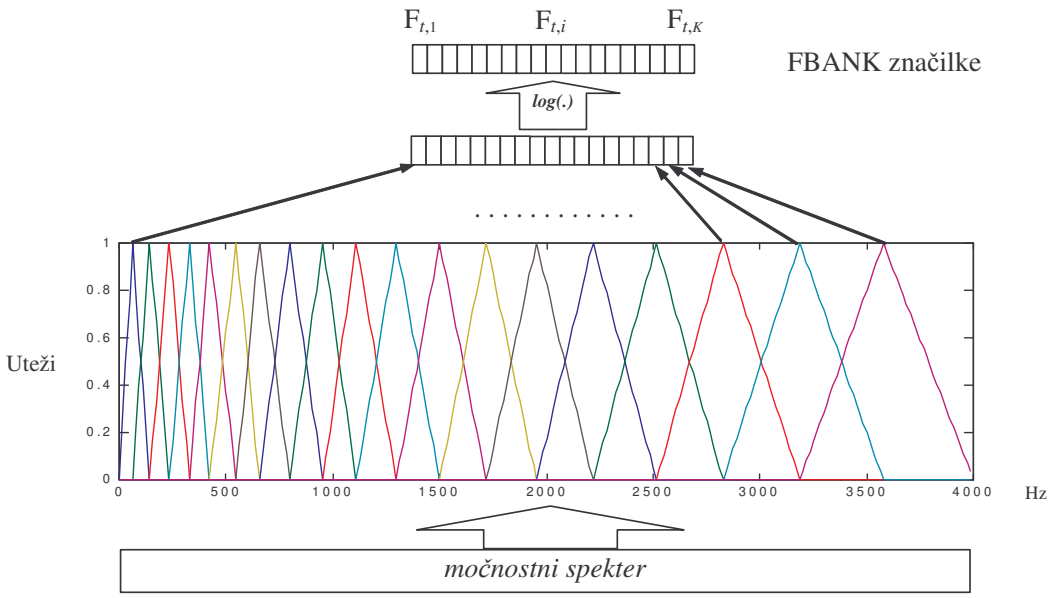
Kratkočasovni spekter je prostorsko zahtevna in za potrebe razpoznavanja preveč podrobna časovno-frekvenčna predstavitev govornega signala. Prav tako vsebuje precej informacij, ki izražajo že prej omenjeno variabilnost, trenutne lastnosti posameznega govorca in akustičnega okolja; vse so za razpoznavanje običajno celo moteče. Večja razsežnost vhodnih podatkov pa poleg računske kompleksnosti povzroči še drugo težavo. SRG namreč posledično potrebuje večje število prostih parametrov, s čimer se zmanjša možnost generalizacije oziroma posploševanja pri učenju. To lahko negativno vpliva na končno uspešnost delovanja. Vsled teh argumentov je potrebno oblikovati bolj kompakten in informacijsko precej bolj splošen zapis. Postopek frekvenčne analize se zato običajno nadaljuje še v drugem koraku, kjer se izvede t.i. podatkovna redukcija; čim več za razpoznavanje pomembnih informacij se skuša predstaviti s čim manjšo količino podatkov.

2.2.2.1 FBANK značilke

Računanje kompaktnejših predstavitev (Slika 2.5) se začne z združevanjem[1] posameznih delov spektra v širše frekvenčne pasove. Njihova širina in središčne frekvence so običajno

[1] *Ta operacija se običajno izvede na amplitudnem oziroma močnostnem spektru.*

razporejene v skladu s katero od frekvenčnih lestvic[1], ki posnemajo neenakomerno frekvenčno ločljivost človeškega sluha. Pri tem ima pomembno vlogo koncept t.i. kritičnega pasu. Ta je formalno definiran kot frekvenčno območje, znotraj katerega se lahko ob konstantni jakostni gostoti spreminja sestava tonskega kompleksa[2] brez subjektivne zaznave spremembe glasnosti. Če so k takemu kompleksu dodane še frekvenčne komponente izven kritičnega pasu, se kljub ohranjeni jakostni gostoti zaznava glasnosti spremeni. Torej se v posameznem kritičnem pasu nahajajo vse tiste frekvenčne komponente govornega signala, ki jih človekov sluh, kadar nastopajo v tonskem kompleksu, zelo slabo ali pa sploh ne loči od celote.



Slika 2.5: Izračun FBANK značilk.

Kritični pasovi so običajno modelirani s pomočjo trikotnih utežnostnih funkcij. Njihove središčne točke so razporejene po melodični lestvici, ki je v skladu z neenakomerno frekvenčno ločljivostjo človekovega sluha [68]. Utežnostne funkcije se med seboj prekrivajo, zato vsaka frekvenčna komponenta "prispeva" k vrednosti v dveh sosednjih frekvenčnih pasovih. To je podobno človekovem sluhu, kjer je prekrivanje še bolj izrazito in tipično sega čez nekaj sosednjih frekvenčnih pasov [68, 71].

Utežnostne funkcije se zaradi oblike in namena imenujejo tudi trikotni filtri. V obeh referenčnih sistemih je teh filtrov 22. V splošnem je to število kompromis. Če je preveliko, se

[1] *Najbolj znani sta t.i. melodična (angl. "Mel Scale") in Barkova lestvica.*

[2] *Tonski kompleks je signal, sestavljen iz tonov več različnih frekvenc.*

preveč zmanjša količina informacij v posameznem frekvenčnem pasu. Če je premajhno, pride do pretiranega združevanja sicer neodvisnih informacij v ožjih frekvenčnih pasovih.

Po "trikotnem" uteževanju se dobljene vrednosti logaritmirajo v skladu s povezavo med dejansko amplitudo zvočnega valovanja in ustrezno subjektivno zaznano glasnostjo pri človeku. Nastane t.i. pasovni spekter, ki se shranjuje v zaporedje vektorjev standardnih FBANK značilk $F_{t,i}$

$$F_{t,i} = \log(\int_0^{\pi} T_i(e^{j\omega}) \left|S(n_t, e^{j\omega})\right|^2 d\omega), \quad i = 1..K, \; n_t = n_0 + tN, \; t = 0,1,2,..., \tag{2.5}$$

kjer je t trenutni okvir, K število vseh trikotnih filtrov, $T_i(e^{j\omega})$ i-ti trikotni filter in $|S(n_t, e^{j\omega})|^2$ močnostni odziv. Osnovna slabost te predstavitve je precejšnja koreliranost med sosednjimi značilkami. To predstavlja potencialni problem predvsem pri SRG, ki temeljijo na uporabi statističnih HMM modelov in s tem povezanih implicitnih predpostavk. Zato se običajno vektorji značilk dodatno dekorelirajo s postopkom, opisanim v naslednjem podpoglavju.

2.2.2.2 MFCC značilke

Koreliranost značilk v sosednjih pasovih se običajno zmanjšuje z nekaterimi dodatnimi transformacijami (npr. DCT[1], LDA[2]). V obeh referenčnih SRG je bila uporabljena diskretna kosinusna transformacija, ki je najbolj razširjena. Izračuna se iz FBANK značilk $F_{t,j}$

$$c_{t,i} = \sum_{j=1}^{K} F_{t,j} \cos(\frac{\pi i}{K}(j - 0.5)), \qquad 1 \le i \le L, \; 1 \le t \le T, \tag{2.6}$$

kjer L pomeni število značilk, K število frekvenčnih pasov, T število okvirjev, $c_{t,i}$ pa MFCC značilke. Izkaže se, da se običajno lahko razsežnost dobljenega vektorja celo zmanjša ($L<K$) brez zaznavnega negativnega vpliva na končno uspešnost SRG.

Glavna prednost DCT v primerjavi z ostalimi sorodnimi postopki je manjša računska kompleksnost, posledica pa nekoliko slabša končna dekoreliranost MFCC značilk.

---

[1] *DCT (angl. "Discrete Cosine Transform") pomeni diskretno kosinusno transformacijo.*

[2] *LDA (angl. "Linear Discriminant Analysis") pomeni linearno diskriminantno analizo.*

2.2.2.3 Dinamične (DELTA) značilke

Prvotna spoznanja o človekovi govorni percepciji so kot osnovo razpoznavanja omenjale le frekvenčno analizo govornega signala. Vendar je danes že znano, da je v kratkočasovni analizi govornih signalov zelo pomembna tudi časovna razsežnost. Prav tako je iz obsežnih psihoakustičnih raziskav človekovega sluha [68, 58] jasno, da je percepcija nekega trenutnega stanja precej odvisna tudi od prejšnjih stanj oziroma akustičnih dogodkov v predhodnem časovnem intervalu. Vse to napeljuje na potrebo po značilkah, ki opisujejo dinamiko vektorjev značilk v relativno daljšem časovnem obdobju (npr. v vsaj nekaj sosednjih okvirjih).

Prvotni SRG so uporabljali večinoma samo osnovni nabor MFCC značilk, pridobljenih na že opisan način (podpoglavje 2.2.2.2). Tovrstne značilke so bile tudi najprimernejše za takratne sisteme; ti so namreč večinoma temeljili na zelo enostavnem šablonskem pristopu oziroma ustreznem razpoznavalnem postopku dinamičnega časovnega izkrivljanja[1]. Sodobnejši SRG lahko informacijo o časovni dinamiki izkoristijo že precej bolj učinkovito.

V novejših SRG se praviloma ob statičnih značilkah pojavijo še t.i. dinamične značilke, ki ponazarjajo trende spreminjanja vrednosti v nekaj sosednjih vektorjih. Ta pristop se je v praksi pokazal za zelo uspešnega, še posebej v šumnih okoljih [77]. Omenjene značilke imajo namreč relativni značaj in so zato na linearne spremembe absolutnih vrednosti, ki jih največkrat povzroči vpliv motenj, dokaj neobčutljive. Ker je eden najbolj popularnih načinov za njihov izračun prav časovni odvod, se imenujejo tudi delta značilke.

Za ponazoritev hitrosti spreminjanja vrednosti značilke je na voljo več izračunov razlik med njenimi časovno razmaknjenimi vrednostmi. V SRG se običajno uporabi t.i. prirejena regresijska analiza, ki osnovnemu vektorju MFCC značilk doda še enako število delta značilk prvega reda. Njihova vrednost se izračuna po naslednjem izrazu, ki upošteva vrednosti v $2P$ sosednjih okvirjih[2]:

$$d_{t,i} = \frac{\sum_{k=1}^{P} k\,(c_{t+k,i} - c_{t-k,i})}{2\sum_{k=1}^{P} k^2}. \qquad (2.7)$$

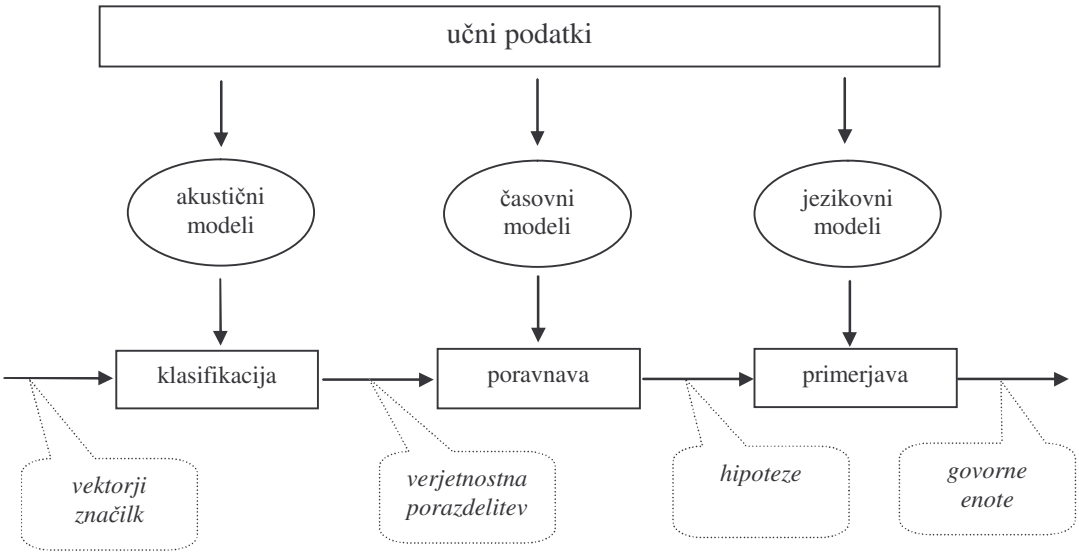
[1] *Angl. "Dynamic Time Warping", krajše tudi DTW.*

[2] *V praktičnih testiranjih je bila vrednost parametra P=2, kar pomeni vrednosti štirih sosednjih okvirjev.*

V hibridnih SRG se časovna dinamika značilk lahko ponazori tudi na nekoliko drugačen način - več sosednjih vektorjev MFCC značilk se združi v enoten vektor oziroma opis signala.

2.3 RAZPOZNAVANJE

V procesu razpoznavanja je potrebno rešiti tri podprobleme (Slika 2.6). Prvi je klasični klasifikacijski problem trenutnega opisa signala, ki se razvršča v enega od možnih razredov referenčnih opisov. V drugem koraku se časovno zaporedje teh klasifikacij oziroma t.i. verjetnostna porazdelitev preslika v ustrezne hipoteze o razpoznanih govornih enotah. S tem se podanemu zaporedju trenutnih opisov signala določi informacijska vsebina.



Slika 2.6: Proces razpoznavanja govora.

Pri razpoznavanju so lahko v veliko pomoč informacije, vsebovane v jezikovnih modelih. Pri večjih slovarjih se namreč zelo hitro pojavi problem prevelikega števila hipotez, ki bi se sicer morale enakovredno ovrednotiti. Omejitev tega prostora pa omogočijo prav podatki o verjetnostih pojavitev zaporedij posameznih govornih enot. Na ta način se prostor možnih hipotez v tretji fazi razpoznavanja preiskuje bolj selektivno in učinkoviteje. Po ovrednotenju hipotez se določi končni izid razpoznavanja.

V disertaciji so preizkusi omejeni na razpoznavanje elementarnih govornih enot - besed, zato so jezikovni modeli trivialni - število hipotez je v vsakem trenutku majhno in predvidljivo.

3. REFERENČNO OKOLJE ZA VREDNOTENJE NESIMETRIČNIH OKEN

Za proučevanje vpliva oken na uspešnost sistemov za razpoznavanje govora je potrebno izvesti veliko število praktičnih preizkusov. Pri tem je za večjo splošnost dobljenih rezultatov pomembna raznolikost uporabljenih govornih zbirk ter zasnov SRG. Zato se v referenčnem okolju nahajajo dve govorni zbirki in dva različno zasnovana referenčna sistema.

Prvi sistem je t.i. HMM razpoznavalnik ločenih besed[1], ki uporablja prikrite Markove modele in je po zasnovi statističen. Drugi je t.i. CSLU[2] razpoznavalnik nizov števk[3], ki je po zasnovi hibriden. V njem sta združena oba najbolj razširjena pristopa: nevronska mreža kot akustični in postopek Viterbijevega iskanja (znan iz statističnega pristopa) kot časovni model. Glavne prednosti obeh sistemov oziroma uporabljenih konceptov v njunih zasnovah so naslednje:

- *prednosti HMM razpoznavalnika oziroma HMM modelov:*
    - *uporaba normalnih porazdelitev zmanjša število parametrov in s tem poveča možnost posplošitve*
    - *enostavnejša prilagoditev spremenjenim pogojem med delovanjem*
    - *uporabnost naučenih parametrov za druge namene (npr. sinteza govora)*
    - *možnost uporabe besednih modelov*
- *prednosti CSLU razpoznavalnika oziroma nevronskih mrež:*
    - *brez omejitev glede porazdelitve vhodnih podatkov*
    - *implicitno zagotovljen diskriminatorni način učenja*
    - *manjša časovna in prostorska kompleksnost*
    - *večja primernost za aparaturne, vzporedne realizacije*
    - *možnost verifikacije govorcev, ocene zanesljivosti razpoznavanja*

V naslednjem podpoglavju sta najprej opisani uporabljeni govorni zbirki in v ločenih podpoglavjih še oba referenčna razpoznavalnika.

---

[1] V nadaljevanju označen kot "HMM razpoznavalnik" ali krajše "HMM sistem".

[2] Angl. okrajšava za "Center for Spoken Language Understanding".

[3] V nadaljevanju označen kot "CSLU razpoznavalnik" ali krajše "CSLU sistem".

3.1 REFERENČNI GOVORNI ZBIRKI

V referenčnem okolju sta bili uporabljeni dve govorni zbirki z različnih jezikovnih področij. Obe vsebujeta večje število izgovorjav različnih govorcev, ki so bile posnete preko telefonskih zvez. Govorna zbirka NUMBERS vsebuje izgovorjave angleških števk[1] in številk[2]. Njena osnovna značilnost je dokaj spontan, tekoč način govora z zelo kratkimi presledki. Govorna zbirka ŠTEVKE v slovarju poleg slovenskih števk vsebuje še tri ukazne besede (ja, ne, stop). V primerjavi z zbirko NUMBERS ima naslednje posebnosti:

- *slabša povprečna kvaliteta telefonskih zvez pri snemanju*
- *večja podobnost izgovorjave posameznih števk*[3]
- *večja narečna raznolikost izgovorjave*
- *manjše število govorcev oziroma posnetkov*
- *daljši premori med besedami*
- *manj spontan način govora*

Snemanje obeh zbirk je potekalo pred približno 10 leti. V tem času so se običajni pogoji delovanja SRG spremenili. Ob snemanju so sicer zaradi omejitev obstoječe tehnologije prevladovali nekoliko slabši pogoji. Prisotnih pa je bilo precej manj ostalih, nepredvidljivih motenj, ki so danes pogostejše. Sodobna tehnologija namreč omogoča uporabo najrazličnejše komunikacijske opreme, večjo izbiro prenosnih poti in možnost komuniciranja v praktično vseh akustičnih okoljih.

Zbirka NUMBERS je bila uporabljena le na CSLU sistemu, ŠTEVKE pa na obeh; v CSLU sistemu kot niz besed, v HMM sistemu pa so bili posnetki razdeljeni na izgovorjave posameznih besed. Zaradi večje primerljivosti so bile motnje dodane najprej na posnetkih celotnih izgovorjav, ki so se zatem razdelile na posamezne odseke za uporabo v HMM razpoznavalniku. Prisotnost motenj je torej v obeh primerih enaka.

V nadaljevanju so opisane osnovne značilnosti obeh referenčnih govornih zbirk in postopek simulacije motenj v govornih signalih, potreben za vrednotenje inherentne robustnosti obeh referenčnih SRG.

---

[1] *Števke pomenijo števila med 0 in 9.*

[2] *V referenčnem okolju so bile uporabljene le števke.*

[3] *Posebnost slovenskega jezika – npr. dve, devet, pet, …*

3.1.1 Govorna zbirka ŠTEVKE

Govorna zbirka je nastala v letih 1992-1993 v - takrat še - Mikroračunalniškem laboratoriju Fakultete za elektrotehniko in računalništvo. V njej se nahajajo izgovorjave 780 govorcev iz vse Slovenije, posnete preko analognih telefonskih zvez. Vsak izbran govorec je v naključnem vrstnem redu izgovoril 13 besed (številke od "0" do "9" ter besede "ja", "ne" in "stop"). Postopek je vodil operater, ki je večje napake tudi sproti odpravljal.

V govorni bazi so sorazmerno zastopani govorci obeh spolov, vseh starosti in večjih narečnih skupin. Podrobnejši podatki o postopku zajemanja, sestavi množice govorcev in drugih značilnostih zbirke se nahajajo v [46, 49, 77].

Postopku snemanja je sledilo označevanje izgovorjenih besed – labeliranje. Vsaka beseda je bila praviloma posneta tako, da se je pred in za njo nahajal še daljši signal tišine oziroma okolja. Premor pred besedo je dolg 200ms, za njo pa 1s. Izgovorjavam so bili "ročno" označeni začetki in konci posameznih besed.

Del govorne zbirke je bil leta 2001 označen še na nivoju manjših govornih enot – fonemov. Od takrat je zbirka uporabna tudi za fonemsko zasnovane SRG (npr. CSLU sistem).

Zbirka je bila v referenčnem okolju razdeljena na 3 različno velike dele:

- *3/5 posnetkov za proces učenja (učna množica – 468 izgovorjav[1])*
- *1/5 posnetkov za proces vmesnega vrednotenja (validacijska množica - 156 izgovorjav)*
- *1/5 posnetkov za proces končnega vrednotenja robustnosti (testna množica - 156 izgovorjav)*

V procesih učenja obeh referenčnih sistemov sta bili uporabljeni le podmnožici učne množice; delno zaradi omejevanja računske kompleksnosti učenja, delno zaradi posebnosti samih sistemov. CSLU sistem namreč pri učenju uporabi le naključno izbran vzorec posameznih vektorjev značilk iz celotne učne množice. HMM sistem pa uporabi celotne izgovorjave, zato zanj zadošča že manjša podmnožica posnetkov.

---

[1] *Vsaka izgovorjava vsebuje vseh 13 besed iz slovarja.*

3.1.2 Govorna zbirka NUMBERS

Zbirka je nastajala v ZDA v 90. letih. Vsebuje približno 10000 izgovorjav, v katerih govorci dokaj spontano (s kratkimi premori) sporočajo številske podatke (poštna, telefonska in hišna številka). Posnetki že vsebujejo realne motnje (glasba v ozadju, dihanje, poki, …).

Približno 6600 posnetkov je bilo označenih tudi na fonemskem nivoju s strani poklicnih označevalcev. Vse transkripcije so zapisane v Worldbet standardu [52].

Zbirka je bila v referenčnem okolju razdeljena na 3 različno velike dele :

- *3/5 posnetkov za proces učenja (učna množica – 6087 izgovorjav)*
- *1/20 posnetkov za proces vmesnega vrednotenja (validacijska množica - 555 izgovorjav)*
- *1/10 posnetkov za proces končnega vrednotenja robustnosti (testna množica – 1168 izgovorjav)*

V vsaki izgovorjavi (datoteki) je v primerjavi z govorno zbirko ŠTEVKE precej manj govornih enot - besed (povprečno 6, ŠTEVKE pa 13).

3.1.3 Vrednotenje robustnosti SRG

V skladu z osnovnimi izhodišči je bila uspešnost referenčnih SRG merjena v simuliranih težjih pogojih delovanja, kjer je inherentna robustnost bolj izrazita. Simulacija teh pogojev je bila izvedena z dodajanjem motenj osnovni testni množici. Uporabljene so bile motnje iz dveh glavnih zvrsti: aditivne motnje (7 posnetkov) in konvolutivne spremembe (4 postopki).

V primeru kombinacije aditivnih in konvolutivnih motenj so bile slednje dodane kar množicam z že vsebovanimi aditivnimi motnjami. Ob upoštevanju vseh možnih kombinacij je nastalo veliko število novih testnih množic:

- *1 nespremenjena testna množica*
- *3 skupine po 7 testnih množic za aditivne motnje ob 3 izbranih vrednostih ciljnega razmerja energij signal/šum[1] (3*7=21 množic)*

[1] *Angl. "Signal to Noise Ratio" ali krajše SNR.*

- *4 skupine kombinacij aditivnih in konvolutivnih sprememb (4*21=84 množic)*
- *1 skupina s kombinacijo konvolutivnih sprememb in prvotne testne množice (4 množice)*

Zaradi velikega števila so bile množice v praktičnih preizkusih uporabljane selektivno v sorazmernem obsegu s pričakovano pomembnostjo rezultatov. Podrobnejša specifikacija vseh testnih množic in ustreznih oznak se nahaja v dodatku D.

3.1.3.1 Aditivne motnje

Nastanejo predvsem zaradi akustičnih okolij, v katerih se govorci nahajajo. To so pravzaprav dodatni zvoki, ki se zajamejo skupaj z govorom ali se prištejejo kasneje pri prenosu oziroma reprodukciji. Za tovrstno simulacijo so bili izbrani realni posnetki nekaterih značilnih akustičnih okolij iz angleške govorne zbirke NOISEX [45]. V skupini aditivnih motenj je bilo uporabljenih sedem posnetkov:

- *govor več govorcev v zaprtem prostoru (angl. oznaka "Babble") - oznaka "govor"*
- *hrup v osebnem avtomobilu Volvo (angl. oznaka "Volvo") - oznaka "volvo"*
- *hrup v tovarni (angl. oznaka "Factory") - oznaka "tovarna"*
- *hrup v vojaškem letalu F-16 (angl. oznaka "F-16") - oznaka "F-16"*
- *beli šum (angl. oznaka "White") - oznaka "beli šum"*
- *roza šum (angl. oznaka "Pink") - oznaka "roza šum"*
- *pasovno omejen šum (angl. oznaka "BP 900Hz") - oznaka "pas 900Hz"*

Pasovno omejen šum (*"pas 900Hz"*) je nastal "umetno"; posnetek belega šuma je bil s pomočjo pasovno prepustnega filtra omejen na ožje frekvenčno področje med 800 in 1000Hz. Po tej značilnosti se dobljena motnja razlikuje od ostalih v skupini.

Vsi posnetki motenj so bili najprej s pasovno prepustnim filtrom omejeni na frekvenčni pas prepustnosti povprečne analogne telefonske linije (300Hz - 3400Hz), po kateri je bila posneta večina izgovorjav v obeh zbirkah. Nato so motnje bile dodane k izgovorjavam v testni množici z različnimi ciljnimi razmerji energij obeh signalov (SNR): *12dB*, *6dB* in *0dB*. Posnetki v govornih zbirkah že vsebujejo določeno količino motenj, zato je izračun razmerja lahko le približen. Za "čist" signal se je namreč upošteval dejanski posnetek. Poleg tega so med posameznimi besedami v zbirki ŠTEVKE daljši premori, ki "znižujejo" skupno energijo

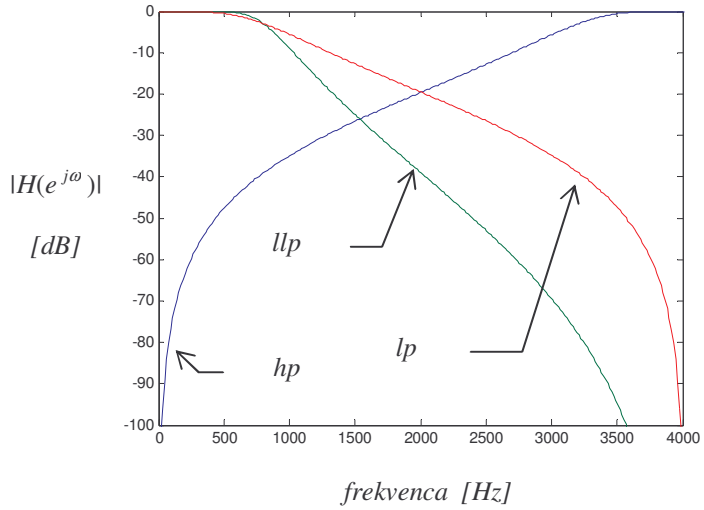
signala in s tem tudi energijo dodanih motenj. Vendar ta dejstva niso moteča; v skladu z osnovnimi izhodišči gre namreč za relativni značaj vrednotenja uspešnosti SRG pod vplivom izbire okna.

Z dodajanjem opisanih aditivnih motenj je testni del v obeh zbirkah pridobil skupino 7 novih testnih množic za vsako od treh izbranih vrednosti SNR. Posamezne množice so na grafičnih prikazih imenovane s prej navedenimi oznakami, povprečne vrednosti množic v skupini pa skupaj s ciljnim razmerjem SNR (npr. *"aditivne motnje 0dB"*).

3.1.3.2 Konvolutivne motnje

Poleg aditivnih motenj se v govornih signalih najpogosteje pojavijo tudi konvolutivne spremembe. Te so največkrat posledica karakteristik naprav za zajem in reprodukcijo govora ter akustičnih lastnosti prostorov, v katerih se komunikacija odvija. Pogosto lahko konvolutivne spremembe nastanejo še kot posledica lastnosti prenosnih poti, še posebej pri analognih telefonskih zvezah.

V referenčnem okolju so bile implementirane štiri tipične konvolutivne spremembe govornih signalov. Prve tri so preprosti NEO filtri (Slika 3.1), četrta pa je simulacija odjeka[1] v akustičnem prostoru (Slika 3.2). Vse so bile aplicirane samostojno ali v kombinaciji z aditivnimi motnjami.

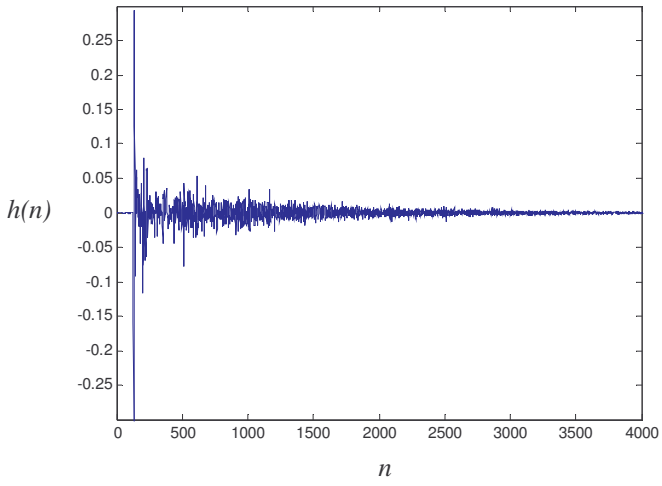


Slika 3.1: Amplitudni odzivi NEO filtrov za simulacijo konvolutivnih sprememb.

[1] *Angl. "reverberation"; pojav je še posebej značilen za prostoročno telefoniranje.*

Prvi konvolutivni motnji predstavljata nizkoprepustna filtra (*"lp"* in *"llp"*)[1] z različno izrazitim dušenjem v višjem frekvenčnem pasu. Vpliv obeh je zelo podoben primeru, ko je med govorcem in mikrofonom zvočna ovira ali ko pride do splošnega relativnega zmanjšanja amplitud višjih frekvenc v govornih signalih. V obeh oblikah je to dokaj pogost pojav.

Tretja konvolutivna motnja je visokoprepustni filter (*"hp"*[2]), ki ima komplementarni potek amplitudnega odziva. V praksi sprememba ni tako pogosta, bo pa v pomoč pri interpretaciji rezultatov. Četrta motnja je podana v obliki končnega odziva na enotin impulz, izmerjenega v realnem akustičnem prostoru (Slika 3.2). V njem se poleg osnovnega signala pojavi še cela vrsta odbojev – sicer precej manjših, a za zmanjšanje uspešnosti SRG dovolj velikih amplitud. V uporabljenem odzivu na enotin impulz sta energiji osnovnega signala in zaznanih odbojev približno enaki.



Slika 3.2: Odziv na enotin impulz akustičnega prostora z odjekom.

3.2 HMM RAZPOZNAVALNIK LOČENIH BESED

Prikriti Markovi modeli (HMM) so trenutno najbolj razširjen koncept na področju ARG. Uporabni so kot akustični in časovni modeli. Kot prvi aproksimirajo dejansko porazdelitev posameznih značilk s pomočjo mešanice Gaussovih normalnih porazdelitev (krajše GMM[3]). S tem konceptom se bistveno zmanjša število potrebnih parametrov, kar je s stališča računske

[1] *Okrajšavi pomenita angleško oznako "lowpass" za nizkoprepustne filtre.*

[2] *Pomeni okrajšavo angleške oznake "highpass" za visokoprepustne filtre.*

[3] *Angl. "Gauss Mixture Model".*

učinkovitosti zelo dobrodošlo. Hkrati pa pomeni omejitev pri aproksimaciji dejanskih porazdelitev vrednosti značilk, ki v splošnem niso podobne normalni porazdelitvi. Zato se kot boljša rešitev vse pogosteje pojavljajo hibridni SRG, ki temeljijo na uporabi nevronskih mrež; te omenjeno slabost presežejo.

HMM modeli so bolj nepogrešljivi pri časovnem modeliranju. V postopku učenja se namreč v parametrih notranjih stanj akumulira opis časovne dinamike značilnosti posamezne govorne enote. Torej so ti podatki zelo uporabni tudi v druge namene (npr. sinteza govornih signalov).

Referenčni HMM SRG je šolski primer realizacije statističnega pristopa. V nadaljevanju so najprej predstavljene teoretične osnove zveznih levo-desnih HMM modelov, ki so za opisovanje govornih signalov še posebej primerni. Sledi še opis implementacije v programskem paketu HTK[1]. Opisi v nadaljevanju so omejeni na najpomembnejše informacije; več podrobnosti se nahaja v [47, 77].

3.2.1 Teorija zveznih prikritih Markovih modelov

Zvezni Markov model prvega reda ima končno množico notranjih stanj $\boldsymbol{S}=\{1, 2, \ldots, N\}$. V vsakem diskretnem časovnem trenutku t se nahaja v enem od teh stanj; označi se s q_t. V diskretnih časovnih trenutkih pa model med notranjimi stanji prehaja glede na vnaprej podane verjetnosti prehodov

$$a_{ij}=P(q_{t+1}=j \mid q_t=i), \qquad 1 \le i, j \le N. \tag{3.1}$$

V izrazu (3.1) a_{ij} določa verjetnost prehoda iz stanja i v stanje j. Osnovna značilnost tovrstnih modelov je predpostavka, da je verjetnost prehoda med stanji odvisna le od predhodnega in ne ostalih stanj, v katerih se je proces nahajal.

Delovanje celotnega procesa se opiše z množico zaporednih notranjih stanj

$$\boldsymbol{q}=\{q_1, q_2, \ldots, q_T\}. \tag{3.2}$$

Ob tem proces v vsakem stanju generira poljuben vektor izhodnih simbolov, ki mu rečemo opazovanje. Verjetnost izhodnega opazovanja $\boldsymbol{o}_t$ v trenutku t in stanju $q_t=i$ je določena z izrazom:

---

[1] *URL: "http://htk.eng.cam.ac.uk/".*

$$b_i(\boldsymbol{o}_t) = P(\boldsymbol{o}_t \mid q_t=i), \quad 1 \le i \le N, \tag{3.3}$$

pri čemer je $\boldsymbol{o}_t$ del celotnega zaporedja opazovanj $\boldsymbol{O}$, ki ga je proces generiral

$$\boldsymbol{O}=\{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}. \tag{3.4}$$

Iz danega zaporedja izhodnih opazovanj ni mogoče enolično rekonstruirati zaporedja notranjih stanj procesa, ker preslikava ni bijektivna. Prehajanje procesa med stanji kljub znanemu zaporedju opazovanj ostane skrito[1].

V referenčnem razpoznavalniku so uporabljeni HMM modeli z zvezno zalogo vrednosti za vsako komponento izhodnega opazovanja oziroma značilko. To lastnost je računsko praktično nemogoče obvladati, zato se modeli pogosto parametrično poenostavijo. Poljubna zvezna porazdelitvena funkcija se tako aproksimira le s končno, uteženo mešanico posameznih gostot Gaussovih normalnih porazdelitvenih funkcij[2]. Verjetnostna gostota v stanju i je določena z izrazom:

$$b_i(\boldsymbol{o}_t) = \sum_{m=1}^{R} c_{im}\, b_{im}(\boldsymbol{o}_t), \qquad 1 \le i \le N. \tag{3.5}$$

V (3.5) je R število gostot Gaussovih normalnih porazdelitev v mešanici in c_{im} ustrezni utežnostni koeficient posamezne gostote $b_{im}(\boldsymbol{o}_t)$, ki mora ustrezati še pogojem

$$\sum_{k=1}^{R} c_{jk} = 1, \quad 1 \le j \le N \quad in \quad c_{jk} \ge 0, \quad 1 \le j \le N,\ 1 \le k \le R. \tag{3.6}$$

Če je izhodno opazovanje sestavljeno iz D komponent, se za njegov opis uporabi D-dimenzionalna gostotna funkcija Gaussove normalne porazdelitve

$$b_{im}(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{im}; \boldsymbol{\Sigma}_{im}). \tag{3.7}$$

Ta je določena s pomočjo kovariančne matrike $\boldsymbol{\Sigma}_{im}$ in vektorja srednjih vrednosti $\boldsymbol{\mu}_{im}$:

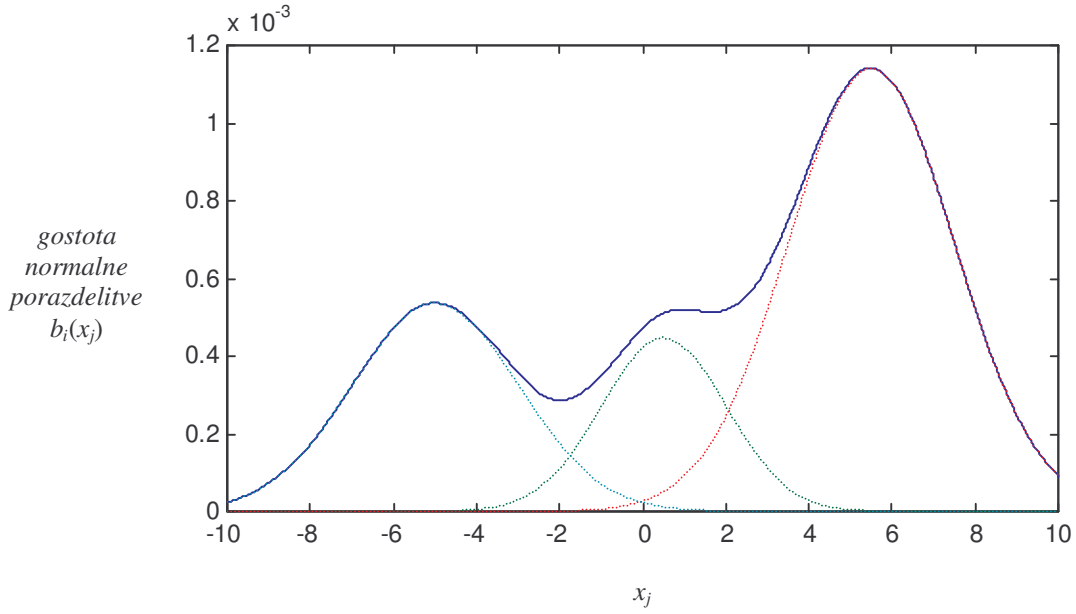
---

[1] *Od tod izvira oznaka "prikriti".*

[2] *Ta se pri razpoznavanju govora v mešanici najpogosteje uporablja, čeprav se lahko v splošnem izbere tudi katera od drugih porazdelitvenih funkcij.*

$$\mathcal{N}(\boldsymbol{o}_t;\boldsymbol{\mu}_{im};\boldsymbol{\Sigma}_{im}) = \frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}_{im}\right|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{o}_t-\boldsymbol{\mu}_{im})^T \boldsymbol{\Sigma}_{im}^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_{im})}. \quad (3.8)$$

Enostavna in gladka porazdelitvena funkcija se lahko učinkovito aproksimira na prikazan način. Zadošča že manjše število komponent v mešanici. V primeru realne porazdelitve pa tovrstna aproksimacija predstavlja njeno zgladitev oziroma posplošitev, ki je s stališča računske učinkovitosti sicer zaželena, a hkrati lahko povzroči veliko napako. Slika 3.3 prikazuje primer gostote verjetnosti posamezne komponente opazovanja kot mešanice treh gostot normalnih porazdelitev.



Slika 3.3: Gostota porazdelitve komponente opazovanja x_j kot mešanica treh gostot normalnih porazdelitev.

Za aproksimacijo bolj razgibane gostote je potrebno večje število komponent, kar zelo hitro povečuje računsko kompleksnost modelov. Ta se lahko učinkovito zmanjša, če se kovariančna matrika $\boldsymbol{\Sigma}_{im}$ nadomesti z vektorjem njenih diagonalnih elementov. Brez večje napake je to mogoče le pri statistično neodvisnih značilkah v opazovanju[1]. Zato sta v tem primeru napaka in s tem tudi vpliv na končno uspešnost sistema zelo odvisna od koreliranosti značilk v opisu signala.

[1] *Nediagonalne vrednosti kovariančne matrike so v tem primeru zanemarljive.*

V splošnem je HMM model določen z množico notranjih stanj $\boldsymbol{S}$, množico izhodnih opazovanj $\boldsymbol{V}$ in tremi verjetnostnimi porazdelitvami

$$\lambda=(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\Pi}). \tag{3.9}$$

Množici $\boldsymbol{A}$ in $\boldsymbol{B}$ opisujeta verjetnostne porazdelitve prehodov med stanji in izhodnih opazovanj za vsako stanje

$$\boldsymbol{A}=\{a_{ij}\}, \tag{3.10}$$

$$\boldsymbol{B}=\{b_i(\boldsymbol{o}_t)\}. \tag{3.11}$$

Množica $\boldsymbol{\Pi}$ opisuje verjetnostno porazdelitev začetnih stanj procesa

$$\boldsymbol{\Pi}=\{\pi_i\}, \tag{3.12}$$

$$\pi_i = P(q_1=i), \qquad 1 \le i \le N. \tag{3.13}$$

Pri razpoznavanju govora se HMM modeli še dodatno omejijo pri prehajanju med notranjimi stanji; v nekem trenutku lahko, ali ostanejo v istem stanju i ali pa preidejo v naslednje $i+1$. Tovrstni modeli se imenujejo tudi "levo-desni"[1], ker je prehajanje možno le v smeri od začetnega stanja z oznako 1 proti končnemu N. Začetno stanje je vedno tudi prvo v množici zaporednih stanj $\boldsymbol{q}$. Za te modele torej velja

$$a_{ij} = \begin{cases} =1 & i=j=N \\ \ge 0 & 1 \le i < N,\ j=i,\ i+1 \\ =0 & sicer \end{cases} \qquad in \quad \pi_i = \begin{cases} 1 & i=1 \\ 0 & sicer \end{cases}. \tag{3.14}$$

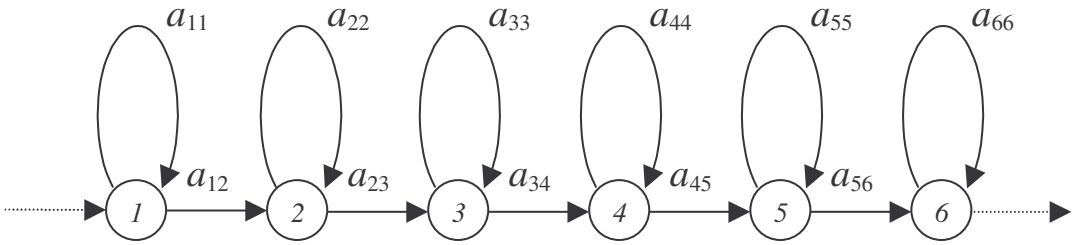
Primer "levo-desnega" modela s 6 notranjimi stanji prikazuje Slika 3.4.

Iz navedenih izrazov se lahko izpelje izraz za verjetnost, da je proces pri znanem zaporedju notranjih stanj $\boldsymbol{q}=\{q_1, q_2, q_3, \ldots, q_T\}$ generiral zaporedje opazovanj $\boldsymbol{O}=\{\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3, \ldots, \boldsymbol{o}_T\}$

$$P(\boldsymbol{O},\boldsymbol{q} \mid \lambda) = \pi_{q_1} b_{q_1}(\boldsymbol{o}_1) \prod_{t=2}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{o}_t). \tag{3.15}$$

[1] *Po avtorju se imenujejo tudi Bakisovi modeli.*

Izraz (3.15) je enostaven, ker je zaporedje notranjih stanj znano. Pri prikritih Markovih modelih je to zaporedje skrito, zato bi bilo teoretično potrebno izračunati vsoto teh verjetnosti po vseh možnih zaporedjih notranjih stanj (3.16). To je računsko zelo zahteven postopek, ki ga je potrebno učinkoviteje rešiti.



Slika 3.4: Levo-desni HMM model s 6 stanji.

Pri praktični uporabi HMM modelov se rešujejo trije osnovni problemi[1]:

- *problem evaluacije: izračun verjetnosti, s katero model* $\lambda=(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\Pi})$ *generira dano zaporedje opazovanj* $\boldsymbol{O}=\{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$.

  Postopek izračuna verjetnost kot statistično mero za podobnost med opisom določene govorne enote, "skritim" v parametrih posameznega modela in dejanskim zaporedjem opazovanj kot trenutnim vhodnim opisom. Na osnovi rezultatov tega postopka se izvrši dejanska izbira med hipotezami v procesu razpoznavanja. Običajno se izbere govorna enota, ki ustreza modelu z največjo verjetnostjo generiranja vhodnega opisa signala.

  Osnova reševanja tega problema je torej izračun verjetnosti:

$$P(\boldsymbol{O}\mid\lambda)=\sum_{q_1,q_2,\ldots,q_T}\pi_{q_1}b_{q_1}(\boldsymbol{o}_1)\prod_{t=2}^{T}a_{q_{t-1}q_t}b_{q_t}(\boldsymbol{o}_t)\,. \tag{3.16}$$

  Ker je zaporedje stanj nedoločljivo, je potrebno izračunati vsoto parcialnih verjetnosti po vseh možnih zaporedjih notranjih stanj. To je računsko kompleksen problem (N^T členov vsote), ki se z uporabo t.i. "Forward – Backward" algoritma zmanjša na kompleksnost N^2T [47].

[1] *Podrobnejši opis se nahaja v [67].*

- *problem razpoznavanja: določanje zaporedja notranjih stanj za model λ pri danem zaporedju opazovanj **O***

  Ker je zaporedje notranjih stanj procesa skrito oziroma nedoločljivo, je mogoče zaporedja stanj le predvidevati. Pri reševanju tega problema se lahko uporabijo različni postopki. Najpogosteje se izbere kar pot od začetnega do končnega stanja sistema, ki maksimira verjetnost $P(\boldsymbol{O}|\lambda)$. Iskano zaporedje stanj se gradi postopoma s pomočjo dinamičnega programiranja. Končni rezultat je najbolj verjetno zaporedje notranjih stanj procesa. Postopek je znan tudi kot "Viterbijevo iskanje".

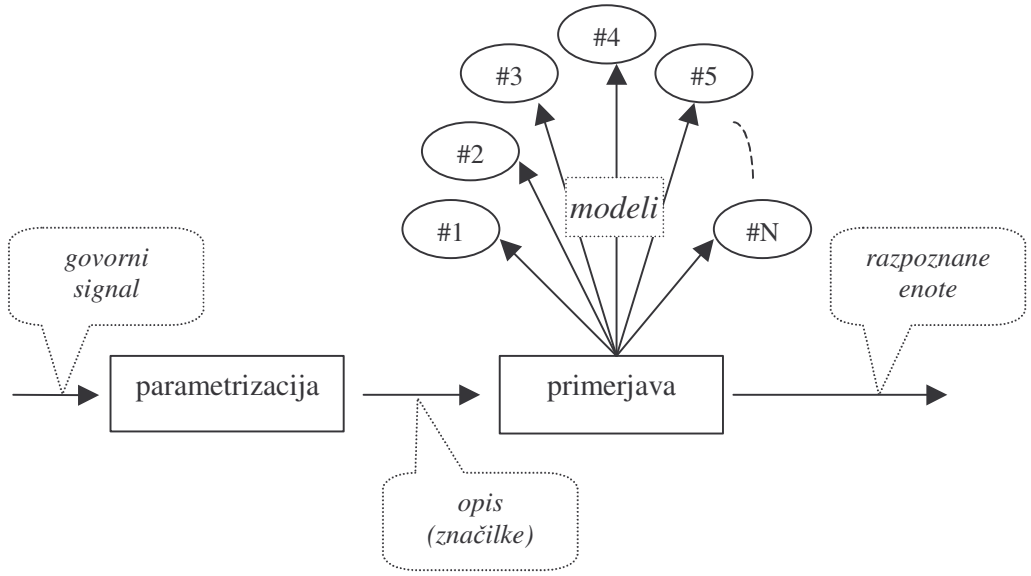
- *problem učenja: določanje parametrov modela $\lambda=(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\Pi})$, da bo verjetnost $P(\boldsymbol{O}|\lambda)$ za podano zaporedje opazovanj **O** največja*

  Problem je zahteven, ker ni znana njegova analitična rešitev. Obstaja pa razmeroma učinkovit iterativni postopek ("Baum-Welchev" algoritem), s katerim se formalno doseže monotono povečevanje želene verjetnosti v vsaki iteraciji do lokalnega maksimuma kriterijske funkcije $P(\boldsymbol{O}|\lambda)$. Glavna slabost tega pristopa je, da ne zagotavlja globalne optimalne rešitve. Zato je uspešnost zelo odvisna od izbire začetne točke optimizacije. Na srečo že enakomerna začetna porazdelitev opazovanj med vsa notranja stanja pripelje do dobrih rešitev.

  Za učenje HMM modelov se lahko uporabijo tudi alternativni postopki. Med njimi so posebej zanimivi tisti z "diskriminatornim" načinom učenja, ki pri klasifikacijskih problemih običajno pripelje do boljših rezultatov.

3.2.2 Implementacija HMM razpoznavalnika

Delovanje in zgradba HMM SRG sta zelo podobni splošnemu prototipu SRG (2. poglavje), zato po analogiji tudi tukaj nastopata dva procesa (Slika 3.5). Prvi (parametrizacija) obsega vsa opravila do končne oblike vektorjev značilk. Drugi (razpoznavanje) pa temelji na izračunavanju verjetnosti posameznih hipotez s pomočjo modelov oziroma njihovih parametrov – postopek Viterbijevega iskanja. Za končni izid razpoznavanja se izbere hipoteza z največjo verjetnostjo. Parametri modelov se določijo že prej v postopku učenja.



Slika 3.5: Zgradba splošnega HMM SRG.

Osnovne atribute HMM referenčnega sistema podaja Tabela 3.1. Glede na te vrednosti in njegov namen se lahko označi kot srednje zahteven. Glede na slovar in način govora je zahtevnost sicer manjša, vendar jo njegovo delovno okolje (klasična telefonija in posebnosti govornih zbirk) ustrezno poveča.

Atributi	***Zaloga vrednosti po naraščajoči zahtevnosti***
časovna povezanost govora	posamezne besede
število govorcev	neznan govorec
velikost slovarja	majhen slovar (13 besed – ŠTEVKE)
opis ciljnega jezika	enostavna gramatika – enakovredna izbira med omenjenimi besedami (ŠTEVKE)

Tabela 3.1: Osnovni atributi referenčnega HMM SRG.

Sistem je implementiran v programskem jeziku ANSI-C s pomočjo paketa HTK[1]. Ta je rezultat večletnega razvoja in raziskovalnega dela na Univerzi v Cambridgeju, trenutno pa je last podjetja Microsoft. Sestavljen je iz vrste posameznih aplikacij, ki izvajajo osnovne operacije v delovanju HMM SRG (pri naštevanju so dodana še imena ustreznih aplikacij):

[1] *Uporabljena verzija* 1.4A. *Podrobnejša dokumentacija je na naslovu: http://htk.eng.cam.ac.uk/.*

- *parametrizacija govornega signala ("HCode")*
- *inicializacija HMM modelov ("HInit")*
- *učenje HMM modelov ("HRest", "HERest")*
- *razpoznavanje ("HVite")*

V postopku priprave sistema se najprej govorna zbirka razdeli na učni in testni del. Nato se izvede parametrizacija učnih in testnih primerov ("*HCode"*). Govorni signal se razdeli na 32 ms dolge okvirje, ki so med seboj razmaknjeni za 10ms. Najprej se izračuna 22 FBANK značilk, iz njih pa še MFCC značilke, ki so bile uporabljene v praktičnih preizkusih:

- *MFCC značilke* (*13 značilk v enem vektorju*)

Vrednosti parametrov modelov se določajo v postopku učenja, ki se po izbranem kriteriju optimalnosti prilagajajo opisom besed v učnem delu baze. V referenčnem sistemu je uporabljen kriterij "maksimalne podobnosti"[1], ki vzpodbuja povečevanje verjetnosti za tisti model, ki ustreza govorni enoti v trenutnemu vhodnemu opisu. Kriterij sicer ni optimalen[2], je pa zaradi razmeroma učinkovitega postopka učenja najbolj razširjen (podpoglavje 3.2.1).

Pri učenju referenčnega sistema se najprej določijo začetne vrednosti parametrov v vseh modelih[3] ("*Hinit"*). Nato se učenje izvaja najprej ločeno po posameznih modelih[4]("*HRest"*), nato pa še v zaporedju po celotnih izgovorjavah ("*HERest"*). Ta korak se običajno iterativno izvede večkrat (v tem primeru 6 iteracij). Nato se poveča število mešanic in ves postopek učenja po celotnih izgovorjavah se ponavlja do končne vrednosti 8 mešanic. S tem se oblikujejo optimalni modeli, ki predstavljajo jedro celotnega sistema. Uspešnost učenja se preveri še z razpoznavanjem ("*HVite"*) na neodvisnih testnih množicah. V njih se nahajajo izgovorjave, ki niso bile v učnem delu.

V procesu razpoznavanja poteka vrednotenje posameznih hipotez s pomočjo Viterbijevega iskanja. Z njim se določijo vsem 13 modelom λ_i verjetnosti $P(\boldsymbol{O}|\lambda_i)$, s katerimi generirajo opis vhodnega signala $\boldsymbol{O}$. Med njimi se izbere tista beseda w_i, kateri ustrezni model kaže največjo verjetnost:

---

[1] *Krajše tudi MP – angl. "ML - Maximum Likehood".*

[2] *Povezava med ML kriterijem in končnim številom napak v razpoznavanju je razmeroma šibka [70].*

[3] *HMM besedni modeli so vsebovali 8 notranjih stanj.*

[4] *V referenčnem sistemu so modeli za izgovorjave vseh možnih besed in model tišine, ki opisuje signal pred in po izgovorjavi.*

$$w_i = \arg\max_{1 \le i \le 13} P(\boldsymbol{O} \mid \lambda_i). \tag{3.17}$$

3.3 CSLU RAZPOZNAVALNIK NIZOV ŠTEVK

Razpoznavalnik je tipičen predstavnik novejšega vala bolj kompaktnih in učinkovitejših SRG. Temelji na kombinaciji dveh trenutno najbolj razširjenih pristopov. Klasični trinivojski perceptron nastopa v vlogi akustičnega modela, medtem ko algoritem Viterbijevega iskanja opravlja funkcijo časovnega modela. V naslednji tabeli so prikazani osnovni atributi sistema.

Atributi	*Zaloga vrednosti po naraščajoči zahtevnosti*
časovna povezanost govora	poljubno dolga zaporedja števk
število govorcev	neznan govorec
velikost slovarja	majhen slovar (13 besed – ŠTEVKE, 10 – NUMBERS)
opis ciljnega jezika	enostavna gramatika – enakovredna izbira med omenjenimi besedami (ŠTEVKE) in nizi besed (NUMBERS)

Tabela 3.2: Osnovni atributi referenčnega CSLU SRG.

Bistvena prednost razpoznavalnika je majhna prostorska in časovna kompleksnost [53]. Zaradi nekaterih poenostavitev je bolj uporaben za majhne in srednje velike slovarje ter omejeno jezikovno domeno. Še nedavno bi bila taka omejitev označena kot pomanjkljivost. Danes pa imajo njegove prednosti precej večji pomen.

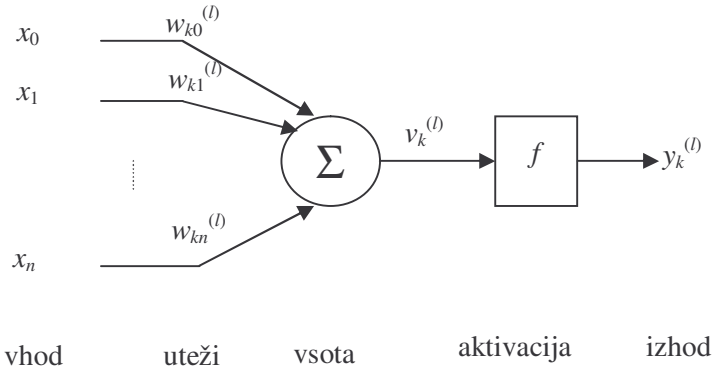
Na področju kompleksnih razpoznavalnikov z velikimi slovarji in kompleksnimi jezikovnimi modeli se je razvoj nekoliko "upočasnil" in se vse bolj nagiba v smeri uporabe metod "grobe sile" (npr. podvajanje aparaturnih elementov in struktur). Že tako kompleksni sistemi so tako še bolj na robu razpoložljivih aparaturnih zmogljivosti. Zato je njihova uporabnost v okolju sodobnih komunikacijskih sistemov z velikim številom uporabnikov (npr. mobilna telefonija) zelo omejena. Povrh vsega je pri zahtevnejših aplikacijah (npr. narek) za zadovoljivo uspešnost potrebna vsaj delna prilagoditev vsakemu govorcu posebej. V tem primeru se zdi alternativna zasnova z večjim številom enostavnejših, učinkovitejših in individualno prilagojenih razpoznavalnikov precej bolj primerna. Še posebej to velja za vse bolj aktualno paradigmo porazdeljenega procesiranja govornih signalov [54]. Zasnova je zanimiva tudi zaradi predvidenega zlitja tehnologij razpoznavanja, sinteze in prenosa govornih signalov po digitalnih omrežjih - t.i. digitalne konvergence. Ob tem se bodo namreč poenotili tudi opisi

govornih signalov, od katerih se pričakuje tako dobra reprodukcija zajetih signalov (pomembno za osebno govorno komunikacijo) kot tudi uspešno razpoznavanje njihove vsebine (npr. avtomatska tvorba pisnih dokumentov). Obe zahtevi sta za obstoječe SRG dokaj kontradiktorni, še posebej za kompleksnejše, od govorca neodvisne razpoznavalnike. Ti namreč uporabljajo splošnejše opise signalov, ki ne omogočajo njihove natančne reprodukcije; zanjo so potrebne tudi informacije o individualnih značilnostih izgovorjave in glasu. Zato je razmišljanje o enotni zasnovi mreže sistemov z individualnimi razpoznavalniki precej bolj logično, vloga enostavnejših SRG pa vse pomembnejša.

V nadaljevanju so najprej predstavljene osnovne značilnosti nevronskih mrež pri akustičnem modeliranju govornih signalov, zatem pa še podrobnosti konkretne izvedbe posameznih postopkov v referenčnem CSLU razpoznavalniku.

3.3.1 Nevronske mreže kot akustični modeli

Nevronske mreže so nastale kot poskus funkcionalnega modeliranja skupkov živčnih celic. V zadnjem času je prišlo do intenzivnega razvoja tega koncepta, ki je zelo popularen na najrazličnejših področjih. Osnovna enota nevronskih mrež je matematični model živčne celice (Slika 3.6).



Slika 3.6: Matematični model živčne celice - nevrona.

Nevron k na nivoju l je sestavljen iz vhodnih uteži $w_{ki}^{(l)}$, ki se pri izvajanju pomnožijo z vhodnimi vrednostmi x_i

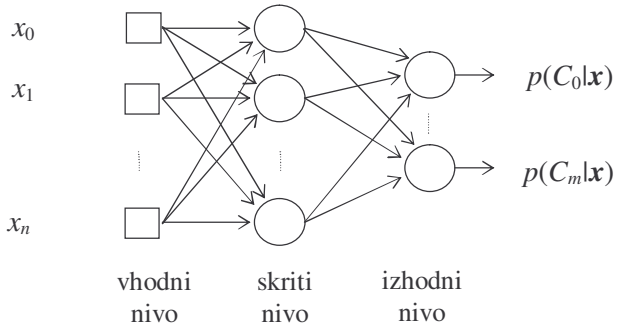
$$v_k^{(l)} = \sum_{i=0}^{n} w_{ki}^{(l)} x_i \ . \tag{3.18}$$

Vsota vseh produktov $v_k^{(l)}$ se preslika še skozi izbrano aktivacijsko funkcijo f v izhodno vrednost oziroma aktivacijo nevrona $y_k^{(l)}$

$$y_k^{(l)} = f(v_k^{(l)}). \tag{3.19}$$

Nevroni se horizontalno združujejo v večje skupine - nivoje. Znanih je več različnih zvrsti mrežnih konceptov. Med najbolj popularnimi je vsekakor trinivojski perceptron, ki sodi med usmerjene nevronske mreže. Zaradi svoje enostavnosti in univerzalnosti je na področju ARG še posebej razširjen. Najpogosteje opravlja vlogo akustičnega modela; posamezne opise signala razvršča v elementarne razrede osnovnih govornih enot – običajno fonemov.

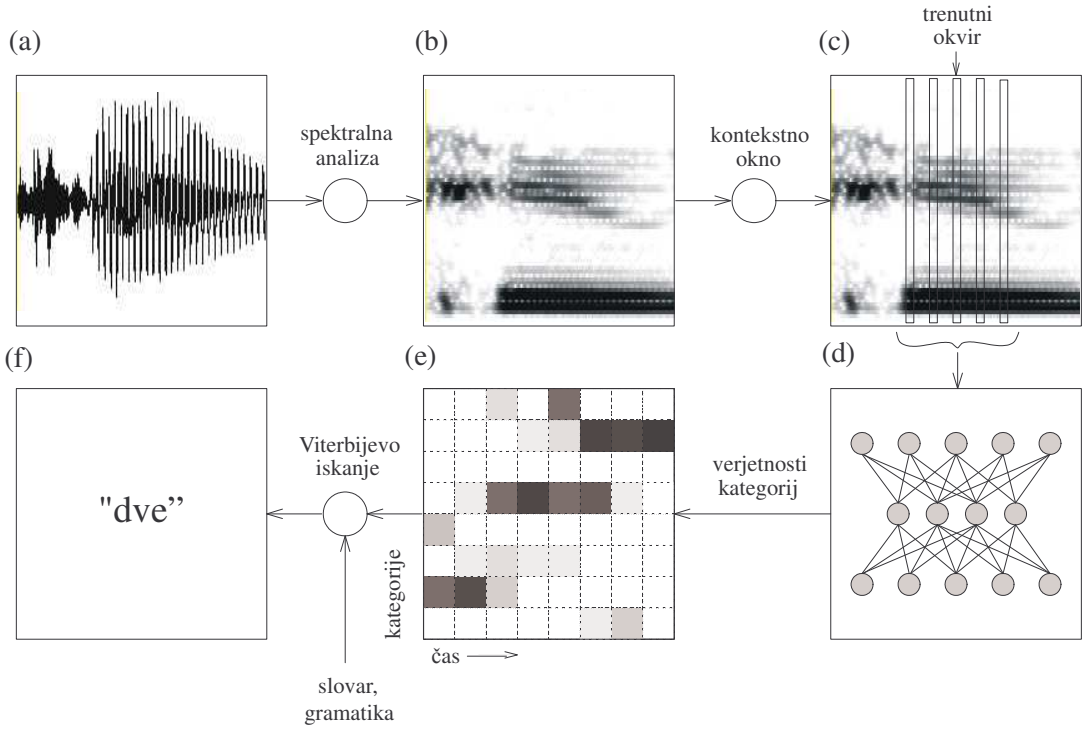
Trinivojski perceptron je mreža s tremi nivoji nevronov in dvema nivojema medsebojnih usmerjenih povezav (Slika 3.7). V vhodni nivo vstopijo vrednosti vektorja značilk $\boldsymbol{x}$, ki predstavlja trenutni akustični opis govornega signala. Nato se izračunajo aktivacije nevronov v "skritem" nivoju. V zadnjem, izhodnem nivoju je toliko nevronov, kot je razredov, ki ustrezajo posameznim govornim enotam v klasifikacijskem problemu. Vrednosti izhodnega vektorja so enake izračunanim aktivacijam nevronov na zadnjem nivoju. Če je perceptron pravilno naučen z zadostnim številom učnih primerov, se na vseh izhodih mreže pojavijo pogojne verjetnosti $p(C_i|\boldsymbol{x})$ [55]; C_i predstavlja govorni enoti i ustrezni razred, $\boldsymbol{x}$ pa vhodni vektor značilk.



Slika 3.7: Trinivojski perceptron kot akustični model.

3.3.2 Implementacija CSLU razpoznavalnika

"CSLU Toolkit"[1] je programski paket, ki je nastal na Univerzi v Oregonu[2]. Vsebuje širok spekter orodij, tako za raziskovalno delo na področju razpoznavanja in sinteze govora kot za snovanje ter izvajanje govornih dialogov. Med drugim omogoča tudi enostavno realizacijo hibridnega SRG, ki je zaradi svoje kompaktne in učinkovite zasnove zelo zanimiv. Osnovno shemo hibridnega sistema prikazuje Slika 3.8[3].



Slika 3.8: Delovanje referenčnega CSLU razpoznavalnika.

Govorni signal (a) se najprej v procesu parametrizacije spremeni v zaporedje vektorjev značilk (b), nato vstopi v fazo akustične klasifikacije (c), ki jo opravi trinivojski perceptron (d). Rezultat klasifikacije je verjetnostna matrika (e), na kateri se izvede postopek Viterbijevega iskanja; ta določi končni rezultat razpoznavanja (f). V nadaljevanju so

[1] *URL: "http://cslu.cse.ogi.edu/toolkit/".*

[2] *"Center for Spoken Language Understanding", Oregon Graduate Institute, Oregon Health & Science University.*

[3] *Povzeto po "http://cslu.cse.ogi.edu/tutordemos/nnet_training/tutorial.html" in [77].*

podrobneje opisani pomembnejši procesi v delovanju sistema, več podatkov pa se nahaja še v dodatku C.

3.3.2.1 Parametrizacija

Postopek parametrizacije je podrobneje opisan že v podpoglavju 2.2. V hibridnem razpoznavalniku se odvija v skladu s tem opisom razen morebitnih malenkostnih razlik, ki so posledica implementacije. Govorni signal se razdeli na 32 ms dolge okvirje z 10 ms razmika. Najprej se izračuna 22 t.i. FBANK značilk, iz njih pa 13 MFCC značilk. Namesto izračuna klasičnih delta značilk po izrazu (2.7), se v tem primeru uporabi t.i. kontekstno okno; več sosednjih okvirjev se preprosto združi v skupni vektor značilk (Slika 3.8-c). Kontekstno okno v referenčnem sistemu vključuje značilke v petih vektorjih (poleg vektorja v času *t*, še vektorje pri *t*-60ms, *t*-30ms, *t*+30ms, *t*+60ms). S tem se učinkoviteje upošteva dinamika daljšega časovnega intervala (v tem primeru 120 ms) v primerjavi s klasičnimi delta značilkami (običajno le 50 ms). Na končnih značilkah se vedno izvede še standardni postopek normalizacije povprečne vrednosti[1] v vsaki izgovorjavi. V praktičnih preizkusih so bile uporabljene MFCC značilke s kontekstnim oknom:

- *MFCC značilke* (*5*13=65 značilk v enem vektorju*)

Glede na željo po preučitvi vpliva oken na bolj elementarni ravni bi bilo smiselno uporabiti še bolj osnovne predstavitve signala. Na končno uspešnost namreč v takem primeru vpliva precej manj dejavnikov. Vendar v tem primeru občutno pade tudi uspešnost SRG in je potemtakem opravičljivost takega vrednotenja vprašljiva. Menim, da je bolj pravilno meriti vpliv oken v pogojih, ki so bližje dejanski uspešnosti delovanja.

3.3.2.2 Kontekstno odvisne govorne kategorije

Nevronske mreže so za modeliranje časovnih zaporedij dokaj omejene. Zato realizacija besednih modelov ni smotrna. Logična izbira so po akustičnih značilnostih bolj homogene enote - fonemi. Besede v slovarju se zato zapišejo kot zaporedja fonemov, ki ponazarjajo njihovo izgovorjavo.

Fonemske zapise besed v slovarjih obeh referenčnih govornih zbirk prikazujeta Tabela 3.3 in Tabela 3.4. Uspešnost CSLU razpoznavalnika bo namreč preizkušena na obeh zbirkah.

---

[1] *Postopek kepstralne normalizacije – angl. "CMN" – Cepstral Mean Normalization.*

Beseda	*Fonemski zapis*	*Beseda*	*Fonemski zapis*
nič	{n i _tS tS}	osem	{o s E m }
ena	{E n a}	osem	{o s @ m }
dve	{_d d v e}	devet	{_d d E v e _t t}
tri	{_t t r i}	devet	{_d d @ v e _t t}
štiri	{S _t t i r i}	ja	{j a}
pet	{_p p e _t t}	ne	{n E}
šest	{S e s _t t}	stop	{s _t t O _p p}
sedem	{s e _d d E m}	premor	{.pau [.garbage .pau] }
sedem	{s e _d d @ m}		

Tabela 3.3: Fonemski zapis besed v zbirki ŠTEVKE.

Iz tabel je razvidno, da imajo nekatere besede več mogočih fonemskih zapisov. To je posledica različnih možnih izgovorjav; npr. "o s E m" ali "o s @ m", kjer govorci dokaj pogosto fonem "E" nadomestijo z mašilom "@".

Beseda	*Fonemski zapis*	*Beseda*	*Fonemski zapis*
zero	{z I 9r oU}	five	{f aI v}
oh	{oU}	six	{s I uc ks}
one	{w ^ n [^3]}	seven	{s E v ^2 n [^3]}
two	{uc th u}	eight	{ei uc [th]}
three	{T 9r i:}	nine	{n aI n [^3]}
four	{f >r}	separator	{.pau[.garbage .pau]}

Tabela 3.4: Fonemski zapis besed v zbirki NUMBERS.

Trenutna izgovorjava je običajno odvisna tudi od predhodnega in naslednjega fonema. Pojav se imenuje koartikulacija in je za govor zelo značilna. To je potrebno upoštevati tudi pri modeliranju fonemov. Običajen pristop v tej situaciji je razdelitev fonema na manjše enote, ki so lahko odvisne od predhodnih in naslednjih govornih enot. Fonemi se glede na njihovo trajanje razdelijo na en, dva ali tri dele; dobljene govorne enote se imenujejo monofon, bifon in trifon (Tabela 3.5).

Problem koartikulacije je na prikazan način učinkovito rešen, vendar za ceno precej večjega števila elementarnih govornih enot - kategorij. Na srečo se izkaže, da je mogoče kontekstne odvisnosti fonemov izraziti z relacijami med njihovimi večjimi skupinami. Trenutna

izgovorjava fonema je namreč bolj odvisna od zvrsti sosednjega fonema kot od njegovih konkretnih značilnosti. Zato sta leva in desna kontekstna odvisnost pogosto izraženi v povezavi s skupinami sorodnih fonemov, kar občutno zmanjša skupno število kategorij v sistemu. Razdelitve fonemov v manjše enote – kategorije – za obe referenčni govorni zbirki prikazujeta Tabela 3.6 in Tabela 3.7.

Fonem	*Št. enot*	*Kontekst*	*Primer*
monofon	1	kontekstno neodvisen	<a>
bifon	2	dve kontekstno odvisni polovici	d<a, a>n
trifon	3	kontekstno odvisni začetek in konec, kontekstno neodvisni koren	d<a, <a>, a>n

Tabela 3.5: Kontekstno modeliranje fonemov.

Fonem	*Št. enot*	*Fonem*	*Št. enot*	*Fonem*	*Št. enot*
.pau	1	v	2	@	2
_p	1	r	2	a	3
_t	1	S	2	o	3
_d	1	t	2	e	3
_tS	1	p	2	E	3
n	2	j	2	i	3
tS	2	s	2	O	3
d	2	m	2		

Tabela 3.6: Kontekstno odvisna delitev fonemov – ŠTEVKE.

Fonem	*Št. enot*	*Fonem*	*Št. enot*	*Fonem*	*Št. enot*
.pau	1	w	2	\>r	3
uc	1	I	2	i:	3
f	2	ks	2	u	3
v	2	^2	2	ei	3
T	2	^3	2	aI	3
s	2	^	2	oU	3
z	2	9r	2	th	r
n	2	E	2		

Tabela 3.7: Kontekstno odvisna delitev fonemov - NUMBERS.

Obstaja še posebna kategorija, katere verjetnost se izračuna implicitno[1] – t.i. Prag (angl. tudi "Garbage"). Njegova vrednost predstavlja osnovo za oceno sprejemljivosti najbolj verjetne hipoteze. Če njena verjetnost ni bistveno višja od Praga, se hipoteza zavrne. Na ta način se dokaj učinkovito reši splošen problem vseh razpoznavalnikov – zavrnitev nezanesljivo razpoznane ali sistemu neznane besede[2]. Enak problem je lahko v drugačnih zasnovah SRG precej težje rešljiv.

3.3.2.3 Učenje trinivojskega perceptrona

V CSLU razpoznavalniku se proces učenja perceptrona izvaja s standardnim postopkom vzvratnega razširjanja napake. Osnovni cilj postopka je zmanjševanje srednje kvadratne napake, ki je za *k*-ti učni primer določena z izrazom:

$$E(k) = \frac{1}{2}\sum_{j=1}^{P} e_j{}^2(k), \tag{3.20}$$

kjer je P število nevronov v izhodnem nivoju. Napaka izhodnega nevrona $e_j(k)$ je razlika med želeno $d_j(k)$ in dobljeno vrednostjo $o_j(k)$:

$$e_j(k) = d_j(k) - o_j(k)\,. \tag{3.21}$$

Postopek vzvratnega razširjanja napake sestavljata dva koraka:

1. ***"prehod naprej"*** *– izvajanje mreže; izračunajo se trenutni izhodi mreže kot odgovor na vhodni vektor značilk:*

   Učni primer k je označen z $[\boldsymbol{x}(k),\boldsymbol{d}(k)]$*;* $\boldsymbol{x}(k)$ *je vhodni vektor. Izračun aktivacij poteka za vse nevrone od vhodnega proti izhodnemu nivoju. Vrednost nevrona j v nivoju l ob učnem primeru k je*

$$v_j^{(l)}(k) = \sum_{i=1}^{Q} w_{ji}^{(l)}(k)\, y_i^{(l-1)}(k)\,, \tag{3.22}$$

   kjer sta $y_i^{(l-1)}(k)$ *izhod nevrona i v predhodnem nivoju in* $w_{ji}^{(l)}(k)$ *sinaptične uteži na povezavi med nevronoma i in j v prejšnjem ter trenutnem nivoju; Q je število*

[1] *Npr. povprečje petih hipotez z najvišjimi verjetnostmi.*

[2] *Angl. OOV - "Out Of Vocabulary rejection".*

nevronov v prejšnjem nivoju. Z uporabo sigmoidne aktivacijske funkcije je končna izhodna vrednost nevrona j v nivoju l

$$y_j^{(l)}(k)=\frac{1}{(1+e^{-v_j^{(l)}(k)})}. \tag{3.23}$$

Za nevron v prvem nivoju (l=1) velja

$$y_j^{(1)}(k)=x_j(k)\,, \tag{3.24}$$

kjer je $x_j(k)$ element vhodnega vektorja $\boldsymbol{x}(k)$. Odgovor mreže na vhodni vektor je enak izhodnim vrednostim nevronov v zadnjem nivoju L

$$o_j(k)=y_j^{(L)}(k). \tag{3.25}$$

2. ***"vzvratni prehod"*** *– učenje mreže; izvrši se prilagoditev uteži v skladu z izbranim pravilom popravljanja napake:*

Postopek se začne z izračunom napake na izhodnem nivoju $e_j(k)$ (3.21) in postopoma napreduje skozi nivoje proti vhodnim vrednostim. Uteži se pri tem popravljajo v skladu z izračunanim lokalnim gradientom kvadratne napake, ki je za nevron j v izhodnem nivoju L in uporabi sigmoidne aktivacijske funkcije določen z

$$\delta_j^{(L)}(k)=e_j^{(L)}(k)\,o_j(k)\;(1-o_j(k))\,, \tag{3.26}$$

kjer je $o_j(k)$ izhodna vrednost mreže in $e_j^{(L)}(k)$ napaka na izhodnem nivoju. Lokalni gradient za nevron j v skritem nivoju l je določen z izrazom

$$\delta_j^{(l)}(k)=y_j^{(l)}(k)\,(1-y_j^{(l)}(k))\quad\sum_{i=1}^{P}\delta_i^{(l+1)}(k)\,w_{ij}^{(l+1)}(k)\,, \tag{3.27}$$

kjer je P število nevronov na višjem nivoju. Sinaptične uteži mreže v nivoju l se spremenijo z upoštevanjem delta pravila

$$\Delta w_{ji}^{(l)}(k) = -\eta\, \delta_j^{(l)}(k)\, y_j^{(l-1)}(k)\,, \tag{3.28}$$

kjer je η učni parameter; ta se običajno med učenjem spreminja v skladu z empirično določenimi pravili.

Z iterativnim izvajanjem postopka vzvratnega razširjanja napake se izhodne vrednosti vse bolj približujejo zadanim učnim vrednostim v skladu s kriterijem srednje kvadratne napake

$$E_{MSE} = \frac{1}{K}\sum_{k=1}^{K} E(k). \tag{3.29}$$

Pri tem je proces učenja potrebno prekiniti, še preden pride do pretirane prilagoditve učnim podatkom, ki povzroči manjšo uspešnost na neodvisni validacijski množici. Zato je slednja na področju ARG običajno merilo za uspešnost postopka učenja. Popravljanje uteži se v CSLU sistemu izvaja na t.i. sprotni (angl. "online") način. To pomeni, da se opisani postopek izvede za vsak vhodni vektor posebej. Celoten postopek učenja je povzet v naslednjem okvirju.

Postopek učenja nevronske mreže

V učni množici se nahajajo pari vhodnih vektorjev $\boldsymbol{x}(k)$ in želenih izhodnih vrednosti mreže $\boldsymbol{d}(k)$:

$$\boldsymbol{U} = \{[\boldsymbol{x}(k), \boldsymbol{d}(k)]\ ;\ k=1,2,...,N\}. \tag{3.30}$$

Osnovni koraki učenja so:

- ***inicializacija*** *(začetne vrednosti vseh parametrov so običajno majhne naključno izbrane vrednosti)*
- *za vsak* ***prehod podatkov***
  - *za vsak* ***učni primer*** *iz* ***U***
    - *prehod naprej*
    - *vzvratni prehod s prilagajanjem uteži*
    - *prilagoditev učnega koraka*

V CSLU sistemu se najprej izvede učenje mreže v 40 iteracijah. Nato se dobljene mreže preizkusijo na validacijski množici in najuspešnejša med njimi še na končni testni množici.

Posebnost postopka učenja v fonemsko zasnovanih SRG (mednje sodi tudi CSLU razpoznavalnik) je t.i. trofazno učenje. Za izvedbo klasičnega postopka je namreč potrebna govorna zbirka, ki je v celoti označena na fonemskem nivoju. V praksi se zaradi visokih stroškov to običajno ne naredi. Fonemsko se označi le majhen del zbirke. Učenje se nato izvede v treh korakih:

1. *učenje nevronske mreže na fonemsko označenih učnih primerih (manjša podmnožica celotne učne množice)*
2. *naučena mreža avtomatsko fonemsko označi preostali del učne množice*
3. *postopek učenja se ponovi z večjim številom učnih primerov*

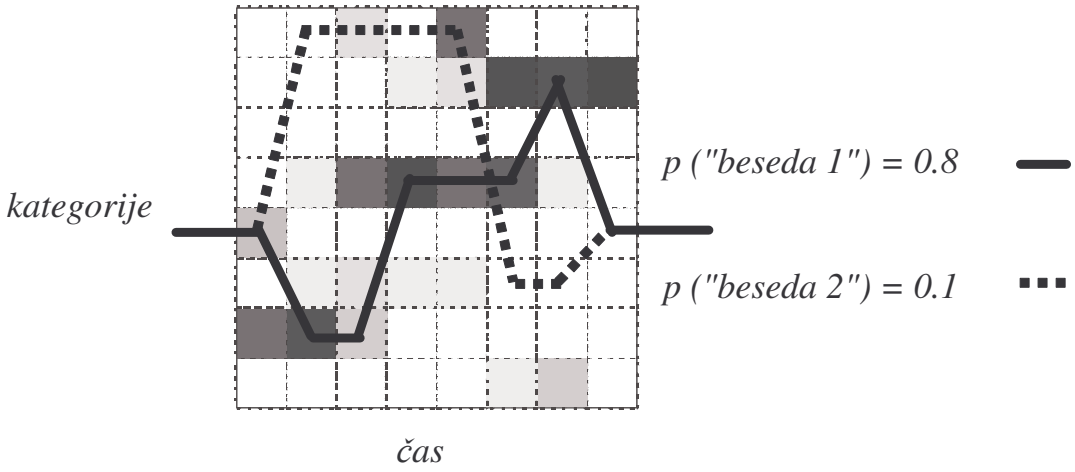
V CSLU paketu je na voljo tudi še postopek učenja po t.i. "Forward-Backward" načelu[1], ki se pri učenju tovrstnih sistemov kljub dodatni kompleksnosti pogosto uporabi. Osnovna posebnost tega postopka je v tem, da zaloga vrednosti želenih izhodnih vektorjev $\boldsymbol{d}(k)$ obsega zaprti interval [0..1]; želene vrednosti se izračunajo s pomočjo že omenjenega "Forward-Backward" algoritma. Pri klasičnem postopku učenja so zahtevane izhodne vrednosti lahko le 0 ali 1, kar je pri govoru pogosto v nasprotju z dejanskim stanjem – še posebej to velja za prehode med sosednjima fonemoma. Praktični preizkusi potrjujejo, da so tako naučene mreže v splošnem uspešnejše od tistih brez tega dodatnega postopka [56]. Pri relativnih primerjavah v tej disertaciji je to povečanje uspešnosti nepomembno, zato "Forward-Backward" postopek učenja ni bil uporabljen.

3.3.2.4 Viterbijevo iskanje

Rezultat procesa klasifikacije je matrika verjetnosti z R vrsticami (R je število kategorij) in T stolpci (T je število vseh vektorjev značilk). Glavna naloga Viterbijevega iskanja je, da matriko preslika v končne verjetnosti za vse besede v slovarju. Izvedbo postopka prikazuje Slika 3.9. Najprej se določijo dovoljena zaporedja kategorij iz zapisov besed v slovarju. Nato se po načelu dinamičnega programiranja "prehodi" pot za vsako besedo v skladu z dovoljenim zaporedjem kategorij. Prestopi med zaporednima kategorijama se izvedejo le takrat, ko je verjetnost naslednje večja od trenutne. Končni rezultat postopka so verjetnosti za vse besede v slovarju. Izbere se tista z največjo, ob pogoju, da je razpoznava dovolj zanesljiva.

---

[1] *Ideja izvira v podobnem postopku pri HMM razpoznavalnikih.*



Slika 3.9: Viterbijevo iskanje optimalne poti.

Postopek Viterbijevega iskanja je bolj učinkovit in uspešen, če upošteva še časovne omejitve trajanja posameznih kategorij. Te se v hibridnem sistemu določijo eksplicitno s pomočjo podatkov v učni množici. Tako se pri napredovanju poleg verjetnosti kategorij upošteva še skladnost z zgornjo in spodnjo mejo trajanja posamezne kategorije. Če je trajanje trenutne kategorije izven dovoljenih meja, se upošteva še kazen, ki je sorazmerna časovni razliki do spodnje ali zgornje časovne meje.

4. NAČRTOVANJE IN UPORABA NESIMETRIČNIH OKEN V SISTEMIH ZA RAZPOZNAVANJE GOVORA

Že v podpoglavju 2.2.1 je bila na kratko obravnavana vloga oken v frekvenčni analizi govornih signalov. V uvodnem poglavju pa so bili podrobneje pojasnjeni razlogi, zaradi katerih je raziskava uporabe nesimetričnih oken v SRG pomembna. Zamenjava okna je minimalni poseg v zgradbo SRG in ne prinaša nobenih dodatnih prostorskih ali računskih zahtev. Če ima zaznaven pozitivni učinek na robustnost SRG in hitrost razpoznave, je sprememba upravičena.

Rezultat frekvenčne analize signalov je kompleksni frekvenčni odziv. Na področju ARG se običajno ohrani le njegova absolutna vrednost – amplitudni odziv. Da bi bil vpliv okna na ta odziv ničen, bi okno moralo imeti idealne lastnosti, ki v praksi niso dosegljive. Podrobnejšem opisu vloge oken v frekvenčni analizi govornih signalov je posvečeno podpoglavje 4.1.

V primeru sprostitve zahteve po simetriji imajo nesimetrična okna nekaj potencialnih prednosti:

- *ožji glavni val ali nižji stranski valovi v amplitudnem odzivu*
- *krajša časovna zakasnitev*

Podrobnejši opis prednosti nesimetričnih oken se nahaja v podpoglavju 4.2. Postopek njihovega načrtovanja je sicer bolj kompleksen, a so pridobitve pomembnejše. V disertaciji je bilo potrebno uporabiti več različnih kriterijev, želenih odzivov in specifičnih metod načrtovanja, zato je smiselna uporaba splošnejših optimizacijskih metod. Te, za razliko od posebej prilagojenih [106-120], dopuščajo večjo svobodo pri določanju želenih lastnosti rešitve in optimizacijskih kriterijev, hkrati pa so lahko časovno zahtevnejše. Vendar so, glede na potreben čas za preučitev in implementacijo množice specializiranih metod, splošne metode v tem primeru boljša izbira.

Uporaba splošnih optimizacijskih metod je zanimiva še z drugačnega vidika. V prihodnje se namreč pričakuje zlitje procesov parametrizacije in razpoznavanja, kar pomeni, da ju bo mogoče bolje prilagajati dejanskim podatkom v poenotenem postopku učenja. Obstajajo že prvi prototipi tovrstnih razpoznavalnikov, ki zaenkrat temeljijo na t.i. globalni nevronski mreži; ta vključuje tako proces parametrizacije kot tudi akustične klasifikacije [80]. V prihodnosti bo takih sistemov vedno več; ustrezni postopki učenja pa bodo najverjetneje bližje

splošnim optimizacijskim metodam, ki bolje obvladujejo raznolikost takšne zasnove. Opisu uporabe splošnih optimizacijskih metod za načrtovanje oken je namenjeno podpoglavje 4.3.

Pri načrtovanju nesimetričnih oken je potrebno najprej določiti lastnosti želene rešitve in ustrezne optimizacijske kriterije. Odločitev ni preprosta, saj se je zaradi neraziskanosti tega področja potrebno opreti na znanja z drugih področij. V podpoglavju 4.4 so podrobneje opisani možni kriteriji za načrtovanje nesimetričnih oken in predstavljene želene lastnosti okna s pričakovanim pozitivnim vplivom na uspešnost SRG.

Proučevanje vpliva na uspešnost SRG zahteva načrtovanje nesimetričnih oken z različnimi lastnostmi in večje število praktičnih preizkusov. V podpoglavju 4.5 je podrobneje opisanih nekaj različnih metod za načrtovanje oken s sprotno analizo lastnosti dobljenih rešitev. Podani so tudi rezultati njihovih praktičnih preizkusov v referenčnem okolju.

4.1 OKNA V FREKVENČNI ANALIZI GOVORNIH SIGNALOV

Kratkočasovna Fourierova transformacija (KČFT) je najbolj razširjen postopek frekvenčne analize na področju ARG. Podrobneje je že opisana v podpoglavju 2.2.1. Njeno bistvo je analiza končnih odsekov govornega signala - okvirjev. Vzorci v vsakem okvirju se najprej pomnožijo z okenskim zaporedjem. Enakovredna operacija v frekvenčnem prostoru je konvolucijski integral med Fourierovima transformoma okna $W(e^{j\omega})$ in neskončno dolgega signala $X'(e^{j\omega})$

$$X(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X'(e^{j\theta})\, W(e^{j(\omega-\theta)})\, d\theta . \tag{4.1}$$

Iz (4.1) se lahko določi idealen frekvenčni odziv okna $W(e^{j\omega})$, pri katerem bi bil dobljeni odziv $X(e^{j\omega})$ enak dejanskemu $X'(e^{j\omega})$. V amplitudnem odzivu okna bi moral biti glavni val neskončno ozek in amplituda stranskih valov neskončno majhna. V praksi se je temu mogoče le bolj ali manj približati.

Na področju splošne frekvenčne analize signalov se običajno izpostavita dva pojava, ki sta neposredno odvisna prav od lastnosti uporabljenega okna in natančneje opredeljena v (4.1):

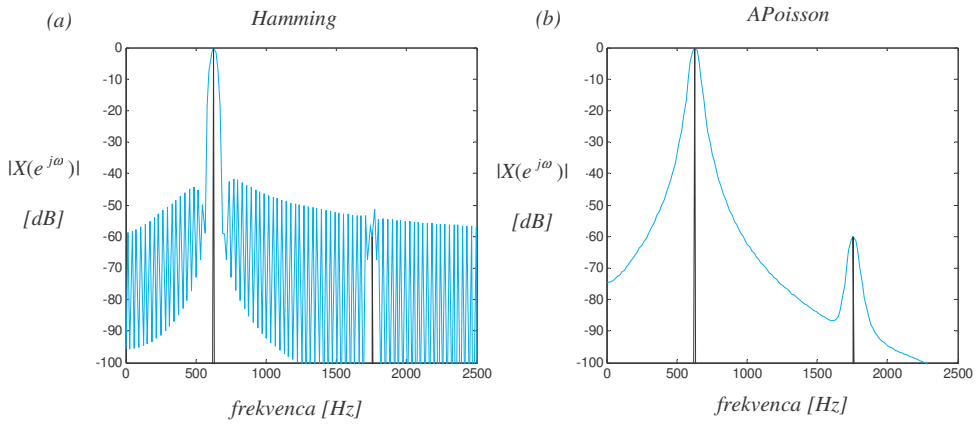
- ***spektralno "razmazanje"***

  Amplitudni odziv okenskega zaporedja ima t.i. glavni val s končno strmino prehoda v področje stranskih valov. Zaradi tega se dobljen amplitudni odziv $X(e^{j\omega})$ "razmaže". S tem se bližnje frekvenčne komponente "zlijejo" in jih ni več mogoče ločiti.

- ***spektralno "puščanje"***

  Amplitudni odziv okenskega zaporedja ima izven glavnega vala le končno dušenje – t.i. stranske valove, ki povzročijo "puščanje" energije posamezne frekvenčne komponente v druge dele amplitudnega spektra.

Slika 4.1 prikazuje vpliv obeh pojavov. Prikazana sta amplitudna odziva signala, sestavljenega iz dveh čistih sinusnih komponent - tonov. S temnejšo črto je označen idealni amplitudni odziv. Slika 4.1-a je nastala z uporabo Hammingovega okna in Slika 4.1-b z modificiranim Poissonovim oknom[1], ki ima sicer zaradi širšega glavnega vala večje razmazanje, a zaradi nižjih stranskih valov precej manjše spektralno puščanje. Na obeh prikazih je vidno razmazanje idealnega amplitudnega odziva vsled vpliva glavnih valov. Na sliki (a) je zaradi spektralnega puščanja oziroma stranskih valov drugi ton "prekrit". Ker ima okno "APoisson" manjše spektralno puščanje, je drugi ton še vedno zaznaven.



Slika 4.1: Primerjava vplivov dveh različnih oken na izračunan amplitudni odziv.

[1] *Označen z "APoisson"; več podrobnosti o tem oknu se nahaja v nadaljevanju (podpoglavje 4.5.3).*

Oba zgoraj opisana pojava sta med seboj povezana. Zato je možnost absolutnih izboljšav v tem primeru omejena, ker ni mogoče poljubno zmanjšati vpliva enega, ne da bi se pri tem povečal vpliv drugega. Tako je za sprejemanje tovrstnih kompromisov potrebno podrobneje spoznati zadani problem in temu ustrezno ovrednotiti relativno pomembnost obeh pojavov.

Z uporabo oken na področju ARG se pojavi še problem določitve njihovega želenega amplitudnega odziva. Njegova optimalnost se namreč v tem primeru vrednoti z vidika robustnosti SRG. Ker medsebojni vpliv glavnih procesov znotraj SRG (parametrizacije in razpoznavanja) še ni dovolj razjasnjen, je določanje želenega amplitudnega odziva težje. Prav tako sta nedorečeni izbiri kriterija oziroma funkcije napake in nenazadnje še ustreznega postopka načrtovanja; pri nesimetričnih oknih je namreč lahko optimizacijski problem precej kompleksnejši. Posebna pozornost je v nadaljevanju posvečena iskanju odgovorov na ta osnovna vprašanja.

4.2 PREDNOSTI NESIMETRIČNIH OKEN

Trenutno v SRG prevladujejo simetrična okenska zaporedja z linearnim faznim odzivom[1], ki je bolj pomemben pri uporabi zaporedja kot enotinega odziva LČI[2] sistema; vse frekvenčne komponente signala se namreč pri prehodu skozi tak sistem enako zakasnijo. V primeru uporabe zaporedja kot okna v frekvenčni analizi SRG pa je linearnost faze manj pomembna. Znano je, da je človekova slušna percepcija dokaj neobčutljiva na nelinearnost faze. Sprostitev zahteve po simetriji okna bi zato morala omogočiti načrtovanje primernejših oken za razpoznavanje govora.

O pomenu in dejanskih prednostih uporabe nesimetričnih oken v SRG je znanega zelo malo. Precej dejstev z nekaterih drugih področij pa opravičuje podrobnejše raziskave in morebitno uporabo omenjenih oken tudi v SRG. Med njimi sta najmočnejša predvsem dva argumenta:

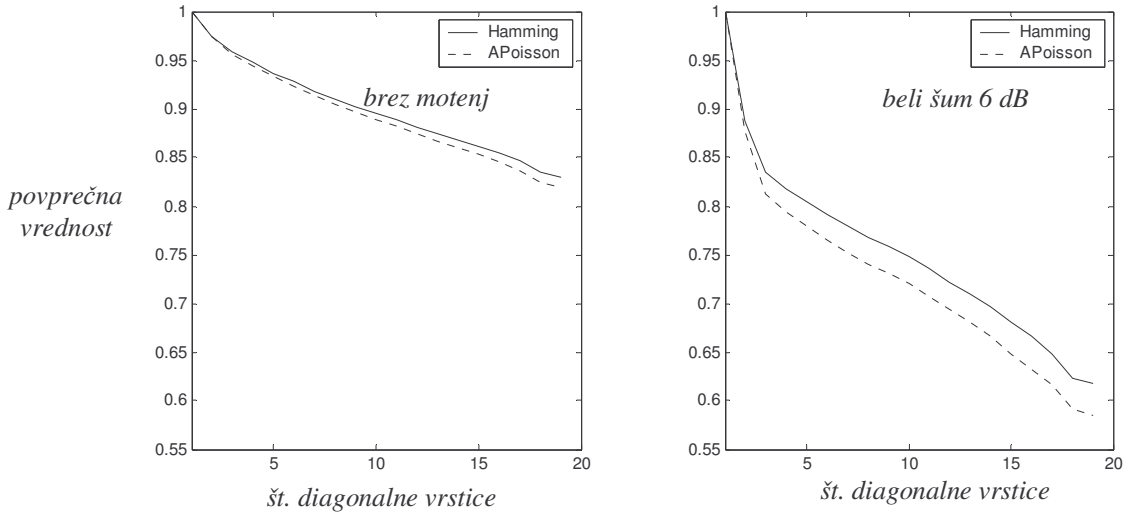
- *s področja načrtovanja KEO filtrov je znano, da se s sprostitvijo zahteve po linearnosti faze lahko dosežejo boljše lastnosti filtra*
- *na področju kodiranja govornih signalov so nesimetrična okna že dokaj razširjena, predvsem zaradi krajših časovnih zakasnitev pri analizi govornih signalov*

---

[1] *Npr. Hammingovo, Blackmanovo, Kaiserjeva okna.*

[2] *Kratica označuje linearen, časovno invarianten sistem.*

Prvo dejstvo je podlaga za načrtovanje nesimetričnih oken, ki imajo nekatere pomembne lastnosti boljše od simetričnih. Tako se npr. lahko dosežejo nižji stranski valovi in s tem zmanjša spektralno puščanje. To je prednost pri detekciji signalov (Slika 4.1), lahko pa tudi pomembno zmanjša koreliranost informacij v kritičnih frekvenčnih pasovih. Ta namreč lahko negativno vpliva na uspešnost HMM razpoznavalnikov, ki uporabljajo le diagonalni del kovariančne matrike (podpoglavje 3.2.1). Slika 4.2 prikazuje rezultate analize koreliranosti FBANK značilk pri skupini 70 govorcev iz govorne zbirke NUMBERS. Prikazana so povprečja vseh diagonalnih vrstic[1] korelacijske matrike FBANK značilk. Povprečna korelacija je manjša ob uporabi okna z manjšim spektralnim puščanjem ("APoisson"). Razlika se v primeru dodanih motenj še poveča (Slika 4.2-b).



Slika 4.2: Vpliv okna na koreliranost FBANK značilk.

Drugi zgoraj naveden argument kaže na dodatno prednost nesimetričnih oken – možnost krajših časovnih zakasnitev pri analizi signalov. Zaradi te lastnosti se nesimetrična okna na področju kodiranja govornih signalov vse intenzivneje uporabljajo. Zanimivo je, da se časovna zakasnitev skrajša že z uporabo preproste modificirane kombinacije dveh simetričnih oken. Na tak način dobljeno okno (t.i. hibridno Hamming-kosinusno okno[2]) je sestavni del standarda ITU[3] za postopek kompresije govornega signala s krajšo časovno zakasnitvijo[4] [31].

---

[1] *1 pomeni glavno diagonalo - elementi (i,i), 2 drugo diagonalo (i+1,i), …. 19.*

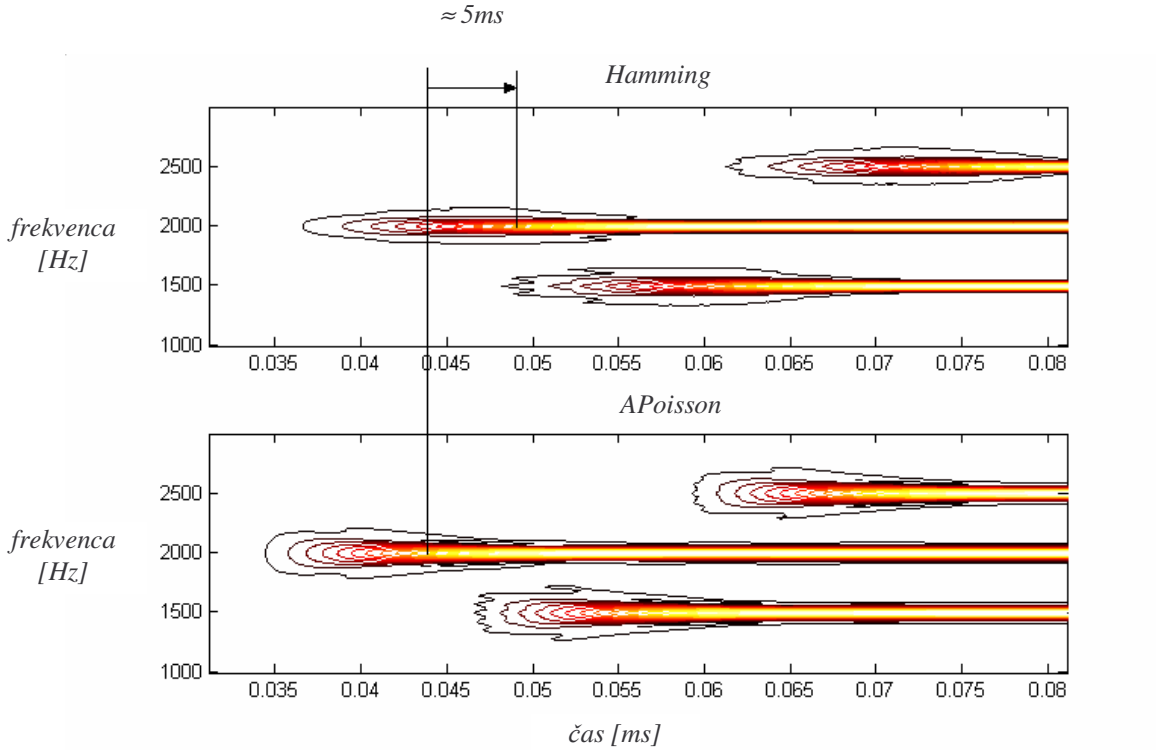
[2] *Podrobnejša obravnava se nahaja v poglavju 4.5.3.*

[3] *Mednarodna koordinacijska organizacija za telekomunikacije - angl. "International Telecommunication Union".*

[4] *Bolj znan pod oznako G.729.*

Pomembno skrajšanje za nekaj ms se v tem primeru doseže brez zaznavnega vpliva na kvaliteto reproduciranega govornega signala.

Lastnost časovne zakasnitve je na področju kodiranja govornih signalov postala bistvenega pomena z razmahom njihovega prenosa po digitalnih omrežjih. S tem se je namreč poenostavila neposredna govorna komunikacija, za katero so večje časovne zakasnitve (značilne za prenos po digitalnih paketnih omrežjih) moteče. Za ilustracijo sta na spodnji sliki prikazana dva spektrograma signala, ki je sestavljen iz treh čistih tonov. Gornji je izračunan z uporabo Hammingovega okna in spodnji z oknom "APoisson". Da bi razlike v časovni zakasnitvi prišle bolj do izraza, je bil uporabljen drseč izračun obeh spektrogramov. Časovna zakasnitev pri uporabi simetričnega okna, v tem primeru Hammingovega, je posebej označena. Razvidno je, da simetrično okno v splošnem povzroča kasnejšo detekcijo posameznih komponent v signalu. Zakasnitev nekaj ms se ne zdi velika, vendar je pri neposredni govorni komunikaciji ključnega pomena. Če je zakasnitev prevelika, se namreč lahko pojavi neželen akustični odmev.



Slika 4.3: Vpliv okna na časovno zakasnitev v kratkočasovni frekvenčni analizi signalov.

Podoben razvoj kot pri kodiranju signalov se pričakuje tudi na področju ARG. Pravzaprav so se prvi večji premiki že zgodili – aktualen je trend konvergence obstoječih govornih tehnologij v enotno zasnovane sisteme. Vzporedno s tem se pojavlja tudi koncept porazdeljenega razpoznavanja in prenosa govornih signalov, ki temelji na njihovi enotni predstavitvi[1] [54].

V omenjenih konvergenčnih zasnovah imajo zelo pomembno vlogo tudi SRG. Zato se vse pogosteje v zvezi z njimi pojavlja tudi zahteva po krajših časovnih zakasnitvah v njihovem delovanju; še posebej to velja za časovno kritične aplikacije - npr. avtomatsko simultano prevajanje. Potencialnim prednostim nesimetričnih oken v amplitudnem odzivu se vse bolj enakovredno pridružuje tudi možnost krajših časovnih zakasnitev pri analizi govornih signalov.

4.3 UPORABA SPLOŠNIH OPTIMIZACIJSKIH METOD

Poglobljena analiza množice obstoječih optimizacijskih metod, ki bi se lahko uporabile za načrtovanje oken v SRG, presega namen tega dela. Poleg tega je raznolikost kriterijev, želenih rešitev in drugih parametrov tako velika, da je uporaba vsaki situaciji prilagojenih metod neučinkovita in zamudna. Prav tako je znano, da so preliminarne raziskave na nekem področju generator pogostih sprememb v kriterijih in drugih parametrih optimizacijskih postopkov. Zato je bila v dani situaciji uporaba splošnih optimizacijskih metod logična izbira.

Poleg splošnih lastnosti, kot so hitrost, stabilnost in robustnost, se v primeru načrtovanja oken od izbrane metode zahteva še enostavno dinamično določanje želene rešitve, funkcije napake in poljubnih omejitev.

Pri izbiri metode je imel odločilno vlogo praktični preizkus v manjšem obsegu [77]. Na klasičnem problemu načrtovanja KEO filtrov je bilo uporabljenih nekaj implementacij znanih optimizacijskih metod. Na osnovi doseženih rešitev, hitrosti in stabilnosti iskanja rešitev ter ugotovljene enostavnosti manipulacije z omejitvami in funkcijo napake je bila izbrana najustreznejša metoda oziroma njena implementacija - SOLVOPT[2]. Slednja se je ves čas po doseženih rezultatih in po subjektivnem vtisu izkazovala za dovolj zanesljivo in učinkovito.

---

[1] *Večja tovrstna združenja so npr. 3GPP, ETSI.*

[2] *"http://www.uni-graz.at/imawww/kuntsevich/solvopt/".*

Ocena globalnosti doseženih rešitev je običajno dokaj zahtevna. Zagotovila, da so bile optimalne točke zares dosežene, namreč ni. Pri samem načrtovanju oken pa je bilo izkazanih kar nekaj dejstev, ki so v prid ugodni oceni primernosti izbrane metode:

- *funkcija napake ima v rešitvah po Čebiševem "minimaks" kriteriju alternirajoče, po absolutni vrednosti enako velike ekstremalne vrednosti*
- *nesimetrične rešitve so konsistentno boljše od simetričnih*
- *metoda je bila uspešno preizkušena tudi na nekaterih referenčnih problemih z znanimi rešitvami*

Najpomembnejše značilnosti izbrane metode SOLVOPT so:

- *temelji na Shorovi minimizacijski metodi s prostorskim širjenjem*[1] *[125]*
- *učinkovita je tudi v primeru minimizacije negladkih funkcij*
- *vsebuje posebne strategije za določanje velikosti začetnega koraka in ustavljanje postopka minimizacije*
- *dobro deluje v večini splošnih optimizacijskih problemov*

Metoda je bila izbrana na podlagi primerjave uspešnosti posameznih implementacij na določenem problemu in subjektivnega vtisa, zato splošnejša ocena uspešnosti teh metod ni mogoča. Podrobnejši opis izbrane metode SOLVOPT se nahaja v [124, 125], njena uporaba pri načrtovanju oken po različnih kriterijih pa je podrobneje predstavljena v dodatku A.

4.4 ŽELENE LASTNOSTI IN KRITERIJI NAČRTOVANJA NESIMETRIČNIH OKEN

Posebnosti področja ARG in specifično vrednotenje optimalnosti oken z vidika robustnosti razpoznavanja pomembno vplivajo na želene lastnosti okna in kriterije za njegovo načrtovanje.

V nadaljevanju sta najprej opisana dva najbolj znana kriterija za načrtovanje digitalnih filtrov oziroma oken. Nato so obravnavane želene lastnosti okna z vidika razpoznavanja govora. Sledi še opis nekaterih alternativnih kriterijev, ki so za dosego želenih lastnosti bolj primerni.

---

[1] *Angl. "space dilation".*

4.4.1 Splošna kriterija načrtovanja končnih zaporedij

Frekvenčni odziv okna $\boldsymbol{h}$= [h(0), h(1), ..., h(N-1)] je določen z izrazom:

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n)\, e^{-j\omega n}. \tag{4.2}$$

Njegovo prileganje želenemu odzivu $D(e^{j\omega})$ se običajno vrednoti po enem od dveh najbolj znanih tovrstnih kriterijev. Prvi je t.i. Čebišev "minimaks" kriterij, pri katerem se za napako $\delta(\boldsymbol{h})$ upošteva maksimalna razlika med obema odzivoma po celotnem področju aproksimacije Ω

$$\delta(\boldsymbol{h}) = \max_{\omega\in\Omega} W(e^{j\omega}) \left| D(e^{j\omega}) - H(e^{j\omega}) \right|, \tag{4.3}$$

kjer je $W(e^{j\omega})$ pozitivna utežnostna funkcija. Izraz (4.3) določa vrednost t.i. Čebiševe napake kompleksnega odziva. Zelo znan pa je tudi kriterij srednje kvadratne napake[1]

$$\delta(\boldsymbol{h}) = \frac{1}{2\pi} \int_{\omega\in\Omega} W(e^{j\omega}) \left| D(e^{j\omega}) - H(e^{j\omega}) \right|^2 d\omega. \tag{4.4}$$

Med optimalnimi rešitvami po obeh kriterijih je nekaj pomembnih razlik. Največja je v porazdelitvi napake glede na celotno področje obravnave Ω. Pri uporabi Čebiševega kriterija je napaka v obliki enakomernega valovanja (angl. "equiripple") porazdeljena po celotnem območju. Pri drugem kriteriju pa se doseže minimalna vrednost celotne srednje kvadratne napake, vendar je ta razporejena neenakomerno po optimizacijskem območju. Praviloma je njena vrednost večja v bližini nezveznih prehodov oziroma mejnih točk pasov v želenem frekvenčnem odzivu.

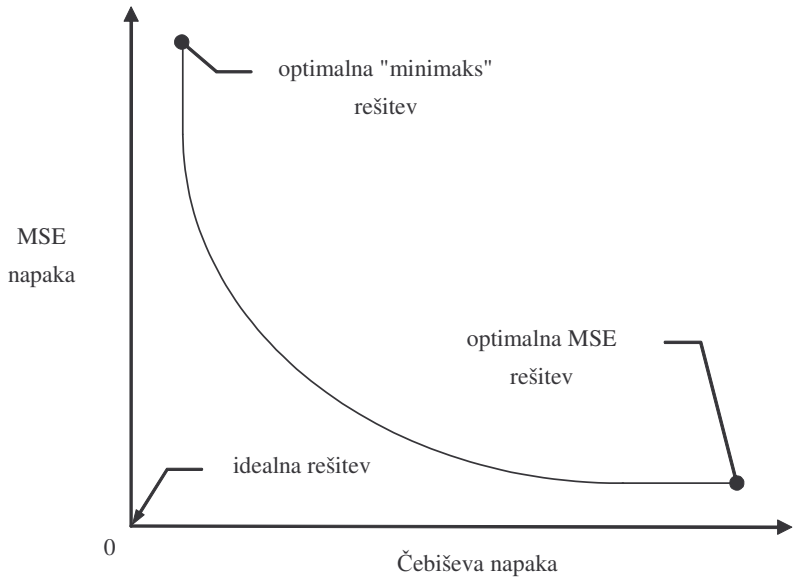
Pri medsebojnih primerjavah kriterijev je večna dilema, kateri je boljši. Običajno se tovrstne primerjave izvedejo na podlagi tretjega kriterija, ki pa v splošnem ni objektivni pokazatelj kakovosti prvih dveh. Zato se v konkretnih primerih običajno izbere kriterij in s tem povezane lastnosti rešitve, ki bolj ustrezajo danim zahtevam. Tako je pri standardnih digitalnih filtrih Čebišev kriterij edina možna izbira, medtem ko je MSE kriterij bolj pogost pri načrtovanju nestandardnih filtrov na nekaterih specifičnih področjih.

---

[1] *Angl. "Mean Square Error" ali krajše MSE.*

Opisane zakonitosti s področja digitalnih KEO filtrov veljajo tudi za okna. Do večjih razlik pride le pri obravnavi spektralnega puščanja. Če so stranski valovi enakomerni (Čebišev kriterij), je enakomerno tudi spektralno puščanje po celem frekvenčnem spektru. Kljub temu je pogosto zaželeno, da omenjeno puščanje pada s spektralno razdaljo. V tem primeru bi okno moralo imeti (ob dopustnem povišanju v neposredni bližini glavnega vala) precej nižje vrednosti amplitudnega odziva v oddaljenih delih spektra. Ta zahteva je nedvomno v tesnejši povezavi z MSE kriterijem.

Med okni je nekaj takih, ki so izraziteje blizu enemu ali drugemu kriteriju. Znano je t.i. Dolph-Čebiševo okno, ki je optimalno po Čebiševem kriteriju. Na drugi strani obstajajo tudi okna, ki so bližje MSE kriteriju – npr. pravokotno okno. Ob tem se pojavi vprašanje, zakaj se matematično sicer optimalne rešitve po obeh kriterijih v praksi le redko uporabljajo. Poleg zahtevnejšega načrtovanja se odgovor skriva tudi v dejstvu, da optimalne rešitve po obeh kriterijih predstavljata skrajni točki krivulje razmerja obeh napak, ki jo prikazuje Slika 4.4. Razvidno je, da sta obe optimalni rešitvi za široko uporabo preveč skrajni. Pri obeh je namreč mogoče z majhnim povečanjem ene napake bistveno zmanjšati vrednost druge [92, 93]. Zato se v praksi bolj uporabljajo okna z enakomernejšimi razmerji vrednosti obeh napak, ki so bližje idealni rešitvi.



Slika 4.4: Medsebojna odvisnost Čebiševe in MSE napake.

4.4.2 Želene lastnosti okna

Eno temeljnih vprašanj pri načrtovanju nesimetričnih oken v SRG so lastnosti želenega amplitudnega odziva okna. Enoznačnega odgovora ni mogoče povsem natančno opredeliti. Na osnovi izkušenj in znanj s področij načrtovanja digitalnih filtrov, frekvenčne analize in razpoznavanja govora se kljub temu lahko izpostavi nekaj ugotovitev, ki so pomembne za oblikovanje želenega amplitudnega odziva:

- *odsekoma konstanten amplitudni odziv (zelo pogost pri načrtovanju filtrov) ni nujno najbolj primeren tudi za načrtovanje oken*
- *parametrizacija v SRG združuje informacije v sosednjih frekvenčnih točkah, zato je bližnje spektralno puščanje manj pomembno*
- *ločljivost frekvenčne analize je v SRG manj pomembna [57]*
- *delovanje SRG pogosto temelji na implicitnih predpostavkah, na katere lahko imajo lastnosti okna pomemben vpliv (npr. spektralno puščanje zmanjšuje neodvisnost informacij v frekvenčnih pasovih oziroma značilkah, kar lahko negativno vpliva na uspešnost HMM razpoznavalnikov)*
- *v govornih signalih so pogoste frekvenčno omejene motnje (npr. brnenje, zvoki, piski) - z večjim spektralnim puščanjem se "prenesejo" tudi v druge dele spektra in zmanjšujejo nivo za razpoznavo koristnih informacij*
- *filtri za frekvenčno analizo v človekovi govorni percepciji imajo širše glavne valove ter monotono padajoče amplitudne odzive*
- *monotoni amplitudni odzivi so zaželeni v nekaterih posebnih postopkih procesiranja zvočnih signalov [66]*

Idealni amplitudni odziv okna je v praksi nedosegljiv. Želeni predstavlja njegovo dosegljivejšo različico, ki se ji postopek načrtovanja želi približati. Če to poteka v skladu z zgoraj opisanimi ugotovitvami, je možnost ugodnega vpliva na uspešnost SRG večja. V tem trenutku je to morda edini možen in učinkovit pristop k reševanju tega problema. Glede na povedano so želene lastnosti amplitudnega odziva naslednje:

- *širši glavni val in nižji stranski valovi*
- *padajoča višina stranskih valov*
- *monoton potek*

4.4.3 Kriteriji za načrtovanje oken

Po določitvi osnovnih značilnosti želene rešitve je potrebna še razprava o možnih kriterijih za njeno načrtovanje. Pri podrobnejšem pregledu literature s področja načrtovanja oken je mogoče zaslediti več možnih alternativnih kriterijev. Glede na časovno kompleksnost iskanja optimalnih rešitev osnovnega problema je smotrno preizkusiti še enostavnejše načine načrtovanja oken s pomočjo parametričnih modelov, ki so zelo blizu značilnostim človekovega sluha. V teh primerih so namreč značilnosti amplitudnega odziva rešitve, določene že z izbiro samega modela in ustreznih parametrov. Zato so za tovrstna načrtovanja zanimivi tudi kriteriji, ki ne vsebujejo eksplicitne definicije želenega amplitudnega odziva; posebej zanimivi so trije:

- *"MSR" - razmerje energij glavnega in stranskih valov [93]*

$$MSR = \frac{\int_{0}^{\omega_s} \left|H(e^{j\omega})\right|^2 \, d\omega}{\int_{\omega_s}^{\pi} \left|H(e^{j\omega})\right|^2 \, d\omega} \tag{4.5}$$

- *"TDI" - indeks časovne zakasnitve [85,86]*

$$TDI = -I^{-1} \int_{-\infty}^{0} t \, h_a^2(t) \, dt \,, \tag{4.6}$$

kjer je $h_a(t)$ zvezna okenska funkcija in I najkrajši interval $[a..b]$, ki vsebuje 95% celotne energije okna

$$I = \min[b-a], \quad da \ \ velja \quad \int_{a}^{b} h_a^2(t) \, dt = 0.95 \int_{-\infty}^{0} h_a^2(t) \, dt \,. \tag{4.7}$$

- *"D" – usmerjenost – angl. "directivity" [86]*

$$D = \frac{\left|H(e^{j0})\right|^2}{\int_{0}^{\pi} \left|H(e^{j\omega})\right|^2 \, d\omega} \tag{4.8}$$

Kriterij MSR je zaradi pozornosti, namenjene nesimetričnim oknom, manj zanimiv, ker vodi v simetrične rešitve[1]. Druga dva pa sta bolj povezana s potencialnimi prednostmi, ki jih nudijo nesimetrična okna. "TDI" izraža mero za časovno zakasnitev[2] pri uporabi okna v KČFT, medtem ko "D" izraža razmerje med energijo najvišje točke glavnega vala in ostalega spektra v celoti. Oba kriterija bosta uporabljena pri načrtovanju nesimetričnih oken v nadaljevanju.

4.5 NAČRTOVANJE NESIMETRIČNIH OKEN

Nekatere naše dosedanje raziskave so pokazale [78-82], da je vpliv okna na robustnost SRG precej večji, kot bi sprva bilo mogoče pričakovati glede na majhno pozornost, posvečeno temu področju.

Najpomembnejšo vlogo pri določanju vpliva okenske funkcije ima njen amplitudni odziv. Ta je najpomembnejši tudi pri KEO filtrih, zato se lahko tovrstne metode uporabijo tudi za načrtovanje oken. Podrobnejši opis se nahaja v prvem podpoglavju (4.5.1).

Psihoakustične in biološke raziskave potrjujejo tezo, da je frekvenčna analiza v človekovem sluhu sestavljena iz množice enostavnejših filtrov, ki jih je mogoče posnemati z ustreznimi modeli. Z njihovo uporabo pri načrtovanju oken se lahko število prostih parametrov bistveno zmanjša, vendar na račun omejevanja oblike amplitudnega odziva. Omenjeni modeli so že v prejšnjih raziskavah pokazali pozitiven vpliv na robustnost SRG [80-82] – enaka pričakovanja pa je mogoče izraziti tudi za rezultate, predstavljene v tem delu. Uporaba NEO parametričnih modelov je podrobneje obravnavana v podpoglavju 4.5.2.

Nesimetrična okna se lahko načrtujejo tudi s pomočjo obstoječih simetričnih. Njihove lastnosti so dobro proučene, vendar običajno nespremenljive. S pomočjo različnih postopkov njihovega kombiniranja se lahko oblikujejo nesimetrična okna z združenimi lastnostmi. V podpoglavju 4.5.3 sta podrobneje predstavljena dva tipična primera tovrstnih postopkov.

4.5.1 Metode za načrtovanje KEO filtrov

Družina KEO filtrov se glede na linearnost faznega odziva deli v dve skupini: v prvi so simetrični KEO filtri z linearno in v drugi nesimetrični KEO filtri z nelinearno fazo. Enaka delitev velja tudi za okna.

---

[1] *Optimalna simetrična rešitev po tem kriteriju so Kaiserjeva okna.*

[2] *Večja vrednost TDI pomeni krajšo časovno zakasnitev.*

Za KEO filtre z linearno fazo oziroma simetrična okna obstaja za načrtovanje učinkovit Parks-McClellanov postopek [117], ki rešuje problem realne aproksimacije po Čebiševem kriteriju. Pri nesimetričnih oknih omejitve linearnosti faze ni. Zato je postopek načrtovanja bolj zahteven – rešuje se namreč nelinearni aproksimacijski problem, pri katerem se želi resnični frekvenčni odziv $H(e^{j\omega})$ kar najbolj približati želenemu $D(e^{j\omega})$

$$min\ (\ D(e^{j\omega}) - H(e^{j\omega})\ ). \tag{4.9}$$

Za oceno aproksimacijske napake se lahko uporabijo različni kriteriji. V nadaljevanju sta opisani dve različici najbolj razširjenega kriterija Čebiševe napake. Prva je t.i. Čebiševa napaka frekvenčnega odziva (podpoglavje 4.5.1.1). Druga, t.i. Čebiševa napaka amplitudnega odziva (podpoglavje 4.5.1.2), je ohlapnejša različica prve, saj upošteva kot funkcijo napake absolutno vrednost razlike obeh amplitudnih odzivov. Razlike v faznih odzivih pri tem zanemari. V zadnjem podpoglavju (4.5.1.3) se nahajajo še analiza dobljenih rešitev in rezultati njihovega praktičnega preizkusa.

4.5.1.1 Kriterij Čebiševe napake frekvenčnega odziva

Kriterij delno sprošča zahtevo po linearnosti faznega odziva oziroma simetriji koeficientov okna; pri napaki namreč upošteva linearnost faznega odziva le v glavnem valu oziroma prepustnem pasu. V zapornem pasu linearnost faze ni tako pomembna, ker ima tam želen amplitudni odziv običajno vrednost nič. Čebiševa napaka je definirana kot absolutna vrednost razlike med obema kompleksnima frekvenčnima odzivoma – želenim in dobljenim.

4.5.1.1.1 Definicija problema

Poiskati želimo optimalni enotin odziv filtra dolžine N, $\boldsymbol{h}^* = [h^*(0), h^*(1), ..., h^*(N\text{-}1)]$, da bo

$$\delta(\boldsymbol{h}^*) = \min_{\boldsymbol{h}}\ \delta(\boldsymbol{h}), \tag{4.10}$$

kjer velja

$$\delta(\boldsymbol{h}) = \max_{\omega\in\Omega}\ W(e^{j\omega})\ |\ E(e^{j\omega})\ |\ , \tag{4.11}$$

$$E(e^{j\omega}) = D(e^{j\omega}) - H(e^{j\omega})\ , \tag{4.12}$$

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n)\, e^{-j\omega n}. \tag{4.13}$$

V gornjih izrazih je $\delta(\boldsymbol{h})$ Čebiševa napaka zaporedja $\boldsymbol{h}$, $D(e^{j\omega})$ želeni, $H(e^{j\omega})$ dobljeni frekvenčni odziv, $W(e^{j\omega})$ pozitivna utežnostna funkcija ter Ω množica frekvenc[1], na kateri se napaka upošteva. Ker je funkcija napake kompleksna, je njena absolutna vrednost enaka:

$$\left| E(e^{j\omega}) \right| = \sqrt{(\mathrm{Re}\{E(e^{j\omega})\})^2 + (\mathrm{Im}\{E(e^{j\omega})\})^2}\,. \tag{4.14}$$

Celotna definicija opisuje nelinearni aproksimacijski problem, katerega direktno reševanje je zahtevno. Zaradi ohlapnejših omejitev je prostor možnih rešitev precej večji kot pri simetričnih oknih.

4.5.1.1.2 Optimalna rešitev problema

Za prej definirani problem (4.10)-(4.14) obstaja enolična optimalna rešitev. Opisuje jo t.i. teorem Rivlina in Shapira [118], ki pravi:

> $H(e^{j\omega})$ je najboljša Čebiševa aproksimacija želenega kompleksnega frekvenčnega odziva $D(e^{j\omega})$ z zveznimi kompleksnimi fukcijami ϕ_i natanko takrat, ko imamo r $(r \le N+1)$ ekstremalnih točk ω_k z enako maksimalno absolutno napako in r ustreznih pozitivnih uteži v vektorju $\boldsymbol{x}$, da velja :
>
> $$\sum_{k=1}^{r} x_k = 1 \tag{4.15}$$
>
> $$\sum_{k=1}^{r} x_k \left[D(e^{j\omega_k}) - H(e^{j\omega_k}) \right] \quad \phi_i(e^{j\omega_k}) = 0 \quad i = 1, 2, ..., \mathrm{N}. \tag{4.16}$$

Teorem določa pogoje, katerim mora zadoščati optimalna rešitev, ne opisuje pa algoritma za njen izračun.

Pri reševanju zastavljenega problema se lahko uporabi več optimizacijskih metod; v nadaljevanju sta v ločenih podpoglavjih podrobneje predstavljeni dve. Prva je znana metoda

---

[1] *Ω je unija kompaktnih neprekrivajočih podintervalov intervala [0 .. π].*

linearnega programiranja, medtem ko je druga ustrezna posplošitev metode za načrtovanje po Čebiševem kriteriju optimalnih KEO filtrov z linearno fazo.

4.5.1.1.3 Načrtovanje s pomočjo linearnega programiranja

Metoda temelji na linearizaciji sicer nelinearnega kompleksnega problema in reševanju s postopkom linearnega programiranja [108, 115]. Ker eksplicitne zahteve po simetriji ni več, fazni odzivi dobljenih oken niso več linearni. Zato pride do odstopanj, ki so v mejah dosežene vrednosti Čebiševe napake.

V izrazih (4.10)-(4.14) je podan nelinearni aproksimacijski problem, katerega reševanje je zahtevno. Lahko pa se problem linearizira ob upoštevanju enakosti, ki velja za vsako kompleksno število:

$$|z| = \max_{-\pi \le \theta < \pi} (\mathrm{Re}\{z\, e^{j\theta}\}). \tag{4.17}$$

Če se gornji izraz vstavi v (4.10)-(4.14), nastane za reševanje primernejša definicija problema:

$$\min_{\boldsymbol{h}} \delta(\boldsymbol{h}) = \min_{\boldsymbol{h}} \max_{\omega \in \Omega} \max_{-\pi \le \theta < \pi} W(e^{j\omega}) \mathrm{Re}\{E(e^{j\omega}) e^{j\theta}\}. \tag{4.18}$$

Problem se lahko sedaj zapiše v obliki primarnega linearnega programa:

Določi $\boldsymbol{h} = [h(0), h(1), ..., h(N\text{-}1), \delta]$, *da bo*

$$\min_{\boldsymbol{h}} (\boldsymbol{h}\,\boldsymbol{b} = \delta), \tag{4.19}$$

pri čemer je $\boldsymbol{b}=[0, 0, \ldots, 0, 1]^{\mathrm{T}}$ *ob omejitvah*

$$Re\{ E(e^{j\omega}) e^{j\theta} \} \le \frac{\delta}{W(e^{j\omega})} \qquad \forall \omega \in \Omega \quad in \quad \forall \theta \in [-\pi .. \pi). \tag{4.20}$$

Po tej definiciji je število spremenljivk končno in množica parametrov (ω,θ) neštevna. Problem je rešljiv z diskretizacijo intervala $[0..\pi]$ in množice Ω, vendar je v tem primeru lažje reševati ustrezni dualni problem v naslednji obliki:

Določi N+1 različnih frekvenc ω_1, ω_2, …, ω_{N+1}, ustreznih kotov θ_1, θ_2, …, θ_{N+1}, ter N+1 pozitivnih uteži x=(x_1, x_2, …, x_{N+1}), da bo

$$\max_{\boldsymbol{x}} \left(\sum_{k=1}^{N+1} x_k \underbrace{\mathrm{Re}\{D(e^{j\omega_k})\, e^{j\theta_k}\}}_{c_k} \right) = \max_{\boldsymbol{x}} (\boldsymbol{c}\ \boldsymbol{x}) \tag{4.21}$$

ob omejitvah

$$\begin{aligned} &\sum_{k=1}^{N+1} x_k \cos(\theta_k - n\omega_k) = 0, \qquad n = 0,1,...,N-1 \\ &\sum_{k=1}^{N+1} \frac{x_k}{W(e^{j\omega_k})} = 1 \quad in \quad x_j \geq 0 \quad za \quad 1 \leq j \leq N+1. \end{aligned} \tag{4.22}$$

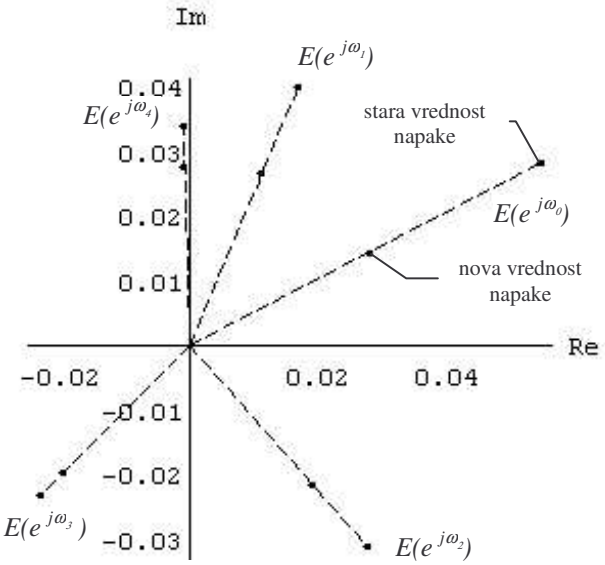
Problem neskončnega števila točk (ω,θ) iz primarnega linearnega programa se v dualnem omeji na obravnavo N+1 "ekstremalnih" točk, katerih uteži so neničelne – imenujejo se tudi bazne spremenljivke. Po predpostavki so uteži vseh ostalih (nebaznih) stolpcev enake 0.

Omenjena redukcija števila omejitev in obvladovanje problema neskončnega števila točk sta glavna razloga, da se dualni program rešuje bolj učinkovito od primarnega. Postopek reševanja je podrobneje opisan v [108] in splošni literaturi.

4.5.1.1.4 Načrtovanje s pomočjo posplošenega Remezovega algoritma

Osnovna ideja te metode je uporaba znanega Remezovega algoritma na splošnejšem problemu aproksimacije poljubnega frekvenčnega odziva. Konvergenca teoretično sicer ni zagotovljena, vendar v praksi postopek večinoma pride do povsem primerljivih rezultatov glede na metodo iz prejšnjega podpoglavja (4.5.1.1.3). Posplošitev Remezovega algoritma je s teoretičnega stališča zanimiva predvsem zaradi ohranitve njegove osnovne ideje v prostoru rešitev z manj omejitvami kot pri linearni fazi. V nadaljevanju so predstavljene le najpomembnejše značilnosti postopka; podrobnejši opis se nahaja v [111].

Postopek načrtovanja poteka v iteracijskih korakih. Absolutne vrednosti napake v ekstremalnih točkah ω_i se postopoma zmanjšujejo in približujejo optimalni rešitvi – najnižji enaki absolutni vrednosti v vseh točkah. Slika 4.5 prikazuje osnovni iterativni korak zmanjšanja napake v ekstremalnih točkah.



Slika 4.5: Zmanjševanje absolutne vrednosti napak.

V vsaki iteraciji se vrednost napake zmanjša s pomočjo faktorja $\rho\,(0<\rho<=1)$, ki s svoje začetne vrednosti narašča s številom iteracij. Zmanjšanje absolutne vrednosti napake v posamezni ekstremalni točki

$$\delta(i) = |\, E(e^{j\omega_i}) \,| \tag{4.23}$$

znaša

$$\delta(i)_{novi} = \delta(i) - \rho(\delta(i) - \delta_{\min}), \quad i = 0, ..., N, \tag{4.24}$$

kjer je δ_{min} minimalna napaka vseh ekstremalnih točk. Faktor ρ določa hitrost in predvsem zanesljivost konvergence celotnega algoritma. Njegove vrednosti se zato med postopkom spreminjajo po vnaprej določenih empiričnih pravilih.

4.5.1.2 Kriterij Čebiševe napake amplitudnega odziva

Kriterij je ohlapnejši od Čebiševe napake frekvenčnega odziva (podpoglavje 4.5.1.1), saj kot napako upošteva le absolutno vrednost razlike v obeh amplitudnih odzivih (4.25). Zaradi manjšega pomena faznega odziva v SRG je kriterij še bolj zanimiv. Funkcija napake je določena z izrazom

$$E'(e^{j\omega}) = | D(e^{j\omega}) | - | H(e^{j\omega}) | , \tag{4.25}$$

kjer je $|D(e^{j\omega})|$ želeni in $|H(e^{j\omega})|$ dobljeni amplitudni odziv. Izraz (4.25) se pomembno razlikuje od definicije napake prejšnjega kriterija v (4.12). Ob upoštevanju (4.25) in pozitivne utežnostne funkcije $W(e^{j\omega})$ se določi Čebiševa napaka amplitudnega odziva

$$\delta'(\boldsymbol{h}) = \max_{\omega \in \Omega} \; W(e^{j\omega}) \; | E'(e^{j\omega}) |, \tag{4.26}$$

ki se v postopku optimizacije skuša čim bolj zmanjšati.

Iskanje optimalne rešitve je bolj zahtevno kot v prejšnjem podpoglavju (4.5.1.1). Izbira optimalnega postopka za reševanje zadanega problema presega namen te obravnave, zato je bila za tovrstno načrtovanje izbrana splošnejša optimizacijska metoda SOLVOPT. Podrobnosti dejanske izvedbe načrtovanja so v dodatku A.

4.5.1.3 Primerjava rezultatov obeh metod

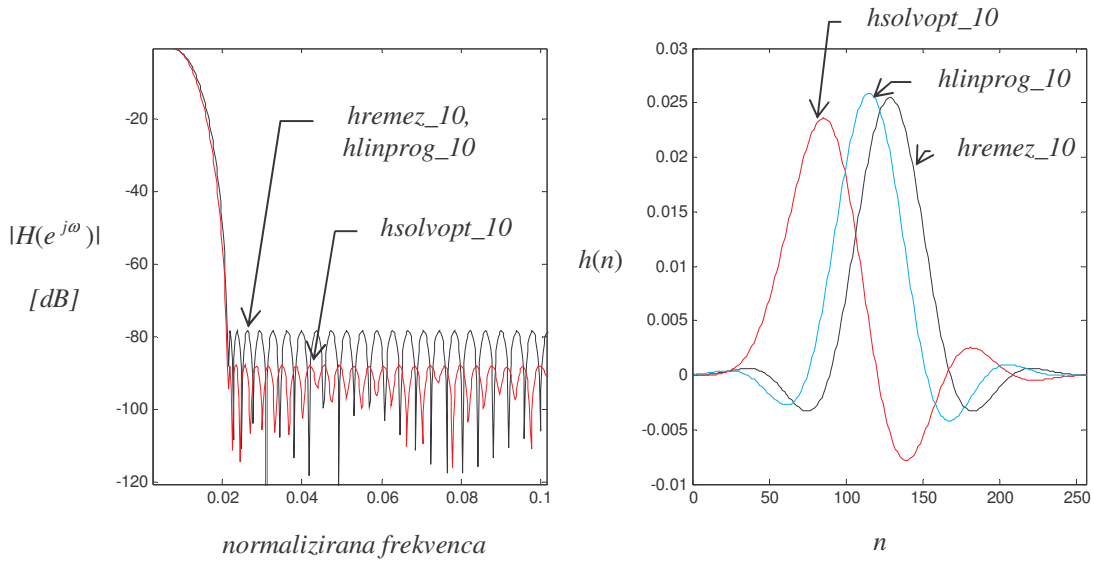
Za primerjavo opisanih metod so bila načrtovana simetrična in nesimetrična okna z enakim želenim amplitudnim odzivom dolžine N=256[1]. Ta je bil odsekoma konstanten z začetkom zapornega pasu v točki 3-kratne širine glavnega vala enako dolgega Hammingovega okna. Naši prejšnji preizkusi so namreč prikazali večjo robustnost SRG pri uporabi oken s to širino glavnega vala [77].

V nadaljevanju (Slika 4.6, Slika 4.7) je prikazana primerjava lastnosti po obeh kriterijih načrtovanih oken z relativnim razmerjem uteži napak v glavnem in stranskih valovih 1:10. Za lažjo primerjavo so okna normalizirana[2]. Prikazani so amplitudni odziv, časovni potek in grupna zakasnitev v glavnem valu. Postopno sproščanje zahtev po linearnosti faznega odziva postopoma zmanjšuje višino stranskih valov in manjša Čebiševo napako. Okno z linearno fazo, načrtovano s Parks-McClellanovim postopkom, ima oznako "hremez_10". Z "hlinprog_10" je označeno okno, dobljeno s pomočjo linearnega programiranja po kriteriju Čebiševe napake frekvenčnega odziva (obe predstavljeni metodi sta imeli zelo podobne rezultate), in s "hsolvopt_10" rešitev, dobljena po kriteriju Čebiševe napake amplitudnega odziva. "Hlinprog_10" je v amplitudnem odzivu le 1-2 dB boljše od okna "hremez_10", zato zaradi preglednosti v amplitudnem odzivu ni prikazano.

---

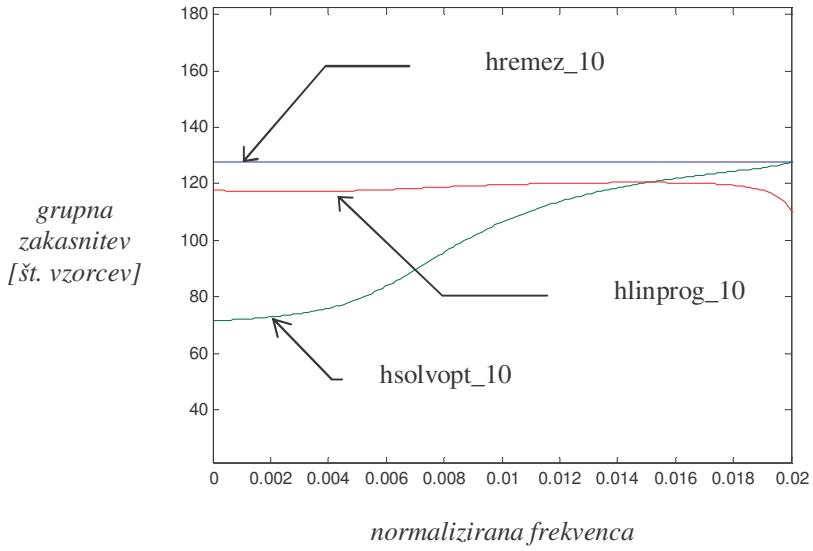
[1] *Pri frekvenci vzorčenja Fs=8000Hz to ustreza oknu dolžine 32 ms.*

[2] *Vsa okna v disertaciji so normalizirana na enotno ojačanje 1 v glavnem valu amplitudnega odziva.*



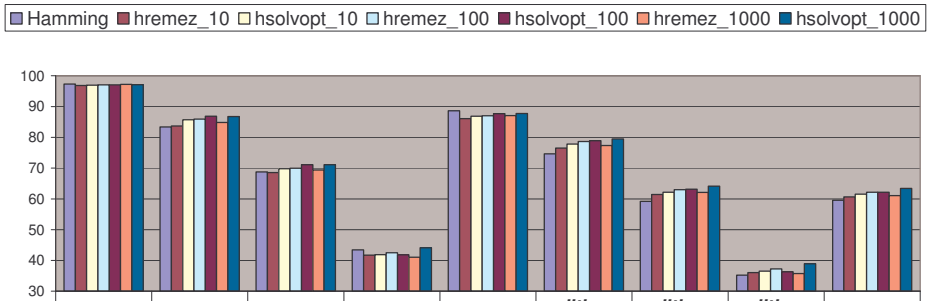
Slika 4.6: Prikaz amplitudnih odzivov in časovnih potekov oken, načrtovanih po kriteriju kompleksne Čebiševe napake.

Zanimiva je tudi primerjava med okni v grupni zakasnitvi (Slika 4.7). Razvidno je, da ima simetrično okno konstantno grupno zakasnitev. Grupna zakasnitev okna "hlinprog_10" se giblje ekvidistančno okrog zadane vrednosti. Pri oknu "hsolvopt_10" pa so bile odpravljene kakršne koli omejitve v faznem odzivu oziroma grupni zakasnitvi; ustrezna nagrada za to so še nižji stranski valovi v amplitudnem odzivu (Slika 4.6).



Slika 4.7: Grupne zakasnitve oken, načrtovanih po kriteriju kompleksne Čebiševe napake.

Okna "hsolvopt" in "hremez" so bila načrtovana za tri možne vrednosti uteži funkcije napake v zapornem pasu: 10, 100, 1000[1]. S tem je napaka v stranskih valovih toliko bolj poudarjena v primerjavi s tisto v glavnem valu (utež 1). Metoda SOLVOPT je dala v vseh primerih konsistentne rezultate. Višina stranskih valov oken "hsolvopt" je bila vedno nižja od ustreznih "hremez" oken; prav tako je večja utež napake vedno pomenila tudi nižje stranske valove – seveda na račun večje napake v glavnem valu, za katero se predvideva, da je v tem primeru manj pomembna. Rezultate praktičnega preizkusa na sistemu CSLU-NUMBERS-MFCC[2] prikazuje Slika 4.8. Razvidne so sicer majhne, a dokaj konsistentne razlike . Zanimivo pa je, da so vsa okna v končnem povprečju boljša od Hammingovega. Po uspešnosti posameznih parov oken v zaporednih vrsticah je torej mogoče sklepati, da že 8 dB razlike pri višini stranskih valov lahko pozitivno vpliva na robustnost SRG. To je najbolj zaznavno pri vrednosti uteži 1000 oziroma oknih "hremez_1000" in "hsolvopt_1000"; pri tem je posebej zanimivo tudi to, da je vpliv konsistenten v vseh navedenih skupinah testnih množic, razen pri osnovni ("čist").



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming	97,36	83,38	68,79	43,43	88,67	74,64	59,20	35,22	59,59
hremez_10	96,88	83,69	68,59	41,74	86,08	76,48	61,44	36,07	60,64
hsolvopt_10	96,98	85,74	69,80	41,89	86,82	77,82	62,18	36,47	61,52
hremez_100	97,06	85,94	69,94	42,50	87,00	78,66	62,99	37,23	62,20
hsolvopt_100	97,07	86,82	71,11	41,81	87,68	78,91	63,15	36,31	62,18
hremez_1000	97,21	84,81	69,37	41,05	87,04	77,36	62,12	35,77	61,08
hsolvopt_1000	97,12	86,75	71,14	44,11	87,75	79,48	64,15	38,90	63,39

Slika 4.8: Uspešnost CSLU-NUMBERS-MFCC SRG ob uporabi oken, načrtovanih s KEO metodami.

[1] *Ustrezne oznake so npr. "remez_1000" ali "solvopt_10", kjer beseda pomeni metodo načrtovanja, številka pa relativno utež napake izven glavnega vala.*

[2] *Praktični preizkusi so označeni z zaporedjem oznak "ime_sistema-govorna zbirka-tip_značilk".*

4.5.2 Družina "NEO" oken

Glede na kompleksnost metod za načrtovanje oken iz podpoglavja 4.5.1 se zdi smiselno obravnavati tudi enostavnejše metode načrtovanja nesimetričnih oken.

Ideja o oknih, ki jih je mogoče načrtovati s pomočjo preprostih NEO sistemov, izvira z več področij:

- *kodiranje govornih signalov – postopek rekurzivnega izračuna LPC značilk [102]*
- *spoznanja o človekovem sluhu [68] – enostavna oblika filtrov v frekvenčni analizi*
- *NEO sistemi nizkega reda – enostavnost, monotonost amplitudnih odzivov*

Predstavniki družine "NEO oken" so končni odseki odzivov na enotin impulz preprostih NEO sistemov s prevajalno funkcijo

$$H(z) = G\frac{1}{\left(1-\alpha z^{-1}\right)^{M}} \qquad \alpha \in (0..1),\ M \ \textit{pozitivno, celo število,} \tag{4.27}$$

in G normalizacijska konstanta[1]. Prevajalna funkcija v (4.27) ima *M* polov v isti točki ($z=\alpha$) na realni osi *Z* ravnine. S spreminjanjem vrednosti parametrov se lahko tvori več različnih oken, zato je bilo za praktični preizkus izbranih le nekaj najbolj značilnih – predvsem s širšim glavnim valom in manjšim spektralnim puščanjem v amplitudnem odzivu. Parametre NEO sistemov za generiranje izbranih oken prikazuje Tabela 4.1.

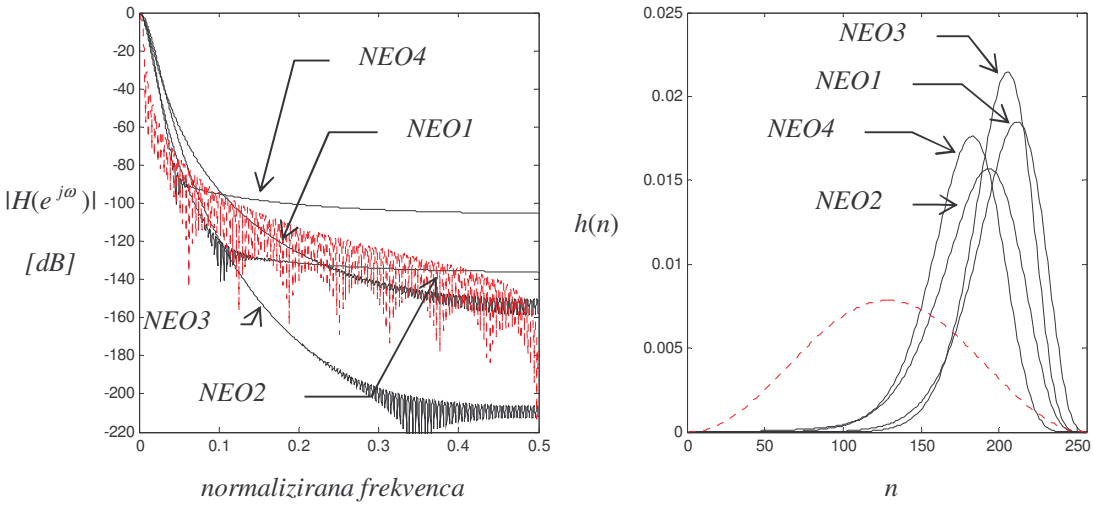
Oznaka	*M*	α	*Opis*
NEO1	6	0.9	podobno NEO3, slabše asimptotično dušenje
NEO2	8	0.9	podobno NEO4, manjše bližnje puščanje, slabše asimptotično dušenje
NEO3	10	0.85	podoben NEO1, boljše asimptotično dušenje
NEO4	14	0.85	podobno NEO2, večje spektralno puščanje

Tabela 4.1: Parametri NEO oken.

Amplitudne odzive in časovne poteke izbranih NEO oken prikazuje Slika 4.9. Okna imajo zaradi krajše zakasnitve prezrcaljene časovne poteke. Za boljšo primerjavo je dodano še

[1] *Vsa okna so bila normalizirana na enotno ojačanje glavnega vala v amplitudnem odzivu.*

Hannovo okno. Razvidno je, da so izbrana NEO okna v časovnem poteku ožja, s širšim glavnim valom v amplitudnem odzivu. Boljše je tudi asimptotično padanje vrednosti amplitudnega odziva, ki je pri oknu NEO3 celo večje kot pri Hannovem. V parih so okna dober testni primer za preizkus pomembnosti lastnosti asimptotičnega dušenja (NEO1, NEO3) in boljšega bližnjega dušenja oziroma asimptotično manj strmo padajočih stranskih valov (NEO2, NEO4).

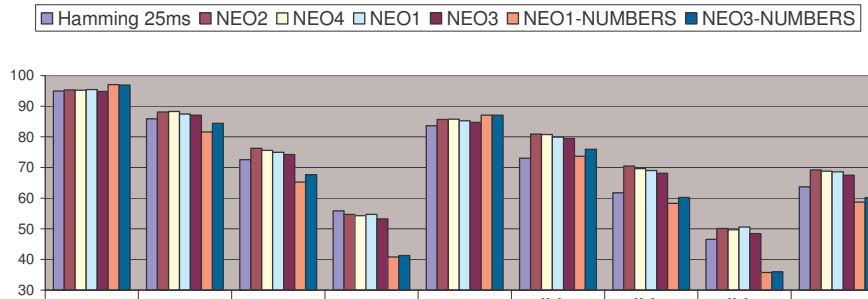


Slika 4.9: Amplitudni odzivi in časovni poteki izbranih NEO oken.

Izbrana NEO okna so bila preizkušena na CSLU-ŠTEVKE-MFCC referenčnem sistemu. Nekoliko presenetljive rezultate prikazuje Slika 4.10. Zaradi manjšega spektralnega puščanja je NEO2 okno v končnem povprečju boljše od NEO4. Podobno bi se moralo zgoditi tudi z NEO3 in NEO1, vendar so rezultati nasprotni.

Očitno lastnosti same po sebi ne zagotavljajo pozitivnega učinka na robustnost SRG. Pri tovrstnih primerjavah je potrebno bolj upoštevati še druge dejavnike. Vsako okno je npr. mogoče obravnavati kot kompromis med širino glavnega vala v časovnem poteku in amplitudnem spektru. NEO okna so v časovnih potekih precej ožja in potemtakem efektivno zajamejo manj podatkov. Znano je, da lahko dolžina okna kar zaznavno vpliva na robustnost SRG. Zato sta v preizkusih poleg 32 ms dolgega Hammingovega sodelovali tudi njegovi

krajši različici (16ms in 25ms); v končnih primerjavah je vedno uporabljeno boljše[1] okno iz te skupine. Tako je vpliv efektivne dolžine okna v primerjavah s Hammingovim zmanjšan.



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	94,97	85,88	72,57	55,85	83,67	73,03	61,73	46,55	63,69
NEO2	95,32	88,15	76,27	54,73	85,65	80,92	70,48	50,09	69,21
NEO4	95,22	88,29	75,59	54,28	85,79	80,75	69,63	49,71	68,80
NEO1	95,41	87,43	74,97	54,71	85,25	79,93	69,01	50,54	68,56
NEO3	94,82	87,10	74,24	53,24	84,70	79,43	68,13	48,44	67,49
NEO1-NUMBERS	97,06	81,64	65,25	40,82	87,12	73,70	58,30	35,74	58,69
NEO3-NUMBERS	96,92	84,44	67,67	41,28	87,12	75,98	60,25	35,98	60,19

Slika 4.10: Uspešnost sistema CSLU-ŠTEVKE(NUMBERS)-MFCC SRG ob uporabi NEO oken.

Iz rezultatov oken NEO1 in NEO3 sledi tudi ugotovitev, da z željo po čim boljšem asimptotičnem dušenju ne gre pretiravati. Relativno hitro so namreč vrednosti amplitudnega odziva in s tem spektralnega puščanja pod slišnim pragom ter na razpoznavanje ne vplivajo več.

Nenazadnje pa je potrebno upoštevati še tezo, ki govori o odvisnosti doseženih rezultatov od dejanskih podatkov oziroma pogojev delovanja. Zato sta bili obe "problematični" NEO okni preizkušeni še na govorni zbirki NUMBERS (Slika 4.10 – zadnji vrstici), kjer so rezultati v skladu s pričakovanji; boljše lastnosti okna NEO3 so se odrazile tudi v večji uspešnosti SRG.

4.5.2.1 NEO eksponentna okna

Pri podrobnejši analizi NEO oken se izkaže, da je mogoče v primeru NEO sistema (4.27) z dvema poloma (M=2) in dodatnim členom αz^{-1} v števcu prevajalne funkcije $H(z)$ odziv na

[1] *Rezultati Hammingovih oken vseh treh dolžin so bili zelo podobni (Slika 5.1). Nekoliko je boljše 25ms dolgo Hammingovo okno, kar sovpada z najbolj razširjeno dolžino okna 25.6 ms, ki je privzeta vrednost v domala vseh SRG.*

enotin impulz eksplicitno določiti. Tako nastane definicija novih okenskih zaporedij[1] s precej enostavnejšim izrazom:

$$h(n) = n\,\alpha^n, \quad n = 0 .. N\text{-}1,\ 0<\alpha<1 . \tag{4.28}$$

Ker se uporabi le končni izsek neskončno dolgega zaporedja $h(n)$, je to ekvivalentno uporabi pravokotnega okna; temu ustrezno se spremeni tudi amplitudni odziv. Ker je želena lastnost hitro asimptotično padanje višine stranskih valov, je bolje uporabiti nepravokotno okno. Izbrano je bilo Hannovo, ki ima strm asimptotični padec višine stranskih valov - 18 dB/dekado. NEO eksponentno okno je določeno z izrazom

$$h'(n) = G\,h(n)\,w_{hann}(n) , \quad n = 0 .. N\text{-}1 , \tag{4.29}$$

kjer je G normalizacijska konstanta in $w_{hann}(n)$ Hannovo okno. S spreminjanjem obeh parametrov (lega pola α in dolžina N) se lahko generirajo različna okna. Da bi bila med seboj primerljiva, je dolžina vseh (tudi ostalih[2]) enaka (N=256). Časovni poteki vseh oken, razen posebej označene izjeme[3], so zaradi krajše časovne zakasnitve prezrcaljeni.

Pri načrtovanju NEO eksponentnih oken z metodo SOLVOPT se pokaže zanimiva lastnost; s spreminjanjem lege polov se napaki po obeh alternativnih kriterijih (TDI in D)[4] spreminjata po značilni krivulji (Slika 4.11), podobni tisti, ki prikazuje relacijo med "minimaks" in MSE napako pri splošnem načrtovanju končnih zaporedij (Slika 4.4). Vendar se alternativna kriterija v tem primeru bolj razlikujeta; eden izraža lastnosti časovnega poteka, drugi pa lastnosti amplitudnega odziva. Kljub temu je tudi v tem primeru mogoče sklepati, da ekstremne rešitve na eni ali drugi osi ne bodo najbolje vplivale na uspešnost SRG. Za praktični preizkus v referenčnem okolju je bilo izbranih 6 eksponentnih oken v značilnih točkah krivulje. Njihove podatke prikazuje Tabela 4.2.

Zanimivo je tudi dejstvo, da sta bili dve eksponentni okni ("Exp3" in "Exp4") uporabljeni že pri naših prejšnjih tovrstnih preizkusih [78-80]. Takrat sta okni bili izbrani na osnovi analize

---

[1] V nadaljevanju imenovana "eksponentna" okna.

[2] Če ima okno drugačno dolžino, je ta posebej navedena v oznaki – npr "Hamming 25ms". Okna, dolžine 32ms (N=256), v oznaki nimajo navedene dolžine.

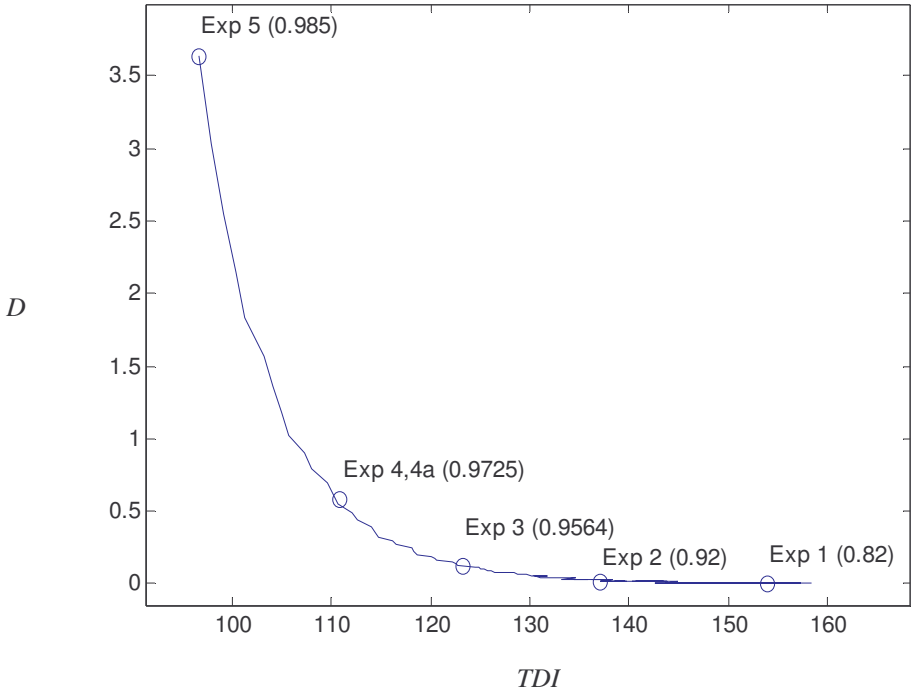
[3] "Exp 4a" - okno je opisano v nadaljevanju.

[4] Kriterija sta opisana v podpoglavju 4.4.3.

njunih časovnih potekov in amplitudnih odzivov, vendar še brez zrcaljenja časovnega poteka zaradi skrajšanja zakasnitve. Zaradi primerljivosti s prejšnjimi raziskavami je bilo dodano še v časovnem poteku neprezrcaljeno okno "Exp 4a". Slika 4.11 potrjuje, da ta izbira ni bila naključje, ampak predstavlja sprejemljiv kompromis. Ob dodatni izbiri še treh značilnih oken je bilo mogoče njihov vpliv na uspešnost SRG še bolje opredeliti.

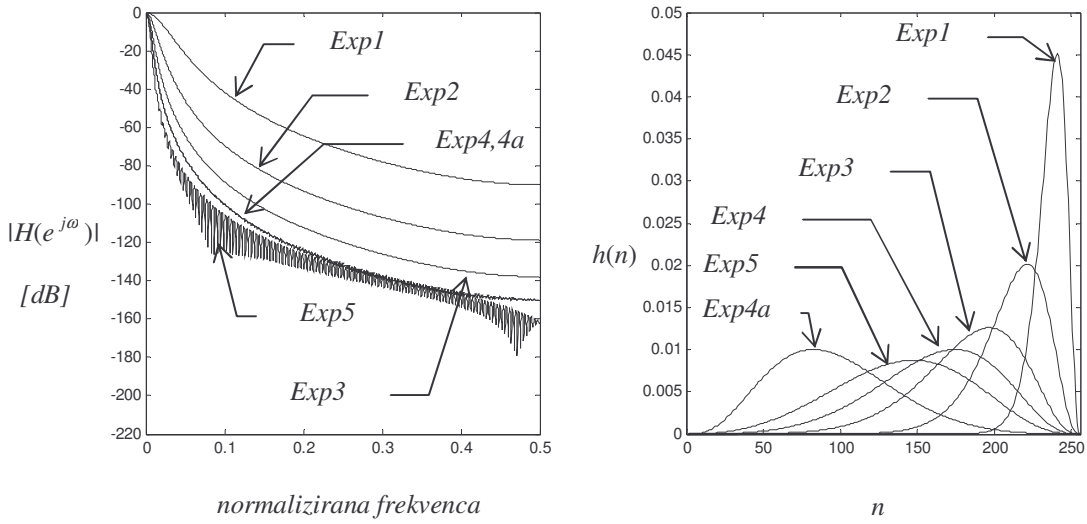
Oznaka	*M*	α	*Opis*
Exp 1	2	0.82	Za naraščanjem parametra α se: - povečuje asimptotično padanje višine stranskih valov - zmanjšuje TDI vrednosti (povečuje časovno zakasnitev) - veča efektivna dolžina okna v časovnem prostoru
Exp 2	2	0.92	
Exp 3	2	0.9564	
Exp 4, Exp 4a	2	0.9725	
Exp 5	2	0.985	

Tabela 4.2: Parametri eksponentnih NEO oken.



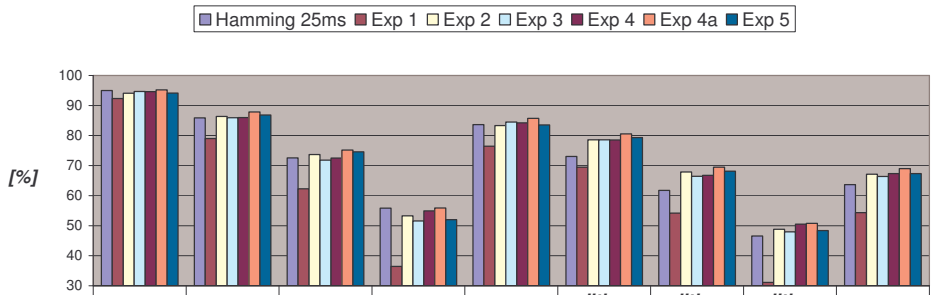
Slika 4.11: Odvisnost parametra α eksponentnih oken od napak po kriterijih TDI in D.

Slika 4.12 prikazuje amplitudne odzive in časovne poteke izbranih eksponentnih oken. Vidna je podobnost z ostalimi NEO okni. Prav tako je asimptotično padanje višine stranskih valov dokaj strmo, kar je glede na množenje s Hannovim oknom pričakovano – množenje v časovnem prostoru namreč pomeni operacijo konvolucije amplitudnih odzivov v frekvenčnem prostoru.



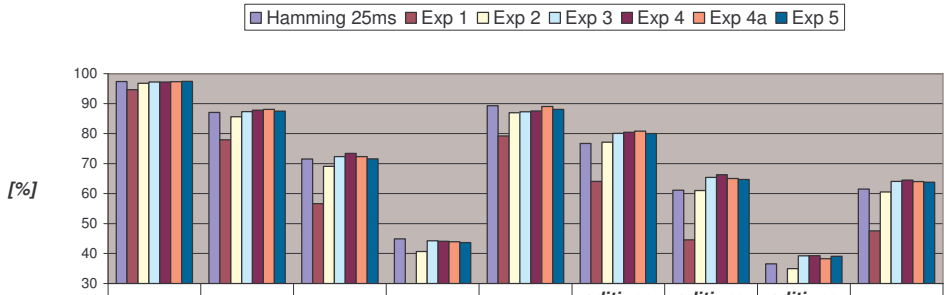
Slika 4.12: Amplitudni odzivi in časovni poteki izbranih eksponentnih oken.

Izbrana eksponentna okna so bila preizkušena na referenčnem CSLU SRG ob uporabi obeh govornih zbirk. Rezultati (Slika 4.13, Slika 4.14) pokažejo zaznavno povečanje robustnosti v primerjavi s Hammingovim oknom. Potrjeno je pričakovanje, da boljše asimptotično dušenje v amplitudnem odzivu okna prinese večjo robustnost SRG. To v celoti velja za govorno zbirko NUMBERS, medtem ko pri ŠTEVKAH naraščanje robustnosti s parametrom α ni povsem konsistentno. Očitno je vpliv okna odvisen še od značilnosti uporabljene govorne zbirke, kar je v prid tezi o odvisnosti vpliva okna od dejanskih pogojev delovanja (govorna zbirka, motnje, …).



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	94,97	85,88	72,57	55,85	83,67	73,03	61,73	46,55	63,69
Exp 1	92,31	78,99	62,27	36,46	76,46	69,48	54,17	31,14	54,33
Exp 2	94,08	86,38	73,67	53,27	83,32	78,59	67,86	48,79	67,16
Exp 3	94,67	85,91	71,81	51,55	84,48	78,60	66,46	47,94	66,38
Exp 4	94,63	86,01	72,53	54,95	84,25	78,54	66,75	50,48	67,34
Exp 4a	95,22	87,83	75,16	55,88	85,70	80,56	69,46	50,77	69,02
Exp 5	94,13	86,84	74,61	52,00	83,56	79,34	68,16	48,37	67,34

Slika 4.13: Uspešnost CSLU-ŠTEVKE-MFCC SRG ob uporabi NEO eksponentnih oken.



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	97,37	87,04	71,56	44,85	89,29	76,73	61,12	36,54	61,47
Exp 1	94,67	77,96	56,65	29,21	79,23	64,04	44,52	22,61	47,56
Exp 2	96,79	85,59	69,06	40,67	86,96	77,15	60,98	34,98	60,53
Exp 3	97,21	87,34	72,35	44,23	87,26	80,07	65,37	39,21	64,03
Exp 4	97,21	87,82	73,41	44,09	87,52	80,47	66,28	39,26	64,48
Exp 4a	97,36	88,08	72,33	43,91	89,02	80,84	65,00	38,26	63,99
Exp 5	97,45	87,51	71,62	43,61	88,09	80,03	64,69	39,05	63,77

Slika 4.14: Uspešnost CSLU-NUMBERS-MFCC SRG ob uporabi NEO eksponentnih oken.

4.5.3 Nesimetrične modifikacije obstoječih oken

Značilnost enostavnih, eksplicitno določenih simetričnih oken[1] so vnaprej znane karakteristične lastnosti, ki jih ni mogoče prilagajati specifičnim potrebam. Zato se v praksi

[1] *Npr. Hammingovo, Hannovo, Blackmannovo, trikotno okno…*

pogosto med seboj kombinirajo na najrazličnejše načine. Dobljene kompromisne rešitve združujejo lastnosti večih oken.

Na področju ARG nesimetrična okna praktično še niso bila preizkušena. Imajo pa zelo pomembno vlogo na področju kodiranja govornih signalov. Med njimi je trenutno najbolj znano hibridno "Hamming-kosinusno" okno[1], ki ga organizacija ITU priporoča za uporabo v sistemih za kodiranje govora z nizko časovno zakasnitvijo [31]. ITU okno je klasični primer nesimetričnega zlitja dveh znanih simetričnih oken; sestavljata ga daljša prva polovica Hammingovega in krajša druga polovica kosinusnega okna

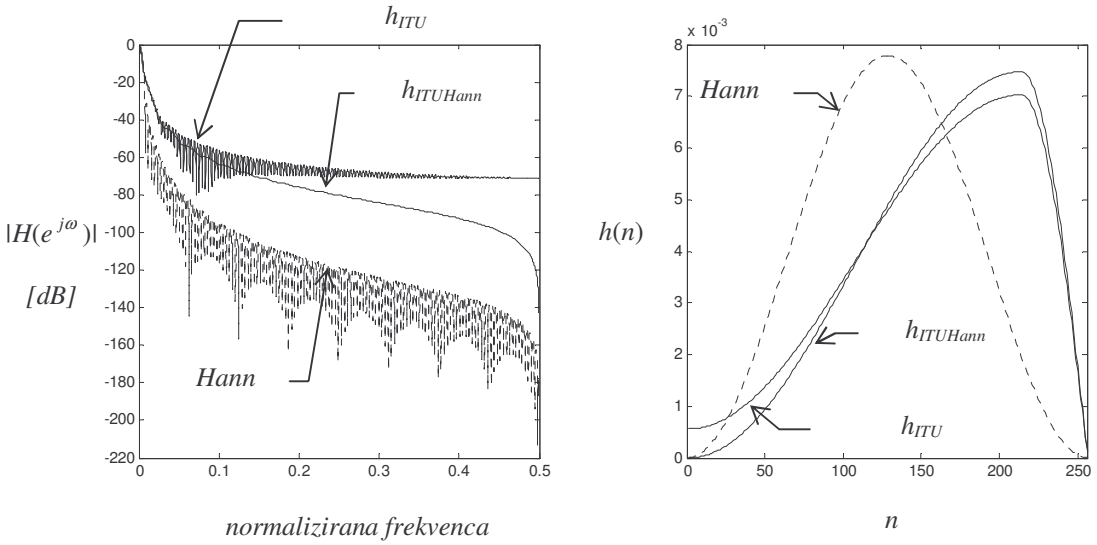
$$h_{ITU}(n) = G \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{415}\right), & n = 0\,..\,207, \\ \cos\left(\dfrac{2\pi(n-208)}{191}\right), & n = 208\,..\,255. \end{cases} \tag{4.30}$$

Iz sicer skopih informacij je razvidno, da je bilo okno izbrano predvsem na podlagi časovnega poteka oziroma s tem povezanega doseganja krajše časovne zakasnitve pri kodiranju govornega signala. ITU okno namreč v okvirju bolj poudari "novejše" vzorce govornega signala, medtem ko simetrično okno (v tem primeru "Hann") poudari "starejše" vzorce okoli sredine okvirja (Slika 4.15).

Iz amplitudnega odziva ITU okna (Slika 4.15) je razvidno, da nekatere njegove lastnosti (predvsem nižje asimptotično dušenje) niso povsem v skladu z že navedenimi spoznanji o potrebnih lastnostih okna za robustno razpoznavanje govora (podpoglavje 4.4.2). Pri tem je potrebno upoštevati, da je primarni namen uporabe ITU okna postopek kodiranja, ki temelji na LPC analizi; ta velja za manj občutljivo na spremembe v amplitudnem odzivu. Kljub temu je mogoče v preizkus na obeh področjih predlagati dve alternativni možnosti. Prva je t.i. Hann-kosinusno okno (oznaka "$h_{ITUHann}$"), ki se dobi z zamenjavo Hammingovega okna s Hannovim v definiciji (4.30)

$$h_{ITUHann}(n) = G \begin{cases} 0.5 - 0.5\cos\left(\dfrac{2\pi n}{415}\right) & n = 0\,..\,207, \\ \cos\left(\dfrac{2\pi(n-208)}{191}\right) & n = 208\,..\,255. \end{cases} \tag{4.31}$$

[1] *V nadaljevanju "ITU okno".*



Slika 4.15: Amplitudni odzivi in časovni poteki kombiniranih kosinusnih oken.

Hannovo okno bolj ustreza deklariranim lastnostim želenega amplitudnega odziva in predstavlja resno alternativo najširše uporabljanemu Hammingovemu oknu. Drugo predlagano okno, t.i. modificirano Poissonovo okno, pa je z množenjem pridobljena kombinacija časovno premaknjenega Poissonovega in simetričnega Hannovega okna:

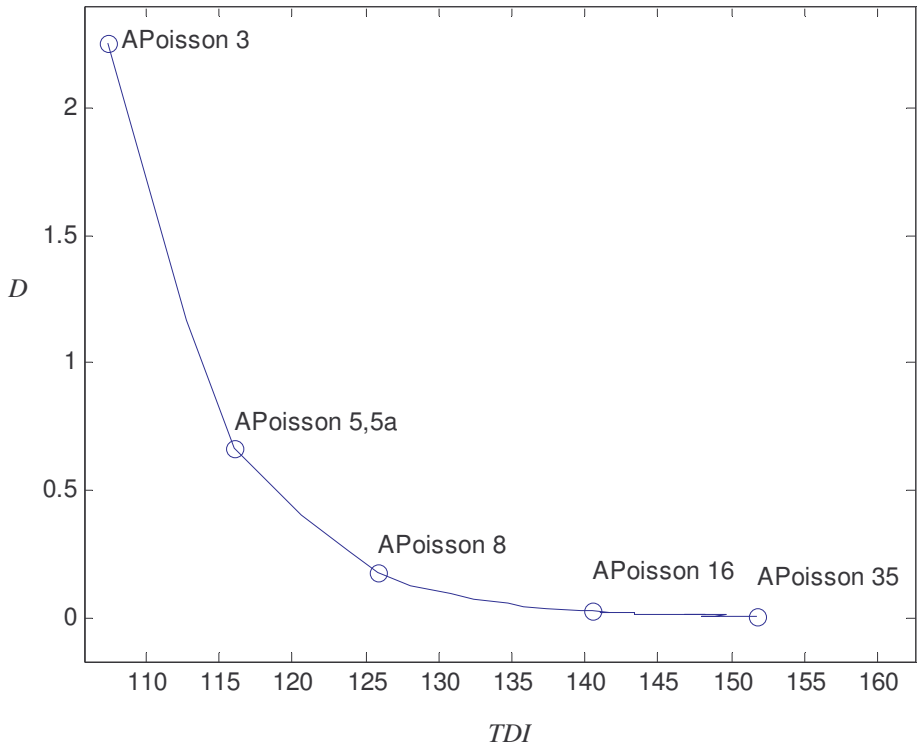
$$h_{AP}(n) = G\ e^{\left(-\alpha\frac{n}{N}\right)}\ \left(0.5 - 0.5\cos\left(\frac{2\pi n}{N}\right)\right), \quad n = 0\,..\,N-1, \tag{4.32}$$

kjer je G normalizacijska konstanta. Simetrično Poissonovo okno je določeno z naslednjim izrazom:

$$h_P(n) = e^{\left(-\alpha\frac{|n|}{N/2}\right)}, \qquad 0 \le |n| \le N/2. \tag{4.33}$$

Modificirano Poissonovo okno združuje dobre lastnosti dveh oken:

- *krajša časovna zakasnitev – časovno premaknjeno Poissonovo okno*
- *dobro asimptotično dušenje – Hannovo okno*



Slika 4.16: Prikaz modificiranih Poissonovih oken v odvisnosti od napak po kriterijih TDI in D.

Opisano okno je v nadaljevanju imenovano kot modificirano nesimetrično Poissonovo okno ali krajše "APoisson" oziroma "AP". Njegova osnovna značilnost v primerjavi z ITU oknom je ob sicer daljši časovni zakasnitvi, precej boljši amplitudni odziv. Za uporabo v SRG je to potencialna prednost. Za uporabo pri kodiranju govora pa je potrebno učinkovitost obeh predlaganih oken še dodatno raziskati. Področji se namreč pomembno razlikujeta, zato splošne ocene niso mogoče.

S spreminjanjem parametra α se lahko generira večje število sorodnih oken (dolžina oken je zaradi primerljivosti enaka kot pri ostalih N=256). Časovni poteki vseh izbranih oken, razen posebej označene izjeme ("*APoisson 5a*"), so zaradi krajše časovne zakasnitve prezrcaljeni. Slika 4.16 prikazuje odvisnost vrednosti parametra α modificiranih Poissonovih oken od napak po obeh alternativnih kriterijih (TDI in D). Dobljena krivulja je zelo podobna tistima na slikah 4.11 in 4.4. Za praktični preizkus je bilo izbranih 6 oken v značilnih točkah omenjene krivulje, katerih časovni poteki so bili zaradi krajše zakasnitve prezrcaljeni.

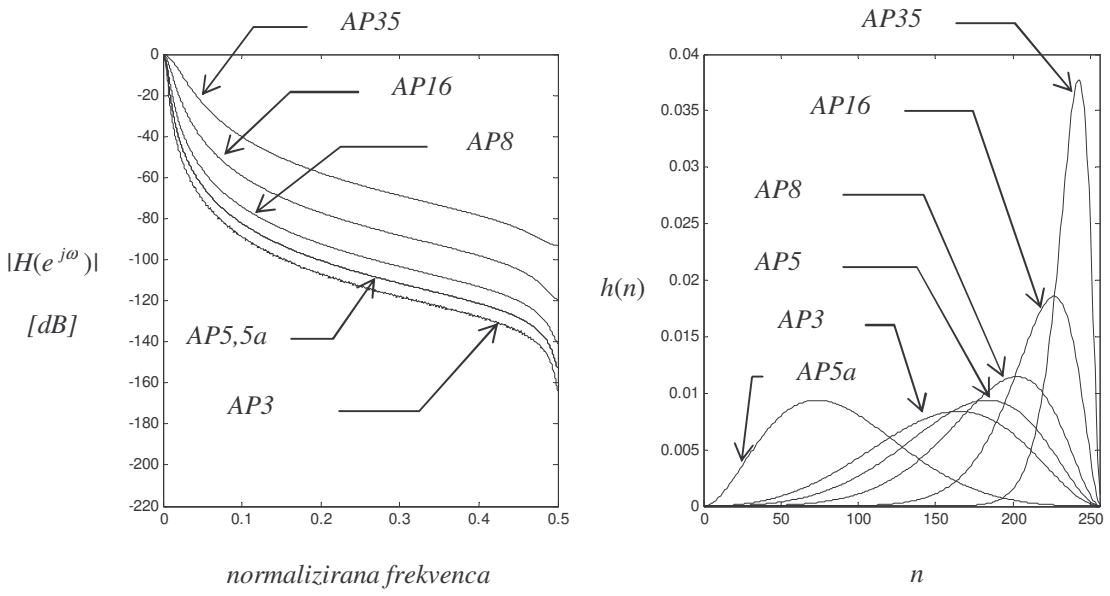
Zanimivo je dejstvo, da je bilo okno z vrednostjo parametra α=5 prisotno že pri naših prejšnjih praktičnih preizkusih [78-80]. Takrat je izbor temeljil na analizi časovnega poteka in

amplitudnega odziva; zaradi primerljivosti je bilo prvotno okno med izbrane dodano kot neprezrcaljeno z oznako – "*APoisson 5a*". Podrobnejši podatki o izbranih oknih se nahajajo v Tabeli 4.3.

Oznaka	α	*Opis*
APoisson3	3	Za naraščanjem parametra α se: - zmanjšuje asimptotično padanje vrednosti amplitudnega odziva - povečuje TDI (krajša časovna zakasnitev) - manjša efektivna dolžina okna
APoisson5, APoisson5a	5	
APoisson8	8	
APoisson16	16	
APoisson35	35	

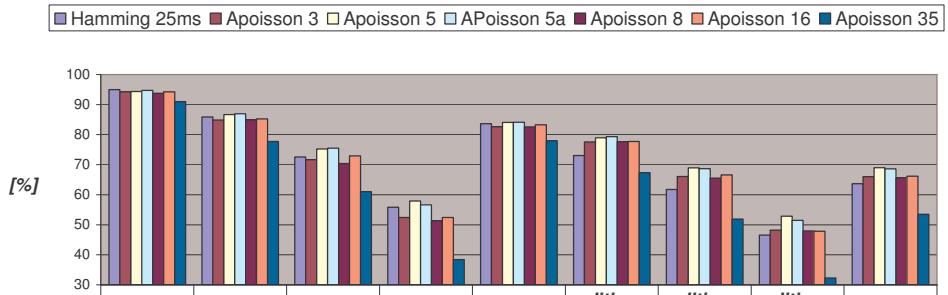
Tabela 4.3: Parametri modificiranih Poissonovih oken.

Slika 4.17 prikazuje amplitudne odzive in časovne poteke izbranih modificiranih Poissonovih oken. Opazen je gladek potek amplitudnih odzivov. Zaradi množenja s Hannovim imajo okna pričakovano asimptotično padanje vrednosti amplitudnega odziva. V ostalih značilnostih so podobna ostalim NEO oknom.



Slika 4.17: Amplitudni odzivi in časovni poteki modificiranih Poissonovih oken.

Rezultati izbranih oken na CSLU-MFCC sistemu in obeh govornih zbirkah (Slika 4.18, Slika 4.19) na prvi pogled niso povsem v skladu s pričakovanji. Pri obeh govornih zbirkah je namreč viden značilni potek uspešnosti v obliki zvona. To pomeni, da je asimptotično padajoča vrednost amplitudnega odziva sicer pomembna lastnost, še zdaleč pa ne tudi edina. Pri tovrstnih oknih (najverjetneje pa tudi pri vseh ostalih) se namreč lahko boljše asimptotično padanje doseže le z večanjem efektivne dolžine okna v časovnem prostoru, kar prav gotovo pomembno vpliva na uspešnost SRG. Zanimivo je tudi dejstvo, da sta največji uspešnosti pri obeh zbirkah doseženi pri vrednosti α=5. Efektivna dolžina tega okna, določena z izrazom (4.7), je namreč enaka efektivni dolžini 25ms dolgega Hammingovega okna, ki se je med svojimi tremi različnimi dolžinami (16ms, 25ms, 32ms) izkazalo za najbolj robustno[1]. Očitno je kompromis pri α=5 optimalen glede na spektralno puščanje in časovno širino okna.

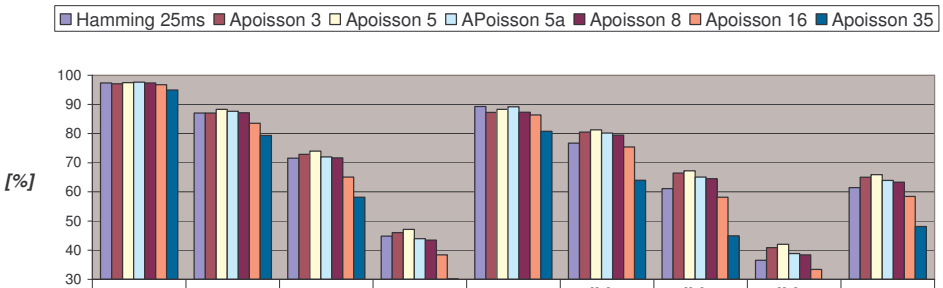


	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	94,97	85,88	72,57	55,85	83,67	73,03	61,73	46,55	63,69
Apoisson 3	94,23	84,85	71,68	52,42	82,66	77,57	66,05	48,23	65,99
Apoisson 5	94,33	86,67	75,21	57,94	84,05	78,98	68,93	52,85	69,01
APoisson 5a	94,72	86,95	75,51	56,61	84,12	79,33	68,68	51,48	68,64
Apoisson 8	93,79	84,93	70,41	51,32	82,56	77,63	65,54	47,98	65,66
Apoisson 16	94,23	85,26	72,92	52,44	83,30	77,71	66,59	47,85	66,20
Apoisson 35	90,98	77,74	61,01	38,42	77,92	67,35	51,89	32,28	53,50

Slika 4.18: Uspešnost CSLU-ŠTEVKE-MFCC SRG ob uporabi modificiranih Poissonovih oken.

V prid tej ugotovitvi je tudi znano dejstvo s področja ARG, da je dolžina uporabljenega okna kompromis med obema bolj skrajnima željama: daljšim oknom za uspešno razpoznavanje stacionarnih glasov (samoglasniki) in krajšim za razpoznavanje krajših, nestacionarnih glasov (npr. zaporniki). Ta dejstva postavljajo pod vprašaj primernost frekvenčne analize govornih signalov z enako dolgim oknom, kar pa je predmet drugih raziskav [57].

[1] *Rezultati so podani v podpoglavju 5.1.*



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	97,37	87,04	71,56	44,85	89,29	76,73	61,12	36,54	61,47
Apoisson 3	97,10	87,08	72,87	46,05	87,29	80,49	66,47	40,87	64,98
Apoisson 5	97,45	88,28	73,95	47,10	88,30	81,28	67,23	42,01	65,91
APoisson 5a	97,64	87,63	71,99	43,93	89,16	80,19	65,06	38,81	63,93
Apoisson 8	97,33	87,18	71,68	43,41	87,30	79,45	64,49	38,39	63,34
Apoisson 16	96,76	83,53	65,04	38,38	86,35	75,43	58,16	33,40	58,42
Apoisson 35	94,90	79,34	58,17	30,13	80,76	63,99	44,94	23,09	48,07

Slika 4.19: Uspešnost CSLU-NUMBERS-MFCC SRG ob uporabi modificiranih Poissonovih oken.

4.6 UGOTOVITVE

Na osnovi analize načrtovanih oken in rezultatov delnih praktičnih preizkusov se ugotovi naslednje:

- *Asimptotično padanje višine stranskih valov in s tem spektralno puščanje pozitivno vpliva na robustnost SRG. Vendar je vpliv povezan še z efektivno dolžino okna v časovnem prostoru. Najuspešnejša okna so zato kompromis med obema lastnostima.*
- *Problem efektivne dolžine okna ima širše ozadje. Za ARG vemo, da je dinamična dolžina okna najboljša rešitev, saj je konstantna dolžina precejšnja poenostavitev. S tega vidika je zelo pomemben koncept NEO modelov, ki omogoča enostavno parametrično spreminjanje lastnosti okna (podpoglavje 4.5.2).*
- *Metode za načrtovanje KEO filtrov za uporabo v SRG niso posebej primerne. Izkazalo se je namreč, da je uporaba parametričnih modelov boljši in enostavnejši pristop.*
- *Med družinama NEO oken in modificiranih Poissonovih oken večjih razlik ni. Zato sta lahko približno enakovredna izbira; še posebej v primeru, ko je pomembna časovna zakasnitev pri delovanju sistema. Zanimivo je, da točki najbolj uspešnih oken iz obeh skupin približno sovpadata na TDI-D krivulji (Slika 4.11 in Slika 4.16).*
- *Kriterij TDI je poleg časovne zakasnitve povezan še s širino okna v časovnem prostoru. Zato odvisnost uspešnosti SRG in vrednosti TDI ni premočrtna. Obstaja*

optimalna točka, ki je najboljši kompromis širin glavnih valov v frekvenčnem in časovnem prostoru. Okna v bližini te točke so uspešnejša. Enaka ugotovitev velja tudi za kriterij usmerjenosti – "D" (Slika 4.11 in Slika 4.16).

- *Medsebojna primerjava obeh kriterijev samo na osnovi točk na karakteristični krivulji je poenostavitev, vendar pri danih parametričnih modelih druge možnosti ni. Za pravilno vrednotenje bi namreč bilo potrebno točke premikati v vzporednih smereh z obema osema; le tako bi se lahko spreminjala vrednost ene napake ob konstantni vrednosti druge.*
- *Opazi se (sicer majhno) razhajanje v uspešnosti parov oken z zrcalnimi časovnimi poteki*[1]*. Glede na to, da imata obe okni v paru enak amplitudni odziv, očitno nastopi še dodaten vpliv krajšanja časovne zakasnitve, ki je glede na rezultate zelo odvisen od uporabljene govorne zbirke. Rezultati namreč pokažejo, da sta razliki nekonsistentni v odvisnosti od uporabljene govorne zbirke (pri NEO eksponentnih oknih je npr. nezrcaljeno okno "Exp 4a" v primeru zbirke ŠTEVKE boljše od zrcaljene različice "Exp 4" – Slika 4.13, v primeru NUMBERS pa slabše – Slika 4.14).*

[1] *Zaradi krajše časovne zakasnitve so časovni poteki večine nesimetričnih oken v podpoglavjih 4.5.2 in 4.5.3 prezrcaljeni. Zaradi naših prejšnjih raziskav, kjer zrcaljenje časovnih potekov ni bilo izvedeno, v primerjavah nastopijo tudi takrat uporabljena okna, ki imajo nezrcaljen časovni potek – njihovi oznaki je na koncu dodana še črka "a".*

5. REZULTATI PREIZKUSOV ROBUSTNOSTI SRG

Pričujoča disertacija ima več osnovnih ciljev. Poleg načrtovanja nesimetričnih oken je pomembna še proučitev njihovega vpliva na uspešnost in robustnost SRG. V tem poglavju bo pozornost namenjena prikazu in analizi rezultatov praktičnih preizkusov v referenčnem okolju ob uporabi izbranih oken.

Osnovno izhodišče praktičnih preizkusov je zagotovitev zadostne raznolikosti referenčnega okolja, ki omogoča splošnejšo oceno vpliva okna na končno uspešnost SRG. S tem se lahko razkrijejo še morebitni vplivi ostalih dejavnikov. Vendar je želeno raznolikost glede na časovno kompleksnost preizkusov težko zagotoviti. Zato so bili slednji izvedeni selektivno glede na pričakovano pomembnost rezultatov. Potrebno je še poudariti, da je pozornost vseskozi bolj usmerjena v relativni značaj medsebojnih primerjav. Podrobnejše "uglaševanje" obeh referenčnih sistemov za dosego absolutno najboljših kvantitativnih rezultatov namreč presega namen disertacije.

Že v 4. poglavju je navedenih nekaj praktičnih rezultatov v povezavi s sprotno analizo lastnosti načrtovanih nesimetričnih oken. Zato bo interpretacija rezultatov v nadaljevanju bolj usmerjena k drugim dejavnikom: npr. vplivu uporabljenih referenčnih sistemov, govornih zbirk in ostalih parametrov.

Že v podpoglavju 1.3 so bili pojasnjeni razlogi za preizkušanje inherentne robustnosti SRG. Podrobnosti simulacije neznanih motenih pogojev se nahajajo v podpoglavju 3.1.3, vsi ostali podatki o izvedenih praktičnih preizkusih pa so v dodatkih B, C in D. Potrebno je še opozoriti, da so bile pri preizkusih uporabljene normalizirane MFCC značilke.

V nadaljevanju so navedeni rezultati[1] in sprotne analize v štirih smiselnih celotah - podpoglavjih. Najprej so v podpoglavju 5.1 prikazani rezultati osnovne primerjave izbranih oken. Temu sledijo zanimivi rezultati t.i. zaporednih preizkusov, ki so dodatno potrdili že ugotovljene prednosti nesimetričnih oken (podpoglavje 5.2). V podpoglavju 5.3 se nahaja primerjava uspešnosti ITU okna in predlaganih alternativnih rešitev. Poglavje se zaključi še s povzetkom končnih ugotovitev.

---

[1] *Uspešnost SRG je v tabelah izražena v odstotku uspešno razpoznanih besed. V nadaljevanju je oznaki okna, vkolikor ni dolgo 32ms, dodan še zapis dolžine okna.*

5.1 PRIMERJAVA IZBRANIH OKEN

Za praktično primerjavo uspešnosti sta bili izbrani dve večji skupini oken. V prvi sta najbolj popularni simetrični Hammingovo in Hannovo, v drugi pa se nahajajo izbrana nesimetrična okna. Njihovo načrtovanje je opisano v podpoglavju 4.5, kjer je njihov vpliv na uspešnost referenčnih SRG tudi že deloma ovrednoten.

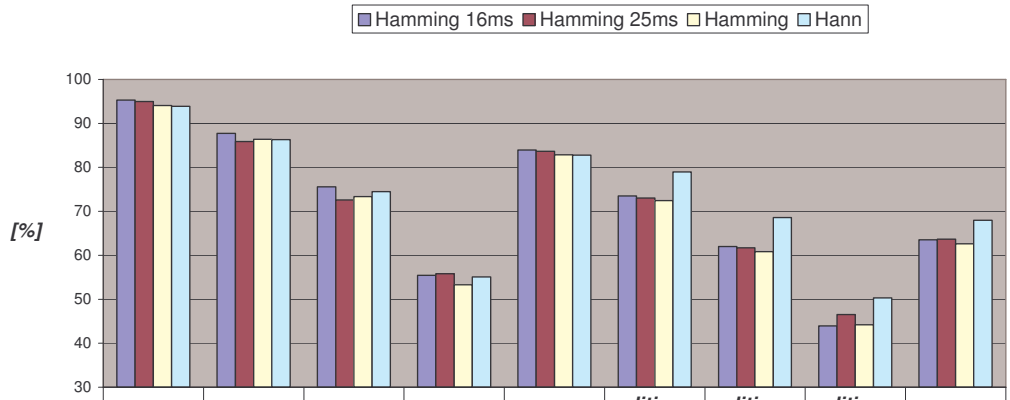
V nadaljevanju so najprej za vsako od kombinacij referenčnih SRG in govornih zbirk navedeni rezultati skupine simetričnih oken. Za razliko od podpoglavja 5.2 so bila uporabljena enaka okna v procesih učenja in razpoznavanja. Hammingovo okno je zaradi že v podpoglavju 4.5.3 omenjenega problema efektivne širine okna v časovnem poteku ovrednoteno v več možnih dolžinah: 16ms, 25ms, 32ms.

Opisanemu prikazu sledi še primerjava izbranih simetričnih oken z dvema najbolj značilnima predstavnikoma nesimetričnih: "APoisson 5a" in "Exp 4a". Obe sta v časovnem prostoru neprezrcaljeni različici ustreznih oken iz podpoglavja 4.5. V primerjavah so navedene povprečne vrednosti po posameznih skupinah testnih množic in na koncu še povprečje po vseh testnih množicah. Podrobnejša specifikacija uporabljenih oznak skupin testnih množic se nahaja v dodatku D. Na koncu vsakega podpoglavja sledi še primerjava uspešnosti na manjšem številu izbranih testnih množic, pri katerih je prišlo do večjih razlik v uspešnosti posameznih oken. Za razliko od podpoglavja 5.2 so tukaj sistemi naučeni in preizkušeni z identičnim procesom parametrizacije.

5.1.1 "CSLU-ŠTEVKE-MFCC" SRG

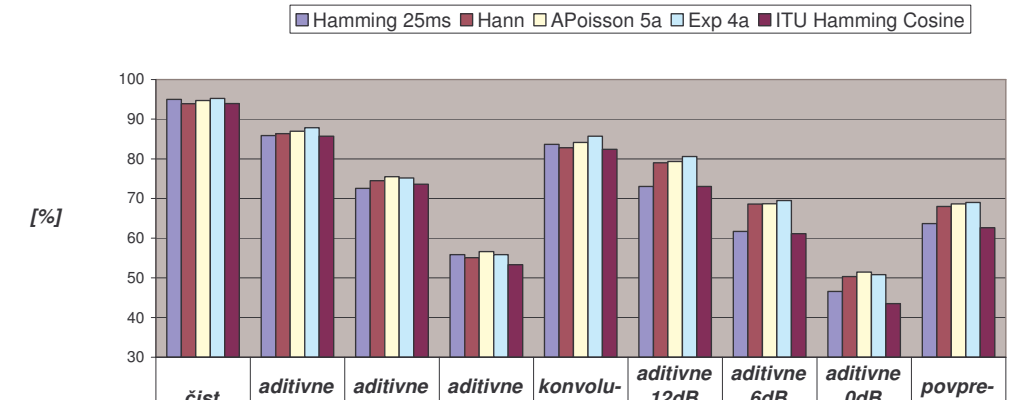
Primerjava uspešnosti CSLU razpoznavalnika ob uporabi govorne zbirke ŠTEVKE je prikazana na slikah 5.1-5.3. Za predstavitev govornega signala so bile uporabljene kepstralno normalizirane MFCC značilke; po 13 v vsakem od 5 okvirjev t.i. kontekstnega okna, kar skupaj pomeni 65 značilk v vektorju.

Med simetričnimi okni (Slika 5.1) v povprečju precej izstopa Hannovo; pozitivna razlika nastane predvsem v kombinaciji aditivnih in konvolutivnih motenj. Še nekoliko bolj sta uspešni obe nesimetrični okni (Slika 5.2), ki sta poleg pričakovane uspešnosti pri konvolutivnih, najuspešnejši še pri aditivnih motnjah. Podrobnejša primerjava uspešnosti na izbranih testnih množicah (Slika 5.3) pokaže največja razhajanja v primeru konvolutivnih motenj; še posebej pri nizkoprepustni konvolutivni spremembi testnih množic (oznaka "llp"). Dokaj nevzpodbudne rezultate ima ITU okno, ki je v primerjavi s Hammingovim še nekoliko slabše; glede na izrazitejše spektralno puščanje je tak izid pričakovan.



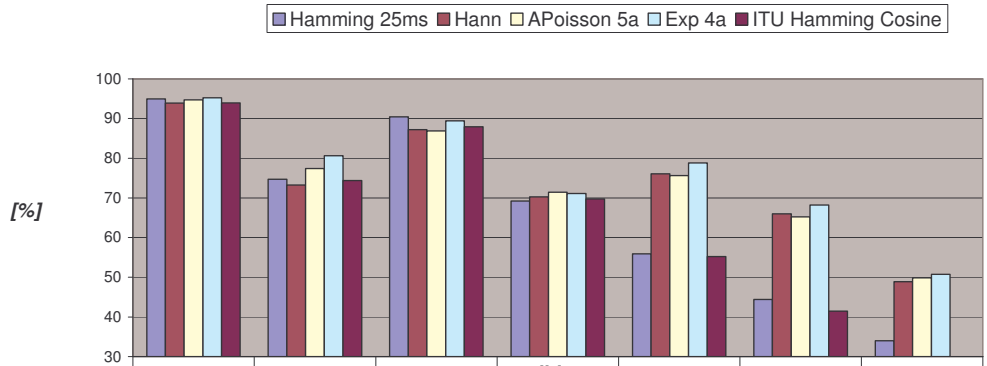
	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 16ms	95,32	87,71	75,61	55,46	83,94	73,50	62,00	43,95	63,52
Hamming 25ms	94,97	85,88	72,57	55,85	83,67	73,03	61,73	46,55	63,69
Hamming	94,08	86,36	73,34	53,28	82,85	72,43	60,86	44,20	62,60
Hann	93,89	86,35	74,48	55,09	82,78	78,99	68,57	50,32	67,97

Slika 5.1: Uspešnost CSLU-ŠTEVKE-MFCC SRG s simetričnimi okni.



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	94,97	85,88	72,57	55,85	83,67	73,03	61,73	46,55	63,69
Hann	93,89	86,35	74,48	55,09	82,78	78,99	68,57	50,32	67,97
APoisson 5a	94,72	86,95	75,51	56,61	84,12	79,33	68,68	51,48	68,64
Exp 4a	95,22	87,83	75,16	55,88	85,70	80,56	69,46	50,77	69,02
ITU Hamming Cosine	93,93	85,72	73,56	53,32	82,39	73,01	61,13	43,48	62,59

Slika 5.2: Uspešnost CSLU-ŠTEVKE-MFCC SRG z izbranimi simetričnimi in nesimetričnimi okni.

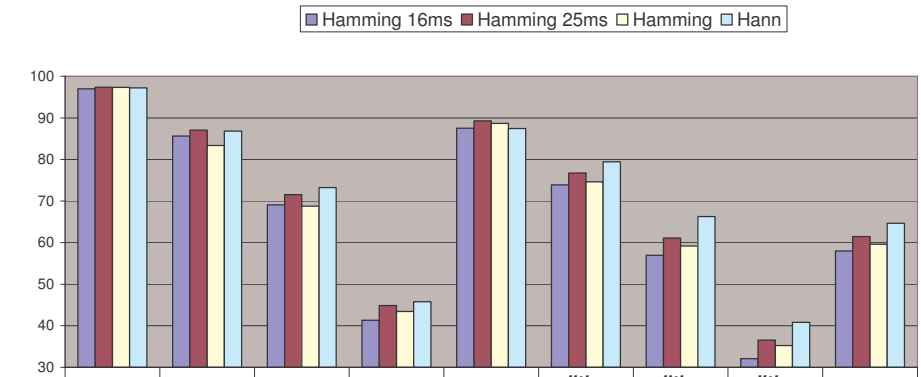


	čist	llp	hp	aditivne 12dB +odjek	aditivne 12dB +llp	aditivne 6dB +llp	aditivne 0dB +llp
Hamming 25ms	94,97	74,70	90,43	69,22	55,91	44,42	34,00
Hann	93,89	73,22	87,18	70,24	76,05	65,97	48,89
APoisson 5a	94,72	77,42	86,88	71,44	75,61	65,19	49,86
Exp 4a	95,22	80,62	89,40	71,11	78,83	68,21	50,74
ITU Hamming Cosine	93,93	74,41	87,92	69,73	55,21	41,46	27,74

Slika 5.3: Uspešnost CSLU-ŠTEVKE-MFCC SRG z izbranimi okni – primerjava na izbranih testnih množicah.

5.1.2 "CSLU-NUMBERS-MFCC" SRG

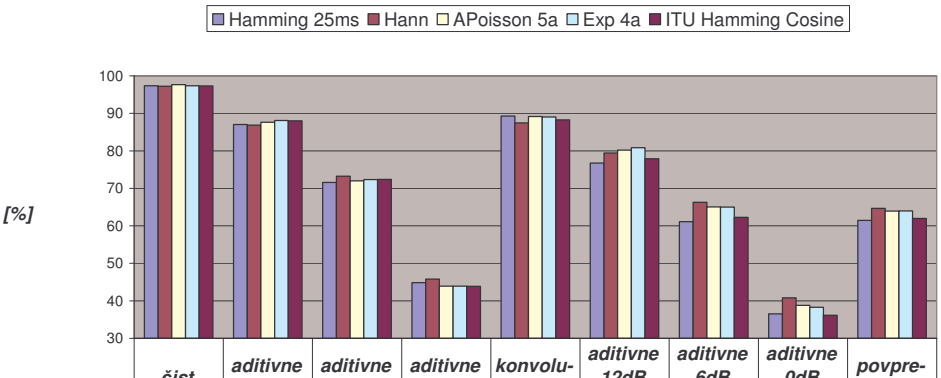
Slike 5.4-5.6 prikazujejo uspešnost CSLU razpoznavalnika ob uporabi govorne zbirke NUMBERS. Proces parametrizacije in predstavitev signala sta enaka tistim pri zbirki ŠTEVKE v podpoglavju 5.1.1. Bistveno se razlikujejo le velikosti učne in testnih množic; te so v zbirki NUMBERS približno trikrat večje. Kljub temu pa so rezultati zelo podobni. Med Hammingovimi je v povprečju najbolj robustno okno dolžine 25ms. V povprečju med vsemi simetričnimi okni izstopa Hannovo. Kljub podobnim rezultatom je mogoče zaslediti tudi nekaj razlik. Hannovo okno je tukaj v povprečju celo boljše od obeh nesimetričnih. Razlika je sicer majhna, vendar imajo zaradi večje velikosti testnih množic rezultati večjo težo. Prav tako je razvidno, da je CSLU razpoznavalnik uspešnejši kot pri ŠTEVKAH v primeru nespremenjene testne množice (oznaka "čist"), v vseh ostalih primerih pa slabši. Čeprav veljajo ŠTEVKE s stališča razpoznavanja za težje obvladljivo govorno zbirko, se temu CSLU razpoznavalnik očitno dobro prilagodi. Zanimiv pojav je težko natančneje opredeliti, ker je zmožnost interpretacije naučenega znanja pri nevronskih mrežah omejena. Potrebno pa je poudariti, da so bile motnje dodane zbirkama ločeno. To pomeni, da je njihov nivo določen v relativnem razmerju SNR z obstoječimi signali in nanj lahko vpliva že dolžina premorov v izgovorjavah, ki so v zbirki NUMBERS precej krajši.



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 16ms	97,01	85,63	69,10	41,31	87,53	73,88	56,94	32,09	58,01
Hamming 25ms	97,37	87,04	71,56	44,85	89,29	76,73	61,12	36,54	61,47
Hamming	97,36	83,38	68,79	43,43	88,67	74,64	59,20	35,22	59,59
Hann	97,21	86,84	73,24	45,81	87,45	79,45	66,28	40,82	64,65

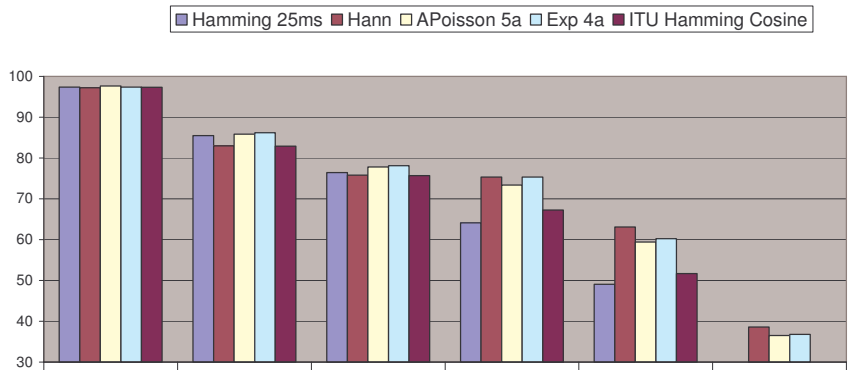
Slika 5.4: Uspešnost CSLU-NUMBERS-MFCC SRG s simetričnimi okni.

Slika 5.6 prikazuje podrobnejšo primerjavo uspešnosti na nekaj izbranih množicah. Razlike so v primerjavi z zbirko ŠTEVKE manjše, vendar še vedno dokaj konsistentno v prid oknom z manjšim spektralnim puščanjem (Hann, Exp, APoisson).



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	97,37	87,04	71,56	44,85	89,29	76,73	61,12	36,54	61,47
Hann	97,21	86,84	73,24	45,81	87,45	79,45	66,28	40,82	64,65
APoisson 5a	97,64	87,63	71,99	43,93	89,16	80,19	65,06	38,81	63,93
Exp 4a	97,36	88,08	72,33	43,91	89,02	80,84	65,00	38,26	63,99
ITU Hamming Cosine	97,33	88,01	72,38	43,87	88,28	77,90	62,27	36,16	61,98

Slika 5.5: Uspešnost CSLU-NUMBERS-MFCC SRG z izbranimi simetričnimi in nesimetričnimi okni.



	čist	odjek	aditivne 12dB+odjek	aditivne 12dB+llp	aditivne 6dB+llp	aditivne 0dB+llp
Hamming 25ms	97,37	85,49	76,42	64,11	49,06	28,32
Hann	97,21	82,97	75,79	75,32	63,10	38,58
APoisson 5a	97,64	85,85	77,82	73,36	59,39	36,50
Exp 4a	97,36	86,21	78,12	75,35	60,25	36,77
ITU Hamming Cosine	97,33	82,89	75,68	67,25	51,70	28,23

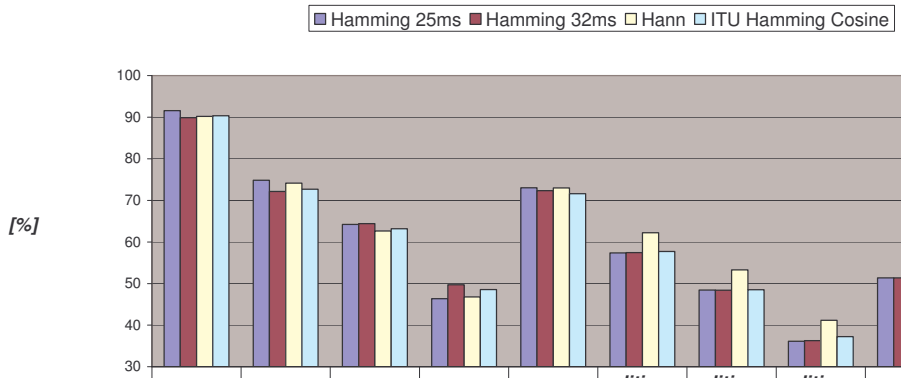
Slika 5.6: Uspešnost CSLU-NUMBERS-MFCC SRG z izbranimi okni – primerjava na izbranih testnih množicah.

5.1.3 "HMM-ŠTEVKE-MFCC" SRG

Slike 5.7-5.9 prikazujejo uspešnost HMM razpoznavalnika ob uporabi govorne zbirke ŠTEVKE. Nivo motenj v testnih množicah je bil enak kot pri "CSLU-ŠTEVKE-MFCC" razpoznavalniku (podpoglavje 5.1.1). Za predstavitev signala je bilo uporabljenih 13 MFCC značilk za vsak okvir signala. Delta značilke pri preizkusih na HMM razpoznavalniku v želji po večji raznolikosti niso bile uporabljene, zato so rezultati v primerjavi s "CSLU-ŠTEVKE-MFCC" razpoznavalnikom pričakovano slabši. Vsi rezultati HMM razpoznavalnika so bili dobljeni z uporabo 8 mešanic normalnih porazdelitev. V podobnih primerih se običajno HMM modeli z različnim številom mešanic ovrednotijo na validacijski množici; za preizkus na testni množici se izberejo najuspešnejši med njimi. Ker število mešanic lahko bistveno vpliva na robustnost SRG, so bili vsi preizkusi izvedeni ob enakem številu uporabljenih mešanic. Na ta način je bil vpliv tega dejavnika na robustnost razpoznavanja precej zmanjšan.

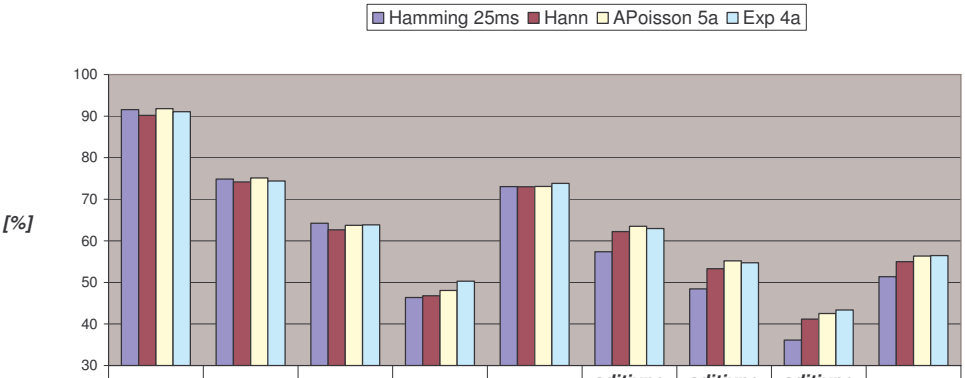
Slika 5.7 prikazuje uspešnost simetričnih oken. Med njimi veljajo podobna razmerja kot v obeh CSLU sistemih. Hannovo okno je precej boljše od Hammingovega. Izbrani nesimetrični okni sta konsistentno najboljši v vseh skupinah testnih množic (Slika 5.8). Slika 5.9 vsebuje podrobnejšo primerjavo izbranih testnih množic z največjimi razlikami; med njimi tudi tukaj prevladujejo konvolutivne spremembe. Zanimivo je, da so razlike v odstotnih točkah pri enakih motnjah in različnih vrednostih razmerja SNR zelo podobne. To je najbrž odraz

pasivne vloge oken v SRG z vidika odpravljanja motenj. Okna namreč zmanjšujejo nivo motenj le posredno z omejevanjem lastnosti spektralnega puščanja. Na področju ARG je znanih kar nekaj aktivnejših načinov odpravljanja motenj (npr. spektralno odštevanje, adaptivno filtriranje), vendar so ti časovno in prostorsko precej bolj zahtevni.



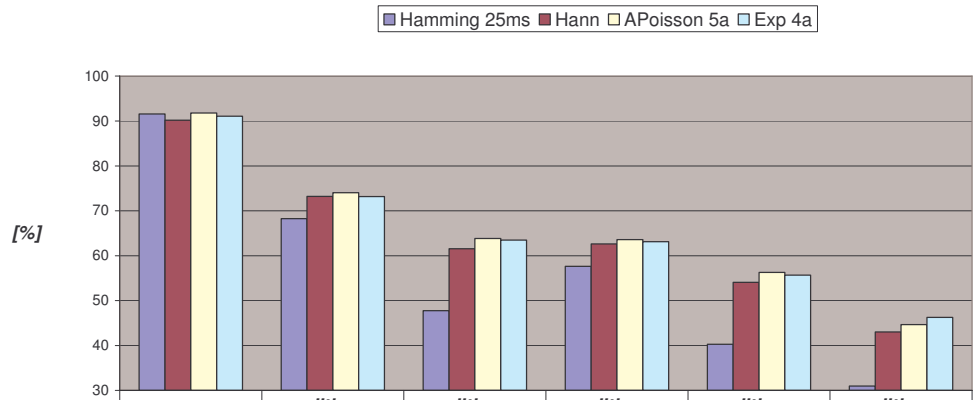
	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	91,57	74,85	64,24	46,37	73,05	57,39	48,46	36,13	51,38
Hamming 32ms	89,89	72,15	64,43	49,70	72,34	57,45	48,40	36,28	51,37
Hann	90,19	74,17	62,65	46,78	73,02	62,24	53,28	41,16	54,97
ITU Hamming Cosine	90,34	72,72	63,20	48,58	71,62	57,71	48,51	37,23	51,63

Slika 5.7: Uspešnost HMM-ŠTEVKE-MFCC SRG s simetričnimi okni.



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	91,57	74,85	64,24	46,37	73,05	57,39	48,46	36,13	51,38
Hann	90,19	74,17	62,65	46,78	73,02	62,24	53,28	41,16	54,97
APoisson 5a	91,81	75,11	63,74	48,04	73,08	63,49	55,20	42,52	56,36
Exp 4a	91,07	74,39	63,83	50,28	73,83	62,97	54,74	43,38	56,45

Slika 5.8: Uspešnost HMM-ŠTEVKE-MFCC SRG z izbranimi simetričnimi in nesimetričnimi okni.



	čist	aditivne 12dB+lp	aditivne 12dB+llp	aditivne 6dB+lp	aditivne 6dB+llp	aditivne 0dB+llp
Hamming 25ms	91,57	68,27	47,77	57,65	40,28	30,95
Hann	90,19	73,23	61,56	62,65	54,08	43,01
APoisson 5a	91,81	74,03	63,86	63,59	56,29	44,67
Exp 4a	91,07	73,20	63,48	63,12	55,67	46,27

Slika 5.9: Uspešnost HMM-ŠTEVKE-MFCC SRG z izbranimi okni – primerjava na izbranih testnih množicah.

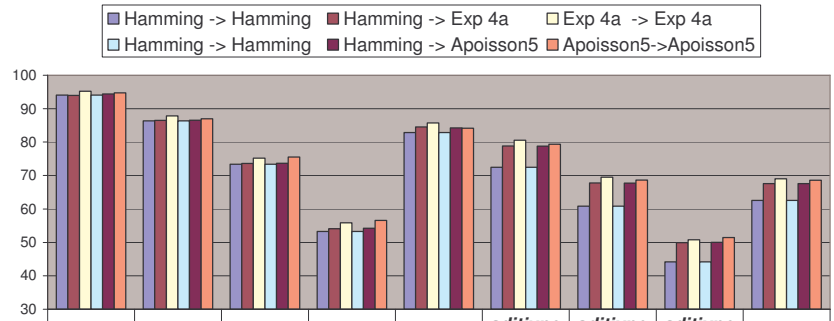
5.2 ZAPOREDNI PREIZKUS

Rezultati tega preizkusa so za končne ugotovitve o vplivu nesimetričnih oken še posebej pomembni. V disertaciji je precej govora o objektivnem vrednotenju vpliva okna na opis govornega signala. Če je ta vpliv res splošen in informacija v dobljenih opisih res objektivno boljša, potem bi to moralo pozitivno delovati tudi na uspešnost že naučenih SRG. Odgovor na to vprašanje je osnovni cilj izvedbe zaporednih preizkusov. Zaradi časovne kompleksnosti je bil uporabljen le del vseh testnih množic[1], zato navedeni rezultati v absolutnem smislu niso primerljivi z ostalimi v disertaciji.

V nadaljevanju so na slikah 5.10-5.12 predstavljeni rezultati preizkusov. Vsako okno je opisano s tremi vrsticami. V prvi vrstici je podana referenčna uspešnost SRG, ki je bil naučen in preizkušen ob uporabi Hammingovega okna; v 3. vrstici je podobno podan še rezultat sistema, ki je bil naučen in preizkušen ob uporabi izbranega okna. Obe vrstici predstavljata referenčni meji za dejanski zaporedni preizkus v 2. vrstici. V tem primeru je bil sistem naučen na učni množici s Hammingovim oknom, nato pa se je pri praktičnem preizkusu uporabilo

[1] *Pri CSLU razpoznavalniku je bilo uporabljenih 89 množic (vse kombinacije brez "lp" motnje), pri HMM pa 45 (kombinacije brez aditivnih motenj pri SNR=0dB in konvolutivnih motenj "llp" in "hp").*

izbrano okno. Na ta način je bilo naučenemu sistemu okno spremenjeno šele v fazi delovanja . V primerjavi izidov vseh treh vrstic je skrit odgovor na vprašanje, ali uporaba "boljšega" okna brez ponovnega učenja povzroči večjo robustnost sistema. Če to res velja, je rezultat v 2. vrstici boljši od tistega v 1. in blizu rezultatu v 3. To hkrati pomeni, da se s preprosto zamenjavo Hammingovega okna (verjetno tudi katerega drugega) v SRG lahko poveča njegova robustnost že med delovanjem, brez potrebe po ponovnem učenju.



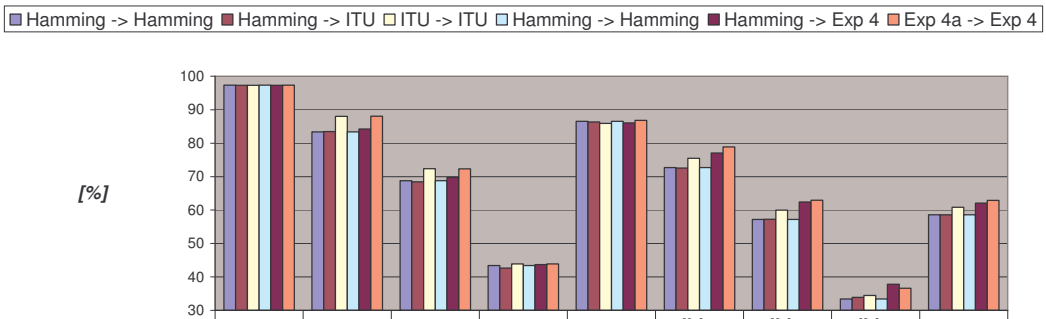
	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming -> Hamming	94,08	86,36	73,34	53,28	82,85	72,43	60,86	44,20	62,60
Hamming -> Exp 4a	93,98	86,50	73,58	54,08	84,51	78,81	67,81	49,95	67,59
Exp 4a -> Exp 4a	95,22	87,83	75,16	55,88	85,70	80,56	69,46	50,77	69,02
Hamming -> Hamming	94,08	86,36	73,34	53,28	82,85	72,43	60,86	44,20	62,60
Hamming -> Apoisson5	94,38	86,53	73,63	54,22	84,22	78,78	67,77	50,03	67,60
Apoisson5->Apoisson5	94,72	86,95	75,51	56,61	84,12	79,33	68,68	51,48	68,64

Slika 5.10: Uspešnost CSLU-ŠTEVKE-MFCC SRG z izbranimi okni v zaporednem preizkusu.

Pri govorni zbirki ŠTEVKE (Slika 5.10) je zboljšanje opazno skoraj v vseh stolpcih. Nekoliko manj izrazito je v primeru zbirke NUMBERS (Slika 5.11), vendar v povprečju še vedno dovolj dobro. Razvidno je tudi, da je uspešnost pri zamenjavi Hammingovega z nesimetričnim oknom že primerljiva z uspešnostjo SRG, ki je bil s tem oknom tudi naučen. Rezultati dokazujejo, da ponovno učenje sistema pri uvedbi nesimetričnega okna ni potrebno, kar predstavlja pomemben dosežek. Stroški ponovnega učenja so namreč lahko zelo visoki.

Enake ugotovitve veljajo tudi v primeru HMM SRG (Slika 5.12), le da so razlike v uspešnosti oken manjše; ustrezno manjše je tudi izboljšanje v 2. vrstici.

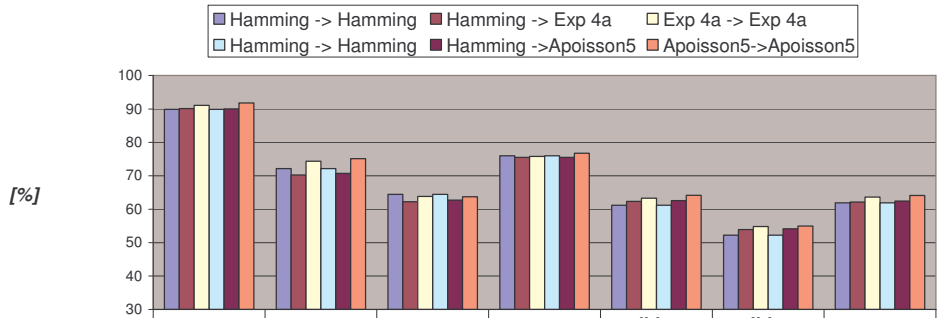
Za vse preizkuse je zanimivo še dejstvo, da se izboljšanje pokaže tudi v primeru aditivnih motenj. Bolj pričakovan je namreč ta pojav pri konvolutivnih motnjah, kjer je vpliv spektralnega puščanja bolj izrazit. Za aditivne motnje pa ni tako značilen, ker je njihov amplitudni spekter v povprečju bolj enakomeren; seveda pa se aditivne motnje med seboj



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming -> Hamming	97,36	83,38	68,79	43,43	86,56	72,70	57,17	33,40	58,58
Hamming -> ITU	97,30	83,47	68,44	42,66	86,35	72,56	57,25	33,89	58,59
ITU -> ITU	97,33	88,01	72,38	43,87	85,93	75,47	59,95	34,43	60,82
Hamming -> Hamming	97,36	83,38	68,79	43,43	86,56	72,70	57,17	33,40	58,58
Hamming -> Exp 4	97,30	84,26	69,72	43,72	86,07	77,05	62,42	37,82	62,08
Exp 4a -> Exp 4	97,36	88,08	72,33	43,91	86,84	78,90	62,94	36,61	62,90

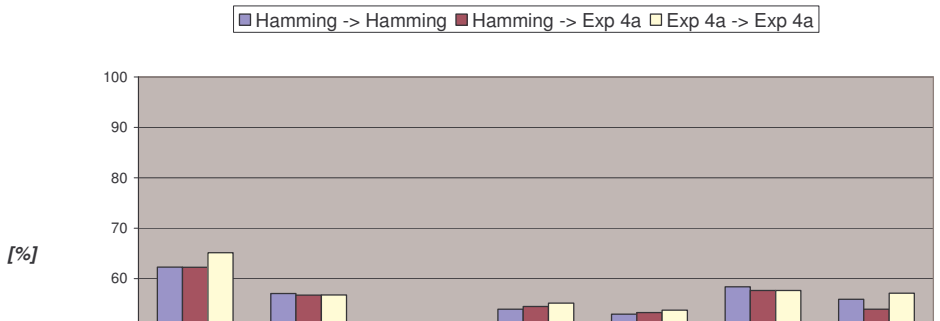
Slika 5.11: Uspešnost CSLU-NUMBERS-MFCC SRG z izbranimi okni v zaporednem preizkusu.

precej razlikujejo, zato je vpliv spektralnega puščanja v primeru nekaterih motenj bolj in drugih manj izrazit. Slika 5.13 prikazuje še podrobnejšo primerjavo po posameznih testnih množicah iz skupine aditivnih motenj z razmerjem SNR 0 dB (oznaka *"aditivne 0dB"*). Občutno izboljšanje je doseženo le v primeru treh motenj ("govor", "volvo", "tovarna"). Pri podrobnejši analizi prikazanih motenj se je pokazala le ena izrazitejša skupna lastnost imenovanih motenj – večje vrednosti amplitudnega odziva v nižjem frekvenčnem pasu. Če bi bilo izboljšanje povezano zgolj s spektralnim puščanjem, bi moralo biti še najbolj izrazito v primeru frekvenčno omejene motnje ("pas 900Hz"), vendar se to ni zgodilo. Očitno je oddaljeno spektralno puščanje okna pomembnejše, saj pada s spektralno razdaljo, ki je največja prav med obema skrajnima točkama amplitudnega spektra. Zato je vpliv okna bolj izrazit pri motnjah z večjimi razlikami na obeh koncih spektra in manj, če se motnja nahaja bližje sredini amplitudnega odziva ("pas 900Hz").



	čist	aditivne 12dB	aditivne 6dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	povprečje
Hamming -> Hamming	89,89	72,15	64,43	76,02	61,19	52,21	61,92
Hamming -> Exp 4a	90,14	70,26	62,26	75,54	62,33	53,90	62,13
Exp 4a -> Exp 4a	91,07	74,39	63,83	75,79	63,30	54,82	63,64
Hamming -> Hamming	89,89	72,15	64,43	76,02	61,19	52,21	61,92
Hamming ->Apoisson5	90,04	70,71	62,72	75,52	62,57	54,14	62,42
Apoisson5->Apoisson5	91,81	75,11	63,74	76,78	64,15	54,96	64,11

Slika 5.12: Uspešnost HMM-ŠTEVKE-MFCC SRG z izbranimi okni v zaporednem preizkusu.



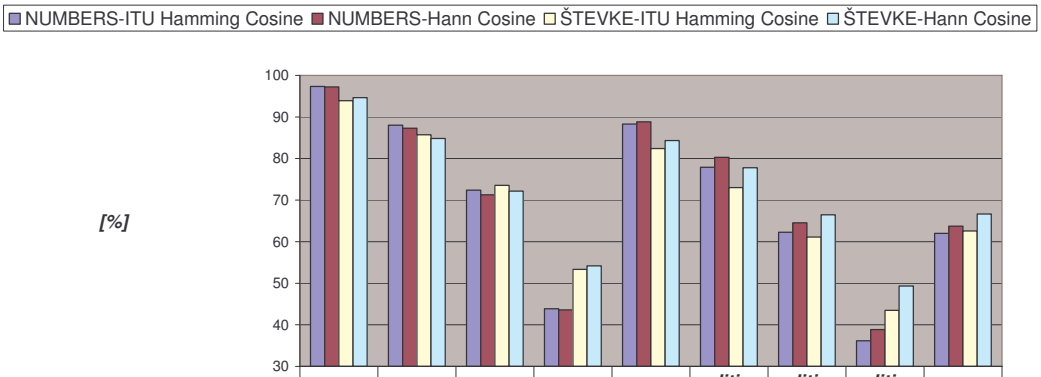
	beli šum 0dB	roza šum 0dB	govor 0dB	volvo 0dB	tovarna 0dB	F-16 0dB	pas 900hz 0dB
Hamming -> Hamming	62,28	57,05	32,50	53,94	52,91	58,38	55,87
Hamming -> Exp 4a	62,23	56,71	40,38	54,44	53,25	57,64	53,94
Exp 4a -> Exp 4a	65,14	56,76	45,61	55,13	53,75	57,64	57,10

Slika 5.13: Uspešnost CSLU-ŠTEVKE-MFCC SRG v zaporednem preizkusu – podrobnejša primerjava v aditivnih motnjah.

5.3 PRIMERJAVA KOMBINIRANIH KOSINUSNIH OKEN

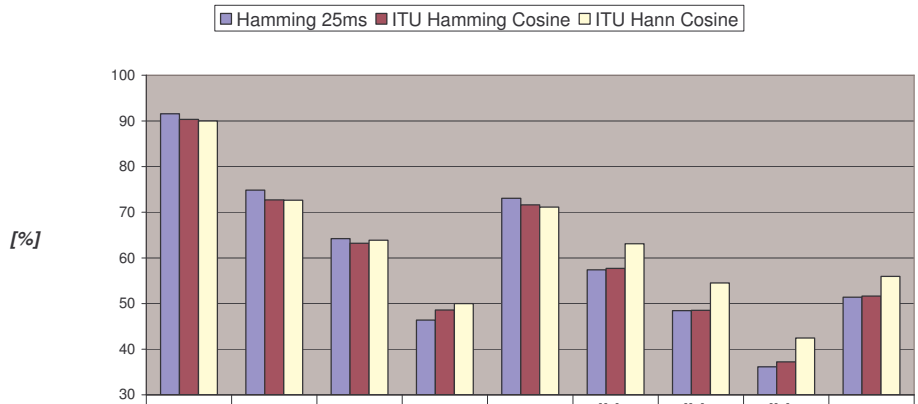
Lastnosti ITU okna so že podrobneje predstavljene v podpoglavju 4.5.3. Strokovne utemeljitve uporabe Hammingovega okna v definiciji ITU okna nismo uspeli zaslediti. Zato je potrebno ovrednotenje primernosti tega okna za uporabo v SRG. Glede na analizo njegovega amplitudnega odziva ni mogoče pričakovati izrazitejših pozitivnih učinkov, saj pri pomembnejših lastnostih ni zaznavno boljši od Hammingovega okna, ki glede na rezultate ne velja za dobro izbiro. To so potrdili tudi prejšnji poskusi njegove uporabe na področju ARG [82].

Rezultati praktičnih preizkusov so prikazani na slikah 5.14 in 5.15. V primerjavo je vključeno še predlagano alternativno "Hann–kosinusno" okno, ki je pokazalo vzpodbudnejše rezultate. Zato ni nobenega razloga, da to okno ne bi nadomestilo ITU okna na področju ARG. Potencial predlagane zamenjave na področju kodiranja signalov bi bilo potrebno posebej raziskati; ob tem bi bilo nedvomno zanimivo obravnavati še druga uspešna nesimetrična okna, predstavljena v disertaciji.



	čist	aditiv. 12dB	aditiv. 6dB	aditiv. 0dB	konvo-lutivne	aditiv. 12dB +konvo	aditiv. 6dB +konvo	aditiv. 0dB +konvo	povpre-čje
NUMBERS-ITU Hamming Cosine	97,33	88,01	72,38	43,87	88,28	77,90	62,27	36,16	61,98
NUMBERS-Hann Cosine	97,21	87,29	71,27	43,59	88,83	80,29	64,54	38,85	63,73
ŠTEVKE-ITU Hamming Cosine	93,93	85,72	73,56	53,32	82,39	73,01	61,13	43,48	62,59
ŠTEVKE-Hann Cosine	94,63	84,85	72,17	54,17	84,35	77,79	66,47	49,31	66,64

Slika 5.14: Uspešnost CSLU-MFCC s kombiniranimi kosinusnimi okni.



	čist	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolu-tivne	aditivne 12dB +konvol.	aditivne 6dB +konvol.	aditivne 0dB +konvol.	povpre-čje
Hamming 25ms	91,57	74,85	64,24	46,37	73,05	57,39	48,46	36,13	51,38
ITU Hamming Cosine	90,34	72,72	63,20	48,58	71,62	57,71	48,51	37,23	51,63
ITU Hann Cosine	89,99	72,66	63,86	49,98	71,14	63,11	54,52	42,45	55,95

Slika 5.15: Uspešnost HMM-ŠTEVKE-MFCC SRG s kombiniranimi kosinusnimi okni.

5.4 UGOTOVITVE

Na osnovi prikazanih rezultatov se ugotovi naslednje:

- *Zaporedni preizkusi dokazujejo, da so obstoječi SRG že sposobni dokaj učinkovito uporabiti boljše opise govornih signalov iz procesa parametrizacije. To dejstvo je vzpodbuda za nadaljnje delo.*
- *Zaradi precejšnje razlike v robustnosti kombiniranih kosinusnih oken (podpoglavje 5.3) bi bila potrebna podrobnejša raziskava učinkovitosti predlaganih alternativnih oken tudi v obstoječih sistemih za kodiranje govornih signalov.*
- *Pri kombinaciji aditivnih in nizkoprepustnih konvolutivnih motenj pride do zelo velikih razlik v uspešnosti SRG. Največja zaznana razlika v uspešnosti je relativno zelo velika (Slika 5.3 in Slika 5.9). To dejstvo je glede na majhno pozornost temu segmentu prav gotovo presenetljivo in si zasluži nadaljnjo obravnavo.*
- *Vloga oken v doseganju večje robustnosti SRG je pasivna. Za razliko od drugih aktivnejših metod lahko okno le posredno vpliva na manjše puščanje motenj v druge*

dele amplitudnega spektra. Glede na to omejitev so dosežene izboljšave robustnosti SRG z nesimetričnimi okni prav gotovo nad pričakovanji.

- *Razlike v uspešnosti med obema govornima zbirkama so sicer zaznavne, vendar zaradi ločenega dodajanja motenj in velikih razlik med zbirkama niso absolutno primerljive. Ločeno dodajanje motenj je namreč potreben pogoj za primerljivost z drugimi raziskavami.*
- *Primerjava uspešnosti obeh referenčnih sistemov je že v izhodiščih opredeljena za manj pomembno. Zagotovitev raznolikosti referenčnega okolja je namreč v nasprotju z objektivnostjo medsebojnih primerjav glavnih dejavnikov uspešnosti razpoznavanja. V disertaciji je imela prednost raznolikost. Kljub temu sta robustnosti obeh referenčnih sistemov primerljivi – sploh če se upošteva dejstvo, da ima CSLU referenčni sistem zaradi kontekstnega okna dostop tudi do informacij o časovni dinamiki MFCC značilk. HMM referenčni sistem namreč v svojem delovanju uporablja le vektor 13 statičnih MFCC značilk. Zato je njegova nekoliko slabša robustnost pričakovana.*

6. ZAKLJUČEK

Disertacija vsebuje dva večja tematska sklopa. Prvi obsega proučitev možnosti načrtovanja nesimetričnih oken in analizo njihovih lastnosti; preizkušenih je bilo nekaj različnih postopkov. Drugi sklop je povezan z oceno pomembnosti lastnosti nesimetričnih oken in vpliva na robustnost SRG; opraviti je bilo potrebno veliko praktičnih preizkusov v referenčnem okolju in pri tem upoštevati še morebitne vplive drugih dejavnikov.

Problem proučitve vpliva okna na robustnost SRG je večplasten. Obravnavan je bil z dveh dokaj različnih vidikov. S pomočjo znanj s področja frekvenčne analize signalov so se bolj natančno opredelile razlike v posameznih lastnostih oken. Tovrstne ugotovitve so sicer splošnejše, vendar bolj merljive in dokazljive. Vrednotenje vpliva z vidika robustnosti razpoznavanja pa je precej bolj odvisno od dejanskih pogojev delovanja in značilnosti SRG.

Na opravljene raziskave sta pomembno vplivala dva dejavnika:

- *časovna zahtevnost praktičnih preizkusov na referenčnih SRG*

  Zaradi časovne zahtevnosti klasična, podatkovno orientirana optimizacija lastnosti oken ni bila izvedljiva. Uporabljeni so bili le praktični preizkusi končne uspešnosti razpoznavanja ob uporabi različnih oken. Zaradi časovne kompleksnosti je pri tem bilo potrebno omejevati množico preizkusov in jo uporabljati karseda selektivno. Že vnaprej pa je bilo znano, da vseh preizkusov ne bo mogoče izvesti. Kljub temu pa lahko njihov dober izbor nudi zadostno empirično podporo splošnejšim ugotovitvam z vidika frekvenčne analize signalov.

- *dualnost vrednotenja lastnosti oken*

  Vrednotenje z dveh (dokaj nepovezanih) vidikov je bilo v našem primeru pravzaprav prednost. Z vidika splošne frekvenčne analize so ugotovitve nedvomno širšega in bolj dolgoročnega značaja. Realno vrednotenje uspešnosti referenčnih sistemov pa precej pove o trenutni učinkovitosti SRG in njihovemu izkoristku kvalitetnejših informacij v opisih govornega signala.

Pri interpretacijah rezultatov sta bila enakomerno zastopana empirični in teoretični vidik. Zato so bile za končne ocene poleg samih rezultatov pomembne še posebnosti in značilnosti

postopkov načrtovanja ter dobljenih oken. Glede na omenjeno dualnost vrednotenja in raznolikost referenčnega okolja se zdi takšen pristop edini sprejemljiv.

V nadaljevanju so najprej podane končne ugotovitve o pomenu posameznih lastnosti okna in priporočili za njihovo uporabo na področju ARG. Poglavje se zaključi z razpravo o splošnejšem pomenu opravljenega dela in smernicami za nadaljnje raziskave.

6.1 POMEN LASTNOSTI OKNA ZA ARG

V podpoglavju 4.4 so hipotetično določene želene lastnosti amplitudnega odziva okna. Ta izhodišča so bila upoštevana pri kasnejšem načrtovanju nesimetričnih oken, ki so se v praktičnih preizkusih v referenčnem okolju pokazale za uspešne. Na osnovi novih izkušenj je mogoče potrditi uresničitev pričakovanj in natančneje opredeliti pomen posameznih lastnosti okna v amplitudnem odzivu:

- *širši glavni val*

  Za to lastnost obstajata dva močna argumenta: podobnost s človekovo slušno percepcijo in manjši pomen frekvenčne ločljivosti[1]. Prav zato se zdi lastnost smiselna, še posebej zaradi s tem povezanega izboljšanja druge želene lastnosti - nižjih stranskih valov.

- *padajoča višina stranskih valov*

  Spektralno puščanje je po moji oceni ena najpomembnejših lastnosti okna pri uporabi v ARG. Zaradi združevanja informacij v kritične pasove je bližnje spektralno puščanje manj pomembno. Precej bolj pomembno pa je oddaljeno spektralno puščanje, ki mora biti čim manjše. Pomen puščanja pa se lahko v povezavi s konkretno zasnovo SRG še potencira – npr. v primeru predpostavke o neodvisnosti značilk, povezane z uporabo diagonalne kovariančne matrike v HMM SRG. Zaradi medsebojne povezanosti lastnosti oken so še posebej zanimive tiste kombinacije z možnostjo obojestranskih izboljšav. Zato je prava vrednost te lastnosti prav v povezavi s prejšnjo – širšim glavnim valom.

---

[1] *V implementacijah SRG se v postopku parametrizacije običajno izvede združevanje frekvenčnih pasov.*

- *monoton potek*

  Večina vseh načrtovanih nesimetričnih oken (razen načrtovanih z metodami za KEO filtre) ima dokaj gladek potek amplitudnega odziva. Glede na omejeno število izvedenih preizkusov ni mogoče podati natančnejše ocene pomembnosti te lastnosti. Splošni vtis pa ji ne pripisuje prevelikega pomena, saj njen učinek že na samih opisih signalov ni zaznaven. Na področju splošnega procesiranja zvočnih signalov [66] je ta lastnost zaželena, zato jo je priporočljivo upoštevati, če seveda to ne pomeni degradacije ostalih, pomembnejših lastnosti.

6.2 PRIPOROČILA ZA UPORABO OKEN V SRG

Iz ugotovitev v podpoglavju 6.1 in rezultatov praktičnih preizkusov (poglavji 5. in 4.5) sledijo naslednja priporočila za uporabo oken na področju ARG:

- *V splošnem je priporočljiva uporaba oken, ki omogočajo krajše časovne zakasnitve pri parametrizaciji govornih signalov in hkrati še povečujejo robustnost razpoznavanja. Takšni okni sta NEO eksponentno ali modificirano Poissonovo okno. Pri obeh se lahko časovno frekvenčne lastnosti enostavno kontrolirajo s pomočjo le enega parametra. Praktični preizkusi celo kažejo, da je mogoče s temi okni uspešno nadomestiti tiste v že delujočih SRG brez ponovnega učenja.*
- *Uporaba ITU okna, ki je sicer zelo pogosto pri kodiranju govora, na področju ARG ne daje dobrih rezultatov.*

6.3 POMEN OPRAVLJENIH RAZISKAV IN NADALJNJE SMERNICE

Analiza predstavljenih praktičnih rezultatov v 4. in 5. poglavju pokaže precejšnje povečanje inherentne robustnosti SRG ob uporabi nesimetričnih oken. Njihove analitično potrjene boljše lastnosti so se pozitivno odrazile tudi na rezultatih praktičnih preizkusov.

Pri razpravi o pomenu opravljenega dela je potrebno posebej poudariti rezultate zaporednih preizkusov, ki dokazujejo sposobnost že obstoječih SRG, da učinkovito izkoristijo objektivno boljšo informacijo v opisu signala, četudi ta pomeni zaznavno spremembo v primerjavi z učnimi pogoji. To dejstvo prav gotovo opozarja na večjo splošnost pozitivnega vpliva nesimetričnih oken na robustnost SRG.

Iz predstavljenih dejstev je mogoče potrditi obstoj lastnosti okna, ki v splošnem prispevajo k večji robustnosti SRG, vendar vedno v povezavi z ostalimi lastnostmi. Če so namreč okna v določenih lastnostih ekstremna, se lahko pozitivni vpliv izgubi, največkrat ravno zaradi s tem povezanega poslabšanja katere od ostalih pomembnih lastnosti.

Večina vseh primerjav v disertaciji je narejenih na osnovi povprečnih vrednosti skupin testnih množic. Odstopanja na nivoju posameznih izgovorjav pa so pri razpoznavi govora običajno še precej večja. To dejstvo opozarja ne samo na to, da je vpliv okna odvisen tudi od realnih podatkov, ampak tudi na to, da se verjetno največji razvojni potencial nahaja prav v možnosti dinamičnega spreminjanja parametrov okna in parametrizacije nasploh. To bi omogočalo individualno obravnavo vsakega govorca posebej, kar bi vsaj na načelni ravni moralo uspešnost še občutno izboljšati. Vsekakor pa bo za učinkovito tovrstno implementacijo potrebno še bolj uskladiti delovanje obeh glavnih procesov v SRG in opraviti še celo vrsto drugih izboljšav.

Nesimetrična okna oziroma raziskave v disertaciji imajo tudi širši pomen. Predlagani postopki načrtovanja nesimetričnih oken so zelo učinkoviti in uporabni tudi na drugih področjih. Zelo pomembna (ne samo za načrtovanje oken) je tudi izpeljava preprostih NEO modelov. Njihov odziv na enotin impulz je bil v disertaciji uporabljen za določitev robustnejših oken. Prava moč omenjenih modelov pa se skriva v njihovi parametrični fleksibilnosti; s pomikanjem polov po osi proti koordinatnem izhodišču se npr. širi oziroma oži glavni val in spreminja dušenje v amplitudnem odzivu ter efektivna dolžina odziva na enotin impulz. S pomikanjem polov v smeri enotske krožnice v Z ravnini pa se spreminja še središčna točka glavnega vala na frekvenčni osi. NEO modeli so zato uporabni tudi kot del množice filtrov, ki lahko opravlja alternativni postopek parametrizacije. Glavna potencialna prednost takšne zasnove je možnost dinamičnega spreminjanja lastnosti glede na trenutne značilnosti vhodnega govora. S tem konceptom bi se lahko še bolj približali idealni rešitvi – človekovi govorni percepciji.

Rezultati v disertaciji so lahko pomembni še z drugačnega vidika. V SRG je namreč opazno zmanjšanje uspešnosti, če naraste prisotnost motenj. Zato bi bilo zanimivo preizkusiti okna v disertaciji še na sistemih z velikim slovarjem, ki so zaradi svoje zahtevnosti na besednem nivoju podobno uspešni kot referenčna sistema v prisotnosti motenj. S tem bi bil dan odgovor na vprašanje, ali je vpliv oken enako zaznaven tudi pri bolj zahtevnem razpoznavanju in manj motenih signalih ali pa je izključno povezan le s pojavom izrazitejših motenj.

Nenazadnje je pri opredelitvi pomena nesimetričnih oken zelo pomembna tudi njihova nesporna prednost - krajša časovna zakasnitev pri frekvenčni analizi signalov. Če se k temu dodajo še prednosti v amplitudnem odzivu, imajo ta okna velik potencial za ARG, najverjetneje pa tudi še za katero od drugih sorodnih področij.

A. UPORABA SPLOŠNE OPTIMIZACIJSKE METODE ZA NAČRTOVANJE NESIMETRIČNIH OKEN

Za načrtovanje nesimetričnih oken z različnimi želenimi lastnostmi in kriteriji napake je najprimernejša uporaba splošnih optimizacijskih metod. V disertaciji je uporabljena metoda SOLVOPT [124, 125], ki je primerna za nelinearno optimizacijo. V osnovi je to gradientna metoda z nekaterimi dodatnimi izboljšavami.

Postopek načrtovanja je v primeru uporabe omenjene metode SOLVOPT dokaj enostaven. Najprej se določi optimizacijski kriterij, po potrebi še omejitve, in vzpostavi začetne pogoje za izvedbo optimizacijske metode. Po končani optimizaciji se dobljena rešitev še analizira in preveri skladnost z zadanimi zahtevami. Vsa opravila so skupaj z ustrezno programsko kodo v okolju MATLAB podrobneje opisana v nadaljevanju.

A.1 Postopek načrtovanja

Pred optimizacijo je potrebno določiti še nekaj začetnih konstantnih vrednosti. Za načrtovanje po Čebiševem kriteriju se izračuna želen amplitudni odziv ("*optim_amp*") in pozitivno utežnostno zaporedje ("*optim_utezi*"). Izbrana spodnja meja zapornega pasu je točka trikratne širine glavnega vala enako dolgega Hammingovega okna. Ta mejna točka je bila uporabljena že v naših prejšnjih raziskavah [77]. Za uporabo kriterijev časovnega indeksa ("TDI") in usmerjenosti ("D") želenega amplitudnega odziva ni potrebno določiti.

Po izračunu konstantnih vrednosti je potrebno določiti še začetno rešitev (običajno kar naključno zaporedje) in funkcijo napake. Ta se nahaja v podprogramu, ki predstavlja vhodni parameter pri klicu optimizacijske metode. Slednja po končanem postopku vrne najboljšo rešitev v vektorju $\boldsymbol{x}$. Za izračun gradienta se po potrebi doda še ločena funkcija. V nasprotnem primeru gradient izračuna metoda sama. Pri načrtovanju oken gradient ni bil uporabljen. Na koncu se izvede še optimizacijska metoda. Programsko kodo izvedbe celotnega postopka prikazuje Izpis A.1.

```
N=256;
global optim_freq frekvence optim_amp optim_utezi Ntock pass trans stop
Ntock=2^(floor(log2(N*64)));
tocke=linspace(0.,0.5, Ntock);
frekvence=2 * pi * tocke;
ratio=272/256;
bands=[0. 3*0.002 3*0.007080 0.5];
bands_freq=bands * 2 * pi;
amp_desired=[1 0];
pass=[]; trans=[]; stop=[];
for ii=1:length(tocke),
  if (tocke(ii) <= bands(2))
    pass=[pass ii] ;
  elseif (tocke(ii) <= bands(3))
    trans=[trans ii] ;
  else stop=[stop ii] ;
  end,
end
weights=[1 10];
optim_freq=[pass stop]*2*pi;
optim_utezi=utezi(bands_freq,weights,frekvence);
optim_amp=zeljeni(bands_freq,amp_desired,frekvence);
[resitev,vrednost]=solvopt(rand(1,N),'funkcija_napake');
```

Izpis A.1: Izvedba postopka načrtovanja nesimetričnih oken.

A.2 Definicija funkcije napake

V nadaljevanju je predstavljenih nekaj različnih funkcij napak, uporabljenih za načrtovanje oken v disertaciji. V izpisih je podana tudi programska koda njihovih konkretnih implementacij.

A.2.1 Čebiševa napaka amplitudnega odziva

Definicija funkcije napake je sestavni del opisa kriterija Čebiševe napake amplitudnega odziva (podpoglavje 4.5.1.2). - Izpis A.2.

```
function napaka=cebisev(x);
global optim_amp optim_utezi Ntock
act_amp=abs(fft(x,2*Ntock));
act_amp=act_amp(1:Ntock);
napaka=max(optim_utezi.*abs((optim_amp-act_amp)));
```

Izpis A.2: Funkcija za izračun Čebiševe napake amplitudnega odziva.

A.2.2 Kriterij časovnega indeksa – "TDI"

Mera za časovno zakasnitev okna (časovni indeks oziroma TDI) je določena z izrazom (4.6). Pri tem je potrebno poudariti, da višja vrednost TDI pomeni krajšo zakasnitev. Ker je vrednost TDI odvisna tudi od absolutnih vrednosti v zaporedju, je potrebno vrednosti vektorja ***x*** še pred tem normalizirati. V disertaciji so vsa okna normalizirana na enotno ojačanje glavnega vala.

```
function napaka=tdi(x);
napaka=tdi(x);

%-----------------------------
function ind=tdi(win);
[a b]=max(size(win));
if b==1
        win=win';
end
n=1:length(win);
ind=sum(n.*(win.^2));
[len,pos_l,pos_h]=effective_length(win,0.95);
if len > 0
        ind=ind/len;
end

%-----------------------------
function [len,pos_l,pos_h]=effective_length(win,factor);
% ozracuna efektivno dolzino okna
[N,cols]=max(size(win));
if cols == 1
        win=win';
end
energy=sum(win.^2)/N;
ener1=0;
[maksi,pos]=max(win);
ener1=ener1+maksi^2/N;
pos_l=pos;
pos_h=pos;
while ener1 < factor*energy,
        if win(pos_l-1)^2 >= win(pos_h+1)^2
                pos_l=pos_l-1;
                ener1=ener1+(win(pos_l)^2)/N;
        else
                pos_h=pos_h+1;
                ener1=ener1+(win(pos_h)^2)/N;
        end
end
len=pos_h-pos_l;
```

Izpis A.3: Funkcija za izračun časovnega indeksa – "TDI".

A.2.3 Kriterij usmerjenosti – "D"

Vrednost usmerjenosti okna je določena z izrazom (4.8). Še pred tem je potrebno izračunati vrednosti amplitudnega odziva ("*act_amp*") zaporedja $\boldsymbol{x}$.

```
function napaka= directivity(x);
act_amp=abs(fft(x,2*Ntock));
act_amp=act_amp(1:Ntock);
napaka=(act_amp(1)^2)/sum(act_amp(stop).^2);
```

Izpis A.4: Funkcija za izračun usmerjenosti – "D".

B. HMM SPECIFIKACIJE PARAMETROV PRAKTIČNIH PREIZKUSOV

V dodatku so podane podrobnosti implementacije referenčnega HMM sistema in na njem izvedenih praktičnih preizkusov. Sistem je implementiran s standardnimi orodji programskega paketa HTK[1]. Orodja so bila modificirana za branje okenskih zaporedij iz zunanjih datotek.

V nadaljevanju sta dve tabeli s podrobnejšimi podatki. Tabela B.1 opisuje sestavo uporabljenih množic izgovorjav[2] in druge pomembnejše podatke HMM SRG .

Tip značilk	*Število značilk*	*Govorna zbirka*	*Št. Gaussovih mešanic*	*Učna množica (stavki, besede)*	*Testna množica (stavki, besede)*
MFCC	13	ŠTEVKE	8	234, 3042	156, 2028

Tabela B.1: Osnovni parametri izvedenih praktičnih preizkusov v HMM razpoznavalniku.

Tabela B.2 vsebuje vhodne podatke pri uporabi orodij za izvedbo praktičnih preizkusov. Navedeni so naslednji podatki:

- *oznaka uporabljenega tipa značilk (1. stolpec)*
- *HTK oznaka uporabljenih značilk (1. vrstica, 2. stolpec)*
- *vhodni parametri orodja za parametrizacijo HCode1 (2. vrstica, 2. stolpec)*
- *začetni del prototipne specifikacije HMM modelov (3. vrstica, 2. stolpec)*
- *vhodni parametri orodja za razpoznavanje HVite (4. vrstica, 2. stolpec)*

Splošen opis referenčnega HMM sistema se nahaja v podpoglavju 3.2, več podrobnosti pa v literaturi na spletnem naslovu "*http://htk.eng.cam.ac.uk/* ".

---

[1] *Uporabljena verzija* 1.4A. *Podrobnejša dokumentacija je na naslovu: http://htk.eng.cam.ac.uk/.*

[2] *Podrobnejša predstavitev uporabljenih testnih množic je v dodatku D.*

MFCC	MFCC_E
	-A -M -g -H ime_datoteke_okna.bin -W -e -m -n 12 -p 22 -f 10.0 -w 32.0
	<NumStates> 10 <StreamInfo> 1 13 <VecSize> 13 <diagC> <nullD> <MFCC_E>
	HVite -d $testhmm -S testset.features labels.lis1 grammar.net1

Tabela B.2: Uporaba programskih orodij v referenčnem HMM razpoznavalniku.

C. CSLU SPECIFIKACIJE PARAMETROV PRAKTIČNIH PREIZKUSOV

V tem dodatku so podane podrobnosti implementacije in praktičnih preizkusov, izvedenih na CSLU referenčnem razpoznavalniku. Sistem je implementiran z orodji programskega paketa "CSLU Speech Toolkit". Uporabljena CSLU orodja so bila modificirana za branje vrednosti okenskega zaporedja iz zunanje datoteke.

V nadaljevanju se nahajajo podrobnejši podatki o sami implementaciji sistema in praktičnih preizkusih.

Tabela C.1 prikazuje najpomembnejše parametre preizkusov CSLU sistema na obeh govornih zbirkah. V prvem podpoglavju (C.1) se nahaja izvorna koda postopka izračuna MFCC značilk v procesu parametrizacije. V ločenih podpoglavjih (C.2 in C.3) sledita še osnovni konfiguracijski datoteki za praktične preizkuse CSLU razpoznavalnika na obeh govornih zbirkah. Pomembnejši parametri so obrazloženi v sprotnih komentarjih.

Tip značilk	*Vektor značilk*	*Zbirka*	*Učna množica št. kategorij (1., 3. korak)*	*Validacijska množica (stavkov, besed)*	*Testna množica (stavkov, besed)*
MFCC	5*13=65	NUMBERS	2000,4000	555, 3211	1168, 6663
MFCC	5*13=65	ŠTEVKE	2000,4000	78,1014	156, 2028

Tabela C.1: Specifikacija izvedenih preizkusov CSLU razpoznavalnika.

Splošen opis referenčnega CSLU sistema se nahaja v podpoglavju 3.3, več podrobnosti pa v literaturi na spletnem naslovu "*http://cslu.cse.ogi.edu/toolkit*".

C.1 Parametrizacija z MFCC značilkami

MFCC

```
proc compute_feat {wave up_feat samplerate} {
upvar $up_feat feat
if {[wave info $wave -rate] != $samplerate} {
       puts "Warning: Converting waveform to $samplerate Hz"
       set wave [wave sampleconvert $wave $samplerate]
}

set nodc      [prep !dc initialize -rate $samplerate]
set wave_nodc [prep !dc $nodc $wave]
set winsize   32.0
set framesize 10.0
set filters   22
set num_feat  13
set fname "ime_okna.bin"
set fbank  [customfeature initialize {window_from_file powerspec fbankmel } \
                        -windowsize $winsize windowfile $fname -filters $filters \
                        -samplerate $samplerate -preemphasis 0.0 -emphasis 0.0 -exp 0.0 ]
set fcep      [customfeature initialize {logspec invdct } -filters $filters   -output $num_feat -scale mel -
preemphasis 0.0 -emphasis 0.0 -exp 0.0 ]
set f1      [customfeature $fbank $wave_nodc]
set f2      [customfeature $fcep $f1]
set f21       [mx zeromean $f2]
set feat   [mx join col [list $f21]]
nuke $nodc $fname $wave_nodc $fbank $fcep $f1 $f2 $f21
return 0
}
#-------------------------------------------------------------------
proc compose_vec {wave feat samplerate up_vector} {
   upvar $up_vector vector
   set collectob [collect initialize -frames {{-6 1} {-3 1} {0 1} \
         {3 1} {6 1}} -coeffs 13]
   if {[wave info $wave -rate] != $samplerate} {
      set wave [wave sampleconvert $wave $samplerate]
      }
   set vector [collect $collectob $feat -flush]
   nuke $collectob
   return 0
   }
#-------------------------------------------------------------------
```

Izpis C.1: Parametrizacija govornega signala - izračun vektorjev MFCC značilk.

C.2 Konfiguracija preizkusov na zbirki NUMBERS

Parameter	*Vrednost*	*Komentar*
name:	mfcc_customfeature_basic_final	ime preizkusa
corpus:	numbers	zbirka
sampling_rate:	8000	frekvenca vzorčenja
filter_train:	2+1	parametri za določanje datotek v vseh množicah
filter_trainfa:	2+1	
filter_trainfb:	2+1	
filter_dev:	4+1	
filter_test:	2+1	
tcl_path:	e:/custom_scripts/training	pot do izvajalnih skript
remap:	e:/custom_scripts/nnet/remap_digits.tcl	skripta za preslikavo fonemov
features:	e:/custom_scripts/nnet/features_mfcc_customfeature_basic_final.tcl	skripta, ki izvede parametrizacijo signala
want_train:	2000	želeno število kategorij - 1. korak učenja
want_trainfa:	4000	želeno število kategorij - 3. korak učenja
iter:	40	zgornje število prehodov pri učenju
from:	20	spodnje število prehodov pri učenju
garbage:	5	vrednost Praga
clean_start:		
vec_sample_size:	65	velikost vektorja značilk
#[1]nonnet:		izvedi 1. korak učenja
#nofa:		izvedi 3. korak učenja
nofb:		ne izvedi »forward-backward« učenja
training:	nntrain	
copy_file:	ime_okna.bin	vzorci okna so zapisani v tej datoteki
#browse_all:		specifikacije testnih množic
#notest:		
#nonoisetest:		
#nomufflecombine:		
noisedir:	e:/baza/telephone_noises_numbers	pot do posnetkov motenj
noises	muffle_hp	specifikacije konvolutivnih sprememb
noises:	muffle_lp	
#noises:	muffle_hhp	
noises:	muffle_llp	
noises:	muffle_reverb	
noises:	white	specifikacije aditivnih motenj
noises:	pink	
noises:	babble	
noises:	volvo	
noises:	factory1	
noises:	f16	
noises:	pass900hz	
snr:	12.0	specifikacije ciljnih razmerij signal – šum (SNR)
snr:	6.0	
snr:	0.0	

Izpis C.2: Konfiguracijska datoteka praktičnih preizkusov CSLU sistema na zbirki NUMBERS.

[1] *"#" označuje komentar – vrstica s komentarjem se ne upošteva.*

C.3 Konfiguracija preizkusov na zbirki ŠTEVKE

Parameter	*Vrednost*	*Komentar*
name:	mfcc_customfeature_basic_final	ime preizkusa
corpus:	stevke	zbirka
sampling_rate:	8000	frekvenca vzorčenja
filter_train:	2+1	filtri za določanje datotek v vseh množicah
filter_trainfa:	2+1	
filter_trainfb:	2+1	
filter_dev:	4+1	
filter_test:	2+1	
tcl_path:	e:/custom_scripts/training	pot do izvajalnih skript
remap:	e:/custom_scripts/nnet/remap_stevke.tcl	skripta za preslikavo fonemov
features:	e:/custom_scripts/nnet/features_mfcc_customfeature_ final.tcl	skripta, ki izvede parametrizacijo signala
want_train:	2000	želeno število kategorij - 1. korak učenja
want_trainfa:	4000	želeno število kategorij - 3. korak učenja
iter:	40	zgornje število prehodov pri učenju
from:	20	spodnje število prehodov pri učenju
garbage:	5	vrednost Praga
clean_start:		
vec_sample_size:	65	velikost vektorja značilk
#[1]nonnet:		izvedi 1. korak učenja
#nofa:		izvedi 3. korak učenja
nofb:		ne izvedi »forward-backward« učenja
training:	nntrain	
copy_file:	ime_okna.bin	vzorci okna so zapisani v tej datoteki
#browse_all:		specifikacije testnih množic
#notest:		
#nonoisetest:		
#nomufflecombine:		
noisedir:	e:/baza/telephone_noises_numbers	pot do posnetkov motenj
noises	muffle_hp	specifikacije konvolutivnih sprememb
noises:	muffle_lp	
#noises:	muffle_hhp	
noises:	muffle_llp	
noises:	muffle_reverb	
noises:	white	specifikacije aditivnih motenj
noises:	pink	
noises:	babble	
noises:	volvo	
noises:	factory1	
noises:	f16	
noises:	pass900hz	
snr:	12.0	specifikacije ciljnih razmerij signal – šum (SNR)
snr:	6.0	
snr:	0.0	

Izpis C.3: Konfiguracijska datoteka praktičnih preizkusov CSLU sistema na zbirki ŠTEVKE.

[1] *"#"označuje komentar – vrstica s komentarjem se ne upošteva.*

D. SESTAVA TESTNIH MNOŽIC V REFERENČNEM OKOLJU

V tabelah je prikazana podrobnejša sestava skupin testnih množic, uporabljenih v praktičnih preizkusih. Vsaka testna množica zaseda eno vrstico v ustrezni tabeli. V vseh tabelah je prikazanih 110 testnih množic. Opis uporabljenih motenj v posamezni množici se nahaja v 1. in 2. stolpcu vseh tabel. V 3. stolpcu je oznaka testne množice, v ostalih pa še uporabljene oznake v prikazih uspešnosti referenčnih SRG v disertaciji. Če oznaka zajema več množic, je ustrezen podatek v prikazu enak povprečju uspešnosti sistema na množicah v skupini. Oznaka "povprečje", ki se običajno pojavi v zadnjih stolpcih tabel za prikaz uspešnosti sistemov, zaradi preglednosti ni prikazana; izraža pa povprečje rezultatov v vseh uporabljenih testnih množicah.

D.1 Aditivne in konvolutivne motnje posamezno

Aditivne motnje	*Konvolutivne motnje*	*Oznaka množice*	*Oznake v tabelah*
		testna množica	***čist***
govor 12.0dB		*govor 12.0dB*	
volvo 12.0dB		*volvo 12.0dB*	
tovarna 12.0dB		*tovarna 12.0dB*	
f-16 12.0dB		*f-16 12.0dB*	***aditivne 12dB***
beli šum 12.0dB		*beli šum 12.0dB*	
roza šum 12.0dB		*roza šum 12.0dB*	
pas 900Hz 12.0dB		*pas 900Hz 12.0dB*	
govor 6.0dB		*govor 6.0dB*	
volvo 6.0dB		*volvo 6.0dB*	
tovarna 6.0dB		*tovarna 6.0dB*	
f-16 6.0dB		*f-16 6.0dB*	***aditivne 6dB***
beli šum 6.0dB		*beli šum 6.0dB*	
roza šum 6.0dB		*roza šum 6.0dB*	
pas 900Hz 6.0dB		*pas 900Hz 6.0dB*	
govor 0.0dB		*govor 0.0dB*	
volvo 0.0dB		*volvo 0.0dB*	
tovarna 0.0dB		*tovarna 0.0dB*	
f-16 0.0dB		*f-16 0.0dB*	***aditivne 0dB***
beli šum 0.0dB		*beli šum 0.0dB*	
roza šum 0.0dB		*roza šum 0.0dB*	
pas 900Hz 0.0dB		*pas 900Hz 0.0dB*	
	llp	*llp*	
	lp	*lp*	***konvolutivne***
	hp	*hp*	
	odjek	*odjek*	

Tabela D.1: Specifikacija skupin testnih množic posameznih aditivnih in konvolutivnih motenj.

D.2 Aditivne in konvolutivne motnje skupaj – SNR=12dB

Aditivne motnje	*Konvolutivne motnje*	*Oznaka množice*	*Oznake v tabelah*	*Oznake v tabelah*
govor 12.0dB	*llp*	*govor 12.0dB+llp*	***aditivne 12dB+llp***	***aditivne 12dB +konvolu-tivne***
volvo 12.0dB	*llp*	*volvo 12.0dB+llp*		
tovarna 12.0dB	*llp*	*tovarna 12.0dB+llp*		
f-16 12.0dB	*llp*	*f-16 12.0dB+llp*		
beli šum 12.0dB	*llp*	*beli šum 12.0dB+llp*		
roza šum 12.0dB	*llp*	*roza šum 12.0dB+llp*		
pas 900Hz 12.0dB	*llp*	*pas 900Hz 12.0dB+llp*		
govor 12.0dB	*lp*	*govor 12.0dB+lp*	***aditivne 12dB+lp***	
volvo 12.0dB	*lp*	*volvo 12.0dB+lp*		
tovarna 12.0dB	*lp*	*tovarna 12.0dB+lp*		
f-16 12.0dB	*lp*	*f-16 12.0dB+lp*		
beli šum 12.0dB	*lp*	*beli šum 12.0dB+lp*		
roza šum 12.0dB	*lp*	*roza šum 12.0dB+lp*		
pas 900Hz 12.0dB	*lp*	*pas 900Hz 12.0dB+lp*		
govor 12.0dB	*hp*	*govor 12.0dB+hp*	***aditivne 12dB+hp***	
volvo 12.0dB	*hp*	*volvo 12.0dB+hp*		
tovarna 12.0dB	*hp*	*tovarna 12.0dB+hp*		
f-16 12.0dB	*hp*	*f-16 12.0dB+hp*		
beli šum 12.0dB	*hp*	*beli šum 12.0dB+hp*		
roza šum 12.0dB	*hp*	*roza šum 12.0dB+hp*		
pas 900Hz 12.0dB	*hp*	*pas 900Hz 12.0dB+hp*		
govor 12.0dB	*odjek*	*govor 12.0dB+odjek*	***aditivne 12dB +odjek***	
volvo 12.0dB	*odjek*	*volvo 12.0dB+odjek*		
tovarna 12.0dB	*odjek*	*tovarna 12.0dB+odjek*		
f-16 12.0dB	*odjek*	*f-16 12.0dB+odjek*		
beli šum 12.0dB	*odjek*	*beli šum 12.0dB+odjek*		
roza šum 12.0dB	*odjek*	*roza šum 12.0dB+odjek*		
pas 900Hz 12.0dB	*odjek*	*pas 900Hz 12.0dB+odjek*		

Tabela D.2: Specifikacija skupin testnih množic kombinacije aditivnih in konvolutivnih motenj pri razmerju SNR 12dB.

D.3 Aditivne in konvolutivne motnje skupaj - SNR=6dB

Aditivne motnje	*Konvolutivne motnje*	*Oznaka množice*	*Oznake v tabelah*	*Oznake v tabelah*
govor 6.0dB	*llp*	*govor 6.0dB+llp*	***aditivne 6dB+llp***	***aditivne 6dB +konvolu-tivne***
volvo 6.0dB	*llp*	*volvo 6.0dB+llp*		
tovarna 6.0dB	*llp*	*tovarna 6.0dB+llp*		
f-16 6.0dB	*llp*	*f-16 6.0dB+llp*		
beli šum 6.0dB	*llp*	*beli šum 6.0dB+llp*		
roza šum 6.0dB	*llp*	*roza šum 6.0dB+llp*		
pas 900Hz 6.0dB	*llp*	*pas 900Hz 6.0dB+llp*		
govor 6.0dB	*lp*	*govor 6.0dB+lp*	***aditivne 6dB+lp***	
volvo 6.0dB	*lp*	*volvo 6.0dB+lp*		
tovarna 6.0dB	*lp*	*tovarna 6.0dB+lp*		
f-16 6.0dB	*lp*	*f-16 6.0dB+lp*		
beli šum 6.0dB	*lp*	*beli šum 6.0dB+lp*		
roza šum 6.0dB	*lp*	*roza šum 6.0dB+lp*		
pas 900Hz 6.0dB	*lp*	*pas 900Hz 6.0dB+lp*		
govor 6.0dB	*hp*	*govor 6.0dB+hp*	***aditivne 6dB+hp***	
volvo 6.0dB	*hp*	*volvo 6.0dB+hp*		
tovarna 6.0dB	*hp*	*tovarna 6.0dB+hp*		
f-16 6.0dB	*hp*	*f-16 6.0dB+hp*		
beli šum 6.0dB	*hp*	*beli šum 6.0dB+hp*		
roza šum 6.0dB	*hp*	*roza šum 6.0dB+hp*		
pas 900Hz 6.0dB	*hp*	*pas 900Hz 6.0dB+hp*		
govor 6.0dB	*odjek*	*govor 6.0dB+odjek*	***aditivne 6dB +odjek***	
volvo 6.0dB	*odjek*	*volvo 6.0dB+odjek*		
tovarna 6.0dB	*odjek*	*tovarna 6.0dB+odjek*		
f-16 6.0dB	*odjek*	*f-16 6.0dB+odjek*		
beli šum 6.0dB	*odjek*	*beli šum 6.0dB+odjek*		
roza šum 6.0dB	*odjek*	*roza šum 6.0dB+odjek*		
pas 900Hz 6.0dB	*odjek*	*pas 900Hz 6.0dB+odjek*		

Tabela D.3: Specifikacija skupin testnih množic kombinacije aditivnih in konvolutivnih motenj pri razmerju SNR 6dB.

D.4 Aditivne in konvolutivne motnje skupaj - SNR=0dB

Aditivne motnje	*Konvolutivne motnje*	*Oznaka množice*	*Oznake v tabelah*	*Oznake v tabelah*
govor 0.0dB	*llp*	*govor 0.0dB+llp*	***aditivne 0dB+llp***	***aditivne 0dB +konvolu-tivne***
volvo 0.0dB	*llp*	*volvo 0.0dB+llp*		
tovarna 0.0dB	*llp*	*tovarna 0.0dB+llp*		
f-16 0.0dB	*llp*	*f-16 0.0dB+llp*		
beli šum 0.0dB	*llp*	*beli šum 0.0dB+llp*		
roza šum 0.0dB	*llp*	*roza šum 0.0dB+llp*		
pas 900Hz 0.0dB	*llp*	*pas 900Hz 0.0dB+llp*		
govor 0.0dB	*lp*	*govor 0.0dB+lp*	***aditivne 0dB+lp***	
volvo 0.0dB	*lp*	*volvo 0.0dB+lp*		
tovarna 0.0dB	*lp*	*tovarna 0.0dB+lp*		
f-16 0.0dB	*lp*	*f-16 0.0dB+lp*		
beli šum 0.0dB	*lp*	*beli šum 0.0dB+lp*		
roza šum 0.0dB	*lp*	*roza šum 0.0dB+lp*		
pas 900Hz 0.0dB	*lp*	*pas 900Hz 0.0dB+lp*		
govor 0.0dB	*hp*	*govor 0.0dB+hp*	***aditivne 0dB+hp***	
volvo 0.0dB	*hp*	*volvo 0.0dB+hp*		
tovarna 0.0dB	*hp*	*tovarna 0.0dB+hp*		
f-16 0.0dB	*hp*	*f-16 0.0dB+hp*		
beli šum 0.0dB	*hp*	*beli šum 0.0dB+hp*		
roza šum 0.0dB	*hp*	*roza šum 0.0dB+hp*		
pas 900Hz 0.0dB	*hp*	*pas 900Hz 0.0dB+hp*		
govor 0.0dB	*odjek*	*govor 0.0dB+odjek*	***aditivne 0dB +odjek***	
volvo 0.0dB	*odjek*	*volvo 0.0dB+odjek*		
tovarna 0.0dB	*odjek*	*tovarna 0.0dB+odjek*		
f-16 0.0dB	*odjek*	*f-16 0.0dB+odjek*		
beli šum 0.0dB	*odjek*	*beli šum 0.0dB+odjek*		
roza šum 0.0dB	*odjek*	*roza šum 0.0dB+odjek*		
pas 900Hz 0.0dB	*odjek*	*pas 900Hz 0.0dB+odjek*		

Tabela D.4: Specifikacija skupin testnih množic kombinacije aditivnih in konvolutivnih motenj pri razmerju SNR 0dB.

E. LITERATURA

E.1 Parametrizacija govornega signala

[1] C. Pan, "Gibbs phenomenon removal and digital filtering directly through the fast Fourier transform," IEEE Trans. Signal Processing, vol. 49, pp. 444 - 448, February 2001.

[2] C.Avendano, "Temporal Processing of Speech in a Time-Feature Space," Ph.D. dissertation, Oregon Graduate Institute of Science & Technology, April 1997.

[3] C.Avendano, S.Tibrewala and H.Hermansky, "Multiresolution Channel Normalization for ASR in Reverberant Environments," Proc. EUROSPEECH '97, Rhodes, Greece, 1997.

[4] C.R.Jankowski Jr., H.D.H.Vo and R.P.Lipmann, "A Comparison of Signal Processing Front Ends for Automatic Speech Recognition," IEEE Trans. on Speech & Audio Processing, vol. 3, no. 4, pp. 286-293, July 1995.

[5] D. J. Thomson, "Spectrum estimation and harmonic analysis," Proc. IEEE, vol. 70, no. 9, pp. 1055-1092, Sept. 1982.

[6] Dan Chazan, Ron Hoory, Gilad Cohen, Meir Zibulski; Speech reconstruction from mel frequency cepstral coefficients and pitch frequency, Proc. ICASSP'00, pp. 1299 - 1302, June 2000.

[7] Dinei A. F. Florêncio, Ronald W. Schafer; Perfect reconstructing nonlinear filter banks, Proc. ICASSP'96, pp. 1815 - 1818, May 1996.

[8] F. Harris,"On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", Proceeding of the IEEE, Vol. 66, No. 1, January 1978, pp. 51-83.

[9] G. Jones and B. Boashash, "Window matching in the time-frequency plane and the adaptive spectrogram," IEEE-SP Int. Sym. Time-Fre. Time-Scale Anal., pp. 87 - 90, October 1992.

[10] H.G.Hirsch, "Estimation of noise spectrum and its application to SNR estimation and speech enhancement," Tech. Report TR-93-012, International Computer Science Institute Berkeley, 1993.

[11] H.Hermansky and N.Morgan, "RASTA Processing of Speech," IEEE Trans. on Speech & Audio Processing, vol. 2, no. 4, pp. 578-589, October 1994.

[12] H.Hermansky, "Speech Beyond 10 miliseconds (Temporal Filtering in Feature Domain)," Invited Keynote Lecture, Proc. of the International Workshop on Human Interface Technology 1994, Aizu, Japan, September 1994.

[13] Kambiz Nayebi, Thomas P. Barnwell III, Mark J. T. Smith; Low delay FIR filter banks: Design and evaluation, IEEE Trans. Signal Processing, vol. 42, pp. 24 - 31, January 1994.

[14] Kambiz Nayebi, Thomas P. Barnwell III, Mark J. T. Smith; On the design of FIR analysis-synthesis filter banks with high computational efficiency, IEEE Trans. Signal Processing, vol. 42, pp. 825 - 834, April 1994.

[15] L.Rabiner and B.H.Juang, "Fundamentals of Speech Recognition," Prentice Hall, Englewood Cliffs, New Jersey.

[16] M. L. Kramer and D. L. Jones, "Improved time-frequency filtering using an STFT analysis-modification-synthesis method," IEEE-SP Int. Sym. Time-Fre. Time-Scale Anal., pp. 264 - 267, October 1994.

[17] M. S. Puckette and J. C. Brown, "Accuracy of frequency estimates using the phase vocoder," IEEE Trans. Speech Audio Processing, vol. 6, pp. 166 - 176, March 1998.

[18] M. Y. Hong, "A sinusoidal signal analysis technique for fast, accurate, and discriminating frequency determination," Proc. ICASSP'00, pp. 249 - 252, June 2000.

[19] M.Cooke, P.Green, C.Anderson and D.Abberley, "Recognition of Occluded Speech by Hidden Markov Models," Tech. Report TR-94-05-01, University of Sheffield, Department of Computer Science, 1994.

[20] Mark J. T. Smith, Thomas P. Barnwell III; A new filter bank theory for time-frequency representation, IEEE Trans. Acoust., Speech, Signal Processing, vol. 35, pp. 314 - 327, March 1987.

[21] N.Kanedera, H.Hermansky and T.Arai, "Desired characteristics of modulation spectrum for robust automatic speech recognition," Proc. ICASSP 1998, Seattle, USA, 1998.

[22] N.Kanedera, T.Arai, H.Hermansky and M.Pavel, "On the Importance of Various Modulation Frequencies for Speech recognition," Proc. EUROSPEECH '97, Rhodes, Greece, 1997.

[23] R. N. Czerwinski and D. L. Jones, "Adaptive short-time Fourier analysis," IEEE Signal Processing Lett., vol. 4, pp. 42 - 45, February 1997.

[24] S. H. Nawab, T. F. Quatieri, and J. S. Lim, "Signal reconstruction from short-time Fourier transform magnitude," IEEE Trans. Acoust., Speech, Signal Processing, vol. 31, pp. 986 - 998, August 1983.

[25] S. Nordebo, S. Nordholm, B. Bengtsson, I. Claesson; Noise reduction using an adaptive microphone array in a car - a speech recognition evaluation, Proc. 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 13 - 15, October 1993.

[26] T. P. Bronez and D. S. Brown, "Alternate windows for multi-window spectral analysis," Proc. ICASSP'92, pp. 429 - 432, March 1992.

[27] T.Arai, M.Pavel, H.Hermansky and C.Avendano, "Intelligibility of Speech with Filtered Time Trajectories of Spectral Envelopes," Proc. ICSLP, p.p. 2490-2493, Philadelphia, October 1996.

[28] W. Chen and N. C. Griswold, "An efficient recursive time-varying Fourier transform by using a half-sine wave window," IEEE-SP Int. Sym. Time-Fre. Time-Scale Anal., pp. 284 - 286, October 1994.

E.2 Kodiranje govornega signala

[29] Alan McCree, Kwan Truong, E. Bryan George, Thomas P. Barnwell, Vishu Viswanathan; An enhanced 2.4 kbit/s MELP coder, IEEE Speech Coding Workshop, pp. 101 - 102, September 1995.

[30] Alan V. McCree, Thomas P. Barnwell III; A mixed excitation LPC vocoder model for low bit rate speech coding, IEEE Trans. Speech Audio Processing, vol. 3, pp. 242 - 250, July 1995.

[31] International Telecommunications Union. Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). ITU-T Recommendation G.729, March 1996.

[32] Kambiz Nayebi, Thomas P. Barnwell III, Mark J. T. Smith; Low delay coding of speech and audio using nonuniform band filter banks, IEEE Speech Coding Workshop, pp. 7 - 8, September 1991.

[33] Kazuhito Koishida, Gou Hirabayashi, Keiichi Tokuda, Takao Kobayashi; A wideband CELP speech coder at 16 kbit/s based on mel-generalized cepstral analysis, Proc. ICASSP'98, pp. 161 - 164, May 1998.

[34] Kazuhito Koishida, Keiichi Tokuda, Takao Kobayashi, Satoshi Imai; CELP coding based on mel-cepstral analysis, Proc. ICASSP'95, pp. 33 - 36, May 1995.

[35] Kazuhito Koishida, Vladimir Cuperman, Allen Gersho; A 16-kbit/s bandwidth scalable audio coder based on the G.729 standard, Proc. ICASSP'00, pp. 1149 - 1152, June 2000.

[36] L. Besacier , C. Bergamini, D. Vaufreydaz, E. Castelli "The effect of speech and audio compression on speech recognition performance " IEEE Multimedia Signal Processing Workshop, Cannes, France, October 2001.

[37] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, F. Pellandini, "GSM Speech Coding and Speaker Recognition", Proc. of ICASSP'00, Istanbul.

[38] Richard C. Rose, Thomas P. Barnwell III; Design and performance of an analysis-by-synthesis class of predictive speech coders, IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, pp. 1489 - 1503, September 1990.

[39] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge and F. Pellandini, "Influence Of GSM Speech Coding On The Performance Of Text-Independent Speaker Recognition", URL = "citeseer.nj.nec.com/grassi00influence.html".

[40] S.Rao and W.A.Pearlman, "Analysis of Linear Prediction, Coding, and Spectral Estimation from Subbands," IEEE Trans. on Information Theory, vol. 42, no. 4, pp. 1160-1178, July 1996.

[41] Takahiro Unno, Thomas Barnwell III, Kwan Truong; An improved mixed excitation linear prediction (MELP) coder, Proc. ICASSP'99, pp. 245 - 248, March 1999.

[42] Thomas P. Barnwell III, Schuyler R. Quackenbush; An analysis of objectively computable measures for speech quality testing, Proc. ICASSP'82, pp. 996 - 999, May 1982.

[43] V. Iyengar and P. Kabal, "A low delay 16 kb/s speech coder," IEEE Trans. Signal Processing, vol. 39, pp. 1049 - 1057, May 1991.

[44] V. Iyengar and P. Kabal, "A low delay 16 kbits/sec speech coder," Proc. ICASSP'88, pp. 243 - 246, April 1988.

E.3 Sistemi za razpoznavanje govora

[45] A.Varga, H.J.M. Steeneken, M.J. Tomlinson and D.Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," CD-ROM available from the Speech Research Unit, DRA Malvern, UK, 1992.

[46] D.Kodek s sodelavci, "Razvoj in izdelava sistema za razpoznavanje izoliranih besed slovenskega govora," Končno poročilo, Fakulteta za elektrotehniko in računalništvo, Univerza v Ljubljani, Maj 1994.

[47] L.R.Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. of the IEEE, vol. 77, no. 2, pp. 257-286, February 1989.

[48] P.J.Moreno and R.M.Stern, "Sources of degradation of speech recognition in the telephone network," Proc. IEEE Intl. Conf. On Acoustics, Speech & Signal Processing, Adelaide, Australia, Vol. I, pp. 109-112, April, 1994.

[49] T.Majnik, "Analiza baze podatkov telefonskega govora," Diplomska naloga, Fakulteta za elektrotehniko in računalništvo, Univerza v Ljubljani, 1993.

[50] V.Valtchev, "Discriminative Methods in HMM-based Speech Recognition," Ph.D. dissertation, University of Cambridge, pp.24, March 1995.

[51] Z.Kačič, "Komunikacija človek – stroj," Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru, Maribor, 1995.

[52] James L. Hieronymus, "Ascii phonetic symbols for the world's languages: Worldbet," AT&T Bell Laboratories Technical Memo, 1994.

[53] P. Vermeulen, E. Barnard, Y. Yan, M. Fanty and R.A. Cole. A Comparison of HMM and neural network approaches to real world telephone speech applications. International Conference on Neural Networks and Signal Processing, Nanjing, P.R. China, December, 1995.

[54] S. H. Maes, G. Cohen, R. Hoory, and D. Chazan, "Conversational Networking: Conversational Protocols for Transport, Coding and Control", in proc. 6th Int. Conf. Spoken Language Processing, Beijing China, Oct. 2000 (ICSLP-2000).

[55] Hampshire, J. and Pearlmutter, B., Equivalence Proofs for Multi-Layer Perceptron Classifiers and the Bayesian Discriminant Function, Proc. of the 1990 Connectionist Summer School, Morgan Kaufman Publishers.

[56] Y. Yan, M. Fanty and R. Cole. Speech Recognition Using Neural Networks with Forward-backward Probability Generated Targets. IEEE ICASSP 1997.

[57] ŠTRANCAR, Andrej. Analiza vpliva aktivacijskih funkcij, akustičnih značilk in dolžine oken na uspešnost avtomatskega razpoznavanja govora : magistrsko delo. Ljubljana: [A. Štrancar], 2001. II, 77 f., ilustr. [COBISS.SI-ID 2273108]

E.4 Analiza lastnosti človekovega slušnega zaznavanja

[58] B.Strope and A.Alwan, "A Model of Dynamic Auditory Perception and its Application to Robust Word Recognition," IEEE Trans. on Speech & Audio Processing, vol. 5, no. 5, pp. 451-464, April 1997.

[59] Bojana Gajic, Kuldip K. Paliwal; Robust speech recognition using features based on zero crossings with peak amplitudes, Proc. ICASSP'03, pp. 64 - 67, May 2003.

[60] C.Darwin, "PERCEPTION: Hearing Lecture Notes," School of Biological Sciences, University of Sussex, Brighton, UK, 1994.

[61] Doh-Suk Kim, Jae-Hoon Jeong, Jae-Weon Kim, Soo-Young Lee; Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments, Proc. ICASSP'96, pp. 61 - 64, May 1996.

[62] Doh-Suk Kim, Moo Young Kim; On the perceptual weighting function for phase quantization of speech, IEEE Speech Coding Workshop, pp. 62 - 64, September 2000.

[63] Doh-Suk Kim, Soo-Young Lee, Rhee M. Kil; Auditory processing of speech signals for robust speech recognition in real-world noisy environments, IEEE Trans. Speech Audio Processing, vol. 7, pp. 55 - 69, January 1999.

[64] Doh-Suk Kim; On the perceptually irrelevant phase information in sinusoidal representation of speech, IEEE Trans. Speech Audio Processing, vol. 9, pp. 900 - 905, November 2001.

[65] Doh-Suk Kim; Perceptual phase redundancy in speech, Proc. ICASSP'00, pp. 1383 - 1386, June 2000.

[66] Ellis, D.P.W. "A perceptual representation of audio," MS thesis, EECS dept., MIT, February 1992.

[67] H.Bourlard, H.Hermansky and Nelson Morgan, "Towards increasing speech recognition error rates," Speech Communication 18, pp. 205-231, 1996.

[68] H.Fletcher, "Speech and Hearing in Communication," New York: Krieger, 1953.

[69] H.Hermansky, "Auditory Modeling in Automatic Recognition of Speech," Proc. Keele Workshop, Keele, Sweden, 1996.

[70] H.Hermansky, "Should Recognizers Have Ears?," Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 1-10, France, 1997.

[71] J.B.Allen, "How Do Humans Process and Recognize Speech ?," IEEE Trans. on Speech & Audio Processing, vol. 2, no. 4, pp. 567-577, October 1994.

[72] O.Ghitza, "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 1, pp. 115-132, January 1994.

[73] Oded Ghitza; Speech analysis/Synthesis based on matching the synthesized and the original representations in the auditory nerve level, Proc. ICASSP'86, pp. 1995 - 1998, April 1986.

[74] R.Baker, "Introduction to hearing for speech," Department of Phonetics and Lingustic, University College London, UK.

[75] Sumeet Sandhu, Oded Ghitza; A comparative study of mel cepstra and EIH for phone classification under adverse conditions, Proc. ICASSP'95, pp. 409 - 412, May 1995.

[76] W.Jesteadt, S.Bacon and J.Lehman, "Forward masking as a function of frequency, masker level, and signal delay," J. Acoust. Soc. Am., vol. 71, pp. 950-962, April 1982.

[77] Robert Rozman, Uporaba spoznanj o človekovi slušni percepciji v sistemu za razpoznavanje govora : magistrsko delo, Ljubljana, [R. Rozman], 1999. [COBISS.SI-ID 1541716]

E.5 Nesimetrična okna v SRG

[78] Robert Rozman, Dušan Kodek, "Uporaba nesimetričnih oken v sistemu za razpoznavanje govora", V: Zbornik osme Elektrotehniške in računalniške konference ERK '99, 23. - 25. september 1999, Portorož, Slovenija, IEEE Region 8, 1999, zv. B, str. 213-216. [COBISS.SI-ID 1627220]

[79] Robert Rozman, Andrej Štrancar, Dušan Kodek, "Analiza vpliva oken na robustnost sistemov za razpoznavanje govora", V: Zbornik devete Elektrotehniške in računalniške konference ERK 2000, 21. - 23. september 2000, Portorož, Slovenija, Ljubljana, IEEE Region 8, 2000, zv. B, str. 177-180. [COBISS.SI-ID 2024020]

[80] Robert Rozman, Andrej Štrancar, Dušan Kodek, "Povečevanje robustnosti sistemov za razpoznavanje govora in optimizacija procesa parametrizacije", V: Zbornik desete Elektrotehniške in računalniške konference ERK 2001, 24. - 26. september 2001, Portorož, Slovenija, Ljubljana, IEEE Region 8, 2001, zv. B, str. 257-260. [COBISS.SI-ID 2414932]

[81] Robert Rozman, Dušan Kodek, "Improving speech recognition robustness using non-standard windows", V: The IEEE Region 8 EUROCON 2003 : computer as a tool : 22-24. September 2003, Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia : proceedings, IEEE, cop. 2003, vol. 2, str. 171-174. [COBISS.SI-ID 3863124]

[82] Robert Rozman, Andrej Štrancar, "Nesimetrična okna v sistemih za procesiranje govora", V: Zbornik trinajste mednarodne Elektrotehniške in računalniške konference ERK 2004, 27. - 29. september 2004, Portorož, Slovenija, (Zbornik ... Elektrotehniške in računalniške konference ERK ..., 1581-4572), Ljubljana, IEEE Region 8, 2004, zv. B, str. 163-166. [COBISS.SI-ID 4420692]

E.6 Načrtovanje oken

[83] Albert H. Nuttall; Some windows with very good sidelobe behavior, IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 84 - 91, February 1981.

[84] B. Fette, R. Gibson, and E. Greenwood, "Windowing functions for the average magnitude difference function pitch extractor," Proc. ICASSP'80, pp. 49 - 52, April 1980.

[85] Dinei A. F. Florêncio; Investigating the use of asymmetric windows in CELP vocoders, Proc. ICASSP'93, pp. 427 - 430, April 1993.

[86] Dinei A. F. Florêncio; On the use of asymmetric windows for reducing the time delay in real-time spectral analysis, Proc. ICASSP'91, pp. 3261 - 3264, May 1991.

[87] Fung I. Tseng, Tapan K. Sarkar, Donald D. Weiner; A novel window for harmonic analysis, IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 177 - 188, April 1981.

[88] G. Thomas, B. C. Flores, and J. Sok-Son, "SAR Sidelobe apodization using the Kaiser window," Proc. ICIP'00, pp. 709 - 712, September 2000.

[89] H. Albrecht, "A family of cosine-sum windows for high-resolution measurements," Proc. ICASSP'01, pp. 3081 - 3084, May 2001.

[90] Ivan W. Selesnick, C. Sidney Burrus; Nonlinear-phase maximally-flat lowpass FIR filter design, The 7th IEEE Digital Signal Processing Workshop, pp. 374 - 377, September 1996.

[91] J. Le Roux, "Optimal design of windows for spectral analysis of mono and bidimensional sampled signals," Proc. ICASSP'92, pp. 501 - 504, March 1992.

[92] John W. Adams, James L. Sullivan; Peak-constrained least-squares optimization, IEEE Trans. Signal Processing, vol. 46, pp. 306 - 321, February 1998.

[93] John W. Adams; A new optimal window, IEEE Trans. Signal Processing, vol. 39, pp. 1753 - 1769, August 1991.

[94] K. M. M. Prabhu and H. Renganathan, "Optimum binary windows for discrete Fourier transforms," IEEE Trans. Acoust., Speech, Signal Processing, vol. 34, pp. 216 - 220, February 1986.

[95] K. M. M. Prabhu and J. P. Agrawal, "Selection of data windows for digital signal processing," Proc. ICASSP'78, pp. 79 - 82, April 1978.

[96] Magdy T. Hanna; Windows with rapidly decaying sidelobes and steerable sidelobe dips, IEEE Trans. Signal Processing, vol. 42, pp. 2037 - 2044, August 1994.

[97] N. C. Geçkinli and D. Yavuz, "Some novel windows and a concise tutorial comparison of window families," IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, pp. 501 - 507, December 1978.

[98] S. Kadambe, "On the window selection and the cross terms that exist in the magnitude squared distribution of the short time Fourier transform," IEEE SSAP Workshop, pp. 22 - 25, October 1992.

[99] S. Kay and D. Smith, "An optimal sidelobeless window," IEEE Trans. Signal Processing, vol. 47, pp. 2542 - 2546, September 1999.

[100] S. Yang and Y. Ke, "On the three-coefficient window family," IEEE Trans. Signal Processing, vol. 40, pp. 3085 - 3088, December 1992.

[101] T. C. Speake and R. M. Mersereau, "A comparison of different window formulations for two-dimensional FIR filter design," Proc. ICASSP'79, pp. 5 - 8, April 1979.

[102] Thomas P. Barnwell III; Recursive windowing for generating autocorrelation coefficients for LPC analysis, IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 1062 - 1066, October 1981.

[103] Thomas P. Barnwell III; Windowless Techniques for LPC Analysis, IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, pp. 421 - 427, August 1980.

[104] Thomas P. Barnwell; Recursive autocorrelation computation for LPC analysis, Proc. ICASSP'77, pp. 1 - 4, May 1977.

[105] W. Kim and M. H. Hayes, "Phase retrieval using a window function," Proc. ICASSP'92, pp. 149 - 152, March 1992.

E.7 Načrtovanje KEO filtrov

[106] A. Dembo and D. Malah, "Generalization of the window method for FIR digital filter design," IEEE Trans. Acoust., Speech, Signal Processing, vol. 32, pp. 1081 - 1083, October 1984.

[107] A. Groth, H.G. Göckler. Design of Equiripple Minimum Phase FIR-Filters. Proceedings of the Eusipco '98, Rhodos, Greece, 08.-11.09.98, Volume III, pp. 1897-1900.

[108] D.Burnside and T.W.Parks, "Optimal Design of FIR Filters with the Complex Chebyshev Error Criteria," IEEE Trans. on Signal Processing, vol. 43, no. 3, pp. 605-616, March 1995.

[109] Fredric J. Harris; On maximally flat low-pass filters with ripple in the stopband, IEEE Transactions on Audio Electroacoust., vol. 20, pp. 345 - 352, December 1972.

[110] H. H. Dam, S. Nordebo, K. L. Teo, A. Cantoni; Design of linear phase FIR filters with recursive structure and discrete coefficients, Proc. ICASSP'98, pp. 1269 - 1272, May 1998.

[111] K. Preuss, “On the Design of FIR filters by Complex Chebyshev Approximation,” IEEE Trans. on Acoustics, Speech & Signal Processing, vol. 37, pp. 702-712, May 1989.

[112] K.Stieglitz, "Optimal Design of FIR Digital Filters with Monotone Passband Response," IEEE Trans. on Acoustics, Speech & Signal Processing, vol. ASSP-27, no. 6, pp. 643-649, December 1979.

[113] Niranjan Damera-Venkata, Brian L. Evans, Shawn R. McCaslin; Design of optimal minimum-phase digital FIR filters using discrete Hilbert transforms, IEEE Trans. Signal Processing, vol. 48, pp. 1491 - 1495, May 2000.

[114] Niranjan Damera-Venkata, Brian L. Evans; Optimal design of real and complex minimum phase digital FIR filters, Proc. ICASSP'99, pp. 1145 - 1148, March 1999.

[115] R.Rozman in D.Kodek, "Načrtovanje optimalnih KEO filtrov po kompleksnem Čebiševem kriteriju napake," Zbornik šeste Elektrotehniške in računalniške konference ERK '97, zv. A, str. 175-178, 25. - 27. september 1997, Portorož, Slovenija.

[116] Shao-Po Wu, William Putnam; Minimum perceptual spectral distance FIR filter design, Proc. ICASSP'97, pp. 447 - 450, April 1997.

[117] Thomas W. Parks, James H. McClellan; A program for the design of linear phase finite impulse response digital filters, IEEE Transactions on Audio Electroacoust., vol. 20, pp. 195 - 199, August 1972.

[118] X. Chen and T. W. Parks, “Design of FIR filters in the Complex Domain,” IEEE Trans. on Signal Processing, vol. ASSP-35, no. 2, pp. 144-153, February 1987.

[119] Zhuquan Zang, Sven Nordholm, Sven Nordebo, Antonio Cantoni; Design of digital filters with amplitude and group delay specifications, IEEE SSAP Workshop, pp. 357 - 360, August 2001.

[120] Robert Rozman, Dušan Kodek, "Posplošeni Remezov algoritem za načrtovanje optimalnih KEO filtrov po kompleksnem Čebiševem kriteriju napake", V: Zbornik sedme Elektrotehniške in računalniške konference ERK '98, 24. - 26. september 1998, Portorož, Slovenija, Baldomir Zajc, ur., Ljubljana, IEEE Region 8, Slovenska sekcija IEEE, 1998, zv. A, str. 245-248. [COBISS.SI-ID 1317972]

E.8 Uporaba optimizacijskih metod za načrtovanje oken

[121] Sven Nordebo, Ingvar Claesson, Zhuquan Zang; Optimum window design by semi-infinite quadratic programming, IEEE Signal Processing Lett., vol. 6, pp. 262 - 265, October 1999.

[122] Sven Nordebo, Ingvar Claesson, Sven Nordholm; Weighted Chebyshev approximation for the design of broadband beamformers using quadratic programming, IEEE Signal Processing Lett., vol. 1, pp. 103 - 105, July 1994.

[123] D.Yuret, "From Genetic Algorithms To Efficient Optimization," M.Sc. thesis, MIT, May 1994.

[124] A.Kuntsevich and F.Kappel, " SOLVOPT – The Solver for Local Nonlinear Optimization Problems," Manual, Institute for Mathematics, Karl Fransenz University of Graz, June 1997.

[125] N.Z. Shor, "Minimization Methods for Non-Differentiable Functions," Springer Series in Computational Mathematics, Vol. 3, Springer-Verlag, Berlin 1985.

ZAHVALA

Na tem mestu bi se predvsem želel zahvaliti tistim, ki so mi na dosedanji poti pomagali, me bodrili in predvsem vztrajali tudi takrat, ko vse ni bilo tako lepo in enostavno.

Po strokovni in človeški plati se iskreno zahvaljujem prof. dr. Dušanu Kodeku, ker mi je pomagal že od samega začetka – takrat, ko se je moja pot prekrižala z načrti mojega štipenditorja. Na srečo so se stvari uredile tako, kot so se, in zato sem mu še danes za minulo ter tudi trenutno podporo posebej hvaležen. Predvsem pa se zahvaljujem za razumevanje v tistih trenutkih, ko zaradi drugih pomembnih življenjskih "korakov" – predvsem moje mlade družine - nisem bil sposoben slediti vsemu z želenim oziroma pričakovanim tempom.

V isti sapi se zahvaljujem vsem v Laboratoriju za arhitekturo in procesiranje signalov, še posebej gospodoma Igorju Škrabi in Zvonetu Petkovšku za prijetno atmosfero, nesebično pomoč in podporo, sodelavcema Andreju Štrancarju in Damjanu Šoncu pa za razbremenitev in podporo v najbolj kritičnih trenutkih nastajanja disertacije ter za izmenjavo izkušenj in znanja.

Vse, kar danes imam in znam, vsekakor dolgujem predvsem staršema, ki sta me na vsej poti podpirala in pokazala obilo razumevanja tudi takrat, ko sem se odločil za odhod v Ljubljano. Zelo težko bi lahko z besedami izrazil zahvalo za tisto, kar sta mi nudila. Iskreno upam, da je ta trenutek vsaj delno plačilo za vse, čemur sta se zaradi mene morala odpovedati. Verjetno se jima nikdar ne bom uspel povsem oddolžiti, vsekakor pa sem in jima bom večno hvaležen. Hvala tudi Bredi in Cvetani.

Nenazadnje pa gre največja zahvala moji mladi družini: Marjani, Timu in Nii. Neizmerna podpora, ljubezen in sreča, v kateri živimo, so mi nudile vse potrebno, da disertacijo, sicer malce pozno, a kljub temu, zaključim. Vsi so "aktivno" sodelovali pri mojem delu; Marjana kot lektorska svetovalka, Tim in Nia pa sta bila poleg vloge radovednih in živahnih otrok še izvor nekaterih idej o razvoju sposobnosti razpoznavanja govora pri človeku.

IZJAVA

Spodaj podpisani Robert Rozman izjavljam, da sem doktorsko disertacijo izdelal samostojno. Strokovna gradiva, ki sem jih pri tem uporabljal, sem v celoti navedel v literaturi, pomoč sodelavcev pa v zahvali. Moje delo je potekalo pod mentorstvom prof. dr. Dušana Kodeka.

mag. Robert Rozman, univ. dipl. inž.

IZVIRNI PRISPEVKI

Izvirni prispevki k znanosti so povzeti v naslednjih točkah:

- *izpeljava splošnega postopka za načrtovanje nesimetričnih oken*

  Izpeljan je bil postopek, s pomočjo katerega se problem načrtovanja nesimetričnih oken pretvori v obliko, ki jo je mogoče rešiti s pomočjo splošne gradientne optimizacijske metode. Njegova glavna prednost je enostavno določanje poljubnega optimizacijskega kriterija in drugih parametrov. Postopek je v disertaciji uporabljen za načrtovanje nesimetričnih oken po kriteriju Čebiševe napake amplitudnega odziva.

- *izpeljava alternativnega postopka za načrtovanje nesimetričnih oken na osnovi digitalnih filtrov z neskončnim enotinim odzivom*

  Glavna slabost načrtovanja nesimetričnih oken s splošnimi optimizacijskimi metodami je velika časovna kompleksnost. Zato je bil izpeljan hitrejši alternativni postopek, ki temelji na uporabi preprostih parametričnih modelov v obliki NEO filtrov. Dobljena okna imajo zaradi uporabe modelov sicer specifične, vnaprej določene amplitudne odzive, ki pa se izkažejo kot zelo primerni za razpoznavanje govora.

- *prikaz uspešnosti uporabe nesimetričnih oken v realnih sistemih za razpoznavanje govora*

  Praktični preizkusi načrtovanih nesimetričnih oken so bili izvedeni na dveh različno zasnovanih SRG ob uporabi dveh jezikovno različnih govornih zbirk. Merjena je bila uspešnost razpoznavanja sistemov na večji skupini testnih množic z motnjami, ki niso bile prisotne v postopku učenja (inherentna robustnost). Rezultati so potrdili povečanje robustnosti SRG ob uporabi nesimetričnih oken.

- *ocena pomembnosti posameznih lastnosti oken pri razpoznavanju govora*

  Lastnost z največjim vplivom na robustnost razpoznavanja je hitrost padanja višine stranskih valov. Vendar je ta lastnost povezana še z drugim pomembnim dejavnikom – efektivno širino okna v časovnem prostoru, ki prav tako vpliva na robustnost SRG. Zato večja hitrost padanja višine stranskih valov pomeni večjo robustnost le do točke, pri kateri prevlada vpliv drugega dejavnika in se robustnost prične zmanjševati.

- *izpeljava družine NEO oken*

  Predstavniki družine se generirajo s pomočjo NEO parametričnega modela (končni izsek njegovega enotinega odziva), ki je blizu lastnostim človekovega sluha. Model ima majhno število parametrov, ki omogočajo enostavno optimizacijo po različnih kriterijih. Okna imajo tudi značilen potek amplitudnega odziva z dobrim asimptotičnim padanjem višine stranskih valov. Z vrednostjo enega parametra (lega pola α) se lahko določa širina glavnega vala v časovnem in frekvenčnem prostoru ter višina stranskih valov.

- *izpeljava postopka nesimetrične modifikacije Poissonovega okna*

  V postopku modifikacije se simetrično Poissonovo okno v časovnem prostoru najprej pomakne in nato še pomnoži s Hannovim oknom. Dobljeno nesimetrično okno ima boljše asimptotično padanje višine stranskih valov in krajšo časovno zakasnitev.