UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

ANDREJ ŠTRANCAR

DINAMIČNO PRILAGAJANJE FREKVENČNO-ČASOVNE LOČLJIVOSTI PRI SPEKTRALNI ANALIZI GOVORNEGA SIGNALA

DOKTORSKA DISERTACIJA

MENTOR: PROF. DR. DUŠAN KODEK

LJUBLJANA, 2006

POVZETEK

Razpoznavanje govora je pravzaprav razpoznavanje zaporedja krajših govornih odsekov, ki ustrezajo govornim enotam. Govorne enote so po trajanju različne, mednje spadajo deli fonemov, fonemi, besede, stavki, itd. Govorni signal je pri razpoznavanju potrebno razrezati na kratke odseke. Te odseke se s pomočjo akustičnih modelov klasificira v krajše govorne enote, kot so deli fonemov ali fonemi. Akustični modeli so lahko parametrični ali neparametrični, kot so nevronske mreže. Daljše govorne enote so predstavljene z zaporedjem krajših, npr. besede z zaporedjem fonemov. Zaporedja kratkih govornih enot se preslika v daljše s pomočjo časovnih modelov. Najpogosteje se za časovno modeliranje uporabljajo prikriti Markovi modeli. V disertaciji je bil pri vseh poskusih uporabljen hibridni sistem za razpoznavanje govora (SRG), pri katerem je akustično modeliranje izvedeno z nevronsko mrežo, časovno modeliranje pa se zgleduje po prikritih Markovih modelih. Določen odsek govornega signala je torej s pomočjo akustičnih modelov potrebno klasificirati v kratko govorno enoto glede na njegov akustični pomen. Ker človekovo uho deluje kot spektralni analizator, je akustični pomen odseka določen s spektrom signala na tem odseku.

Pri veliki večini SRG je vsak kratek odsek signala predstavljen z množico značilk, katere izračunamo v postopku parametrizacije. Na področju razpoznavanja govora se kot značilke najpogosteje uporablja MFCC koeficiente (Mel-Frequency Cepstral Coefficients). MFCC koeficienti temeljijo na spektru signala, ki se ga izračuna s kratkočasovno diskretno Fourierovo transformacijo (KČDFT). Pri računanju KČDFT se govorni signal razreže na kratke odseke oziroma okvirje. Nato se z uporabo diskretne Fourierove transformacije izračuna spekter za vsak okvir posebej. Spekter govornega signala je torej zaporedje spektrov okvirjev. Dolžina okvirja je pri tem zelo pomembna. Z njo sta določeni frekvenčna in časovna ločljivost izračunanega spektra. Ker se z diskretno Fourierovo transformacijo izračuna povprečni spekter signala v okvirju, je s stališča časovne ločljivosti spektra potrebno uporabiti kratek okvir. Spremembe spektra signala, ki so krajše od dolžine okvirja, bodo v izračunanem spektru namreč zabrisane. Vendar pa uporaba kratkega okvirja pomeni slabšo frekvenčno ločljivost v izračunanem spektru. To pomeni, da gre pri dolžini okvirja vedno za kompromis med časovno in frekvenčno ločljivostjo. Zaradi nestacionarnosti govornega signala, je določitev primerne dolžine okvirja lahko težavna.

V postopku računanja KČDFT se pri spektralni analizi govornega signala običajno uporabljajo 15 - 35 ms dolgi okvirji. Tako izračunani spekter ima enakomerno frekvenčno-časovno ločljivost, kar se ne sklada z lastnostmi človekovega sluha. Ena izmed značilnosti sluha je namreč nelinearna frekvenčna ločljivost, ki je pri nizkih frekvencah dobra, z višanjem frekvence pa se poslabšuje. Nelinearno frekvenčno ločljivost je mogoče z uporabo večločljivostne spektralne analize posnemati že pri spektralni analizi. V disertaciji je bila preizkušena zvezna valčna transformacija, pri kateri se frekvenčna ločljivost, podobno kot pri sluhu, s frekvenco slabša. To po drugi strani pomeni, da je časovna ločljivost pri visokih frekvencah boljša kot pri nizkih. Zato je spremembe spektra, ki so prisotne (tudi) pri visokih frekvencah, mogoče časovno natančneje opredeliti, pri tem pa je frekvenčna ločljivost pri nizkih frekvencah kljub temu dobra. Primerjava rezultatov, ki so bili doseženi z MFCC koeficienti, pri katerih je bil spekter izračunan z uporabo zvezne valčne transformacije z rezultati, doseženimi z MFCC koeficienti na osnovi KČDFT, ni pokazala prednosti valčne transformacije pred KČDFT. Uspešnost razpoznavanja je bila celo nekoliko slabša (razlika je bila zelo majhna). Ker je računanje zvezne valčne transformacije računsko zahtevno, so bili preizkušeni tudi MFCC koeficienti na osnovi diskretne valčne transformacije, s katerimi pa so bili doseženi slabši rezultati. Uporaba valčne transformacije pri računanju MFCC koeficientov glede na rezultate poskusov torej ni smiselna. Težava je najverjetneje v tem, da frekvenčna ločljivost pri valčni transformaciji s frekvenco pada linearno, torej bistveno hitreje kot pri sluhu, kjer pada z logaritmom frekvence. Različna dinamika spreminjanja spektra govornega signala na različnih odsekih z valčno transformacijo ni upoštevana, saj ima izračunani spekter na vseh odsekih enako frekvenčno-časovno ločljivost.

V disertaciji so predstavljene tri metode za dinamično prilagajanje frekvenčno-časovne ločljivosti dinamiki spreminjanja spektra. Pri vseh je potrebno dinamiko na nek način oceniti. Pri prvi metodi so bila uporabljena spoznanja o zgradbi govornega signala. V naših poskusih smo se oprli na dejstvo, da je govor zaporedje kratkih govornih enot, torej fonemov oziroma delov fonemov. Osnovne lastnosti spektrov posameznih fonemov glede dinamike spreminjanja spektra so znane. Tako je spekter npr. pri samoglasnikih in še nekaterih dolgih fonemih skoraj stacionaren. Za druge, npr. zapornike, so značilne hitre spremembe. Frekvenčno-časovno ločljivost spektra je fonemski zgradbi mogoče prilagajati s spreminjanjem dolžine okvirja pri računanju KČDFT. Ker fonemska zgradba razpoznavanih vzorcev ni znana vnaprej, jo je potrebno določiti s pomočjo SRG, pri katerem se uporabi okvir fiksne dolžine. Nato se izvede še drugi prehod razpoznavanja, v katerem se dolžina okvirja prilagodi fonemski zgradbi, ki je bila določena v prvem prehodu. Tak pristop ima

pomanjkljivosti, tudi če zanemarimo potrebo po razpoznavanju v dveh prehodih in posledično večjo računsko zahtevnost. V prvem prehodu določena fonemska zgradba je pri napačno razpoznanih besedah nepravilna. Metoda je kljub temu uporabna, ker pri napakah največkrat pride do zamenjave besed s podobno fonemsko zgradbo. Če se pri podobnih fonemih uporablja enaka dolžina okvirja, bo ta v drugem prehodu primernejša tudi pri besedah, ki so bile v prvem prehodu zamenjane z besedami s podobno fonemsko zgradbo. Ker je torej pri, s stališča dinamike spreminjanja spektra podobnih fonemih, potrebno uporabiti enako dolg okvir, se zelo zmanjša število različno dolgih okvirjev, ki jih je smiselno uporabiti. V naših poskusih sta bila uporabljena le dva okvirja različnih dolžin. Foneme je bilo torej potrebno razvrstiti v dve veliki skupini; na foneme, pri katerih se pri računanju KČDFT uporabi dolg okvir in na foneme, pri katerih se uporabi kratek okvir. Za razvrščanje fonemov v samo dve skupini je fonemska zgradba preveč podrobna in z opisanim pristopom ni v celoti izkoriščena.

Pri drugi predstavljeni metodi se frekvenčno-časovno ločljivost spektra prilagaja glede na spremembe jakosti zvoka. Večje spremembe jakosti zvoka so namreč povezane z odseki, na katerih je časovna ločljivost spektra pomembnejša od frekvenčne. Pojavijo se npr. pri izgovorjavi zapornikov in na prehodih med fonemi. Pri tej metodi se spekter govornega signala izračuna dvakrat, pri čemer se uporabi okvirja različnih dolžin. Pri računanju MFCC koeficientov se nato uporabi obteženo vsoto obeh spektrov. Na odsekih, kjer so spremembe jakosti velike, se poudari spekter z boljšo časovno ločljivostjo, drugod pa spekter z boljšo frekvenčno ločljivostjo. Računanje sprememb jakosti je bistveno manj zapleteno od določanja fonemske zgradbe, zato je postopek razpoznavanja hitrejši. Dodaten prehod razpoznavanja ni potreben.

Tretja v tej disertaciji predlagana metoda za dinamično prilagajanje frekvenčno-časovne ločljivosti temelji na določanju zvenečih odsekov govora. Na zvenečih odsekih govora morajo biti glasilke napete, da vibrirajo. Ker glasilk ni mogoče napeti za zelo kratek čas, so zveneči odseki dolgi, spekter pa se, z nekaj izjemami, na zvenečih odsekih spreminja počasi. Zato je bil na zvenečih odsekih uporabljen daljši okvir kot na nezvenečih.

Vsi tri opisane metode za dinamično prilagajanje frekvenčno-časovne ločljivosti so bile preizkušene z enakim SRG in dvema govornima zbirkama. Z uporabo aditivnih in konvolutivnih motenj je bil preverjen tudi vpliv na robustnost razpoznavanja. Prilagajanje dolžine okvirja fonemski zgradbi besed se je izkazalo za prezapleteno. Zaradi računske zahtevnosti je bila ta

metoda preizkušena le z manjšo govorno zbirko. Vpliv na uspešnost razpoznavanja je bil zelo majhen, robustnost pa se je celo nekoliko poslabšala. Najboljši rezultati so bili doseženi s prilagajanjem frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Izboljšala se je tako uspešnost razpoznavanja kot robustnost. S prilagajanjem dolžine okvirja zvenečim in nezvenečim odsekom govora se je izboljšala robustnost in v nekaterih poskusih tudi uspešnost razpoznavanja.

Glede na rezultate poskusov, je z dinamičnim prilagajanjem frekvenčno-časovne ločljivosti mogoče izboljšati robustnost in uspešnost razpoznavanja. Ker izboljšanje ni bilo dovolj veliko, da bi odtehtalo veliko povečanje računske zahtevnosti razpoznavanja, se nameravamo v prihodnosti posvetiti enostavnejšim postopkom, kot je npr. prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Pri tej metodi bi bilo za določanje sprememb jakosti mogoče uporabiti parametre, ki so potrebni že za izračun MFCC koeficientov. Namesto uporabe obtežene vsote dveh spektrov, bi bilo mogoče dolžino okvirja glede na prisotnost sprememb jakosti spreminjati v majhnih korakih. Tako bi odpadla tudi potreba po dvakratnem računanju spektra, ki je največja slabost predlagane metode.

DYNAMIC ADAPTATION OF TIME-FREQUENCY RESOLUTION IN SPECTRAL ANALYSIS OF SPEECH SIGNALS

ABSTRACT

Speech recognition can be viewed as a search process in which a sequence of short speech segments is mapped into speech units, such as phonemes, words or sentences. To recognize the sequence, the speech signal is first divided in short segments. These segments are then mapped into short speech units (usually phonemes or parts of phonemes) with the help of acoustic models. Acoustic models can be parametrical or non-parametrical, such as artificial neural networks (ANNs). To map sequences of short subword speech units to words, the temporal models are used. Currently the most dominant method for temporal modeling is hidden Markov model (HMM) approach. In this dissertation a hybrid HMM/ANN speech recognition system was used in all experiments. Short speech segments are classified with acoustic models according to their acoustical meaning. Human ear is essentially a spectrum analyzer, therefore acoustical meaning of a speech segment is determined by its spectrum.

In most recognition systems, each short speech segment is represented by a set of features which are extracted from the signal in the parametrization process. Mel frequency cepstral coefficients (MFCCs) are currently the most popular form of features in speech recognition systems. MFCCs are based on the signal spectrum, which is computed with short-time discrete Fourier transform. First, the signal is divided into short segments called frames, and the spectrum of each frame is computed using discrete Fourier transform. Therefore, the spectrum of the speech signal can be considered as a sequence of frame spectra. The frame length determines the time-frequency resolution of the speech signal spectrum. The product of time and frequency resolution is bounded. If longer frame is used the resulting spectrum has better frequency resolution at the expense of the

temporal resolution or vice versa. This means that frame length is always a compromise between temporal and frequency resolution. Because of nonstationary nature of the speech signal, determining an appropriate frame length can be problematic.

In MFCC extraction the speech signal is analyzed with a frame of typical length between 15 and 35 msec. The resulting spectrum has fixed time-frequency resolution that does not conform well to the properties of human hearing. One of the hearing properties is nonlinear frequency resolution which is finer at lower and coarser at higher frequencies. Nonlinear frequency resolution can be approximated with multiresolution spectral analysis. In this thesis the continuous wavelet transform was tried as a possible approach for MFCC extraction. The resulting spectrum, similar to the human hearing, has the frequency resolution that is better at lower frequencies and gets coarser with increasing frequency. This results in better time resolution at higher frequencies which allows spectral changes to be detected more precisely if they are (also) present at higher frequencies. Lower frequency band can be analyzed with fine frequency resolution at the same time. The comparison of success rate achieved with the same speech recognition systems did not show any advantages of wavelet transform based MFCCs over the short-time Fourier transform based MFCCs. Because computing continuous wavelet transform is computationally quite intensive, discrete wavelet transform based MFCCs were also tested. The use of discrete wavelet transform resulted in a significantly decreased success rate when compared to discrete Fourier transform based MFCCs. In our experiments, the use of wavelet transform for computing speech signal spectrum did not match the discrete Fourier transform approach. The reason could be that the frequency resolution of wavelet spectrum decreases linearly with frequency which does not conform to nonlinear frequency resolution of hearing where frequency resolution decreases with logarithm of frequency. Of course, the use of wavelet transform does not solve the problems related to nonstationarity of speech signal, as the time-frequency resolution of its spectrum is not time dependent.

In this dissertation, three approaches for adapting time-frequency resolution were presented. In every approach, one has to estimate how rapidly the spectrum is changing at a given time. This estimation can be based on known facts about the structure of speech or about production of speech. In the first presented approach, the adaptive time-frequency was achieved by varying the frame length based on the phonetic structure of the speech. For each phoneme, the basic properties of spectrum are known. The spectrum of vowels and some other long phonemes is almost

stationary, but spectrum of other phonemes, such as stops changes rapidly. If phonetic structure of speech is known, the time-frequency can be adapted by using appropriate frame length for each phoneme. In speech recognition, the phonetic structure is not known. Therefore, speech recognition needs to be done in two passes. Phonetic structure is unknown in the first pass and a fixed frame length is used for parametrization. In the second pass, the phonetic structure from the first pass is known, and the frame length is selected on its basis. Besides the need for additional pass, this approach has some disadvantages. The phonetic structure from the first pass is not always correct. Therefore, the approach is based on an assumption that the incorrectly recognized words are most often substituted with words with similar phonetic structure. If the same frame length is used in parametrization of similar phonemes in the second pass, a more appropriate frame length will be used even if the word was recognized wrong in the first pass, but was substituted with a word with similar phonetic structure. This means that the same frame length should be used for spectral analysis of many phonemes and the number of different frame lengths is very limited. In our experiments only two different frame lengths were used. Therefore, phonemes need to be grouped in only two groups, a group which is analyzed with the short frame, and a group which is analyzed with the long frame.

In the second presented approach, the time-frequency resolution was adapted according to Moore's formula, which describes human's perception of intensity changes in speech signal. Most of intensity changes are related to sections of speech where temporal resolution is more important than frequency resolution. Larger intensity changes are related to short phonemes, such as burst release in plosives. Intensity changes are also related to phoneme transitions. Therefore, when intensity changes are high, the wideband spectrum is emphasized and when they are low narrowband spectrum is emphasized. Computing intensity changes is far less computationally intensive than determining the phonetic structure in an additional pass.

The third approach is based on recognition of voiced and unvoiced speech segments. When voiced speech is produced, the vocal folds need to be closed to obstruct the airflow. Because voiced and unvoiced segments are determined by opening or closing the vocal folds, a voiced segment cannot be very short. Most of voiced phonemes are long and have almost stationary spectrum. In feature extraction longer frame was used on voiced segments and shorter frame on unvoiced segments.

All of the three above-mentioned approaches to dynamic time-frequency resolution adapting were

tested with the same speech recognition system with two speech databases. Several additive and two convolutive distortions were used to test the robustness.

In our experiments, adapting frame length based on phonetic structure of speech proved to be too complicated. It is computationally demanding, and was only tested with the smaller speech database. The success rate was almost unchanged, and robustness decreased slightly in comparison to the original speech recognition system which uses standard MFCCs. Adapting time-frequency resolution to intensity changes resulted in increased success rate and robustness. The improvement was quite large and very consistent. Adapting the frame length according to voiced and unvoiced speech segment improved the robustness and in some experiments the success rate.

The results show that adapting time-frequency resolution can result in improved success rate and robustness. However, the improvement is not large enough to compensate computationally intensive approaches. Therefore, our future work will be focused on approaches, such as adapting the time-frequency resolution according to intensity changes. This approach can be further improved since most of the parameters for computing intensity changes are also needed for computing MFCCs. Instead of weighted sum of two spectrums the frame length could be adapted in small steps. This way it is only necessary to compute spectrum once which would further reduce the computing load.

KAZALO

1. UVOD

Sposobnost govornega sporazumevanja je ena izmed naravnih značilnosti človeka. Govor omogoča učinkovito komunikacijo, izmenjavo občutkov, misli in idej. Z uveljavljanjem računalnikov na različnih področjih postaja, poleg komuniciranja z drugimi ljudmi, vse bolj pomemben tudi način komuniciranja z računalniki. Tu se kljub izredno hitremu povečevanju zmogljivosti računalnikov še vedno opiramo na tipkovnico, kar je lahko moteče. Možnost govorne komunikacije z računalniki bi bila zelo dobrodošla, še posebej, ker so se računalniki uveljavili tudi na področjih, kjer človek tipkovnice ne more uporabljati.

Morda so bile prav do nedavnega premajhne zmogljivosti računalnikov ovira in glavni krivec, da se z razpoznavanjem govora ukvarjamo šele v zadnjem času. Področje razpoznavanja govora je razmeroma mlado. Je izrazito multidisciplinarno, saj vključuje fiziologijo nastajanja govora, akustične značilnosti govora in proces zaznavanja govora pri poslušalcih. Za začetek dela na tem področju lahko štejemo prispevke, ki segajo v devetnajsto stoletje [1, 11, 22]. Sodobne raziskave so se začele hkrati z uveljavljanjem telefona v letih po 1930 [6]. Prvi sistemi za razpoznavanje govora[1] so nastali približno 20 let kasneje. Raziskave so v tem času potekale hkrati z gradnjo SRG in so bile precej inženirske, večinoma usmerjene k reševanju konkretnih problemov pri izdelavi posameznih SRG. Mnoga spoznanja iz tega časa so strnjena v Flanaganovi knjigi [8].

Do večjih premikov je na področju razpoznavanja govora prišlo v sedemdesetih letih prejšnjega stoletja. Na osnovi temeljnih raziskav [13, 20, 26] so se pojavili SRG, s katerimi je bilo mogoče razpoznavati posamezne besede. V tem času so bile v razpoznavanje govora uvedene rešitve s področja razpoznavanja vzorcev. Sem spadajo npr. dinamično programiranje in linearna predikcija. V naslednjem desetletju so se raziskovalci preusmerili k povezanemu govoru. Cilj je bil izgradnja SRG, ki bi bil s povezovanjem modelov posameznih besed sposoben razpoznavati tekoče izgovorjena zaporedja besed. Nastalo je več SRG za razpoznavanje tekoče izgovorjenih zaporedij

[1] *v nadaljevanju SRG*

števk [16, 21]. Pomemben korak, ki se je pri tem zgodil, je bil prehod na metode statističnega modeliranja. Uveljavili so se prikriti Markovi modeli [7]. Nova spoznanja o možnostih in omejitvah nevronskih mrež so omogočila njihovo uporabo pri razpoznavanju govora [14]. Oba pristopa sta v SRG močno prisotna tudi v današnjem času.

Kljub precejšnjim vloženim naporom, razpoznavanje govora danes še zdaleč ni rešen problem. Razlog bi lahko bil ravno preveč inženirski pristop v preteklosti. Morda je bilo preveč raziskav usmerjenih v izboljševanje obstoječih rešitev in premalo v uvajanje novih. Na ta način so nastali SRG, ki so ob določenih omejitvah (samo en govorec, zelo omejeno število besed) zelo uspešni. Taki SRG lahko celo predstavljajo oviro nadaljnjim raziskavam, saj zbujajo vtis, da je dodatno izboljševanje skoraj nesmiselno. Vendar v bolj realnih pogojih, kot je razpoznavanje večjega števila besed v realnem okolju, v katerem so prisotni šumi, in ob večjem številu govorcev, SRG na nivoju posameznih besed ne presežejo uspešnosti okoli 85%. To pa je za širšo uporabnost bistveno premalo, zato se razpoznavanje govora uspešno uporablja zgolj na področjih, kjer je potrebno razpoznavati manjše število besed. Sem spada npr. govorno krmiljenje naprav, kjer je potrebno razpoznati samo nekaj različnih ukazov.

Pri človeku je razpoznavanje govora razdeljeno v zaznavanje govora, ki poteka s pomočjo sluha in samo razpoznavanje, ki se odvija v možganih. Podobno je pri SRG. Analogna zaznavanju govora je parametrizacija govornega signala. Ta del je razmeroma dobro raziskan. Rešitve se zgledujejo po lastnostih človekovega sluha. Obratno je razpoznavanje govora, ki poteka v možganih, skoraj popolna neznanka. Pri SRG je modelirano z razmeroma preprostimi statističnimi modeli. V večini so to še vedno prikriti Markovi modeli, ki so bili na področje razpoznavanja govora vpeljani pred približno 25 leti. Ta del razpoznavanja je bolj problematičen, saj statistični modeli ne ustrezajo procesu razpoznavanja v možganih.

Disertacija je usmerjena na področje parametrizacije govornega signala, katere bistvo je računanje značilk na osnovi spektra signala. Spektralna analiza je pomemben del parametrizacije in največkrat vključuje računanje kratkočasovne Fourierove transformacije. Pri računanju značilk so običajno upoštevana spoznanja o lastnostih človekovega sluha in spoznanja o načinu nastajanja govornega signala [18]. Zaradi narave govora ima govorni signal na različnih časovnih odsekih zelo različne lastnosti. Presenetljivo je, da se v večini obstoječih SRG v postopku parametrizacije tega ne upošteva. To je glavna motivacija za nastanek disertacije. Njen glavni del predstavlja

proučevanje vpliva prilagajanja frekvenčno-časovne ločljivosti pri spektralni analizi na uspešnost razpoznavanja. Frekvenčno-časovno ločljivost je na različnih odsekih govornega signala mogoče prilagajati s spreminjanjem dolžine okvirja, ki se uporabi pri računanju kratkočasovne Fourierove transformacije.

1.1 Spektralna analiza govornega signala

Spektralna analiza je pri veliki večini SRG bistveni del parametrizacije govornega signala. Največ se uporablja kratkočasovna Fourierova transformacija. Glavni razlog za to je obstoj učinkovitega algoritma FFT[2] za računanje diskretne Fourierove transformacije. Znano je, da je spekter govornega signala dinamičen, torej se s časom spreminja. Zato je govorni signal potrebno razrezati na kratke časovne odseke (okvirje) in za vsak okvir posebej izračunati diskretno Fourierovo transformacijo. Razrez govornega signala na okvirje se doseže z množenjem signala z okensko funkcijo (oknom), ki jo predstavlja končno dolgo utežnostno zaporedje. Spekter načeloma neskončno dolgega govornega signala je torej zaporedje spektrov okvirjev, na katere se razreže signal. Izbrana dolžina okvirja oziroma okna ima pomemben vpliv na izračunani spekter. Za vsak okvir namreč izračunamo povprečni spekter signala, ki ga okvir zajema. Posledično to pomeni, da izračunani spekter slabo opisuje dinamične spremembe spektra govornega signala, ki so krajše od dolžine okvirja. S stališča časovne ločljivosti spektra govornega signala je torej smiselno uporabljati kratke okvirje. Vpliv dolžine okvirjev na frekvenčno ločljivost signala je obraten; izračunani spekter signala znotraj okvirja je podan v končno mnogo frekvenčnih točkah. Število frekvenčnih točk je sorazmerno s številom vzorcev znotraj okvirja, torej kratek okvir pomeni manj vzorcev in slabšo frekvenčno ločljivost. Zato gre pri določitvi dolžine okvirja za kompromis med časovno in frekvenčno ločljivostjo spektra. Kot je opisano v nadaljevanju, je pri frekvenčni analizi govornega signala ta problem še posebej izrazit.

Govorni signal nastane zaradi toka iz pljuč iztisnjenega zraka skozi vokalni trakt. Zveneči zvoki nastanejo v glasilkah. Takrat so glasilke napete in zaradi toka zraka začnejo vibrirati. Če so glasilke sproščene, mora zračni tok teči skozi oviro in se zato zvrtinči. Tako nastajajo nezveneči glasovi.

[2] *angl. Fast Fourier Transform*

Signal govora je zaporedje glasov, ki nastanejo glede na napetost glasilk in nadaljnjo artikulacijo v vokalnem traktu (žrelu, ustni in nosni votlini) in v grobem ustrezajo fonemom. Govorni signal človek oblikuje s spreminjanjem oblike vokalnega trakta. Bistvo govornega signala je torej ravno spreminjanje oblike njegovega, s prevajalno funkcijo vokalnega trakta določenega, spektra s časom.

Pri tem je dinamika spreminjanja spektra govornega signala na različnih odsekih govora različna. Pri nekaterih fonemih je spekter skoraj stacionaren, pri drugih se spreminja zelo hitro. Prav tako se največkrat spekter hitro spreminja na prehodih med fonemi. Presenetljivo malo je člankov, ki se ukvarjajo s problemom različne dinamike spreminjanja spektra govornega signala oziroma z dinamičnim prilagajanjem frekvenčno-časovne ločljivosti spektra, še posebej na področju razpoznavanja govora. Večinoma spadajo na sorodno področje kodiranja govora [87-96]. Prilagajanje frekvenčno-časovne ločljivosti pri spektralni analizi je razmeroma pogosto tudi, če v signalu prevladuje ena frekvenčna komponenta, npr. na področju radarjev [99].

Kljub spremenljivi dinamiki spreminjanja spektra govornega signala, danes skoraj vsi SRG temeljijo na spektralni analizi s fiksno frekvenčno-časovno ločljivost; ta je določena z dolžino okvirja, ki se uporabi pri računanju kratkočasovne Fourierove transformacije. Pri obstoječih SRG se uporablja okvirje fiksne dolžine 15 - 35 ms. Dolžino se običajno določi s preizkušanjem uspešnosti razpoznavanja določene množice besed oziroma stavkov pri različno dolgih okvirjih. Vendar je, glede na opisano, težko govoriti o optimalni dolžini okvirja. To bi bilo potrebno sproti čim bolje prilagoditi trenutni dinamiki spreminjanja spektra govornega signala. Prilagajanje dolžine okvirja dinamiki spektra govornega signala, ob upoštevanju spoznanj o njegovem nastanku, predstavlja jedro disertacije.

1.2 Možnosti dinamičnega prilagajanja dolžine okvirja

Ker je mehanizem nastajanja govornega signala razmeroma dobro poznan, je ta spoznanja mogoče uporabiti za oceno, kako hitro se na določenem časovnem odseku spreminja njegov spekter. Ker je govor sestavljen iz zaporedja fonemov, za katere je znano na kakšen način se spekter pri njihovi izgovorjavi spreminja, je mogoče dolžino okvirja temu prilagoditi. Da to lahko storimo, je potrebno poznati fonemsko zgradbo besed, ki jih govorni signal predstavlja. To zgradbo je mogoče dobiti

tako, da se govorni signal najprej razpozna s SRG s fiksno dolžino okvirja. Nato se izvede še drugi prehod, v katerem se govorni signal razpozna še enkrat, pri čemer se dolžino okvirja ustrezno prilagodi fonemski zgradbi. Težava tega pristopa je, da v prvem prehodu pri napačno razpoznanih besedah dobimo napačno fonemsko zgradbo. Ker razpoznavanje sloni na značilnostih spektra fonemov, pride pri napakah največkrat do zamenjave fonemov s fonemi, ki imajo podobne lastnosti spektra. Če pa sta si spektra obeh fonemov glede dinamike spreminjanja spektra podobna, je pri obeh fonemih pri računanju spektra smiselno uporabiti enako dolg okvir. Če je ta okvir primernejši od okvirja fiksne dolžine, ki je bil uporabljen v prvem prehodu, se s tem možnosti za pravilno razpoznavo povečajo. Foneme je torej potrebno razvrstiti v skupine s podobnimi lastnostmi glede dinamike spreminjanja spektra in pri vseh fonemih iz iste skupine uporabiti enako dolg okvir. To pa pomeni, da uporaba večjega števila različno dolgih okvirjev ni smiselna. V naših poskusih, ki so opisani v nadaljevanju disertacije, sta bila vedno uporabljena samo dva različno dolga okvirja.

Pri razvrščanju fonemov v skupine se je torej potrebno osredotočiti predvsem na dinamiko spreminjanja spektra pri fonemih. To pa je pogosto mogoče oceniti tudi brez poznavanja fonemske zgradbe. Hitre spremembe spektra so mnogokrat povezane s spremembami jakosti govornega signala. Tako npr. zaporniki nastanejo s kratkim zaprtjem vokalnega trakta. To pomeni, da se tok zraka skozi vokalni trakt za kratek čas ustavi in je posledično jakost govornega signala zelo majhna. Ob izgovorjavi zapornika se vokalni trakt nenadoma odpre, kar se odrazi v hitrem povečanju jakosti. Torej lahko zapornike povežemo s hitrimi spremembami jakosti govornega signala. Ker so za njih značilne hitre spremembe spektra, je pri računanju spektra smiselno poudariti časovno ločljivost. Za razliko od razmejitve na posamezne foneme, so spremembe jakosti zvezna mera. Zato pri računanju spektra ni potrebno izbirati med dvema možnima dolžinama okvirja. Spekter signala se lahko izračuna z dvema različno dolgima okvirjema, pri računanju značilk pa se uporabi obteženo vsoto obeh spektrov. Ob večjih spremembah jakosti govornega signala je v vsoti smiselno poudariti s krajšim okvirjem izračunani spekter. Tak pristop je dal dobre rezultate, tudi če so bile govornemu signalu dodane različne motnje.

Foneme je mogoče razvrstiti v skupine tudi glede na to, ali je določen fonem zveneč ali ne. Pri zvenečih fonemih so glasilke napete in vibrirajo, spekter signala pa se oblikuje z obliko vokalnega trakta. Glasilk človek ne more napeti za zelo kratek čas, zato so zveneči fonemi večinoma dolgi (razen pri zvenečih zapornikih, pri katerih pride do zaprtja vokalnega trakta), spekter pa se pri njih spreminja počasi, ker tudi oblike vokalnega trakta ni mogoče spreminjati zelo hitro. Zaradi

počasnih sprememb spektra govornega signala, je na zvenečih odsekih smiselno uporabiti daljši okvir kot na nezvenečih. Podobno kot za spremembe jakosti zvoka, tudi za določanje zvenečih odsekov ni potrebno poznati fonemske zgradbe besed. Za zveneče odseke je značilno, da je govorni signal zaradi vibriranja glasilk periodičen. Zaznavanje periodičnosti signala in iskanje osnovne frekvence (frekvence s katero vibrirajo glasilke) je mogoče z uporabo avtokorelacije in je dobro raziskano [83-85]. Prav periodičnost predstavlja še dodaten razlog za uporabo daljših okvirjev. Osnovna frekvenca govornega signala je lahko zelo nizka, običajno znaša od 60 Hz do 300 Hz. Če pri računanju spektra uporabimo prekratek okvir, ki ne zajame vsaj ene periode osnovne frekvence, bodo v spektru prisotne oscilacije, ki so povezane z osnovno frekvenco. Te oscilacije v nadaljnjem postopku razpoznavanja niso zaželene in predstavljajo motnjo.

Frekvenčno-časovno ločljivost spektra je ob upoštevanju spoznanj o nastajanju govornega signala torej mogoče prilagajati na več načinov. Pri tem ni nujno potrebno poznati fonemske zgradbe. Prilagajanje je mogoče uspešno izvesti z uporabo akustičnih značilk, med katere spadata določanje sprememb jakosti in periodičnosti govornega signala. Če torej izberemo primerne akustične značilke, lahko z njimi ocenimo dinamiko spreminjanja spektra govornega signala in ob hitrejših spremembah uporabimo krajši okvir kot na odsekih, kjer se spekter spreminja počasi.

1.3 Pregled vsebine

Preizkus omenjenih postopkov za dinamično prilagajanje dolžine okvirja je osnovni motiv za nastanek disertacije. Področje je namreč slabo raziskano, saj člankov, v katerih bi frekvenčno-časovno ločljivost spektra prilagajali s pomočjo akustičnih značilk, nismo zasledili. To je deloma presenetljivo, saj se podobne akustične značilke uspešno uporabljajo v sistemih za označevanje govora [82].

V poskusih je bil uporabljen hibridni SRG, izdelan z orodjem *CSLU Toolkit* [49]. Postopek parametrizacije je bil prilagojen tako, da pri spektralni analizi omogoča uporabo različno dolgih okvirjev glede na fonemsko zgradbo, prisotnost sprememb jakosti zvoka in glede na zveneče oziroma nezveneče odseke govora. Postopke prilagajanja frekvenčno-časovne ločljivosti smo primerjali s klasičnim SRG s fiksno dolžino okvirja. Sistemi so bili preizkušeni z dvema govornima zbirkama z različnih jezikovnih območij. Primerjava je pri vseh poskusih narejena z uporabo istih

učnih vzorcev in z istimi testnimi množicami. Uspešnost je podana z odstotkom pravilno razpoznanih besed v vzorcih, v katerih je več besed. Besede v vzorcih so bile izgovorjene samostojno in povezano. Poleg vpliva na samo uspešnost razpoznavanja, nas je zanimal tudi vpliv dinamičnega prilagajanja frekvenčno-časovne ločljivosti na robustnost oziroma odpornost na motnje. Zato so bile testnim vzorcem dodane različne aditivne motnje v različnih razmerjih signal/šum. Samostojno in v kombinaciji z aditivnimi motnjami sta bili preizkušeni še dve konvolutivni motnji: slabljenje visokih frekvenc in odjek. Še posebej s prilagajanjem frekvenčno-časovne ločljivosti glede na spremembe jakosti zvoka so bili doseženi boljši rezultati kot z uporabo okvirja fiksne dolžine.

V nadaljevanju disertacije je v 2. poglavju predstavljeno avtomatsko razpoznavanje govora. Opisana je osnovna zgradba SRG. Iz nje je razviden postopek razpoznavanja govora in način modeliranja govora. Poudarek je na parametrizaciji govornega signala. Podrobneje je predstavljen postopek računanja MFCC koeficientov, ki že daljši čas prevladujejo v obstoječih SRG. Sledi opis nevronskih mrež, s katerimi je pri hibridnih SRG izvedena klasifikacija v akustične modele.

V poglavju 3 sta opisani govorni zbirki, ki sta bili uporabljeni pri poskusih. To sta zbirki ŠTEVKE, v kateri so izgovarjave slovenskih števk in NUMBERS, v kateri so izgovorjave angleških števk. Obe sta posneti preko telefonske linije. Razen po jeziku, se razlikujeta še po načinu izgovorjave. Besede so v zbirki ŠTEVKE izgovorjene ločeno (med besedami so razmeroma dolgi premori), v zbirki NUMBERS pa se poleg ločeno izgovorjenih besed pojavljajo tudi vezana zaporedja števk.

Referenčni SRG je predstavljen v poglavju 4. Podrobno je opisana zgradba in izvedba vseh stopenj sistema. Prikazan je postopek parametrizacije govornega signala in lastnosti nevronske mreže, s katero je izvedeno razvrščanje značilk v akustične modele. Podrobno je opisano modeliranje fonemov s kontekstno odvisnimi kategorijami. Na kratko so predstavljeni prikriti Markovi modeli, po katerih se zgleduje časovno modeliranje v referenčnem sistemu. Navedene so tudi podrobnosti o učenju SRG, kamor spada učenje nevronske mreže in tvorba časovnih modelov.

V poglavju 5 je podrobneje predstavljena kratkočasovna Fourierova transformacija, ki se pri večini obstoječih SRG uporablja za spektralno analizo. Poudarek je na pomenu dolžine okvirja, ki se ga pri tem uporabi. Kot alternativa je predstavljen še večločljivostni pristop k spektralni analizi z zvezno in diskretno oziroma paketno valčno transformacijo. Na koncu poglavja so podani rezultati

primerjave vpliva uporabe kratkočasovne Fourierove transformacije, zvezne valčne transformacije in paketne valčne transformacije v postopku parametrizacije na uspešnost razpoznavanja.

Poglavje 6 predstavlja jedro disertacije. V treh podpoglavjih so opisane omenjene metode za dinamično prilagajanje dolžine okvirja govornemu signalu: prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed, prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka in prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora. V vsakem od podpoglavij je podrobno predstavljena metoda za prilagajanje frekvenčno-časovne ločljivosti in rezultati poskusov, ki so bili z njo doseženi. Rezultati so vedno primerjani z rezultati SRG, pri katerih se uporablja okvir fiksne dolžine. Opisani poskusi so bili izvedeni z govorno zbirko ŠTEVKE in, razen v poglavju o prilagajanju dolžine okvirja na osnovi fonemske zgradbe, tudi z delom govorne zbirke NUMBERS.

V poglavju 7 sledi medsebojna primerjava najuspešnejših poskusov iz poglavja 6. Pri preverjanju vpliva na robustnost razpoznavanja so glede na poskuse iz prejšnjih poglavij uporabljene dodatne oblike aditivnih motenj in dve konvolutivni motnji. Pri poskusih z govorno zbirko NUMBERS je uporabljena celotna zbirka.

8. poglavje je zaključek, v katerem so podani zaključki in končne ugotovitve, ki izhajajo iz analize rezultatov vseh opisanih raziskav v pričujoči disertaciji, ter smernice za nadaljnje raziskave na področju dinamičnega prilagajanja frekvenčno-časovne ločljivosti pri spektralni analizi govornega signala.

2. SISTEMI ZA AVTOMATSKO RAZPOZNAVANJE GOVORA

Pri razpoznavanju govora gre pravzaprav za razpoznavanje govornih enot v govornem signalu [9]. Govorne enote lahko enačimo z različno dolgimi časovnimi odseki, kot so npr. deli fonemov, fonemi, besede in stavki. Ko razpoznavamo govor, preslikujemo odseke govornega signala v zaporedje govornih enot. Zato lahko razpoznavanje govora opišemo kot razpoznavanje vzorcev, ki je razdeljeno v več nivojev in je zaradi lastnosti govornega signala zapleteno. Vokalni trakt in artikulatorji (jezik, mehko nebo, itd.), s pomočjo katerih človek oblikuje govor, so namreč biološki organi z nelinearnimi lastnostmi [18]. Poleg tega je oblika govornega signala odvisna še od drugih dejavnikov, kot sta npr. spol govorca in čustveno stanje govorca [31, 32]. Posledično je izgovorjava in s tem oblika govornega signala lahko različna in se pri istem govorcu razlikuje po lastnostih kot so naglas, artikulacija, osnovna frekvenca, glasnost, hitrost govora, itd. Dodatno so v govornem signalu pogosto prisotni tudi vplivi okolja. Sem spada odmev, šumi in večkrat tudi vplivi medija, po katerem se govor prenaša (telefonska linija, lastnosti opreme za zajemanje govornega signala).

Lastnosti SRG so odvisne od njihovega namena. Enostavnejši so namenjeni enemu govorcu in lahko razpoznajo le nekaj različnih besed. Drugi so namenjeni razpoznavanju velikega števila besed, ki jih izgovarjajo različni govorci, in morajo biti odporni na različne motnje. Zaradi tega so lahko SRG po kompleksnosti zelo različni. Na kompleksnost in uspešnost SRG vplivajo atributi kot so:

- **Velikost slovarja in podobnost besed v slovarju.** Med manjšim številom besed je lažje razlikovati. Po velikosti slovarja se SRG delijo na majhne (nekaj besed), srednje (do 5000 besed) in velike (nad 5000 besed). Tudi če je slovar majhen, so lahko besede v njem podobne, kar poslabša uspešnost razpoznavanja

- **Neodvisnost od govorca.** Od govorca odvisni sistemi so namenjeni enemu samemu govorcu. Če je sistem namenjen različnim govorcem, je od govorca neodvisen. Od govorca neodvisni sistemi so pri razpoznavanju manj uspešni, tipično je število nepravilno razpoznanih besed

nekajkrat večje kot pri od govorca odvisnih SRG.

- **Povezanost govora.** SRG so lahko namenjeni (naraščajoča kompleksnost) razpoznavanju posameznih besed, zaporedij besed, ki so ločene s premori, ali razpoznavanju naravnega, povezanega govora. Dodatna težava povezanega govora je koartikulacija; izgovorjava neke besede je odvisna od predhodne besede.

- **Spontanost govora.** Spontan govor je težje razpoznavati kot npr. branje ali narekovanje.

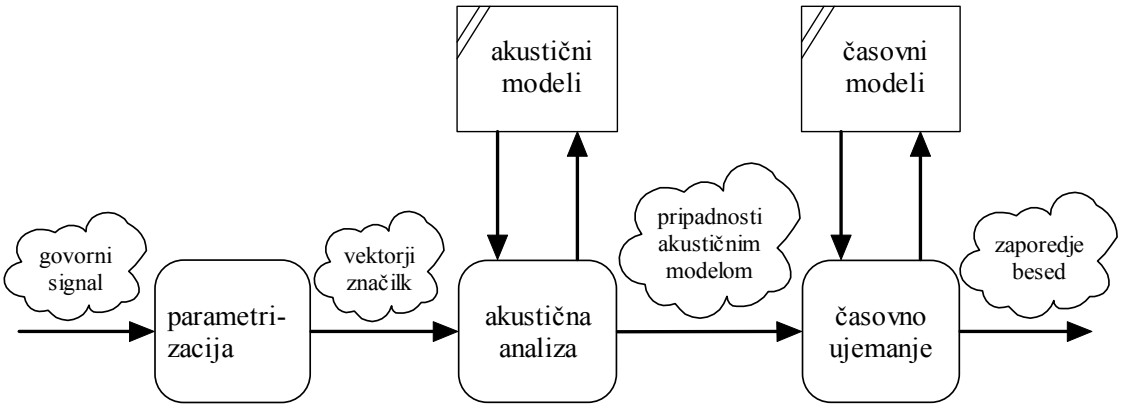
- **Ostali pogoji.** Mednje spada npr. šum okolja (razpoznavanje v avtu, govoru ljudi v ozadju), različni odmevi in akustika prostorov, kvaliteta mikrofona (usmerjeni, telefonski), frekvenčne omejitve (telefonska linija, GSM telefonija), itd.

Ne glede na različno kompleksnost SRG, je njihova osnovna zgradba zelo podobna. Izjema so le SRG, ki so namenjeni razpoznavanju zelo kratkih govornih enot (fonemov). Pri teh zadošča, da posamezne odseke govora klasificiramo glede na njihov akustični pomen. Pri bolj zapletenih SRG je potrebno še časovno modeliranje, s katerim je opisano časovno zaporedje krajših odsekov govora, npr. zaporedje fonemov tvori besedo, zaporedje besed je stavek, itd. V naslednjem poglavju je predstavljena osnovna zgradba SRG.

2.1 Zgradba sistemov za razpoznavanje govora

Razpoznavanje govora lahko opišemo kot razpoznavanje vzorcev, ki je izvedeno v več nivojih. Nivoji oziroma stopnje razpoznavanja govora so pri večini sistemov enaki, kar je razlog za podobno osnovno zgradbo večine SRG. Do razlik prihaja pri načinu razpoznavanja vzorcev na različnih nivojih, kjer se lahko uporabljajo različni statistični modeli. Tako lahko npr. zelo kratke časovne odseke govornega signala klasificiramo glede na njihov akustični pomen na različne načine. Uporabimo lahko različne statistične modele, pogosto pa se uporabljajo tudi parametrični modeli, kot so Gaussove mešanice (GMM)[3].

[3] *angl. Gaussian Mixture Model*



Slika 1:Osnovna zgradba sistemov za razpoznavanje govora.

Na sliki 1 je prikazana osnovna zgradba SRG. Prikazani so:

- **Govorni signal.** Zaporedje vzorcev, ki ga dobimo z vzorčenjem analognega govornega signala. Vzorčevalna frekvenca običajno znaša vsaj 8000 Hz, kar da zelo dolgo zaporedje. Pred začetkom postopka razpoznavanja je zato potrebno dobiti kompaktnejšo predstavitev govornega signala.

- **Parametrizacija.** Računanje kompaktne predstavitve govornega signala - *značilk*. Značilke morajo dobro opisovati značilnosti govornega signala. Pri tem želimo govorni signal opisati s čim manj značilkami, saj to poenostavi postopek razpoznavanja. Parametre oziroma značilke se združi v vektorje značilk. Vsak vektor značilk predstavlja določen časovni odsek govornega signala. Običajno se v postopku parametrizacije govorni signal najprej razreže v krajše časovne odseke - *okvirje,* nato pa se posamezne okvirje preslika v frekvenčni prostor. Postopek je podoben dogajanju v človekovem ušesu. Bazilarna membrana, ki je del notranjega ušesa, namreč deluje kot spektralni analizator. Točke, razporejene vzdolž membrane, imajo različne resonančne frekvence. Nihanje teh točk se preko ušesnih laskov prenese neposredno do slušnega živca [19]. Preko slušnega živca se torej prenašajo informacije o prisotnosti različnih frekvenc v signalu.

  Spekter posameznega okvirja se v nadaljevanju postopka parametrizacije še dodatno obdela. Pri tem se običajno upoštevajo tako spoznanja o lastnostih govornega signala kot tudi lastnosti zaznavanja govornega signala, torej značilnosti človekovega sluha [100-111].

- **Akustična analiza.** Na tej stopnji se vektorje značilk klasificira glede na njihov akustični pomen. Uporabimo *akustične modele*. Za vsak vektor značilk je potrebno izračunati verjetnost, da vektor ustreza določenemu akustičnemu modelu. Verjetnosti se izračunajo za vse akustične modele. Dobimo vektor, v katerem posamezne komponente predstavljajo verjetnost, da določen vektor značilk oziroma ustrezen odsek govora, ustreza določenemu akustičnemu modelu. Akustični modeli predstavljajo akustične značilnosti kratkih govornih enot, ki najpogosteje ustrezajo posameznim fonemom ali delom fonemov. Podobnost oziroma verjetnost, da vektor značilk ustreza določenemu akustičnemu modelu, je mogoče določiti na različne načine, kot sta npr. spektralna razdalja in statistična ustreznost.

- **Časovno ujemanje.** Za določitev, kateri daljši govorni enoti ustreza določen odsek govora, je poleg vektorja ustreznosti akustičnim modelom, pomembno še časovno zaporedje. *Časovni modeli* predstavljajo možna zaporedja krajših govornih enot, s katerimi so predstavljene daljše govorne enote. To so npr. fonemska zgradba besed oziroma zaporedje fonemov v besedah, ki jih SRG razpoznava. Za iskanje najverjetnejšega zaporedja se pogosto uporablja Viterbijev algoritem, ki je podrobneje opisan v poglavju 4.3.1. Rezultat je zaporedje besed, ki jih predstavlja govorni signal. V kompleksnejših SRG se uporabljajo še jezikovni modeli, ki opisujejo najverjetnejša oziroma možna zaporedja besed. Običajno so ti modeli predstavljeni s preprostimi gramatikami.

Iz osnovne zgradbe SRG je razvidno, da razpoznavanje vzorcev na višjih nivojih omejuje prostor možnih hipotez na nižjih nivojih. Na ta način višji nivoji vnašajo dodatne omejitve, npr. možna zaporedja besed, na podlagi katerih se poskuša odpraviti napake, do katerih prihaja na nižjih nivojih.

V preteklosti so se uveljavili SRG, ki temeljijo na prikritih Markovih modelih (HMM)[4]. Klasični sistemi HMM temeljijo na levo-desnih Markovih modelih. Ti so uporabljeni kot časovni modeli, kot akustični modeli pa so uporabljene Gaussove mešanice. Uporaba Gaussovih mešanic predstavlja eno izmed večjih slabosti SRG tega tipa. Verjetnostna porazdelitev v akustičnem prostoru je opisana kot obtežena vsota Gaussovih funkcij, kar resnični porazdelitvi ne ustreza

[4] *angl. Hidden Markov Models*

najbolje. Kasneje so se pojavili SRG, ki temeljijo na nevronskih mrežah (ANN)[5]. Po uspešnosti niso dosegli sistemov HMM, se pa nevronske mreže uporabljajo v hibridnih sistemih HMM/ANN. Hibridni SRG za časovno modeliranje uporabljajo modele HMM, za akustično modeliranje pa nevronske mreže [24]. Nekoliko poenostavljen model tega tipa je tudi v tej disertaciji uporabljen referenčni SRG, ki je podrobno opisan v poglavju 4.

2.2 Parametrizacija govornega signala

V tem poglavju so podrobno predstavljeni MFCC[6] koeficienti [58], ki jih kot značilke uporablja velik del obstoječih SRG. Temeljijo na homomorfni analizi govornega signala, s katero se izvede ločitev vzbujanja od vpliva prenosne funkcije vokalnega trakta. Za razpoznavanje govora je vpliv vokalnega trakta bistvenega pomena, saj je način spreminjanja oblike vokalnega trakta in s tem njegove prenosne funkcije razmeroma neodvisen od govorca. Za MFCC koeficiente velja, da koeficienti z nižjimi indeksi opisujejo prevajalno funkcijo vokalnega trakta, vzbujanje pa je opisano s koeficienti z višjimi indeksi. Zato se kot značilke uporablja koeficiente z nižjimi indeksi. Vsak okvir oziroma kratek časovni odsek govornega signala je običajno opisan z 10 - 15 MFCC koeficienti. Ker je z njimi opisan celoten okvir, mora biti ta dovolj kratek, da lahko spremembe oblike vokalnega trakta znotraj okvirja zanemarimo.

MFCC koeficienti temeljijo na spektru govornega signala, zato kratek okvir hkrati pomeni manjšo frekvenčno ločljivost spektra (poglavje 5), kar pomeni, da se podrobnosti o obliki prenosne funkcije vokalnega trakta izgubijo. Za računanje MFCC koeficientov je torej izbira primerne dolžina okvirja bistvenega pomena. V postopku računanja MFCC koeficientov so na več mestih upoštevana spoznanja o lastnostih sluha, ki so podrobneje opisne npr. v [19]. S tem se poudari tiste značilnosti govornega signala, ki bi jih zaznal človek.

---

[5] *angl. Artificial Neural Networks*

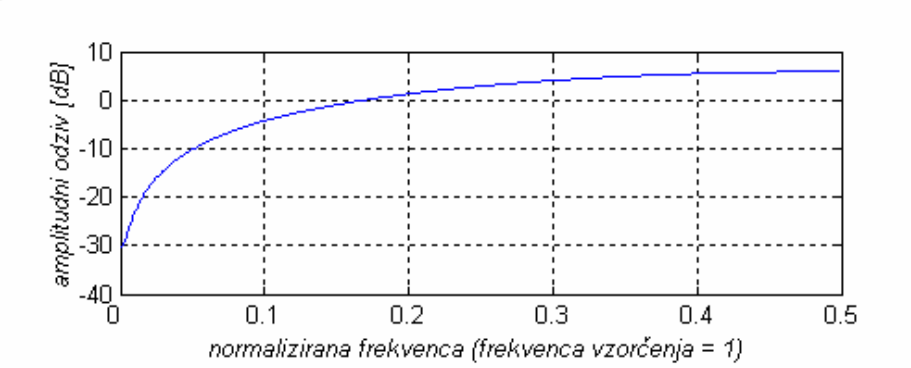
[6] *angl. Mel Frequency Cepstral Coefficients*

2.2.1 *Računanje MFCC koeficientov*

Prvi korak računanja MFCC koeficientov je obdelava govornega signala s predoblikovalnim[7] filtrom. Filter, ki je uporabljen v našem sistemu in se tudi sicer pogosto uporablja, je določen z enačbo

$$x(n) = x'(n) - kx'(n-1). \qquad (1)$$

Z *x'(n)* je označen *n*-ti vzorec v govornem signalu. Vrednost koeficienta *k* je v našem primeru znašala *k*=0,97; običajne vrednosti koeficienta *k* so med 0,9 in 0,98. Osnovni namen uporabe predoblikovalnih filtrov je (bil) zmanjšanje dinamičnega obsega vzorcev v signalu. To je pomembno predvsem pri računanju s celimi števili. Ker večina današnjih SRG uporablja aritmetiko s plavajočo vejico, je vpliv uporabe predoblikovalnega filtra zelo majhen, kljub temu pa se predoblikovalni filtri še vedno uporabljajo. Drug pogled na uporabo predoblikovalnih filtrov je podobnost z lastnostmi človekovega sluha. S predoblikovalnim filtrom se poudarijo visoke frekvence, nizke pa slabijo. Podobne lastnosti ima tudi človekov sluh, ki je najbolj občutljiv na frekvence med 4 in 5 kHz. Amplitudni odziv filtra, ki je določen z (1) in *k*=0,97, je prikazan na sliki 2.



Slika 2: Amplitudni odziv predoblikovalnega filtra, izračunan za vrednost k=0,97.

Po filtriranju s predoblikovalnim filtrom izračunamo spekter *X(l,k)* signala *x(n)* z

[7] *angl. pre-emphasis*

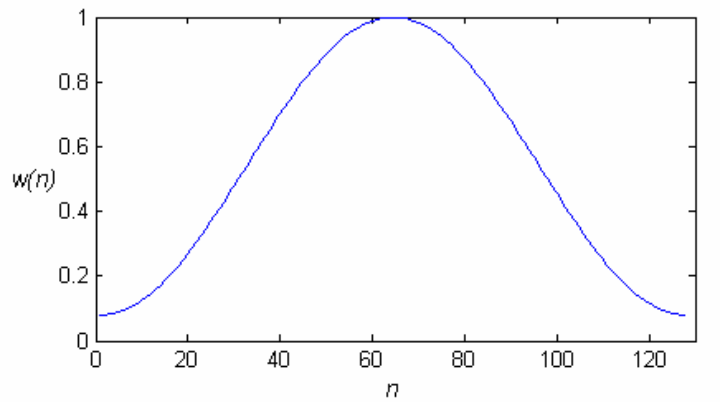
$$X(l,k)=\sum_{n=0}^{\infty}x(n)w(n-lL)e^{-2\pi jkn/N},\qquad k=0,1,...,N-1. \tag{2}$$

Iz enačbe je razvidno, da signal pomnožimo z za *lL* pomaknjeno *okensko funkcijo w(n)*. Osnovna lastnost okenskih funkcij je, da so od 0 različne na končno dolgem intervalu dolžine *N*. Z množenjem z okensko funkcijo torej iz načeloma neskončno dolgega signala dobimo končno dolg *okvir*, v katerem je *N* vzorcev. Vsoto v enačbi (2) lahko zato računamo na končno dolgem intervalu, na katerem je (pomaknjena) okenska funkcija različna od 0. Z *L* je določeno, po koliko vzorcev okvirje pomikamo, s čimer je določena frekvenca okvirjev, ki običajno znaša 100 Hz. Indeks *l* v spektru *X(l,k)* določa številko okvirja. Enačba (2) predstavlja kratkočasovno diskretno Fourierovo transformacijo, ki se zaradi učinkovitega algoritma FFT za računanje spektra najpogosteje uporablja. Spekter je sicer mogoče izračunati tudi z uporabo drugih transformacij. Tako so bili poglavju 5 preizkušeni MFCC koeficienti, pri katerih je bil spekter izračunan z uporabo valčne transformacije.

Pri računanju MFCC koeficientov in tudi sicer se na področju razpoznavanja govora najpogosteje uporablja Hammingovo okno, ki je določeno z enačbo

$$w(n)=\begin{cases}0{,}54-0{,}46\cos\left(\dfrac{2\pi n}{N-1}\right), & n=0,1,...,N-1\\ 0, & sicer\end{cases}. \tag{3}$$

N je dolžina okna, ki je enaka številu vzorcev v okvirju. V enačbi (2) je torej *w(n-lL)* od 0 različna na intervalu [*lL*, *lL*+*N*-1]. Ta interval ustreza *l*-temu okvirju. Na sliki 3 je prikazano Hammingovo okno dolžine *N*=128, kar pri frekvenci vzorčenja f_s=8000 Hz ustreza trajanju 16 ms.



Slika 3: Hammingovo okno dolžine N=128.

V nadaljnjem postopku parametrizacije uporabimo močnostni spekter signala, ki je določen z enačbo

$$P(l,k) = \left|X(l,k)\right|^2, \qquad k = 0,1,...N/2-1. \tag{4}$$

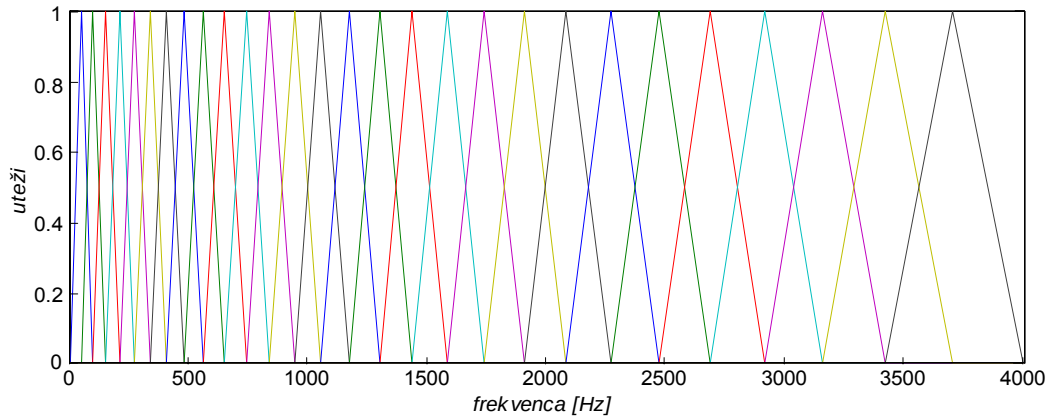
P(l,k) je močnostni spekter *l*-tega okvirja signala *x(n)*. Podan je pri *N*/2 frekvencah (močnostni spekter realnih signalov je sodo simetričen, zato velja *P(l,k)=P(l,N-k)*, *k*=0,1,…,*N*-1), ki so enakomerno razporejene med 0 in $f_s/2$, kjer je f_s vzorčevalna frekvenca.

Človekovo uho je na spremembe frekvence bolj občutljivo pri nižjih frekvencah, pri višjih pa manj. Nelinearno frekvenčno ločljivost človekovega sluha se pri parametrizaciji upošteva tako, da se močnostni spekter pomnoži z množico v frekvenčnem prostoru nelinearno razporejenih filtrov različnih širin. Pri računanju MFCC koeficientov se uporabljajo trikotni filtri, določeni z *Mel-lestvico:*

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{700}\right). \tag{5}$$

Enačba frekvenco *f* preslika v frekvenco *Mel(f)*, ki je sorazmerna z logaritmom *f*. To pomeni, do frekvenčna ločljivost človekovega ušesa pada z logaritmom frekvence.

Običajno uporabimo 20 - 30 trikotnih filtrov. Pri SRG, s katerim so bili izvedeni v disertaciji opisani poskusi, je bilo uporabljenih *F*=28 filtrov, ki so prikazani na sliki 4.



Slika 4: Po Mel-lestvici razporejeni trikotni filtri. Prikazan je primer za F=28 filtrov in f_s=8000 Hz. Širina in razporeditev filtrov posnema nelinearno frekvenčno ločljivost človekovega sluha.

Centralne frekvence filtrov so določene tako, da za vsak par sosednjih filtrov velja

$$Mel\left(f_c(i+1)\right) - Mel\left(f_c(i)\right) = Mel\left(\frac{f_s}{2(F+1)}\right), \qquad i = 1,2,..,F-1. \qquad (6)$$

V enačbi je s $f_c(i)$ označena centralna frekvenca i-tega filtra, F pa je število vseh trikotnih filtrov. Centralne frekvence filtrov s slike 4 so prikazane v tabeli 1.

Filter	Frekv. [Hz]	Filter	Frekv. [Hz]	Filter	Frekv. [Hz]
1	47	11	741	21	2080
2	98	12	839	22	2268
3	152	13	943	23	2470
4	210	14	1055	24	2685
5	272	15	1174	25	2914
6	338	16	1302	26	3160
7	408	17	1437	27	3422
8	483	18	1582	28	3701
9	564	19	1737		
10	650	20	1903		

Tabela 1: Centralne frekvence po Mel-lestvici razporejenih filtrov. Izračunane so za F=28 filtrov in vzorčevalno frekvenco f_s=8000 Hz.

Vsak filter ima vrednost 1 pri svoji centralni frekvenci. Pri centralnih frekvencah sosednjih filtrov ima vrednost 0. Prenosna funkcija *i*-tega trikotnega filtra za i=1,2,…F je torej določena z enačbo

$$TF(i,f)=\begin{cases}\big(f-f_c(i-1)\big)\big(f_c(i)-f_c(i-1)\big), & f_c(i-1)<f<f_c(i)\\ \big(f-f_c(i+1)\big)\big(f_c(i)-f_c(i+1)\big), & f_c(i)<f<f_c(i+1)\\ 0, & sicer\end{cases}. \qquad (7)$$

Izhod *i*-tega filtra se za *l*-ti okvir izračuna po enačbi

$$m(l,i)=\sum_{j=0}^{N/2-1}P(l,j)TF\left(i,j\frac{f_s}{N}\right), \qquad i=1,2,..,F\,. \qquad (8)$$

Iz enačbe (8) je razvidno, da vsak filter predstavlja obteženo vsoto preko ustreznega frekvenčnega pasu (glede na širino trikotnega filtra). Rezultat je *avditorni spekter m(l,i)*, ki predstavlja moč signala govora v frekvenčnih pasovih, ki so razporejeni skladno z nelinearno frekvenčno ločljivostjo človekovega sluha.

Naslednji korak je logaritmiranje avditornega spektra in računanje *diskretne kosinusne transformacije*

$$c(l,j)=\sum_{i=1}^{F}\log\big(m(l,i)\big)\cos\left(\frac{\pi i}{N}\big(i-0.5\big)\right), \qquad j=0,1,...,J\,. \qquad (9)$$

F ustreza številu trikotnih filtrov, *c(l,j)* pa je *j*-ti MFCC koeficient *l*-tega okvirja. Ker je prevajalna funkcija vokalnega trakta razmeroma gladka, so vrednosti log(*m(l,j)*) sosednjih filtrov (filtrov s podobnimi indeksi) medsebojno zelo korelirane. To je neugodno predvsem pri SRG, ki temeljijo na prikritih Markovih modelih.

Z diskretno kosinusno transformacijo se množico koreliranih vrednosti log(*m(l,j)*) preslika v množico bistveno manj koreliranih koeficientov *c(l,j)*. Pri SRG, s katerim so izvedeni vsi v disertaciji opisani poskusi, se kot značilke uporablja prvih 13 (*J*=12) MFCC koeficientov. Ker imajo MFCC koeficienti z višjimi indeksi v primerjavi s koeficienti z nižjimi indeksi zelo majhne

vrednosti, bi se njihove vrednosti v nadaljnjem postopku razpoznavanja praktično zanemarile. Zato se te MFCC koeficiente poudari (izvede se t.i. *cepstral liftering*) z eksponentno funkcijo

$$MFCC(l, j) = e^{kj} c(l, j), \qquad j = 1,2,...,J\,. \tag{10}$$

V enačbi je k konstanta, ki določa faktor, s katerim se pomnoži koeficient $c(l,j)$. Večja vrednost k pomeni, da se MFCC koeficiente z višjimi indeksi bolj poudari. V našem primeru je bila uporabljena vrednost k=0,6. $MFCC(l,j)$ je j-ti poudarjeni MFCC koeficient l-tega okvirja.

2.3 Nevronske mreže kot akustični modeli

Zamisel o nevronskih mrežah se zgleduje po nevronih in povezavah med nevroni v možganih in živčevju živih bitij. Nevronska mreža je množica preprostih modelov[8] [14], s katerimi je modelirano delovanje posameznega nevrona. Ti modeli so s *povezavami* povezani v strukturo. Izhod nevrona je funkcija obtežene vsote njegovih vhodov. Uporabljajo se navzgor in navzdol omejene zvezne funkcije sigmoidne oblike[9]. Nevronska mreža je sposobna vhodni prostor razmejiti tako, da ga klasificira v dane razrede. Na izhodu mreže dobimo za vsak razred verjetnost, da vhodni vektor pripada temu razredu. Razmejitev vhodnega prostora na razrede določimo z *učenjem* nevronske mreže. Za klasifikacijo značilk v hibridnih SRG se običajno uporabljajo nevronske mreže brez rekurzivnih povezav[10]. Najpogosteje se uporabljajo *večnivojski perceptroni* (MLP[11]), pri katerih so nevroni strukturirani v nivoje. Vsi nevroni določenega nivoja so povezani z vsemi nevroni naslednjega višjega nivoja.

[8] *v nadaljevanju nevronov*

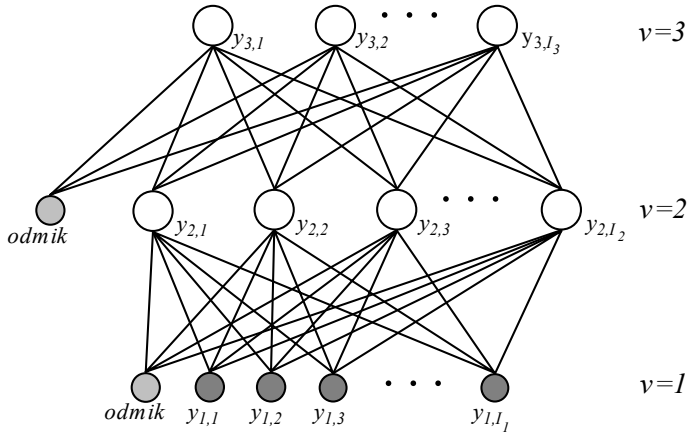
[9] *aktivacijska funkcija*

[10] *angl. feed forward*

[11] *angl. Multi Layer Perceptron*

2.3.1 Večnivojski perceptron

Na področju razpoznavanja govora in v primerih, ko razredi med sabo niso linearno ločljivi, se uporabljajo večnivojski perceptroni z dvema nivojema povezav (slika 5). To pomeni, da ima nevronska mreža tri nivoje nevronov, pri čemer so v prvem nivoju t. i. vhodni nevroni. Ti svoj vhod samo pripeljejo na vhode vseh nevronov drugega nivoja.



Slika 5: Večnivojski perceptron z dvema nivojema povezav.

Prikazana mreža ima V=3 nivoje. Nivo 1 je *vhodni nivo*. Nivo 2 se imenuje *skriti nivo* (želeni izhodi nevronov pri učenju niso znani, ampak se izračunajo - so skriti). Število nevronov v nivoju v je označeno z I_v. Nivo 3 je izhodni nivo; število nevronov I_V v tem nivoju je enako številu razredov, v katere klasificiramo vhodne vektorje. *Odmik* (bias) je dodatni nevron, katerega izhod je vedno enak 1. Uporaba odmika omogoča hitrejše konvergiranje v postopku učenja nevronske mreže.

Vsi nevroni v določenem nivoju v so s povezavami povezani z vsemi nevroni iz predhodnega nivoja v-1. Uteži povezav so označene z w. $w_v(i,j)$ predstavlja utež povezave od nevrona j v nivoju v-1 do nevrona i v nivoju v. Perceptroni s tremi nivoji nevronov ob predpostavki, da je število nevronov dovolj veliko, lahko klasificirajo vhodne vektorje v poljubno določene razrede - sposobni so poljubno kompleksnih preslikav. Zato uporaba večnivojskih perceptronov z več kot tremi nivoji običajno ni smotrna.

Mreža deluje v smeri od nivojev z nižjimi indeksi proti nivojem z višjimi indeksi. Če vhodni vektor označimo z **u**, vektor $\mathbf{y_1}$, ki predstavlja izhod oziroma *aktivacijo* vhodnega nivoja dobimo po enačbi

$$\mathbf{y_1} = \begin{bmatrix} 1 \\ \mathbf{u} \end{bmatrix}. \tag{11}$$

Izhode drugega nivoja izračunamo po enačbah (12) in (13), kjer so z $w_2(i,j)$ označene uteži povezav med nevronoma *i* v skritem in *j* v vhodnem nivoju.

$$x_2(i) = \sum_{j=0}^{I_1} w_2(i,j) y_1(j), \qquad 1 \le i \le I_2 \tag{12}$$

Odmik in ustrezna povezava se v vsoti upoštevata enako kot ostali nevroni. Ker je aktivacija odmika vedno enaka 1, se k obteženi vsoti doda vrednost povezave med odmikom in določenim nevronom. Tako je v postopku učenja mogoče pomakniti obteženo vsoto (12) v bližino vrednosti 0, kjer je odvod aktivacijske funkcije največji. Rezultat je hitrejše konvergiranje mreže v postopku učenja. Aktivacija posameznega nevrona je določena z enačbo

$$y_2(i) = \sigma\big(x_2(i)\big) = \sigma\left(\sum_{j=0}^{I_1} w_2(i,j) y_1(j) \right), \qquad 1 \le i \le I_2 \,. \tag{13}$$

S σ je označena aktivacijska funkcija sigmoidne oblike. Zaradi postopka učenja nevronske mreže je pomembno, da je aktivacijska funkcija odvedljiva in je njen odvod mogoče izraziti z vrednostjo funkcije oziroma aktivacije nevrona. Najpogosteje se uporablja *sigmoidna*[12] funkcija, določena z enačbo (14). Navzdol je omejena z 0, navzgor pa z 1.

$$\sigma\big(x_v(i)\big) = \frac{1}{1+e^{-x_v(i)}}, \quad \begin{matrix} 1 \le v \le V \\ 1 \le i \le I_v \end{matrix}. \tag{14}$$

[12] *tudi logistična*

Aktivacije nevronov izhodnega nivoja se izračunajo na enak način kot pri nevronih skritega nivoja. Vhode nevronov izhodnega nivoja predstavljajo aktivacije nevronov skritega nivoja.

Učenje večnivojskih perceptronov poteka z množico parov vhodnih in izhodnih vektorjev $\{\mathbf{u}, \mathbf{d}\}$. Po opisanem postopku za vhodni vektor **u** izračunamo izhodni vektor $\mathbf{y_V}$, kjer je *V* izhodni nivo. Nato izračunamo *napako* nevronske mreže *E*, ki predstavlja odstopanje izračunanega izhodnega vektorja $\mathbf{y_V}$ od želenega izhodnega vektorja **d**. Če seštejemo napake *E* za vse pare vektorjev v učni množici, dobimo *globalno napako* mreže. Uporablja se več različnih funkcij napake. Najpogostejši sta *srednja kvadratna napaka* in *križna entropija*, določeni z enačbama

$$E = \frac{1}{2}\sum_{i=1}^{I_V}\left(y_V(i) - d(i)\right)^2 \tag{15}$$

in

$$E = -\sum_{i=1}^{I_V}\left[d(i)\log y_V(i) + (1-d(i))\log(1-y_V(i))\right]. \tag{16}$$

Če mrežo učimo kot klasifikator 1-od-*N* (vsak vhodni vektor spada v enega od *N* razredov), aktivacije izhodnih nevronov aproksimirajo pogojne verjetnosti

$$P(r_i \mid \mathbf{u}), \qquad i = 1,2,..,I_V, \tag{17}$$

kjer r_i predstavlja *i*-ti razred, **u** pa vhodni vektor. To velja za vse pogosto uporabljane funkcije napake (srednja kvadratna napaka, križna entropija, McClellandova napaka, itd.) [10].

V postopku učenja iščemo tak nabor uteži $w_v(i,j)$, da je globalna napaka mreže minimalna. Zaradi nelinearnosti, ki jo v mrežo vnašajo nelinearne aktivacijske funkcije, vrednosti uteži ne moremo določiti analitično. Zato smo pri učenju mrež omejeni na gradientne algoritme. To pomeni, da je za vsako povezavo oziroma njeno utež potrebno izračunati gradient $\frac{\partial E}{\partial w_v(i,j)}$ in za določen korak η

popraviti utež $\Delta w_v(i,j)$ tako, da se pomaknemo v nasprotni smeri gradienta (18).

$$\Delta w_v(i,j) = -\eta \frac{\partial E}{\partial w_v(i,j)}, \qquad \begin{matrix} 2 \le v \le V \\ 1 \le i \le I_v \\ 1 \le j \le I_{v-1} \end{matrix} \tag{18}$$

V enačbi je z *v* označen nivo, v katerega vodi povezava, *i* je indeks nevrona v nivoju *v*, *j* pa je indeks nevrona v predhodnem nivoju *v*-1. Z $w_v(i,j)$ je torej označena utež povezave od nevrona z indeksom *j* v nivoju *v-1* do nevrona z indeksom *i* v nivoju *v*.

Tak način popravljanja uteži postopoma minimizira globalno napako mreže *E*. Pri tem postopku je primerno velik korak bistvenega pomena. Korak η je v postopku učenja potrebno postopno zmanjševati [5]. Učenje torej poteka v iteracijah. V vsaki iteraciji se kot vhod mreže vzame vektor **u** in v smeri od vhodov proti izhodom[13] izračuna izhodne vrednosti vseh nevronov. Nato se v smeri od izhodov proti vhodom[14] za vsako utež izračuna parcialni odvod in po enačbi (18) popravi uteži. Za računanje parcialnih odvodov posameznih povezav se uporabi algoritem vzvratnega razširjanja napake [29].

[13] *angl. forward pass*

[14] *angl. backward pass*

3. GOVORNI ZBIRKI

Pri izvedbi poskusov, ki so opisani v nadaljevanju disertacije, sta bili uporabljeni dve govorni zbirki. Prva je zbirka slovenskih števk, ki je bila v našem laboratoriju posneta za potrebe projekta *Razvoj in izdelava sistema za razpoznavanje izoliranih besed slovenskega govora* v letih 1992-1993 [37]. V njej so ločeno izgovorjene slovenske števke. Premori med števkami so dolgi. Druga zbirka vsebuje razmeroma spontane izgovorjave angleških števk. Med števkami so zelo kratki premori, deloma pa so izgovorjene povezano. Zbirki se torej razlikujeta po govornem področju oziroma jeziku, spontanosti in povezanosti govora.

3.1 Govorna zbirka ŠTEVKE

Zbirka je bila posneta preko telefona. Zato vsebuje motnje, ki so bile značilne za analogno telefonsko linijo. Pri snemanju so sodelovali naključno izbrani govorci z območja celotne Slovenije. Pod vodstvom operaterja je vsak govorec izgovoril 13 besed; števke 0 - 9 in besede *"ja"*, *"ne"* in *"stop"*. Posamezen vzorec v zbirki torej sestavlja naključno zaporedje besed, ki jih je izgovoril isti govorec. Vsaka beseda se v vzorcu pojavi natanko enkrat. Med posameznimi besedami je razmeroma dolg premor. V zbirki so enakomerno zastopani govorci različnih starosti in spolov z vseh območij Slovenije. Zato so v izgovorjavah vzorcev prisotna narečja. Zaradi prisotnosti operaterja govorci besed niso izgovarjali spontano, kar je mogoče opaziti pri poslušanju vzorcev.

V zbirki je 780 vzorcev. V vseh vzorcih so označeni začetki in konci besed. Za izvedbo poskusov je bila zbirka razdeljena v 3 množice. 3/5 vzorcev pripada učni množici, po 1/5 pa razvojni in testni množici. V učni množici je na voljo 180 ročno fonemsko označenih vzorcev. Ti vzorci so bili v poskusih uporabljeni za učenje razpoznavalnika, s katerim je bila nato (avtomatsko) fonemsko označena celotna učna množica. Avtomatsko fonemsko označeno učno množico uporabimo za učenje končnega SRG. Z vzorci iz razvojne ovrednotimo uspešnost razpoznavanja z nevronskimi

mrežami, ki jih dobimo v posamezni iteraciji učenja. Izberemo nevronsko mrežo, pri kateri je uspešnost razpoznavanja največja. Testno množico uporabimo za končno oceno uspešnosti sistema. To se oceni z nevronsko mrežo, s katero je bila pri razpoznavanju razvojne množice dosežena največja uspešnost.

Fonemski modeli vseh besed v zbirki so bili določeni skladno s [23]. Dodani so bili premori pred zaporniki, ki so označeni s podčrtajem, npr. *"_t"*. Fonemski modeli so prikazani v tabeli 2.

Beseda	**Model**	**Beseda**	**Model**
nič	{n i _tS tS}	sedem	{s e _d d @ m}
ena	{E n a}	osem	{o s E m }
dve	{d v e}	osem	{o s @ m }
tri	{_t t r i}	devet	{_d d E v e _t t}
štiri	{S _t t i r i}	devet	{_d d @ v e _t t}
pet	{_p p e _t t}	ja	{j a}
sest	{S e s _t t}	ne	{n E}
sedem	{s e _d d E m}	stop	{s _t t O _p p}

Tabela 2: Fonemski modeli besed v zbirki ŠTEVKE.

Učna, razvojna in testna množica so bile izbrane naključno in so disjunktne. Lastnosti govorne zbirke števke so povzete v tabeli 3.

Število vzorcev v zbirki:	780 (v vsakem vzorcu je 13 besed)
Velikost učne množice:	468 vzorcev (180 ročno fonemsko označenih)
Velikost razvojne množice:	156 vzorcev
Velikost testne množice:	156 vzorcev

Tabela 3: Lastnosti govorne zbirke ŠTEVKE.

3.2 Govorna zbirka NUMBERS

V zbirki so na naraven način izgovorjena števila v angleškem jeziku. Izgovorjave števil so dobljene iz posnetkov pogovorov, v katerih so govorci sporočali različne številčne podatke (poštne številke, telefonske številke, itd.). V zbirki so zaporedja izoliranih števk, povezano izgovorjenih števk in izgovorjave večjih števil. Podrobnosti so opisane v [3]. V naših poskusih smo se omejili na razpoznavanje števk 0 - 9.

Tudi ta zbirka je bila posneta v 90. letih preko telefonske linije (digitalne in analogne). Zato so v vzorcih že prisotne za telefonsko linijo značilne motneje. Prav tako je pri nekaterih vzorcih v ozadju slišati motnje kot so glasba, pogovor, dihanje, itd.

V celotni zbirki je 28735 vzorcev. Ker smo se omejili na razpoznavanje števk, je bil uporabljen samo del zbirke. V učni množici je bilo 6428 vzorcev, v katerih je v povprečju po 6 števk. 2705 vzorcev iz učne množice je ročno fonemsko označenih. V razvojni množici je bilo 1111 vzorcev, v testni pa 2287 vzorcev. Zaradi obširnosti zbirke sta bili v večini poskusov (poglavje 6) uporabljeni manjša razvojna in testna množica s po 600 vzorci. Z razvojno in testno množico z večjim številom vzorcev so bili izvedeni poskusi, ki so opisani v poglavju 7.

Fonemski modeli števk v zbirki so povzeti po [3] in so prikazani v tabeli 4.

Beseda	Model	Beseda	Model
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}		

Tabela 4: Fonemski modeli besed v zbirki NUMBERS.

Podatki o velikosti učne, razvojne in testne množice so navedeni v tabeli 5. V primerjavi z govorno zbirko ŠTEVKE je bila uporabljena manjša razvojna množica (merjeno v odstotkih uporabljenih vzorcev). To množico je potrebno razpoznavati z nevronskimi mrežami iz posameznih iteracij učenja, torej večkrat. Zato je bilo zaradi hitrejše izvedbe poskusov, na račun zmanjšanja razvojne

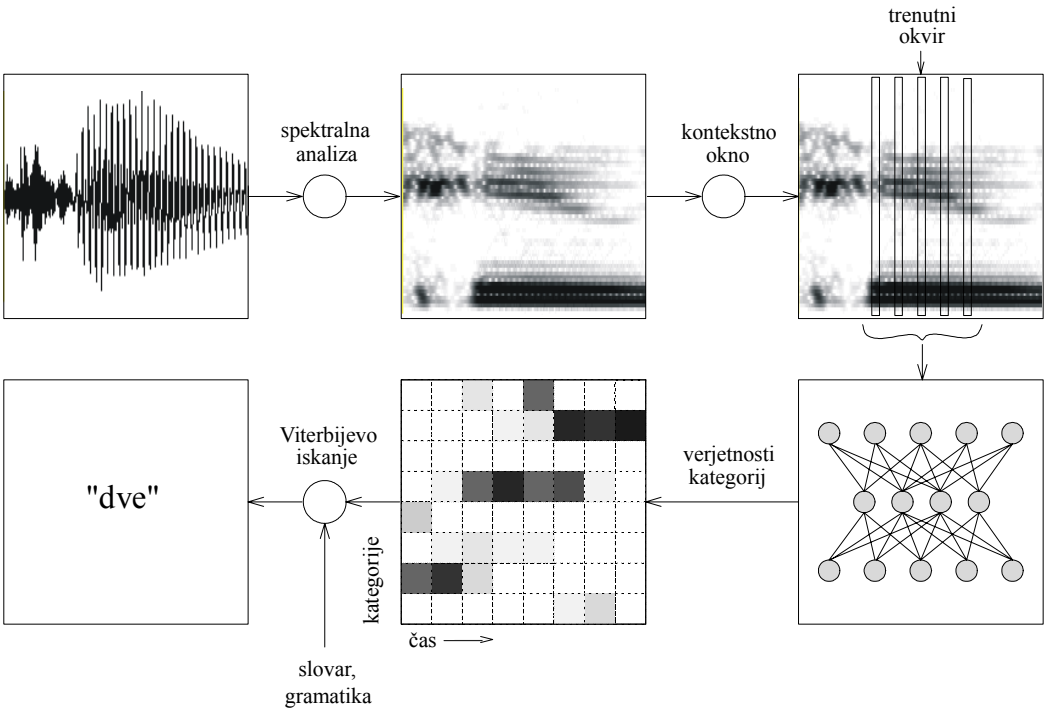
množice, več vzorcev umeščenih v učno in testno množico.

Število uporabljenih vzorcev:	9826 (v vzorcih je v povprečju po 6 števk)
Velikost učne množice:	6428 vzorcev (2705 ročno fonemsko označenih)
Velikost razvojne množice:	1111 vzorcev (v manjših poskusih 600)
Velikost testne množice:	2287 vzorcev (v manjših poskusih 600)

Tabela 5: Lastnosti uporabljenega dela govorne zbirke NUMBERS.

4. REFERENČNI SISTEM ZA RAZPOZNAVANJE GOVORA

V tem poglavju je opisan referenčni SRG, ki je bil uporabljen za izvedbo v disertaciji opisanih poskusov. SRG je narejen z orodjem CSLU-Toolkit [43, 49]. Njegova zgradba je prikazana na sliki 6.



Slika 6: Shematski prikaz referenčnega SRG. Slika je povzeta po [4].

Zgradba SRG se sklada z osnovno zgradbo večine SRG (podobnost z osnovno zgradbo s slike 1). Najprej se izvede spektralna analiza govornega signala. Na osnovi spektra se za vsak okvir izračuna vektor značilk. Vektorji značilk petih okvirjev predstavljajo vhod nevronske mreže, s katero je izvedeno akustično modeliranje. Za vsak okvir se izračunajo verjetnosti, da okvir ustreza določenemu akustičnemu modelu. Nato se v matriki verjetnosti z uporabo Viterbijevega iskanja poišče najverjetnejša beseda, ki jo signal predstavlja.

4.1 Parametrizacija

Sistem kot značilke uporablja MFCC koeficiente, ki opisujejo spektralno ovojnico signala govora (prilagojeno lastnostim govornega signala in človekovega zaznavanja zvoka). Podrobneje so opisani v poglavju 2.2.1.

Razen trenutnih lastnosti spektra, so za uspešno razpoznavanje govora pomembni tudi podatki o načinu oziroma trendu spreminjanja spektra. Zato je poleg MFCC koeficientov v vektor značilk smiselno vključiti tudi t.i. *delta značilke*, ki so časovni odvod MFCC koeficientov. Vektor značilk za *l*-ti okvir predstavlja prvih 13 MFCC koeficientov, določenih z (10) in njihovi delta koeficienti, ki so določeni z enačbo

$$delta(l,i) = \frac{\sum_{j=1}^{2} j\big(MFCC(l+j,i) - MFCC(l-j,i)\big)}{2\sum_{j=1}^{2} j^2}, \qquad i = 0,1,..,12\,. \tag{19}$$

Z *delta(l,i)* je označen delta koeficient *i*-tega MFCC koeficienta v *l*-tem okvirju. Odvisen je od *i*-tih MFCC koeficientov dveh predhodnih in dveh naslednjih okvirjev. Dolžina vektorja značilk je torej 26.

Pri referenčnem SRG so okvirji dolgi 16 ms. Koeficient predoblikovalnega filtra ima vrednost 0,97. Kot okenska funkcija je uporabljeno Hammingovo okno. Razmik med okvirji znaša 10 ms (frekvenca okvirjev 100 Hz), kar pomeni, da se okvirji prekrivajo za 60%. Za računanje avditornega spektra je uporabljenih F=28 trikotnih filtrov, ki so razporejeni po Mel-lestvici. MFCC koeficienti so poudarjeni po metodi *cepstral liftering* s koeficientom 0,6. Koeficientom je odšteta njihova srednja vrednost, ki je izračunana preko celotnega vzorca (CMS[15]). To v določeni meri zmanjša vpliv motenj prenosnega kanala (v našem primeru telefonske linije).

[15] *angl. Cepstral Mean Subtraction*

4.2 Akustično modeliranje

Pri referenčnem SRG je za akustično modeliranje uporabljen trinivojski perceptron. Vhod v mrežo je kontekstno okno značilk. Kontekstno okno, poleg vektorja značilk trenutnega okvirja, sestavljajo še vektorji značilk okvirjev, ki so od trenutnega odmaknjeni za -60, -30, 30 in 60 ms, skupno torej 5 vektorjev značilk. Nevronska mreža kontekstna okna klasificira v razrede, ki ustrezajo kontekstno odvisnim kategorijam oziroma delom fonemov. Število izhodov mreže je enako številu razredov, torej mreža deluje kot klasifikator 1-od-*N*. Aktivacije izhodnih nevronov ustrezajo pogojnim verjetnosti $P(r_i|\mathbf{u})$, kjer r_i predstavlja *i*-ti razred in **u** vhodni vektor (kontekstno okno značilk, ima torej 26*5=130 elementov).

V vzorcih se lahko pojavijo tudi besede in z njimi povezane kategorije, ki jih ni v slovarju. Take kategorije so modelirane z dodatno kategorijo .*garbage*. To ni običajna kategorija, ki ji ustreza izhodni nevron nevronske mreže, pač pa se njena pogojna verjetnost oceni kot povprečna vrednost nekaj najverjetnejših kategorij. Izračuna se torej povprečna aktivacija določenega števila nevronov, pri katerih aktivacija presega določeno mejo. Pristop temelji na predpostavki, da bodo nepoznane kategorije podobne več poznanim kategorijam. Posledično bodo v tem primeru aktivacije večjega števila nevronov razmeroma visoke in bo visoka tudi njihova povprečna vrednost. Število aktivacij, ki so zajete v povprečju za računanje aktivacije .*garbage*, običajno znaša 5 - 10.

4.2.1 Kontekstno odvisne kategorije

Akustične značilnosti določenega fonema so zaradi koartikulacije odvisne od sosednjih fonemov, torej predhodnega in naslednjega sosednjega fonema. Vsak fonem zato lahko nastopa v več različnih oblikah, ki se imenujejo *alofoni*. Pri referenčnem SRG je uporabljen nekoliko drugačen pristop, ki temelji na predpostavki, da od predhodnega fonema ni kontekstno odvisen celoten fonem, ampak samo njegov začetni del. Prav tako na zadnji del fonema vpliva naslednji fonem. Zato se pri referenčnem SRG fonemi glede na njihovo trajanje delijo na *bifone*, ki so modelirani z dvema kontekstno odvisnima kategorijama in *trifone*, ki so modelirani s tremi kategorijami. Pri bifonih sta prva in zadnja kategorija kontekstno odvisni od predhodnega oziroma naslednjega fonema. Trifoni imajo še srednjo kategorijo, ki je kontekstno neodvisna.

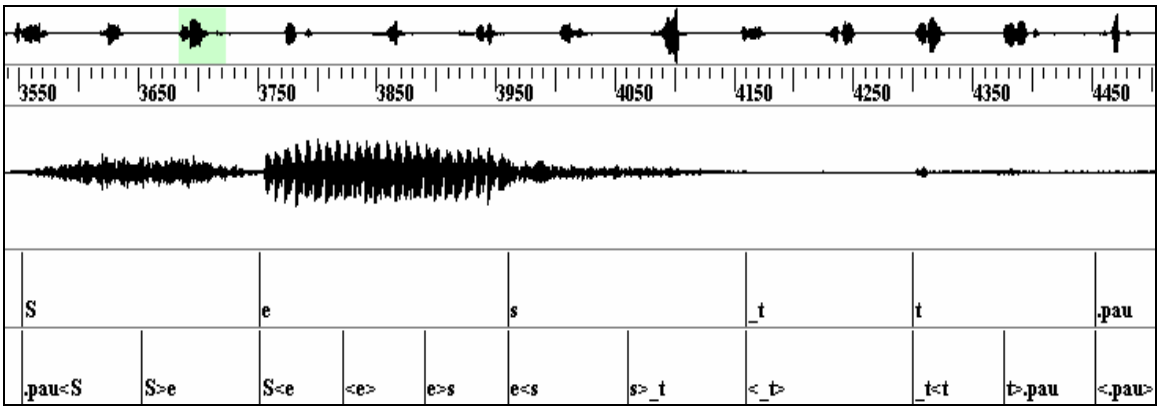
Kot *monofoni* oziroma ena kontekstno neodvisna kategorija so modelirani premori (tišina) med besedami in tišina, ki nastane pri izgovorjavi zapornikov, ko je vokalni trakt za kratek čas zaprt. Z bifoni so modelirani kratki fonemi.

V besedah iz zbirke ŠTEVKE je 18 fonemov. Dodanih je še 5 monofonov, s katerim je modeliran premor med besedami *.pau* in tišina, do katere pride pri izgovorjavi zapornikov /p/, /t/, /d/ in /č/, ko se vokalni trakt za krajši čas zapre (*_p, _t, _d, _ts*). Skupaj s podatki o številu kontekstno odvisnih kategorij, s katerimi so fonemi modelirani, so prikazani v tabeli 6.

Fonem	Št. kategorij	Fonem	Št. kategorij	Fonem	Št. kategorij
.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 6: Razdelitev fonemov iz zbirke ŠTEVKE v kontekstno odvisne kategorije. S številom kategorij je določeno, ali je posamezen fonem modeliran kot monofon(1), bifon (2) oziroma trifon (3).

Na sliki 7 je kot primer prikazana razdelitev v kontekstno odvisne kategorije za besedo *"šest"*. Prvi fonem v besedi je /š/, ki je označen s *S*. Ker je bifon, je modeliran z dvema kategorijama. Leva je kontekstno odvisna od predhodnega fonema *.pau*. Leve kontekstno odvisne kategorije so označene z zaporednima fonemoma, med katerima je znak <. Levo kontekstno odvisno kategorijo fonema S torej označimo s *.pau<S*. Podobno velja tudi za desne kontekstno odvisne kategorije, le da je pri njih uporabljen znak >. Tako dobimo naslednjo kategorijo *S>e*. Drugi fonem v besedi je /e/. Podobno kot pri fonemu označenem s *S*, dobimo levo in desno kontekstno odvisno kategorijo; *S<e* in *e>s.* Med njima je kontekstno neodvisna kategorija *<e>*. Če opisani postopek nadaljujemo do konca besede, dobimo naslednje zaporedje kategorij: {*.pau<S, S>e, S<e, <e>, e>s, e<s, s>_t, <_t>, _t<t, t>.pau*}.



Slika 7: Segmentacija besede "šest". Prikazan je signal, ročno določene meje fonemov in razdelitev fonemov na kontekstno odvisne kategorij.

Če podobno naredimo za vse besede v slovarju, dobimo 85 (v splošnem) kontekstno odvisnih kategorij. Prikazane so v tabeli 7.

<.pau>	d<@	s<@	@>m	@>v	.pau<E	d<E	n<E	s<E
<E>	E>.pau	E>m	E>n	E>v	t<O	<O>	O>_p	.pau<S
S>_t	S>e	<_d>	<_p>	<_t>	<_tS>	j<a	n<a	<a>
a>.pau	_d<d	d>@	d>E	d>v	S<e	p<e	s<e	v<e
<e>	e>.pau	e>_d	e>_t	e>s	n<i	r<i	t<i	<i>
i>.pau	i>_tS	i>r	.pau<j	j>a	@<m	E<m	m>.pau	.pau<n
E<n	n>E	n>a	n>i	.pau<o	<o>	o>s	_p<p	p>.pau
p>e	i<r	t<r	r>i	.pau<s	e<s	o<s	s>@	s>E
s>_t	s>e	_t<t	t>.pau	t>O	t>i	t>r	_tS<tS	tS>.pau
@<v	E<v	d<v	v>e					

Tabela 7: Kontekstno odvisne kategorije, ki nastopajo v zbirki ŠTEVKE.

V zbirki NUMBERS najdemo 22 različnih fonemov (tabela 8). Dodan je monofon *.pau*, s katerim je modeliran presledek med besedami.

Fonem	Št. kategorij	Fonem	Št. kategorij	Fonem	Št. kategorij
.pau	1	w	2	\>r	3
uc	1	l	2	i:	3
f	2	ks	2	u	3
v	2	^2	2	ei	3
T	2	^3	2	al	3
s	2	^	2	oU	3

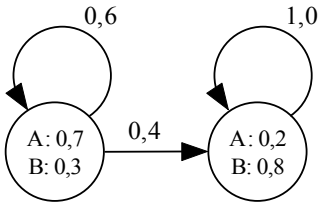
z	2	9r	2	th	r
n	2	E	2		

Tabela 8: Razdelitev fonemov iz zbirke NUMBERS v kontekstno odvisne kategorije. S številom kategorij je določeno, ali je posamezen fonem modeliran kot monofon(1), bifon (2) oziroma trifon (3). Oznaka r pri fonemu /th/ pomeni, da ta fonem ni predstavljen s svojo kategorijo, je pa od njega kontekstno odvisna prva kategorija naslednjega fonema.

V zbirki NUMBERS je 217 kontekstno odvisnih kategorij. Razlog za večje število je v povezanem izgovarjanju števk. V tem primeru med števkami ni vedno kategorije *.pau*, zato se zadnji fonem predhodne števke stika s prvim fonem trenutne števke. Ker je zaporedje števk naključno, se to odrazi v bistveno večjem številu kontekstno odvisnih kategorij.

4.3 Časovni modeli

Časovni modeli, ki so uporabljeni pri referenčnem SRG, se zgledujejo po levo-desnih prikritih Markovih modelih. Prikriti Markov model je množica stanj povezanih z usmerjenimi povezavami. Povezave določajo možne prehode med stanji (slika 8). Vsak model ima neko začetno stanje (pri modelu na sliki je to levo stanje), nato pa v vsakem diskretnem koraku prehajamo med stanji. Pri prehajanju je možen tudi prehod nazaj v isto stanje (zanke na sliki). Takoj po prehodu se v trenutnem stanju z določeno verjetnostjo odda izhodni simbol A ali B. Prikriti Markov model si predstavljamo kot zaprto celoto, pri kateri lahko opazujemo samo zaporedje izhodnih simbolov, kako pri tem prehajamo med stanji pa ne (zaporedje stanj je prikrito).



Slika 8: Levo-desni prikriti Markov model. Ima dve stanji, ki lahko oddata izhodna simbola A in B.

Na področju razpoznavanja govora se prikriti Markovi modeli uporabljajo tako, da stanja predstavljajo odseke govora. Pri prehodu v vsako stanje se z določeno verjetnostjo odda nek izhodni simbol (npr. kontekstno odvisno kategorijo). Prehodi med stanji torej določajo možna zaporedja fonemov oziroma kontekstno odvisnih kategorij. Pri levo-desnih Markovih modelih (slika 8) v vsakem stanju obstajata samo dva možna prehoda: nazaj v isto stanje in prehod v naslednje stanje. Težava takega pristopa je neustrezno modelirano trajanje določenega izhodnega simbola. Ker stanja v modelih predstavljajo odseke govora, je pri daljšem trajanju nekega simbola (npr. kontekstno odvisne kategorije, ki je del daljšega fonema) potrebno v istem stanju ostati več diskretnih korakov. Trajanje fonemov je torej modelirano s prehodi v isto stanje oziroma množenjem z določenim faktorjem (npr. 0,6 v prvem stanju modela na sliki). Verjetnost, da po določenem številu korakov ostanemo v istem stanju zato s časom eksponentno pada, to pa ne ustreza verjetnostni porazdelitvi trajanja posameznega odseka govora.

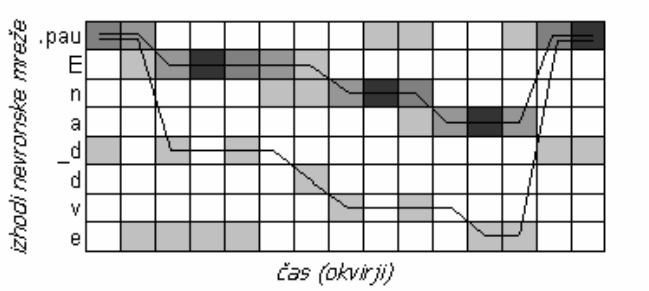
Pri referenčnem SRG je trajanje simbolov oziroma kategorij modelirano glede na najmanjše in največje trajanje te kategorije v učnih vzorcih. Če je trajanja določene kategorije v razpoznavanem vzorcu krajše od spodnje oziroma daljše od zgornje meje, se aktivacijo ustreznega izhodnega nevrona zmanjša za faktor, ki je sorazmeren z odstopanjem od meje. Časovno ujemanje razpoznavanega vzorca s časovnim modelom se oceni z Viterbijevim algoritmom [17, 27].

4.3.1 Viterbijev algoritem

Iz nevronske mreže kot rezultat dobimo matriko dimenzij $O*I_V$, v kateri so za vsak okvir izračunane verjetnosti, da okvir pripada posameznim kategorijam (17). Pri tem je I_V število vseh kategorij in O število okvirjev, na katere se razreže govorni signal. Zaporedja kategorij znotraj posameznih besed so določena s fonemsko zgradbo besed in modeliranjem fonemov (monofoni, bifoni, trifoni). Naloga Viterbijevega iskanja je, da glede na matriko verjetnosti za vse možne besede izračuna verjetnost, da govorni signal, za katerega je matrika izračunana, predstavlja določeno besedo.

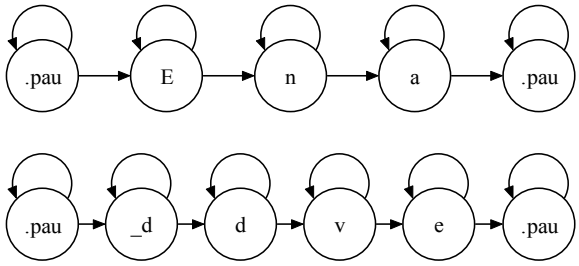
Najprej se glede na fonemsko zgradbo določijo veljavna zaporedja kontekstno odvisnih kategorij. Nato se ta zaporedja primerja z verjetnostmi v matriki. Prehod iz določene kategorije v naslednjo se izvede, ko ima v matriki naslednja kategorija večjo verjetnost od trenutne. Pri tem se upošteva tudi trajanje kategorije. Če je trajanje kategorije v matriki prekratko ali predolgo, se njeno verjetnost

zmanjša sorazmerno z odstopanjem od spodnje oziroma zgornje meje trajanja. Primer Viterbijevega iskanja skozi matriko za preprost razpoznavalnik, je prikazan na sliki 9.



Slika 9: Primer matrike verjetnosti in Viterbijevega iskanja.

Na sliki je skica matrike verjetnosti za primer razpoznavalnika, ki opisuje spreminjaje verjetnosti kontekstno odvisnih kategorij *{E, n, a, _d, d, v, e}* s časom. Poleg matrike potrebujemo še množico veljavnih časovnih zaporedij kategorij, ki so pri referenčnem SRG določene glede na fonemsko zgradbo besed. Za razpoznavanje besed *"ena"* in *"dve"*, ki sta modelirani z zaporedji monofonov *{E n a}* in *{_d d v e}*, sta možni zaporedji kategorij[16] prikazani na sliki 10.



Slika 10: Možni zaporedji kategorij za besedi "ena" in "dve".

V predstavljenem primeru je iskanje preprosto, saj razpoznavamo samo dve besedi, pred in za katerima je premor (*.pau*). Pri iskanju pri prehodu skozi matriko s slike 9 preidemo v novo stanje modela s slike 10, če je v matriki verjetnost nove kategorija večja od vrednosti trenutne kategorije.

---

[16] Pri poskusih opisanih v disertaciji smo razpoznavali zaporedja besed, zato je pri določitvi veljavnih časovnih zaporedjih kategorij poleg fonemskih modelov besed potrebno upoštevati še gramatiko oziroma veljavna zaporedja besed.

Ko pridemo do konca matrike, dobimo verjetnost najverjetnejšega zaporedja kategorij in ustrezno pot skozi matriko. Na sliki 9 je bolj verjetno (temnejši kvadrati na poti) zaporedje kategorij, ki ustrezajo besedi *"ena"*. Podrobnejši opis postopka lahko najdemo v [17], Viterbijev algoritem pa je na kratko povzet v dodatku D.

4.4 Postopek učenja

Za učenje in ocenitev uspešnosti SRG so potrebne tri množice vzorcev: učna, razvojna in testna. S pomočjo vzorcev iz učne množice dobimo časovne in akustične modele. Razvojno množico se uporablja za vrednotenje uspešnosti SRG v postopku učenja. S testno množico se oceni uspešnost dobljenega SRG (SRG, ki je bil najuspešnejši pri razpoznavanju razvojne množice). Pri tem v postopku učenja ni dovoljeno uporabiti nobenega vzorca iz testne množice.

Postopek učenja poteka v dveh korakih. V prvem koraku se učenje uporabi učna množica z ročno fonemsko označenimi vzorci. Izvede se določeno število (30) iteracij učenja, nakar se s pomočjo razvojne množice izbere najuspešnejši SRG. Ta SRG nato uporabimo za avtomatsko označevanje celotne učne množice na nivoju fonemov. Večja učna množica je ročno označena na besednem nivoju, zato se SRG uporabi le za razmejitev posameznih besed na foneme. Rezultat je na nivoju fonemov avtomatsko označena večja učna množica, katero se v drugem koraku uporabi za učenje dokončnega SRG. V postopku učenja torej dobimo dva SRG. Učenje obeh poteka enako, razlika je le v učni množici: pri učenju prvega se uporabi manjša, ročno fonemsko označena učna množica, pri učenju dokončnega pa večja, avtomatsko označena učna množica. Učenje vsakega od obeh SRG se deli v dva dela, ki sta opisana v nadaljevanju.

4.4.1 Tvorba časovnih modelov

Časovni modeli pri referenčnem sistemu so določeni s fonemsko zgradbo besed in preprosto gramatiko, ki dovoljuje poljubno dolga naključna zaporedja števk. Pri tem je upoštevano tudi trajanje posameznih kategorij oziroma delov fonemov. Za vsako kategorijo je v učni množici potrebno poiskati največje trajanje te kategorije t_{max} in takšno trajanje t_{min}, da je v učni množici pri 5% pojavitev te kategorije trajanje krajše od t_{min}. Obe meji se poišče s prehodom skozi fonemske

označbe učne množice. Taka določitev obeh mej je bila izbrana glede na rezultate poskusov s podobnim SRG [12].

4.4.2 Učenje večnivojskega perceptrona

Vhod nevronske mreže, ki je pri SRG uporabljena za akustično modeliranje, predstavlja *kontekstno okno*, v katerem so vektorji značilk petih okvirjev: trenutnega in okvirjev oddaljenih za ±30 in ±60 ms. Ker znaša dolžina posameznega vektorja značilk 26 (podpoglavje 4.1), ima nevronska mreža 130 vhodnih nevronov (in odmik, ki se doda samodejno). V drugem oziroma skritem nivoju je 200 nevronov. S številom nevronov v skritem nivoju je določeno število prostih parametrov mreže. To mora biti primerno veliko in je odvisno tudi od števila primerov, na katerih nevronsko mrežo učimo. Če je število prostih parametrov premajhno, se mreža učnim podatkom preveč prilagodi in ni sposobna posploševanja. S posploševanjem je mišljeno, da se nevronska mreža na nek vhodni vektor (ki ni enak nobenemu vhodnemu vektorju iz učne množice) odzove na način, ki je najbližji glede na "izkušnje", ki jih je mreža dobila v postopku učenja. Število izhodnih nevronov je enako številu kontekstno odvisnih kategorij. Odvisno je od uporabljene govorne zbirke in pri zbirki ŠTEVKE znaša 85, pri zbirki NUMBERS pa 217. Večnivojski perceptron je popolnoma povezan.

Za učenje večnivojskega perceptrona potrebujemo množico parov vhodnih vektorjev in želenih izhodnih vektorjev. Vhodni vektorji so sestavljeni iz kontekstnih oken vektorjev značilk petih okvirjev. V izhodnih vektorjih je pri izhodnem nevronu, ki ustreza pravilni kategoriji vrednost 1, pri ostalih izhodnih nevronih pa vrednost 0 (klasifikacija 1-od-*N*). Pare vhodnih in izhodnih vektorjev dobimo iz učne množice, označene na nivoju kontekstno odvisnih kategorij. Pari vhodnih in izhodnih vektorjev so izbrani naključno, pri čemer je največje število parov za posamezno kategorijo omejeno na 4000 (če je za posamezno kategorijo na voljo več parov, smo jih uporabili samo 4000). Za nekatere kategorije, ki se v učni množici pojavijo zelo redko, je število parov, ki so na voljo, manjše. Problem manjše zastopanosti redkih kategorij se omili s prilagajanjem koraka učenja. Pri redkih kategorijah se uporabi večji korak (podrobnosti o postopku so opisane v [28]).

Učenje poteka v iteracijah. Začetne vrednosti uteži nevronske mreže se izbere naključno. V vsaki iteraciji se izvede prehod preko celotne množice učnih vektorjev in ustrezno popravi uteži. Nato se korak učenja zmanjša in izvede naslednjo iteracijo učenja (v naših poskusih smo izvedli 40 iteracij učenja). Rezultat je torej 40 nevronskih mrež, po ena iz vsake izmed iteracij.

Sledi evaluacija nevronskih mrež s pomočjo razvojne množice. V naših poskusih so bile z razvojno množico preizkušene nevronske mreže iz iteracij 20 - 40. V dokončnem SRG se uporabi nevronsko mrežo iz najuspešnejše iteracije. Uspešnost SRG se določi z razpoznavanjem testne množice. Ko se SRG preizkusi s testno množico, ga ni več dovoljeno spreminjati.

5. FREKVENČNA ANALIZA GOVORNEGA SIGNALA

V tem poglavju je podrobneje predstavljena frekvenčna analiza signalov, ki je bistveni del parametrizacije oziroma računanja MFCC koeficientov. Predstavljeni sta kratkočasovna Fourierova transformacija in *večločljivostna* (multiresolucijska) spektralna analiza z valčno transformacijo. Prednost Fourierove transformacije je, da jo lahko računamo z učinkovitim algoritmom FFT za hitro računanje transformacije, poglavitna slabost pa frekvenčno-časovna ločljivost spektra, ki je določena z dolžino okvirja in je zato fiksna. Težavnost izbire primerno dolgega okvirja je bila predstavljena že v uvodnem poglavju. Razlog je predvsem v zgradbi govornega signala, ki ima na različnih časovnih odsekih zelo različne lastnosti. Tako pri spektralni analizi na nekaterih odsekih prevlada potreba po dobri frekvenčni ločljivosti spektra, na drugih je pomembnejša časovna ločljivost. Pri nekaterih SRG se težavo poskuša omiliti z uporabo večločljivostne spektralne analize, najpogosteje na osnovi valčne transformacije. Pri spektru, ki ga izračunamo na ta način, je frekvenčno-časovna ločljivost odvisna od frekvence. Pri višjih frekvencah je boljša časovna, pri nižjih pa frekvenčna ločljivost. To pomeni, da se časovno kratke spremembe v spektru ne bodo povsem zabrisale, če so prisotne preko celotnega frekvenčnega območja, torej tudi pri višjih frekvencah. Do takih sprememb spektra pride npr. pri izgovorjavi zapornikov, zato jih je mogoče časovno bolje opredeliti kljub temu, da je pri nižjih frekvencah časovna ločljivost spektra slabša. Razlog za uporabo večločljivostne spektralne analize je torej podoben kot za dinamično prilagajanje frekvenčno-časovne ločljivosti spektra. Z obema pristopoma želimo doseči dobro časovno ločljivost za natančno določitev hitrih sprememb spektra, ob tem pa zagotoviti dovolj dobro frekvenčno ločljivost. Pri valčni transformaciji to dosežemo z dobro časovno ločljivostjo pri višjih frekvencah (pri izgovorjavi zapornikov pride do sprememb spektra tudi pri višjih frekvencah) in dobro frekvenčno ločljivost pri nižjih frekvencah, kar je skladno s človekovim sluhom. Pri dinamičnem prilagajanju poskušamo določiti odseke, na katerih je potrebna dobra časovna ločljivost in odseke, na katerih prevlada potreba po dobri frekvenčni ločljivosti. Zato so v tem poglavju opisani MFCC koeficienti, ki temeljijo na valčni transformaciji. Podani so tudi rezultati primerjave tako izračunanih MFCC koeficientov s klasičnimi MFCC koeficienti, pri katerih se spekter izračuna z uporabo kratkočasovne Fourierove transformacije.

Pri frekvenčni analizi govornega signala je smiselno upoštevati spoznanja o človekovem slušnem zaznavanju zvoka. V zvezi s spektralno analizo je s tem mišljena predvsem nelinearna frekvenčna ločljivost človekovega sluha. To se pri računanju MFCC koeficientov upošteva z uporabo po Mel-lestvici razporejenih filtrov. Pri višjih frekvencah, kjer so ti filtri široki, dobra frekvenčna ločljivost izračunanega spektra ni potrebna. To pa pomeni, da je mogoče pri frekvenčni analizi višjih frekvenc uporabiti krajši okvir in izboljšati časovno ločljivost spektra. Frekvenčna ločljivost je zato pri višjih frekvencah slabša, vendar ob dejstvu, da v tem frekvenčnem območju slabša tudi frekvenčna ločljivost človekovega sluha, sprejmemo predpostavko, da slabša frekvenčna ločljivost na uspešnost razpoznavanja govora ne vpliva bistveno. Pri nižjih frekvencah je potrebno uporabiti daljši okvir, da ima spekter zadostno frekvenčno ločljivost. Želimo torej, da bi bila dolžina okvirja odvisna od frekvence. Podoben učinek dosežemo, če spekter izračunamo z uporabo valčne transformacije.

V zadnjem delu poglavja smo pri računanju MFCC koeficientov uporabili kratkočasovno diskretno Fourierovo transformacijo, zvezno valčno transformacijo in paketno valčno transformacijo, ki je različica diskretne valčne transformacije. Za računanje zvezne in paketne valčne transformacije sta bila uporabljena Haarov valček in valček D^4 iz družine Daubechies. Za računanje zvezne valčne transformacijo je bil uporabljen tudi Morletov valček, ki se na področju razpoznavanja govora pogosto uporablja, ni pa primeren za diskretno in paketno valčno transformacijo. MFCC koeficienti na osnovi zvezne valčne transformacije in MFCC koeficienti na osnovi paketne valčne transformacije so opisani v nadaljevanju poglavja. Od MFCC koeficientov na osnovi kratkočasovne diskretne Fourierove transformacije se razlikujejo samo po uporabljeni metodi spektralne analize, vsi ostali parametri so enaki. Podrobnosti o računanju MFCC koeficientov na osnovi kratkočasovne diskretne Fourierove transformacije so bile opisane v že poglavju 2.2.1. Na koncu poglavja je podana primerjava vpliva uporabe različnih tipov MFCC koeficientov na uspešnost razpoznavanja govora v odvisnosti od tipa uporabljenih MFCC koeficientov.

5.1 Kratkočasovna Fourierova transformacija

Zaradi pogostosti uporabe pri frekvenčni analizi govornega signala v postopku parametrizacije, je najprej podrobneje predstavljena kratkočasovna Fourierova transformacija. Poudarek je na vplivu

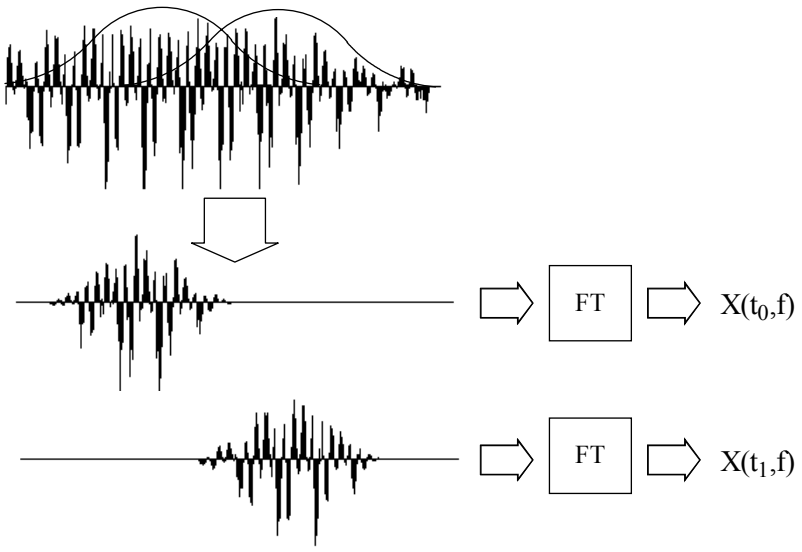
dolžine okvirja na izračunani spekter oziroma na frekvenčno-časovni ločljivosti izračunanega spektra.

Fourierova transformacija je določena z enačbo:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi j f t} dt \,. \tag{20}$$

Z *X(f)* je označen spekter signala oziroma porazdelitev frekvenčnih komponent signala *x(t)* v frekvenčnem prostoru. Ker je rezultat *X(f)* frekvenčna porazdelitev za celoten signal *x(t)*, z njo niso opisane časovne spremembe spektra po posameznih odsekih signala, ki so za razpoznavanje govora bistvene.

Pri razpoznavanju govornega signala nas zanima predvsem, kako se spekter signala spreminja s časom. Da s pomočjo Fourierove transformacije dobimo spekter v odvisnosti od časa, je potrebno signal *x(t)* razdeliti na časovno omejene odseke, ki se imenujejo okvirji. To se doseže z množenjem signala z okensko funkcijo *w(t)*, ki je od nič različna samo na končno dolgem časovnem odseku, ki je enak dolžini okvirja. Okno oziroma okvir se pomika v času in z uporabo Fourierove transformacije izračuna spekter za vsak okvir posebej. Postopek je prikazan na sliki 11.



Slika 11: Računanje kratkočasovne Fourierove transformacije.

Če Fourierovo transformacijo (na sliki označeno s FT) uporabljamo tako kot je prikazano na sliki, dobimo kratkočasovno Fourierovo transformacijo, ki je določena z enačbo

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-2\pi j f t} d\tau \,. \tag{21}$$

Kratkočasovna Fourierova transformacija preslika signal *x(t)* v dvodimenzionalni časovno-frekvenčni prostor. Tako oblika okenske funkcije *w(t)* kot dolžina okvirja imata pomemben vpliv na izračunani spekter. Kot okenske funkcije se najpogosteje uporabljajo simetrične funkcije zvončaste oblike. Med njih spada tudi Hammingovo okno (3), ki je bilo uporabljeno pri vseh poskusih v tej disertaciji. Časovna ločljivost Δt in frekvenčna ločljivost Δf tako izračunanega spektra sta določeni z dolžino okvirja in obliko okenske funkcije, kot je opisano v nadaljevanju.

Ob dani okenski funkciji *w(t)* in njenem Fourierovem transformu *W(f)* se kot mera za frekvenčno ločljivost običajno uporablja *srednja kvadratna pasovna širina*, ki je določena z enačbo

$$\Delta f^2 = \frac{\int_{-\infty}^{\infty} f^2 |W(f)|^2 df}{\int_{-\infty}^{\infty} |W(f)|^2 df} \,. \tag{22}$$

Ustrezna mera za časovno ločljivost je *srednja kvadratna časovna širina:*

$$\Delta t^2 = \frac{\int_{-\infty}^{\infty} t^2 |w(t)|^2 dt}{\int_{-\infty}^{\infty} |w(t)|^2 dt} \,. \tag{23}$$

Frekvenčna in časovna ločljivost sta med seboj povezani. Njun produkt je navzdol omejen s Heisenbergovo nedoločenostjo:

$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi}. \tag{24}$$

To pomeni, da sta frekvenčna in časovna ločljivost obratno sorazmerni; če se ena izboljša (manjša vrednost Δt oziroma Δf), se druga poslabša. Pri izboru okenske funkcije poleg frekvenčno-časovne ločljivosti pomembno vlogo igra še več dejavnikov, med katerimi je zelo pomembno npr. spektralno puščanje. Zato je izbor okenske funkcije odvisen od področja uporabe.

Časovna ločljivost se za Hammingovo okno izračuna tako, da se v enačbi (23) upošteva

$$w(t) = \begin{cases} 0{,}54 - 0{,}46\cos(2\pi t / T), & 0 \leq t \leq T \\ 0, & sicer \end{cases}, \tag{25}$$

kjer je s T označena dolžina okna, ta pa je enaka dolžini okvirja. Ker je Hammingovo okno pozitivna funkcija, ki je različna od 0 samo na intervalu [0,T], iz (23) dobimo enačbo

$$\Delta t^2 = \frac{\int_0^T t^2 \left(0{,}54 - 0{,}46\cos(2\pi t / T)\right)^2 dt}{\int_0^T \left(0{,}54 - 0{,}46\cos(2\pi t / T)\right)^2 dt}, \tag{26}$$

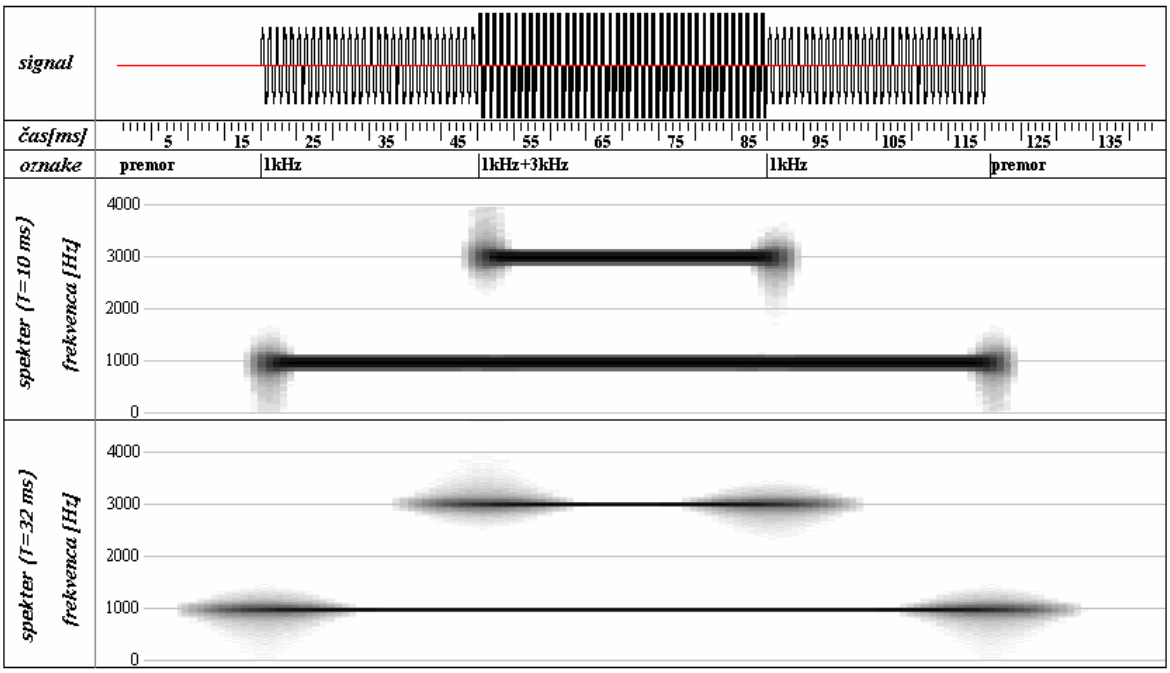
iz katere sledi

$$\Delta t = 0{,}5229T. \tag{27}$$

Časovna ločljivost Δt je torej premo sorazmerna z dolžino okvirja, frekvenčna ločljivost Δf pa obratno sorazmerna.

Vpliv dolžine okvirja na izračunani spekter signala je prikazan na sliki 12. Na zgornjem delu slike je prikazan signal. Na začetku in koncu je v signalu premor (tišina), vmes je prisoten sinusni signal frekvence 1 kHz. Signalu se na sredini (odsek je označen z 1kHz + 3kHz) pridruži še sinusni signal

frekvence 3 kHz. Meje med posameznimi odseki so razvidne (tudi) iz amplitude samega signala. Signalu je dodano časovno merilo in oznake omenjenih odsekov signala. Sledita spektra signala, ki sta bila izračunana z uporabo 10 in 32 ms dolgega Hammingovega okna.



Slika 12: Vpliv dolžine okna na časovno in frekvenčno ločljivost spektra signala. Od zgoraj navzdol so prikazani: signal, časovna os, oznake odsekov in spektra signala, izračunana z uporabo Hammingovega okna dolžine 10 in 32 ms. Spektra sta prikazana dvodimenzionalno (čas, frekvenca) in z uporabo sivin. Temnejša barva pri določeni frekvenci pomeni večjo moč signala pri tej frekvenci.

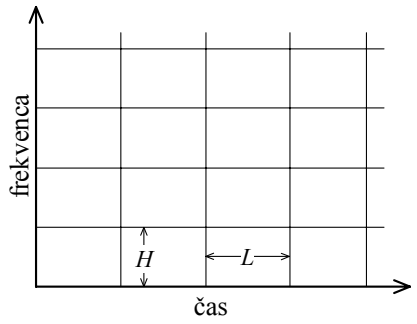
Na sliki so jasno vidne razlike v časovni in frekvenčni ločljivosti spektrov. Spekter, ki je izračunan z 10 ms dolgim oknom, ima razmeroma dobro časovno ločljivost - razmazanje v časovni (horizontalni) osi je majhno. Frekvenčna ločljivost je zato slaba (razmazanje po vertikalni osi). Obratno velja za spekter izračunan z 32 ms dolgim oknom. Frekvenčna ločljivost je dobra - vertikalno razmazanje je majhno. Časovna ločljivost je slaba. To je še posebej opazno na mejah, ko se v signalu pojavi ali izgine določena frekvenca. V spektru se frekvenca v pojavi precej pred tem, kot se v resnici pojavi v signalu. Tudi izgine prepozno.

Za spektralno analizo vzorčenih signalov se uporablja diskretna različica kratkočasovne Fourierove transformacije, kratkočasovna diskretna Fourierova transformacija (KČDFT[17]), ki je bila uporabljena tudi pri izdelavi slike 12. KČDFT je diskretna tako v času kot frekvenci. Za *l*-ti okvir je podana z enačbo

$$X(l,k) = \sum_{n=lL}^{lL+N-1} x(n)w(n-lL)e^{-2\pi jkn/N}, \qquad k = 0,1,...,N-1, \tag{28}$$

kjer je z *x(n)* označen vzorčen signal *x(t)*, *N* pa predstavlja dolžino okenskega zaporedja *w(n)*. Okensko zaporedje (vzorčeno okensko funkcijo) pomikamo v času za določeno število vzorcev *L*. KČDFT je torej diskretna Fourierova transformacija (DFT) z okenskim zaporedjem pomnoženega signala, ki jo izračunamo vsakih *L* vzorcev. Pri frekvenčni analizi govornih signalov običajno velja $L<N$, kar pomeni, da se okna prekrivajo.

Če se spekter signala izračuna s KČDFT, pri kateri okna pomikamo po *L* vzorcev, dobimo razdelitev časovno-frekvenčne ravnine na pravokotnike enakih dimenzij, ki je prikazana na sliki 13.



Slika 13: Razdelitev časovno-frekvenčne ravnine pri KČDFT.

L predstavlja časovni pomik, višina pravokotnikov v frekvenci *H* pa je enaka f_s/N, kjer je s f_s označena vzorčevalna frekvenca. Ploščina pravokotnika v časovno-frekvenčni ravnini znaša f_sL/N. Za uspešno klasifikacijo vzorcev je potrebno izbrati primerno okensko funkcijo in določiti *L* in *N*. Na področju razpoznavanja govora se uporabljajo okna dolžine 15 - 35 ms (*N* se določi temu ustrezno) in časovni pomik *L*, ki običajno znaša približno 10 ms.

---

[17] *angl. Short Time Discrete Fourier Transform (STDFT)*

Glavna prednost KČDFT pred drugimi pristopi za določanje spektra je obstoj učinkovitega algoritma FFT za računanje diskretne Fourierove transformacije. Želeno frekvenčno-časovno ločljivost spektra določimo z uporabo primerno dolgega okna oziroma okvirja ter primerno izbranim pomikom *L*. Poglavitna slabost KČDFT je, da je frekvenčno-časovna ločljivost fiksna (frekvenčno-časovno ravnino razdelimo na pravokotnike enakih dimenzij). To pri realnih signalih ni vedno najbolje. Kot že omenjeno, je človekovo uho pri nižjih frekvencah bistveno bolj občutljivo na spremembe frekvence kot pri višjih. To pomeni, da bi bili (če pri spektralni analizi želimo upoštevati lastnosti sluha) pravokotniki v časovno-frekvenčni ravnini pri visokih frekvencah lahko višji (večji *H*) in ožji kot pri nižjih. Pri višjih frekvencah bi torej lahko dosegli boljšo časovno ločljivost spektra, česar pa z uporabo KČDFT ne moremo izkoristiti.

Spekter z neenakomerno frekvenčno-časovno ločljivostjo dobimo z uporabo večločljivostne frekvenčne analize. Pogosto se jo izvede z uporabo valčne transformacije, za katero je značilno da se časovna ločljivost v spektru z večanjem frekvence izboljšuje [52, 62, 64, 64]. V nadaljevanju so opisane osnove valčne transformacije, pri čemer je poudarek na frekvenčno-časovni ločljivosti z valčno transformacijo izračunanega spektra.

5.2 Valčna transformacija

Odvisnost frekvenčno-časovne ločljivosti od frekvence je osnovna lastnost valčne transformacije (WT[18]). Če bi želeli izračunati podoben spekter z uporabo KČDFT, bi morali pri višjih frekvencah uporabiti krajši okvir kot pri nižjih.

KČDFT (28) je pravzaprav računanje korelacije signala s časovno (za *lL* vzorcev) pomaknjenim oknom, ki je "napolnjeno" z oscilacijami določene frekvence. Ker je dolžina okna konstantna, je število oscilacij v oknu pri višjih frekvencah večje kot pri nižjih. Če pa dolžino okna določimo obratno sorazmerno s frekvenco, dobimo v oknu vedno enako število oscilacij, ki je torej neodvisno od frekvence. V tem primeru je oblika funkcije, s katero računamo korelacijo signala, vedno enaka. To funkcijo pomikamo v času in jo z raztezanjem in krčenjem prilagajamo želenim frekvencam.

[18] *angl. Wavelet Transform*

Tak pogled na spektralno analizo pripelje do valčne transformacije. Zvezna valčna transformacija je določena z enačbo

$$CWT_x(\tau, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-\tau}{a} \right) dt \, . \tag{29}$$

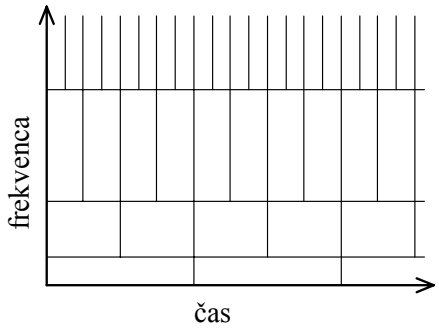
V enačbi je z * označeno kompleksno konjugiranje. Funkcija $\Psi(t)$ se imenuje osnovni valček in ima lastnosti okenske funkcije. To pomeni, da je osnovni valček od 0 različen na končnem časovnem intervalu. Pri spektralni analizi gre za računanje korelacije signala z izpeljanimi valčki, ki so za faktor translacije τ pomaknjene in za faktor dilacije *a* raztegnjene različice osnovnega valčka. S pomočjo τ se torej valček pomika po času, s pomočjo *a* pa se ga širi oziroma oži. Večji *a* pomeni, da je izpeljani valček v času ožji, kar ustreza višjim frekvencam. Ozko funkcijo v času pri višjih frekvencah lahko primerjamo s kratkim oknom pri KČDFT, kar pomeni dobro časovno ločljivost. Za nizke frekvence velja obratno. Zaradi boljše časovne ločljivosti pri višjih frekvencah, bi morala imeti valčna transformacija pri analizi govornega signala prednosti pred KČDFT, še posebej če upoštevamo nelinearno frekvenčno ločljivost človekovega sluha. Vendar pa frekvenčna ločljivost pri valčni transformaciji pada linearno s frekvenco, kar je bistveno hitreje kot pri človekovemu sluhu, kjer frekvenčna ločljivost pada sorazmerno z logaritmom frekvence.

Računanje zvezne valčne transformacije je računsko zahtevno in ustreza filtriranju signala z naborom filtrov. Zato se na področju digitalnega procesiranja signalov običajno uporablja diskretna valčna transformacija (DWT[19]). Dobimo jo, če v enačbi (29) nadomestimo a in τ z $a = a_0^s$ in $\tau = l \cdot a_0^s$. Učinkovito računanje diskretne valčne transformacije je mogoče samo ob izboru $a = 2^s$ in $\tau = l \cdot 2^s$. Dobimo torej enačbo:

$$DWT_x(l, s) = \frac{1}{\sqrt{2^s}} \sum_{n=0}^{\infty} x(n) \Psi^* \left(\frac{n - 2^s l}{2^s} \right). \tag{30}$$

[19] *angl. Discrete Wavelet Tranform*

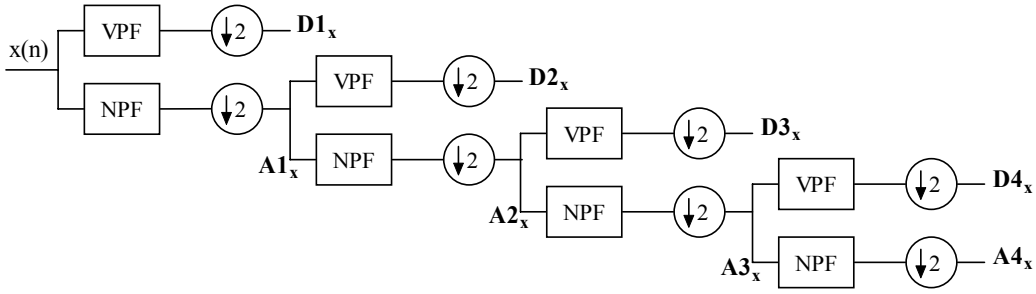
Ob takem izboru parametrov a in τ je določena frekvenčno-časovna ločljivost oziroma oblika pravokotnikov v časovno-frekvenčni ravnini kot je prikazano na sliki 14.



Slika 14: Razdelitev časovno-frekvenčne ravnine pri diskretni valčni transformaciji.

Najmanjši časovni pomik 2^s s frekvenco pada (majhen s ustreza višji frekvenci). Časovna ločljivost pri visokih frekvencah je torej dobra, pri nizkih pa slaba. Za frekvenčno ločljivost velja obratno.

Algoritem za hitro računanje diskretne valčne transformacije temelji na filtriranju signala s pari nizkoprepustnih in visokoprepustnih kvadraturnih zrcalnih filtrov, ki ustrezajo določenemu valčku. Filtrom sledi decimacija[20] za faktor 2. Postopek se iterativno ponavlja nad decimiranimi izhodi nizkoprepustnih filtrov, kot je prikazano na sliki 15.



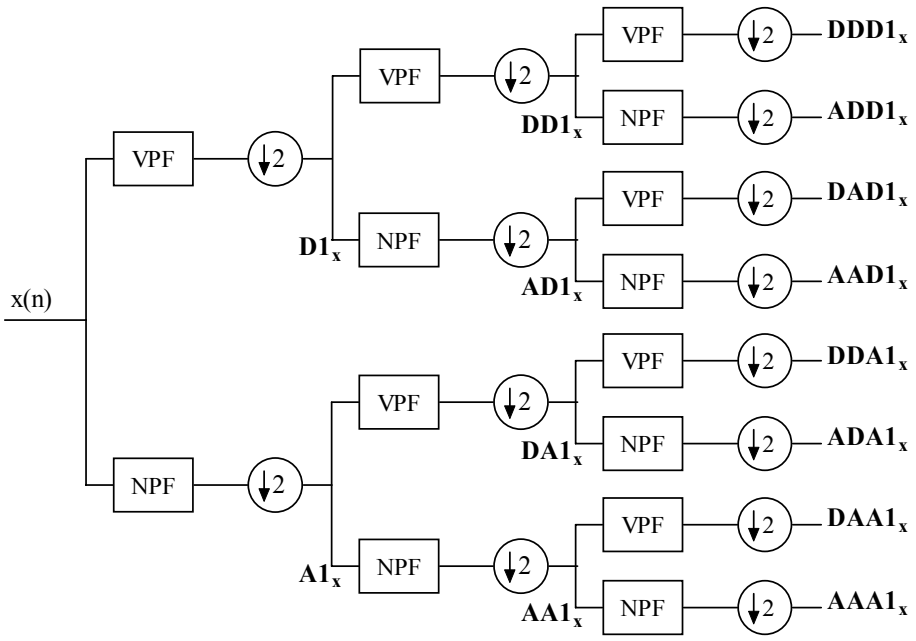
Slika 15: Računanje diskretne valčne transformacije s kvadraturnimi zrcalnimi filtri.

---

[20] decimacija za faktor 2 pomeni, da v nadaljnjem računanju uporabimo vsak drugi koeficient, ki ga dobimo na izhodu ustreznega filtra. To lahko storimo, ker je frekvenčni pas, ki ga opisuje izhod filtra dvakrat ožji od frekvenčnega pasu na vhodu filtra.

Signal *x(n)* se filtrira z visokoprepustnim filtrom VPF in nizkoprepustnim filtrom NPF. Izhoda obeh filtrov se decimira za faktor 2. Koeficienti, ki so označeni z $\mathbf{D1_x}$, opisujejo podrobnosti signala *x(n)*, z $\mathbf{A1_x}$ pa je označena njegova aproksimacija. Postopek se iterativno ponavlja na aproksimacijah, ki jih dobimo v predhodni stopnji. Ker se v vsaki stopnji izvede decimacija, je valčnih koeficientov v vsaki naslednji stopnji manj. Koeficientov v zaporedju $\mathbf{D1_x}$ in $\mathbf{A1_x}$ je veliko, kar pomeni, da posamezni koeficient opisuje kratek časovni odsek signala *x(n)*. Časovna ločljivost je dobra, frekvenčna pa slaba - s koeficienti $\mathbf{D1_x}$ je namreč opisana celotna zgornja, z $\mathbf{A1_x}$ pa spodnja polovica celotnega frekvenčnega območja 0 - $f_s/2$. V nadaljnjih iteracijah se frekvenčna ločljivost vsakič izboljša za faktor 2, časovna pa se ustrezno poslabša.

S tako delitvijo frekvenčnega območja, frekvenčna ločljivost pri visokih frekvencah ne zadošča za uspešno razpoznavanje govora. Koeficienti $\mathbf{D1_x}$ namreč opisujejo celotno zgornjo polovico frekvenčnega pasu (od $f_s/4$ do $f_s/2$). Zato je potrebno tudi koeficiente $\mathbf{D1_x}$ filtrirati s parom kvadraturnih zrcalnih filtrov. Če to naredimo na vseh stopnjah, pridemo do *paketne valčne transformacije* (WPT[21]), kot to prikazuje slika 16.



Slika 16: Računanje paketne valčne transformacije s kvadraturnimi zrcalnimi filtri.

[21] *angl. Wavelet Packet Transform*

Na sliki je prikazano drevo paketne valčne transformacije. Za razpoznavanje govora pri višjih frekvencah ni potrebna tako dobra frekvenčna ločljivost kot pri nižjih (lastnost človekovega sluha), zato se pri višjih frekvencah dekompozicije ne izvede do konca. Ali bomo na določeni stopnji z dekompozicijo nadaljevali ali ne, je običajno odvisno od frekvenčne občutljivosti sluha v ustreznem frekvenčnem pasu. Ker se za modeliranje nelinearne frekvenčno ločljivosti človekovega sluha v celotni disertaciji uporablja Mel-lestvica, ji je smiselno prilagoditi tudi dekompozicijsko drevo pri paketni valčni transformaciji, kot je opisano v nadaljevanju (glej sliko 20).

V nadaljevanju poglavja so bile KČDFT, zvezna valčna transformacija in paketna valčna transformacija uporabljene za frekvenčno analizo govornega signala v postopku računanja MFCC koeficientov. Izvedena je bila primerjava vpliva uporabljenega tipa frekvenčne analize na uspešnost razpoznavanja govora. Podrobnosti o spektralni analizi so podane v nadaljevanju poglavja, vsi ostali parametri za računanje MFCC koeficientov so taki kot je navedeno v poglavju 4.1.

5.3 Parametrizacija govornega signala z uporabo valčne transformacije

V tem poglavju je opisana primerjava MFCC koeficientov na osnovi zvezne in diskretne oziroma paketne valčne transformacije z MFCC koeficienti na osnovi kratkočasovne Fourierove transformacije. Vpliv določenega tipa frekvenčne analize na uspešnost razpoznavanja je bil ocenjen z govorno zbirke ŠTEVKE. Podatki o velikosti učne, razvojne in testne množice so navedeni v poglavju 3.1 (tabela 3). Rezultati primerjave kažejo, da zamenjava kratkočasovne Fourierove transformacije z zvezno oziroma paketno valčno transformacijo uspešnosti razpoznavanja ne izboljša. Do podobnih rezultatov so prišli tudi drugi avtorji, npr. [61, 74].

5.3.1 MFCC koeficienti na osnovi zvezne valčne transformacije

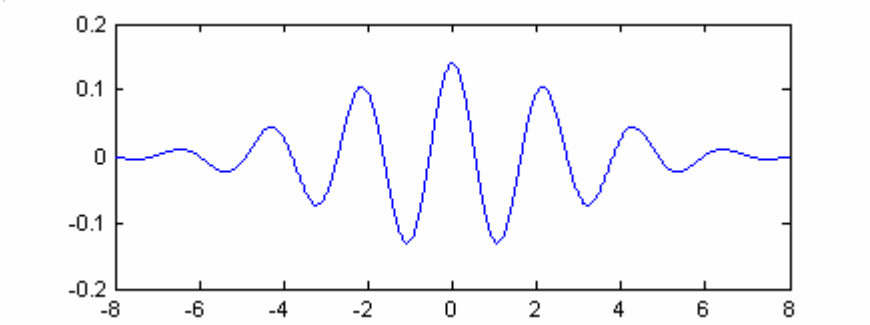
Na področju razpoznavanja govora se pogosto uporabljajo valčki, ki imajo obliko moduliranih okenskih funkcij. Za primerjavo smo zaradi podobnosti s Fourierovo transformacijo izbrali Morletov valček, ki se na področju razpoznavanja govora uporablja najpogosteje [64, 65]. Ker Morletov valček ni primeren za diskretno oziroma paketno valčno transformacijo, sta bila v

primerjavo vključena še dva valčka iz skupine Daubechies, ki sta podrobneje opisana v podpoglavju 5.3.2: Haarov valček in valček D^4. Podrobnosti o uporabi zvezne valčne transformacije pri računanju MFCC koeficientov so povzete po [74].

Morletov valček je določen z enačbo

$$\Psi(t) = \frac{1}{\sqrt{\pi f_b}} e^{2\pi i f_c t} e^{\frac{-t^2}{f_b}} . \qquad (31)$$

Osnovni valček $\Psi(t)$ je Gaussovo okno, napolnjeno z oscilacijami frekvence f_c. S parametroma f_b in f_c sta določeni dolžina okna in centralna frekvenca osnovnega valčka. Realna komponenta osnovnega valčka, ki je bil uporabljen pri računanju spektra, je prikazana na sliki 17.



Slika 17: Morletov valček izračunam za vrednosti koeficientov f_c=0,46 in f_b=16. Prikazana je realna komponenta valčka.

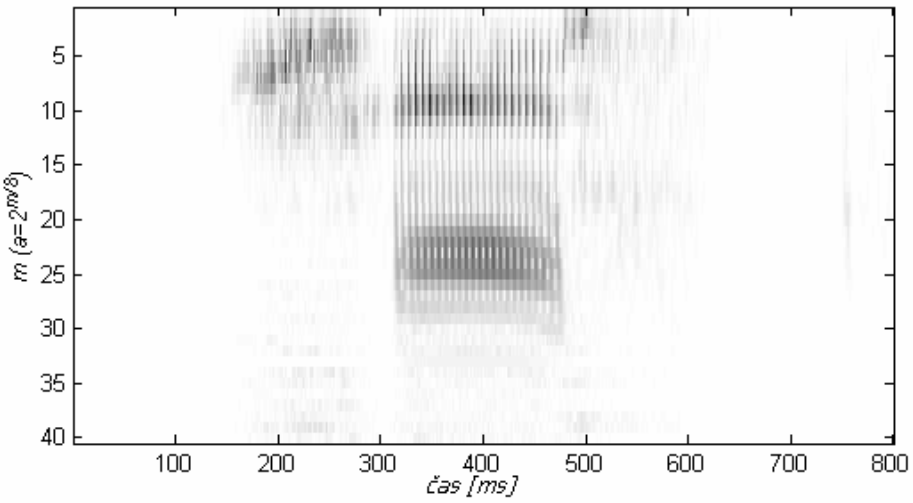
Centralna frekvenca osnovnega valčka je s parametrom f_c=0,46 določena pri 3700 Hz, ki ustreza centralni frekvenci Mel-filtra z najvišjim indeksom (glej sliko 4). Valčna transformacija je bila, podobno kot v [74], izračunana pri vrednostih parametra dilacije *a*, določenih z enačbo

$$a = 2^{m/8}, \qquad m = 0,1,...,51 . \qquad (32)$$

Take vrednosti *a* ustrezajo frekvencah, ki so nelinearno razporejene med 42 in 3700 Hz, kar se dobro ujema s frekvenčnim pasom, ki ga pokrijejo po Mel-lestvici razporejeni trikotni filtri.

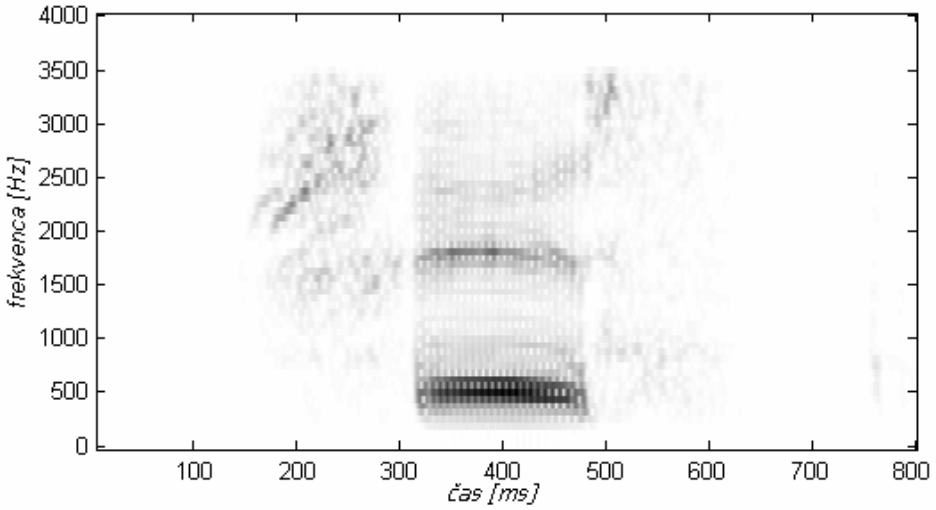
Dolžina osnovnega valčka je bila določena s faktorjem $f_b = 16$. Pri računanju avditornega spektra so bile uporabljene amplitude (absolutne vrednosti) koeficientov valčne transformacije.

Amplitude koeficientov za besedo *"šest"* so prikazane na sliki 18. Ker je vzorec posnet preko telefonske linije, je signal omejen na frekvenčni pas 300 - 3400 Hz. Zato so prikazani samo koeficienti izračunani pri vrednostih *a* določenih z (32) in *m*=0,1,...,40. Na sliki je jasno opazna spremenljiva frekvenčno-časovna ločljivost. Pri visokih frekvencah (majhen *a*) je frekvenčna ločljivost slabša kot pri nizkih, časovna ločljivost pa je celo predobra. To je dobro vidno pri spektru fonema /e/ (med 320 in 480 ms), kjer bi spekter moral biti (skoraj) stacionaren. Predvsem pri visokih frekvencah so v spektru jasno vidne oscilacije (izmenjavanje vertikalnih svetlih in temnih lis), ki so povezane z osnovno frekvenco govornega signala. Težavo je mogoče omiliti z uporabo daljšega osnovnega valčka, vendar bi bil v tem primeru izpeljani valček pri nizkih frekvencah predolg. Torej gre tudi pri izbiri dolžine osnovnega valčka za podoben kompromis kot pri izbiri dolžine okvirja pri KČDFT.



Slika 18: Koeficienti zvezne valčne transformacije izračunane z Morletovim valčkom. Osnovni valček je bil določen s $f_c = 0{,}46$ *in* $f_b = 16$. *Prikazane so absolutne vrednosti koeficientov.*

Za primerjavo je na sliki 19 prikazan spekter signala s fiksno frekvenčno-časovno ločljivostjo, ki je bil izračunan s KČDFT. Pri računanju je bilo uporabljeno 16 ms dolgo Hammingovo okno.



Slika 19: Spekter signala izračunan s KČDFT. Prikazan je amplitudni spekter. Pri računanju je bilo uporabljeno Hammingovo okno dolžine 16 ms.

Oscilacije, do katerih ob uporabi zvezne valčne transformacije pride pri višjih frekvencah, na uspešnost razpoznavanja govora vplivajo negativno. Zato je potrebno pred računanjem avditornega spektra absolutne vrednosti koeficientov obdelati z nizkoprepustnim filtrom. Koeficiente se filtrira posebej pri vsaki vrednosti *a* (časovno zaporedje koeficientov pri določeni frekvenci), pri čemer je bil v našem primeru uporabljen filter s končnim odzivom na enotin impulz dolžine n=128. Enotin odziv filtra ustreza Hammingovemu okenskemu zaporedju ustrezne dolžine (n=128 oziroma 16 ms pri f_s=8000 Hz). Nad izhodi nizkoprepustnega filtra se izvede decimacija iz f_s=8000 Hz (vzorčevalna frekvenca) na 100 Hz (frekvenca okvirjev). Uspešnost SRG, ki uporablja na osnovi tako dobljenega spektra izračunane MFCC koeficiente, je podana v zadnjem delu poglavja.

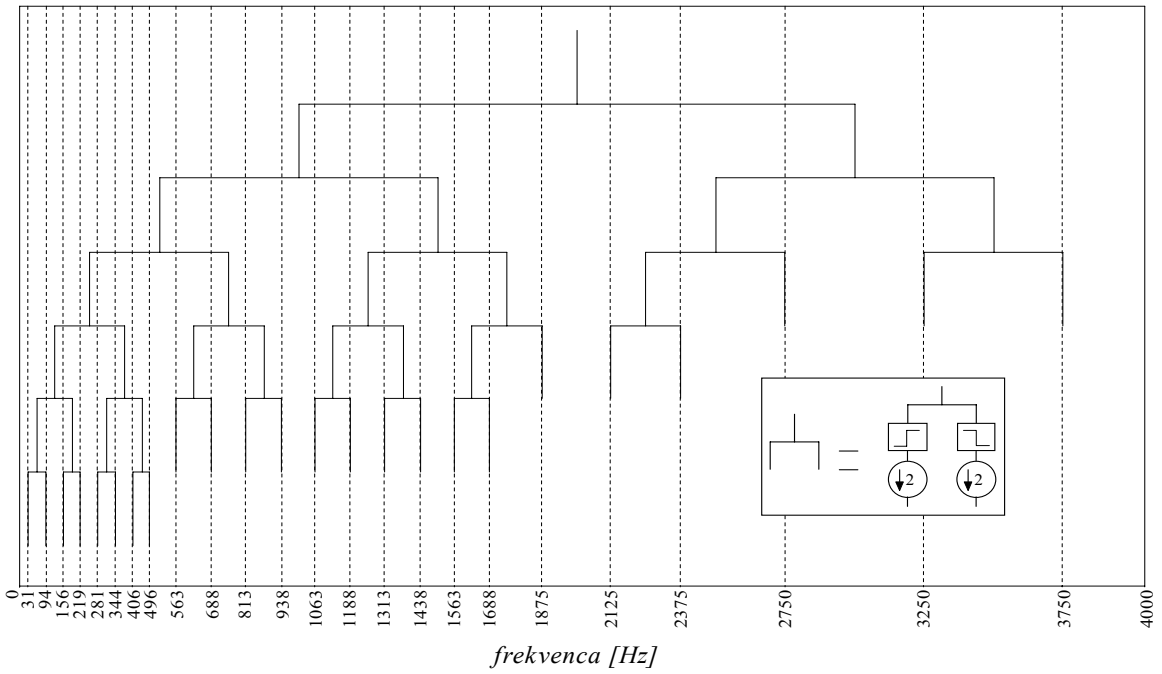
5.3.2 MFCC koeficienti na osnovi diskretne valčne transformacije

Zaradi računske zahtevnosti zvezne valčne transformacije se na področju procesiranja govora pogosteje uporablja različica diskretne valčne transformacije - paketna valčna transformacija. Paketna valčna transformacija je bila v poskusih izračunana z iterativnim filtriranjem s kvadraturnimi zrcalnimi filtri. Postopek je povzet po [69]. Uporabljena sta bila dva valčka: Haarov

valček, ki je določen z enačbo (33) in valček D^4 iz skupine Daubechies[22].

$$\Psi(t)=\begin{cases}1, & 0<\mathrm{t}\le 1/2\\ -1, & 1/2\le \mathrm{t}<1\\ 0, & \text{sicer}\end{cases} \tag{33}$$

Na sliki 20 je prikazano drevo filtrov, ki je bilo uporabljeno za računanje paketne valčne transformacije. Pravzaprav gre za poddrevo dekompozicijskega drevesa paketne valčne transformacije, ki je določeno tako, da frekvenčna ločljivost koeficientov približno ustreza Mel-lestvici.



Slika 20: Poddrevo filtrov, za računanje paketne valčne transformacije tako, da frekvenčna ločljivost koeficientov približno ustreza Mel-lestvici. Levo vejo na vsakem razcepu dobimo s filtriranjem koeficientov iz prejšnjega nivoja z nizkoprepustnim filtrom, desno pa z visokoprepustnim filtrom. Izhoda obeh filtrov se decimira za faktor 2.

---

[22] za funkcije iz skupine Daubechies (razen D^1) eksplicitne enačbe ne obstajajo, funkcije so določene s koeficienti nizkoprepustnega filtra, te pa izračunamo neposredno iz ustreznega sistema enačb.

Govorni signal se najprej razreže na okvirje dolžine 16 ms. Uporabljeno je bilo Hammingovo zaporedje ustrezne dolžine. Pomik okvirjev znaša 10 ms (frekvenca okvirjev 100 Hz). Frekvenčna analiza se z drevesom filtrov s slike 20 izvede za vsak okvir posebej. Izhodi filtrov, ki so predstavljeni z listi drevesa, so koeficienti paketne valčne transformacije, ki opisujejo različno široke frekvenčne pasove. Če filtre (oziroma liste drevesa) oštevilčimo od leve proti desni, izhodi filtrov opisujejo frekvenčne pasove, ki so prikazani v tabeli 9.

Filter	Frekv. [Hz]	Filter	Frekv. [Hz]	Filter	Frekv. [Hz]
1	31	9	563	17	1563
2	94	10	688	18	1688
3	156	11	813	19	1875
4	219	12	938	29	2125
5	281	13	1063	21	2375
6	344	14	1188	22	2750
7	406	15	1313	23	3250
8	496	16	1438	24	3750

Tabela 9: Centralne frekvence filtrov. Izbrane so tako, da približno ustrezajo Mel-lestvici. Frekvence so izračunane za vzorčevalno frekvenco 8000 Hz.

Pri izračunu frekvenc je potrebno upoštevati, da je govorni signal vzorčen z vzorčevalno frekvenco f_s=8000 Hz. Celotno drevo filtrov torej opisuje frekvenčni pas 0 - 4000 Hz. Ko gremo v drevesu po stopnjah od korena proti listom, na vsaki naslednji stopnji koeficienti valčne transformacije opisujejo dvakrat ožji frekvenčni pas kot koeficienti na prejšnji stopnji. Ker se na vsaki stopnji izvede decimacija, se število koeficientov vsakič prepolovi.

Pri parametrizaciji posameznega okvirja se po enačbi (34) za vsak filter posebej izračuna energijo signala v ustreznem frekvenčnem pasu. Tak pristop se pogosto uporablja, npr. [61, 69].

$$P_{WPT}(l,k) = \frac{\sum_{i=1}^{N(k)} W_l^2(k,i)}{N(k)}, \qquad k = 1,2,..,24 \tag{34}$$

$P_{WPT}(l,k)$ predstavlja energijo signala l-tega okvirja v frekvenčnem pasu filtra k. $W_l(k,i)$ so za l-ti okvir izračunani koeficienti valčne transformacije (izhodi filtra k), $N(k)$ pa je število valčnih koeficientov, ki se izračunajo s filtrom k.

5.3.3 *Primerjava MFCC koeficientov na osnovi KČDFT, zvezne in diskretne valčne transformacije*

Vpliv uporabe valčne analize v postopku parametrizacije govornega signala na uspešnost SRG je bil ocenjen z govorno zbirko ŠTEVKE. Rezultati so prikazani v tabeli 10. S CWT so označeni MFCC koeficienti, ki so izračunani z uporabo zvezne valčne transformacije kot je to opisano v podpoglavju 5.3.1 in ustreznega valčka, z WPT pa MFCC koeficienti, ki so izračunani z uporabo paketne valčne transformacije in ustreznega valčka. S KČDFT so označeni MFCC koeficienti, pri katerih je spekter izračunan KČDFT in 16 ms dolgim Hammingovim oknom. S CWT* so označeni koeficienti zvezne valčne transformacije, pri katerih je bilo filtriranje z nizkoprepustnim filtrom izpuščeno.

Tip MFCC koeficientov	Uspešnost - besede [%]
CWT(Morlet)	96,6
CWT(Morlet)*	89,7
CWT(Haar)	95,6
CWT (D^4)	96,3
WPT(Haar)	94,7
WPT (D^4)	95,7
KČFDT	96,7

Tabela 10: Vpliv različnih tipov MFCC koeficientov na uspešnost SRG. Uspešnost je podana z odstotkom pravilno razpoznanih besed v testni množici.

Rezultati primerjave kažejo, da so MFCC koeficienti, ki so izračunani na osnovi paketne valčne transformacije, manj primerni za uporabo v SRG od MFCC koeficientov, pri katerih spektralno analizo izvedemo s KČDFT. Vseeno je zanimiv rezultat, ki je bil dosežen s koeficienti *WPT(Haar)*. Dolžina posameznega filtra v dekompozicijskem drevesu v tem primeru znaša le N=2, kar pomeni, da je taka frekvenčna analiza računsko zelo nezahtevna. Tudi s koeficienti CWT, pri katerih je spekter izračunan z zvezno valčno transformacijo, ni bila presežena uspešnost MFCC koeficientov, pri katerih se spektralna analiza izvede s KČDFT. Ker je računanje zvezne valčne transformacije računsko zahtevnejše od računanja KČDFT, menimo, da uporaba zvezne valčne transformacije za računanje MFCC koeficientov ni smiselna. Poudariti je potrebno tudi, da gre pri filtriranju koeficientov zvezne valčne transformacije z nizkoprepustnim filtrom za povprečenje v času

(podobno zaradi (34) velja tudi za paketno valčno transformacijo). To pa pomeni izgubo boljše časovne ločljivosti pri višjih frekvencah, ki naj bi prinašala prednosti pred računanjem spektra s KČDFT. Če povprečenje izpustimo, je uspešnost SRG še bistveno slabša CWT(Morlet)*.

Težava je v tem, da je pri zvezni valčni transformaciji dolžina izpeljanega valčka obratno sorazmerna s frekvenco. Če želimo, da je dolžina valčka pri nizkih frekvencah primerna, bo ta pri visokih frekvencah premajhna. Poudariti je potrebno tudi, da je dolžino osnovnega valčka oziroma števila oscilacij v osnovnem valčku mogoče določiti samo za nekatere valčne funkcije. Izmed preizkušenih smo to lahko naredili le pri Morletovem valčku. Prednost tako izračunanih MFCC koeficientov na osnovi zvezne valčne transformacije je torej le v nelinearni frekvenčni ločljivosti (podobnost s človekovim sluhom), kar pa ni prineslo izboljšanja uspešnosti razpoznavanja.

6. METODE ZA DINAMIČNO PRILAGAJANJE FREKVENČNO-ČASOVNE LOČLJIVOSTI

Kompromis, ki ga je potrebno narediti pri izbiri dolžine okvirja za frekvenčno analizo oziroma računanje kratkočasovne diskretne Fourierove transformacije (KČDFT) govornega signala, je bil predstavljen že v uvodnem poglavju. Rezultati iz prejšnjega poglavja kažejo, da problema ni mogoče zmanjšati niti z uporabo večločljivostne analize z valčno transformacijo. V tem primeru moramo kompromis narediti pri izbiri dolžine oziroma števila oscilacij v osnovnem valčku, če to za uporabljeno valčno funkcijo sploh lahko naredimo. Zaradi velikih razlik v dinamiki spreminjanja spektra govornega signala na različnih odsekih, je težava pri govornem signalu še posebej izražena. Na določenih odsekih govora je spekter skoraj stacionaren, na drugih pa se spreminja zelo hitro. Zato je na odsekih s stacionarnim spektrom pri računanju KČDFT smiselno uporabiti daljši okvir, ki zagotavlja večjo frekvenčno ločljivost izračunanega spektra. Z računanjem spektra za vsak okvir, dobimo povprečen spekter signala, ki ga okvir zajema. To pomeni, da bodo spremembe spektra, ki so krajše od trajanja okvirja, v izračunanem spektru zabrisane. Zato je na odsekih, kjer se spekter spreminja razmeroma hitro, dolžino okvirja smiselno zmanjšati in izboljšati časovno ločljivost. Ker je frekvenčna ločljivost izračunanega spektra obratno sorazmerna z dolžino okvirja, je ta sedaj slabša. Dinamiko spreminjanja spektra je mogoče oceniti na različne načine ter ustrezno prilagoditi dolžino okvirja in s tem frekvenčno-časovno ločljivost spektra. Splošne lastnosti in način nastajanja govornega signala so znani. Ta spoznanja je mogoče izkoristiti pri določanju odsekov govora, ko je potrebna boljša časovna ločljivost in odsekov, na katerih je pomembnejša frekvenčna ločljivost spektra.

V nadaljevanju poglavja (podpoglavje 6.1) so najprej podrobneje predstavljene skupne lastnosti poskusov, s katerimi je bil ocenjen vpliv dinamičnega prilagajanja frekvenčno-časovne ločljivosti spektra na uspešnost razpoznavanja. Nato sledijo tri podpoglavja, v katerih so predstavljene različne metode za prilagajanje frekvenčno-časovne ločljivosti dinamiki spreminjanja spektra govornega signala. Pri vsaki metodi so podani rezultati, ki so bili doseženi pri razpoznavanju

govornih zbirk ŠTEVKE in NUMBERS vključno s testom vpliva na robustnost razpoznavanja (z metodo prilagajanja dolžine okvirja na osnovi fonemske zgradbe besed je bila uporabljena le zbirka ŠTEVKE). Preizkušene so bile naslednje tri metode:

1. **Prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed.** Fonemska zgradba dela učne množice je znana, za preostali del učne množice, za razvojno množico in testno množico je bilo potrebno fonemsko zgradbo določiti s pomočjo SRG, pri katerem je bil spekter izračunan z 32 ms dolgim okvirjem. Razpoznavanje torej poteka v dveh prehodih. V prvem se pri računanju spektra uporabi okvir fiksne dolžine, v drugem pa se dolžino okvirja prilagodi fonemski zgradbi, ki je bila določena v prvem prehodu. Opisani sta dve skupini poskusov:

    a) **Uporaba daljšega okvirja pri parametrizaciji srednjih delov trifonov.** Pristop temelji na predpostavki, da se spekter srednjih delov dolgih fonemov, ki so modelirani s trifoni, spreminja počasi. Zato časovna ločljivost spektra na teh odsekih ni kritična in je smiselno uporabiti daljši okvir.

    b) **Uporaba daljšega okvirja pri parametrizaciji zvenečih fonemov.** Pri izgovorjavi zvenečih fonemov so glasilke napete. Ker človek glasilk ne more napeti za zelo kratek čas, so zveneči fonemi dolgi. Ker med zveneče foneme spadajo tudi vsi samoglasniki, ki imajo skoraj stacionaren spekter, je pri parametrizaciji smiselno uporabiti daljši okvir.

2. **Prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka.** Večje spremembe jakosti zvoka so značilne za določene foneme, npr. zapornike in prehode med fonemi, pri katerih se spekter hitro spreminja. Spekter se izračuna dvakrat, enkrat s krajšim in enkrat z daljšim okvirjem. Pri računanju značilk se uporabi obtežena vsota obeh spektrov. Spremembe jakosti zvoka se ocenijo z uporabo akustične značilke *spremembe jakosti zvoka*, ki je opisana v nadaljevanju. V primeru, da so te velike, se v obteženi vsoti poudari s krajšim okvirjem izračunani spekter.

3. **Prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora.** Metoda temelji na predpostavki, da se pri zvenečih fonemih spekter spreminja počasi, zato se pri frekvenčni

analizi zvenečih odsekov uporabi daljši okvir kot pri analizi nezvenečih odsekov. Tudi zveneči in nezveneči odseki so bili določeni z uporabo ustrezne akustične značilke.

6.1 Skupne lastnosti poskusov

Vsi v nadaljevanju poglavja opisani poskusi so bili izvedeni z enakimi učnimi, razvojnimi in testnimi množicami, zato so rezultati medsebojno neposredno primerljivi. Pri poskusih z govorno zbirko NUMBERS sta bili uporabljeni manjša razvojna in testna množica s po 600 vzorci. Ostale množice vzorcev so take, kot je navedeno v poglavju 3.

Učenje SRG je bilo izvedeno vedno na enak način, ki je podrobneje opisan v poglavju 4.4. Uspešnost SRG iz posameznih poskusov je vedno podana v odstotkih pravilno razpoznanih besed v testni množici. Poleg uspešnosti razpoznavanja testne množice, v kateri so vzorci posneti pod enakimi pogoji kot celotna govorna zbirka, nas je zanimal tudi vpliv prilagajanja frekvenčno-časovne ločljivosti spektra na robustnost razpoznavanja. Zato so bile testni množici naknadno dodane različne motnje, s pomočjo katerih je bil ocenjen vpliv na robustnost.

Rezultati so pri vseh pristopih prikazani na enak način. Z vsako metodo je bilo izvedeno večje število poskusov, katerih rezultati so podani v tabeli. Poleg uspešnosti razpoznavanja testne množice, so prikazani tudi povprečni rezultati razpoznavanja testnih množic z dodanimi motnjami v določenem razmerju signal/šum. SRG iz štirih poskusov, pri katerih je bila uspešnost razpoznavanja testne množice brez dodanih motenj največja, smo primerjali s SRG, pri katerih se pri spektralni analizi uporablja okvir fiksne dolžine. Ta primerjava je prikazana z grafom, ki mu je dodana tabela s podrobnimi podatki o uspešnosti razpoznavanja testne množice brez dodanih motenj in povprečne uspešnosti razpoznavanja množic z dodanimi aditivnimi motnjami, ki so podrobneje opisane v nadaljevanju.

6.1.1 Ocena robustnosti

Robustnost razpoznavanja je bila ovrednotena z dodajanjem štirih različnih aditivnih motenj v različnih razmerjih signal/šum (SNR[23]). Aditivne motnje so zvoki, ki so poleg govora vsebovani v vzorcih. Testni množici so bili zato prišteti zvoki, ki so značilni za različna okolja. Na ta način so bili simulirani pogoji, ki bi bili prisotni v testni množici, če bi bila posneta v ustreznem okolju. V tem poglavju so bili uporabljeni štirje posnetki iz zbirke NOISEX [25]:

- pogovor ljudi v ozadju (angl. oznaka *Babble*);
- hrup v osebnem avtomobilu (angl. oznaka *Volvo*);
- beli šum (angl. oznaka *White*);
- roza šum (angl. oznaka *Pink*).

Ker sta obe uporabljeni govorni zbirki posneti preko telefonske linije, so bile motnje pred prištevanjem h govornemu signalu omejene na ustrezno frekvenčno območje (300 - 3400 Hz). Motnje so bile posnetkom dodane v treh ciljnih razmerjih signal/šum: 12dB, 6dB in 0dB. Na ta način je nastalo 12 testnih množic, s katerimi je bila ocenjena robustnost SRG.

Razmerje signal/šum je v našem primeru povprečno razmerje v celotnem posnetku. Premori med besedami v zbirki ŠTEVKE so precej daljši kot premori v zbirki NUMBERS. Ker se premori upoštevajo pri računanju energije posnetka, mora biti energija dodanih motenj ustrezno manjša, da dosežemo želeno razmerje signal/šum na nivoju celotnega posnetka. Dolžina premorov torej vpliva na razmerje signal/šum na odsekih, kjer je govor dejansko prisoten. Posledično je bila uspešnost razpoznavanja v prisotnosti motenj pri poskusih z govorno zbirko ŠTEVKE precej večja kot pri poskusih z zbirko NUMBERS. V rezultatih so podane povprečne uspešnosti razpoznavanja posameznih besed v testnih množicah, v katerih so motnje dodane v enakem razmerju signal/šum. Npr. z *"aditivne motnje 6dB"* je označena povprečna uspešnost razpoznavanja testnih množic, v katerih so prisotne aditivne motnje v razmerju signal/šum 6dB.

[23] *angl. Signal to Noise Ratio*

6.2 Prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed

V tem podpoglavju je opisan postopek prilagajanja dolžine okvirja fonemski zgradbi besed. Fonemsko zgradbo se določi s pomočjo klasičnega SRG, pri katerem se pri računanju spektra uporabi okvir fiksne dolžine. Nato se izvede še razpoznavanje s prilagodljivo dolžino okvirja. Pri določanju dolžine okvirja se uporabi fonemska zgradba, ki je bila določena v prvem prehodu. Razpoznavanje torej poteka v dveh prehodih, v prvem se spekter računa z okvirjem fiksne dolžine, v drugem pa se na podlagi rezultatov iz prvega prehoda dolžino okvirja prilagodi glede na fonemsko zgradbo vzorcev.

6.2.1 Prilagajanje dolžine okvirja

Na postopek učenja SRG prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed nima bistvenega vpliva. Že za izvedbo učenja večnivojskega perceptrona je namreč potrebna na nivoju kontekstno odvisnih kategorij označena učna množica. Razmejitev govornega signala na kontekstno odvisne kategorije je torej na voljo in se jo lahko uporabi v postopku parametrizacije tudi za določanje dolžine okvirja pri računanju spektra signala. Dodaten prehod je potreben samo pri razpoznavanju razvojne in testne množice.

Pri poskusih z obema govornima zbirkama je bil za določitev fonemske zgradbe v prvem prehodu uporabljen SRG z 32 ms dolgim okvirjem, ki je izmed vseh SRG s fiksno dolžino okvirja najmanj občutljiv na motnje (slika 21). Razpoznavanje s tem SRG se izvede do točke, ko dobimo razmejitev govornega signala v kontekstno odvisne kategorije. Nato se postopek parametrizacije in razpoznavanja ponovi še enkrat. Pri računanju spektra se pri različnih kontekstno odvisnih kategorijah uporabi različno dolg okvir.

Poudariti je potrebno, da uporaba večjega števila različno dolgih okvirjev ni smiselna. Če želimo z dinamičnim prilagajanjem dolžine okvirja izboljšati uspešnost razpoznavanja, je potrebno v drugem prehodu pravilno razpoznati (vsaj nekatere) besede, ki so bile v prehodu s fiksno dolžino okvirja razpoznane napačno. Če je določena beseda v prvem prehodu napačno razpoznana, je seveda napačna tudi fonemska zgradba, na osnovi katere se v drugem prehodu prilagaja dolžino okvirja. Prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed zato temelji na predpostavki, da pri

razpoznavanju najpogosteje pride do napačnega razpoznavanja oziroma zamenjave besed, ki so si po fonemski zgradbi podobne. Primer sta npr. besedi *"dve"* in *"devet"*, katero mnogo govorcev izgovarja kot *"dvet"*. Če se npr. pri parametrizaciji zapornikov in premora med besedami (oziroma ustreznih kontekstno odvisnih kategorij) uporabi krajši okvir kot pri vseh ostalih fonemih, se bo tudi v primeru, da je v prvem prehodu prišlo do zamenjave besed, v drugem prehodu pri parametrizaciji zapornikov /d/ in /t/ uporabil krajši okvir kot v prvem prehodu, zato bo verjetnost, da bosta razpoznana pravilno, večja. Zahteva, da pri parametrizaciji podobnih fonemov uporabimo enako dolg okvir je razlog, da sta bili pri poskusih v tem poglavju, podobno kot v [98], uporabljeni samo dve dolžini okvirja. V nadaljevanju sta predstavljena dva načina prilagajanja dolžine okvirja na osnovi fonemske zgradbe besed.

6.2.2 *Uporaba daljšega okvirja pri parametrizaciji srednjih delov trifonov*

V prvi skupini poskusov z govorno zbirko ŠTEVKE je bil daljši okvir uporabljen samo pri parametrizaciji kontekstno neodvisnih kategorij (to so srednji deli trifonov). Na ta način se zagotovi boljšo frekvenčno ločljivost spektra pri parametrizaciji srednjih delov dolgih fonemov, kjer se spekter spreminja zelo počasi. Pri parametrizaciji krajših fonemov in kategorij na prehodih med fonemi je bil uporabljen krajši okvir. Preizkušeni so bili okvirji dolžine 10 - 32 ms. Kot krajši okvir so bili preizkušeni okvirji dolžin 10, 12, 14 in 16 ms, kot daljši okvir pa okvirji dolžin 16, 24 in 32 ms (v nekaterih poskusih je bil 16 ms dolg okvir uporabljen kot krajši, v nekaterih pa kot daljši okvir). Kategorije, pri katerih je bil pri računanju spektra uporabljen daljši okvir, so navedene v tabeli 11.

kategorija	okvir	kategorija	okvir
<E>	daljši	<i>	daljši
<O>	daljši	<o>	daljši
<a>	daljši	ostale	krajši
<e>	daljši		

Tabela 11: Kategorije, pri katerih je bil uporabljen daljši okvir.

Ostale, oziroma vse kategorije, ki se pojavijo v zbirki, so navedene v poglavju 3.1 o govorni zbirki ŠTEVKE (tabela 2). Rezultati poskusov so prikazani v tabeli 12.

V oznaki poskusov je najprej navedena dolžina krajšega in nato dolžina daljšega okvirja v milisekundah. Tako je bil npr. pri izvedbi poskusa, ki je označen z *fon-trif_10_16* pri parametrizaciji srednjih delov trifonov uporabljen 16 ms dolg okvir, pri vseh ostalih kategorijah pa 10 ms dolg okvir. V zadnjih dveh vrsticah so rezultati, ki so bili doseženi z okvirjem fiksne dolžine 24 ms (največja uspešnost pri razpoznavanju testne množice brez dodanih motenj) in 32 ms (okvir take dolžine je bil uporabljen za določitev fonemske zgradbe v prvem prehodu).

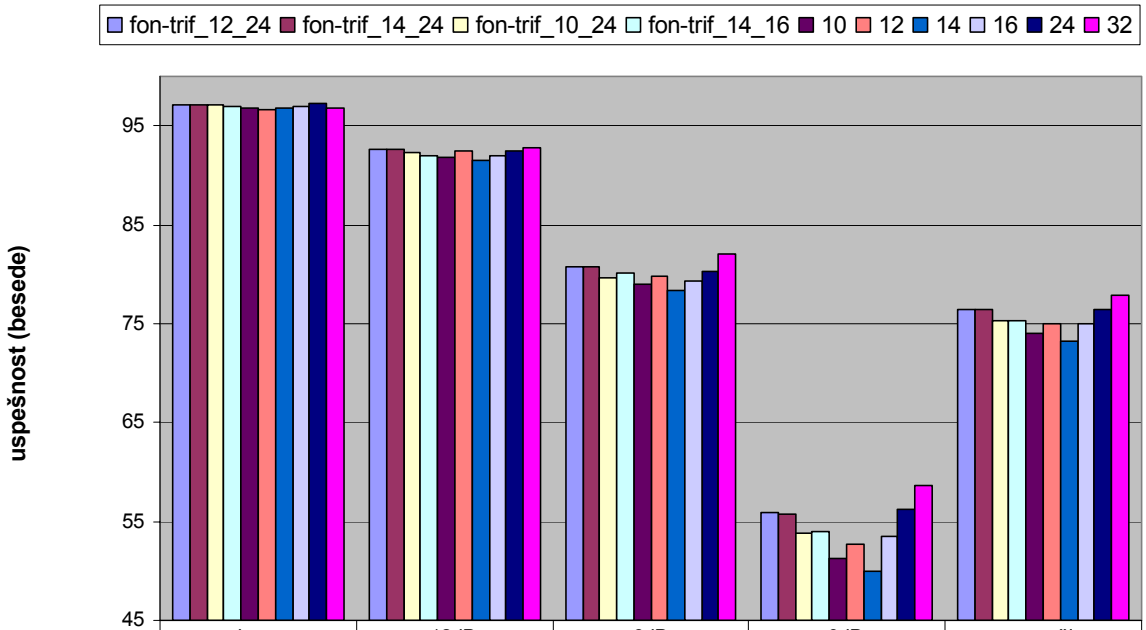
poskus	brez	12dB	6dB	0dB	povprečje
fon-trif_10_16	96,7	91,6	77,8	50,1	73,2
fon-trif_10_24	97,0	92,3	79,7	53,7	75,2
fon-trif_10_32	96,8	92,7	80,5	53,8	75,7
fon-trif_12_16	96,8	92,3	79,1	52,3	74,6
fon-trif_12_24	97,1	92,6	80,8	55,9	76,4
fon-trif_12_32	96,7	92,5	79,3	50,7	74,2
fon-trif_14_16	96,9	92,1	80,1	54,0	75,4
fon-trif_14_24	97,1	92,7	80,8	55,8	76,4
fon-trif_14_32	96,5	92,5	80,4	52,0	74,9
fon-trif_16_24	96,7	92,9	81,7	57,1	77,2
fon-trif_16_32	96,6	92,9	80,9	56,4	76,7
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Tabela 12: Rezultati uporabe daljšega okvirja pri parametrizaciji srednjih delov trifonov. Za primerjavo so dodani rezultati, ki so bili doseženi z uporabo okvirjev fiksne dolžine 24 in 32 ms. Nanašajo se na govorno zbirko ŠTEVKE.

Uspešnost razpoznavanja je izražena z odstotkom pravilno razpoznanih besed v vzorcih testne množice. Podane so uspešnosti razpoznavanja testne množice brez dodanih motenj in testnih množic z dodanimi aditivnimi motnjami v razmerju signal/šum 12dB, 6dB in 0dB, ter povprečna uspešnost razpoznavanja v prisotnosti aditivnih motenj. Opazimo lahko, da so bili najuspešnejši poskusi, pri katerih je bil pri parametrizaciji srednjih delov trifonov uporabljen 24 ms dolg okvir. V primerjavi s SRG iz prvega prehoda, pri katerem se spekter izračuna z okvirjem fiksne dolžine 32 ms (več podatkov o SRG s fiksno dolžino okvirja je na sliki 21), se uspešnost razpoznavanja testne množice brez dodanih motenj skoraj ni spremenila. Povprečna uspešnost razpoznavanja v prisotnosti aditivnih motenj se je v primerjavi s SRG iz prvega prehoda zmanjšala.

Primerjava štirih SRG, pri katerih je bila uspešnost razpoznavanja testne množice brez dodanih

motenj največja (*fon-trif_12_24*, *fon-trif_14_24*, *fon-trif_10_24* in *fon-trif_14_16*) s SRG, pri katerih se uporablja okvir fiksne dolžine, je prikazana na sliki 21.



	brez	12dB	6dB	0dB	povprečje
fon-trif_12_24	97,1	92,6	80,8	55,9	76,4
fon-trif_14_24	97,1	92,7	80,8	55,8	76,4
fon-trif_10_24	97,0	92,3	79,7	53,7	75,2
fon-trif_14_16	96,9	92,1	80,1	54,0	75,4
10	96,8	91,9	79,0	51,3	74,1
12	96,7	92,4	79,8	52,7	75,0
14	96,8	91,6	78,4	49,9	73,3
16	96,9	92,1	79,3	53,5	74,9
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Slika 21: Primerjava uporabe daljšega okvirja pri parametrizaciji srednjih delov trifonov. Rezultati se nanašajo na govorno zbirko ŠTEVKE.

SRG so označeni enako kot v tabeli 12. SRG, pri katerih se uporablja okvir fiksne dolžine, so označeni z dolžino okvirja. SRG, pri katerih se dolžina okvirja prilagaja, so označeni z dolžinama krajšega in daljšega okvirja. Med SRG s fiksno dolžino okvirja močno izstopa SRG, pri katerem je bil uporabljen 24 ms dolg okvir. To odstopanje je nenavadno veliko in ga pripisujemo majhnosti testne množice - do podobnega odstopanja pri večji govorni zbirki NUMBERS ni prišlo. Posledično so bili razmeroma uspešni tudi SRG, pri katerih se dolžino okvirja prilagaja in je daljši

okvir dolg 24 ms. Pri SRG, pri katerih se uporablja okvir fiksne dolžine, ki je krajši od 24 ms , lahko opazimo precej večjo občutljivost na motnje. Iz rezultatov lahko sklepamo, da je dobro robustnost mogoče doseči že, če se pri parametrizaciji srednjih delov trifonov uporabi (vsaj) 24 ms dolg okvir. Uporaba krajšega okvirja pri parametrizaciji ostalih kategorij robustnosti ni bistveno poslabšala. Torej je dovolj, da so v odseke, na katerih uporabimo daljši okvir zajeti srednji deli trifonov. Z uporabo različnih metod za prilagajanje dolžine okvirja na ostalih odsekih je mogoče poskusiti izboljšati uspešnost in/ali robustnost razpoznavanja.

6.2.3 Uporaba daljšega okvirja pri parametrizaciji zvenečih fonemov

Prilagajanje dolžine okvirja zvenečim in nezvenečim fonemom temelji na predpostavki, da se spekter govornega signala pri večini zvenečih fonemov (izjema so zveneči zaporniki) spreminja počasi. Zaradi tega je pri parametrizaciji zvenečih fonemov smiselno uporabiti daljši, pri nezvenečih fonemih pa krajši okvir. Krajši okvir je bil uporabljen tudi pri vseh kontekstno odvisnih kategorijah, ki se pojavijo na meji zvenečih fonemov z nezvenečimi. Preizkušeni so bili okvirji enakih dolžin kot pri poskusih iz prejšnjega podpoglavja. Kategorije, pri katerih je bil uporabljen krajši okvir, so navedene v tabeli 13.

kategorija	**okvir**	**kategorija**	**okvir**	**kategorija**	**okvir**	**kategorija**	**okvir**
<\.pau>	krajši	a>\.pau	krajši	m>\.pau	krajši	s>_t	krajši
\.pau<E	krajši	_d<d	krajši	\.pau<n	krajši	s>e	krajši
s<E	krajši	S<e	krajši	\.pau<o	krajši	_t<t	krajši
t<O	krajši	p<e	krajši	o>s	krajši	t>\.pau	krajši
O>_p	krajši	s<e	krajši	_p<p	krajši	t>O	krajši
\.pau<S	krajši	e>\.pau	krajši	p>\.pau	krajši	t>i	krajši
S>_t	krajši	e>_t	krajši	p>e	krajši	t>r	krajši
S>e	krajši	e>s	krajši	t<r	krajši	_tS<tS	krajši
<_d>	krajši	t<i	krajši	\.pau<s	krajši	tS>\.pau	krajši
<_p>	krajši	i>\.pau	krajši	e<s	krajši	ostale	daljši
<_t>	krajši	i>_tS	krajši	o<s	krajši		
<_tS>	krajši	\.pau<j	krajši	s>@	krajši		

Tabela 13: Kategorije, pri katerih je bil uporabljen krajši okvir.

Rezultati poskusov so prikazani v tabeli 14. Poskusi so označeni z dolžinama krajšega in daljšega okvirja. Rezultati so glede uspešnosti razpoznavanja testne množice brez dodanih motenj podobni

kot pri poskusih iz prejšnjega podpoglavja, v katerih je bil daljši okvir uporabljen pri parametrizaciji srednjih delov trifonov.

poskus	brez	12dB	6dB	0dB	povprečje
fon-zv_10_16	96,9	92,3	80,3	54,2	75,6
fon-zv_10_24	96,9	93,1	82,4	60,7	78,7
fon-zv_10_32	96,9	93,0	82,6	60,1	78,6
fon-zv_12_16	97,2	92,3	79,2	52,8	74,8
fon-zv_12_24	96,6	92,9	81,8	59,4	78,0
fon-zv_12_32	96,4	92,6	81,5	60,1	78,1
fon-zv_14_16	97,0	92,5	79,9	53,3	75,3
fon-zv_14_24	96,5	93,0	82,6	61,8	79,1
fon-zv_14_32	96,3	92,5	81,8	60,7	78,4
fon-zv_16_24	96,4	92,9	82,5	60,4	78,6
fon-zv_16_32	96,8	92,5	82,2	60,8	78,5
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Tabela 14: Rezultati uporabe daljšega okvirja pri parametrizaciji zvenečih fonemov. Dodani so rezultati, ki so bili doseženi z uporabo okvirjev fiksne dolžine 24 in 32 ms. Nanašajo se na govorno zbirko ŠTEVKE.

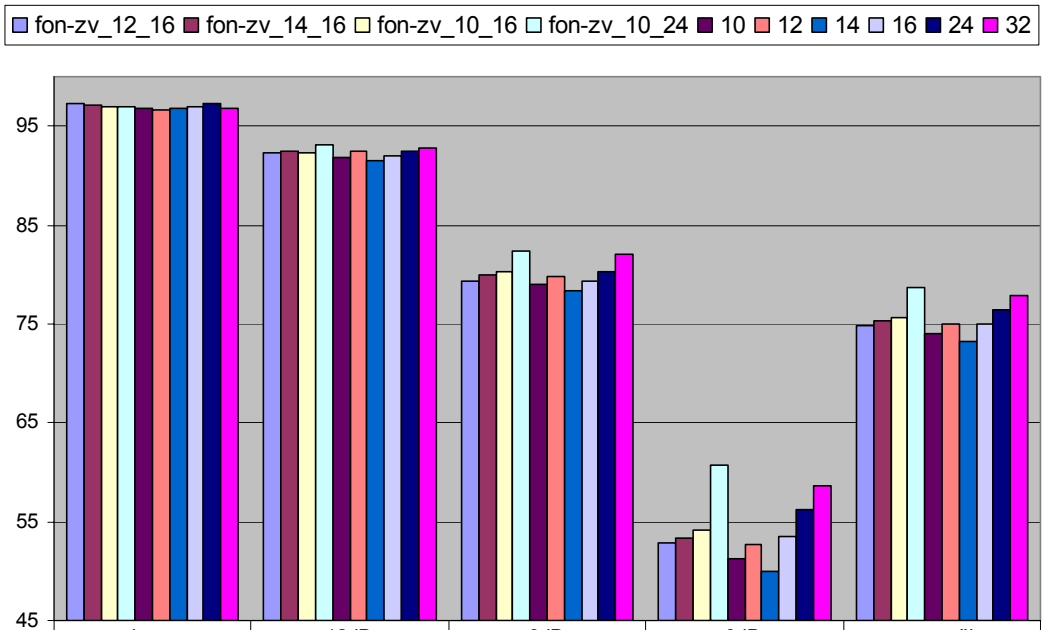
Robustnost se je izboljšala in je bila v več poskusih boljša od SRG s fiksno dolžino okvirja 32 ms (SRG iz prvega prehoda). Žal je bila robustnost največja pri poskusih, pri katerih je bila uspešnost razpoznavanja testne množice brez dodanih motenj najmanjša.

Na sliki 22 je prikazana primerjava štirih SRG, pri katerih je bila uspešnost razpoznavanja testne množice največja (*fon-zv_12_16*, *fon-zv_14_16*, *fon-zv_10_16* in *fon-zv_10_24*) in SRG s fiksno dolgimi okvirji. Pri nekaterih poskusih so rezultati prilagajanja dolžine okvirja (izbiranje ene izmed dveh dolžin okvirja) boljši, kot če se uporabi okvir samo ene dolžine, vendar je razlika zelo majhna. To velja npr. če primerjamo SRG označene s *fon-zv_12-16*, pri katerem je uspešnost večja kot pri SRG, pri katerih je uporabljen okvir fiksne dolžine 12 oziroma 16 ms (SRG označena z *10* oziroma *16*).

6.2.4 Ugotovitve

Oba postopka prilagajanja dolžine okvirja na osnovi fonemske zgradbe besed sta na uspešnost

razpoznavanja testne množice brez dodanih motenj vplivala zelo malo. Največja slabost prilagajanja dolžine okvirja na osnovi fonemske zgradbe besed je, da sta za razpoznavanje potrebna dva prehoda. Ker je pri parametrizaciji podobnih fonemov potrebno uporabiti enako dolg okvir, podrobnih informacij o fonemski zgradbi za prilagajanja dolžine okvirja ni mogoče izkoristiti.



	brez	12dB	6dB	0dB	povprečje
fon-zv_12_16	97,2	92,3	79,2	52,8	74,8
fon-zv_14_16	97,0	92,5	79,9	53,3	75,3
fon-zv_10_16	96,9	92,3	80,3	54,2	75,6
fon-zv_10_24	96,9	93,1	82,4	60,7	78,7
10	96,8	91,9	79,0	51,3	74,1
12	96,7	92,4	79,8	52,7	75,0
14	96,8	91,6	78,4	49,9	73,3
16	96,9	92,1	79,3	53,5	74,9
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Slika 22: Primerjava uporabe daljšega okvirja pri parametrizaciji zvenečih fonemov. Rezultati se nanašajo na govorno zbirko ŠTEVKE.

Težave se pri uporabi daljšega okvirja pri parametrizaciji srednjih delov trifonov pojavijo tudi v zvezi z robustnostjo. Kljub temu, da je bil v prvem prehodu pri parametrizaciji uporabljen 32 ms dolg okvir, ki je glede robustnosti najboljši, v prisotnosti motenj pri razmejitvi na foneme očitno

nastane preveč odstopanj. Posledično se pri parametrizaciji (lahko) uporabi okvir napačne dolžine. Fonemska zgradba je torej za določanje dolžine okvirja preveč podrobna, saj enako dolg okvir uporabljamo pri parametrizaciji velike skupine kontekstno odvisnih kategorij, ki pripadajo fonemom s podobnimi lastnostmi spektra (podobnost glede potrebe po boljši časovni ali frekvenčni ločljivosti). Za razdelitev fonemov v tako velike skupine pa poznavanje podrobne fonemske zgradbe ni potrebno.

Zveneče odseke govora je npr. mogoče določiti z uporabo metod za določanje osnovne frekvence v govornem signalu, ki je razmeroma dobro raziskano področje [81, 83, 84]. Podobno so odseki, na katerih se spekter signala spreminja hitro, povezani s spremembami jakosti govornega signala. Te so zelo značilne npr. za zapornike in tudi za prehode med fonemi. Osnovna ideja prilagajanja frekvenčno-časovne ločljivosti spektra na osnovi zvenečih odsekov govora in glede na spremembe jakosti zvoka je torej primerljiva s pristopoma, ki sta bila predstavljena v tem poglavju, vendar poznavanje fonemske zgradbe in s tem razpoznavanje v dveh prehodih ni potrebno.

6.3 Prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka

Večje spremembe jakosti govornega signala so značilne za prehode med fonemi. Pri samoglasnikih, ki so modelirani s trifoni, to pomeni, da so spremembe jakosti najmanjše ravno pri srednjih delih. Pri njih je smiselno z daljšim okvirjem poudariti frekvenčno ločljivost spektra. Tudi pri izgovorjavi nekaterih fonemov, predvsem zapornikov, pride do velikih sprememb jakosti. Za zapornike je značilno, da se vokalni trakt za kratek čas zapre. Nato pride do nenadnega odprtja, kar se odrazi s skoraj trenutnim povečanjem jakosti govornega signala. Torej je z uporabo sprememb jakosti zvoka, frekvenčno-časovno ločljivost mogoče prilagoditi tudi spremembam spektra pri zapornikih. Zaznavanje sprememb jakosti v govornem signalu je bistveno manj zapleteno od prej opisanega določanja fonemske zgradbe oziroma razpoznavanja v dveh prehodih.

Pri poskusih, ki so predstavljeni v nadaljevanju poglavja, je bila za določanje sprememb jakosti zvoka uporabljena akustična značilka, s katero so upoštevana tudi spoznanja o lastnostnih človekovega sluha [82]. Z občutljivostjo človekovega sluha na spremembe jakosti zvoka se je v

svojih raziskavah ukvarjal Moore [86]. Izvedel je večje število različnih poskusov in vedno prišel do istega zaključka: pomembna značilnost zaznavanja relativnih sprememb jakosti zvoka je odvisnost od absolutne jakosti zvoka. Če je absolutna jakost majhna, človek lahko zazna bistveno manjše spremembe, kot če je absolutna jakost velika. Jakost zvoka je mogoče oceniti z energijo signala v določenem časovnem odseku. Zaznavanje sprememb jakosti je Moore modeliral z naslednjo enačbo:

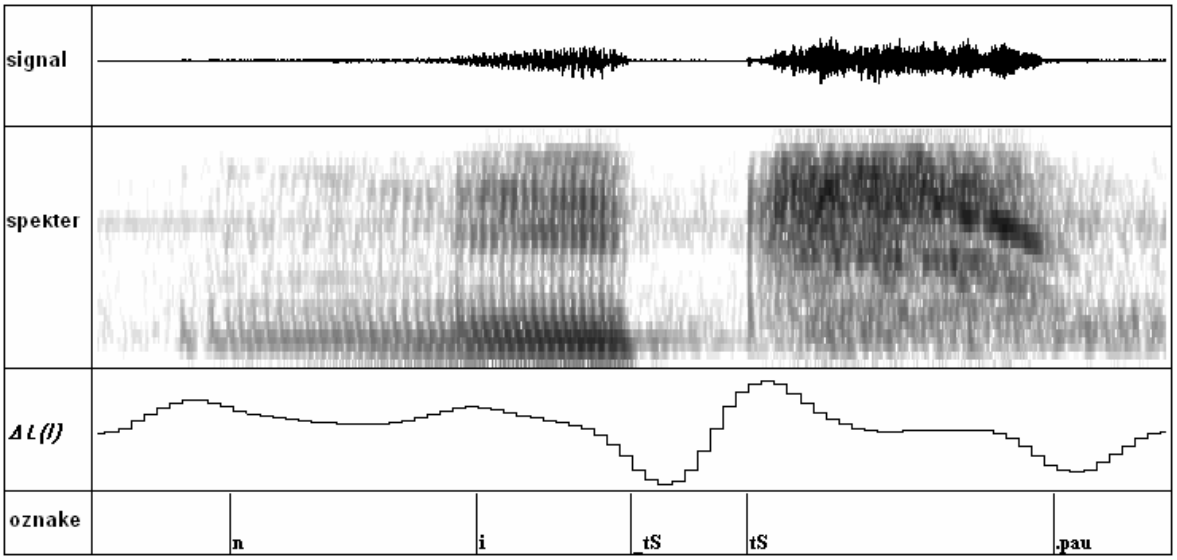
$$\Delta L(l) = 10\log\left(\frac{I_l(l) + \Delta I(l)}{I_l(l)}\right). \tag{35}$$

V enačbi je z $\Delta L(l)$ označena mera za zaznano relativno spremembo jakosti, $I_l(l)$ je absolutna jakost, $\Delta I(l)$ pa predstavlja absolutno spremembo jakosti zvoka, ki se jo izračuna z uporabo enačbe (36).

$$\Delta I(l) = \frac{\sum_{j=1}^{20} j\left(I_s(l+j) - I_s(l-j)\right)}{2\sum_{j=1}^{20} j^2}. \tag{36}$$

Če želimo s pomočjo enačbe (35) zaznati relativne spremembe jakosti, do katerih pride na prehodih med fonemi, je potrebno za računanje $I_l(l)$ in $\Delta I(l)$ uporabiti primerno dolga okvirja. Pri uporabljeni akustični značilki je bil za računanje absolutne jakosti $I_l(l)$ je uporabljen 250 ms dolg okvir. Ta dolžina ustreza povprečni dolžini zlogov. Pri računanju $I_s(l)$ je bil uporabljen 40 ms dolg okvir. Ta dolžina ustreza dolžini najkrajših fonemov. Pri računanju so bile spremembe jakosti z rezanjem omejene na interval [-1, 1]. Če je $\Delta L(l)$ večji kot 1 oziroma manjši kot -1, se v nadaljevanju upošteva vrednost 1 oziroma -1. $I_l(l)$ in $I_l(l)$ predstavljata energijo signala znotraj okvirjev navedenih dolžin.

Primer tako izračunanih relativnih sprememb jakosti zvoka za besedo *"nič"* je prikazan na sliki 23. Razvidno je, da do največjih relativnih sprememb jakosti pride pri prehodih med fonemi, znotraj posameznega fonema pa so relativne spremembe jakosti blizu ničle. Prav tako lahko opazimo, da se tudi spekter signala znotraj fonemov spreminja zelo malo, do največjih sprememb pride na prehodih med fonemi.



Slika 23:Relativne spremembe jakosti zvoka pri besedi "nič". Prikazani so: signal, spekter signala, relativne spremembe jakosti $\Delta L(l)$ *in fonemska zgradba besede.*

6.3.1 Prilagajanje frekvenčno-časovne ločljivosti pri računanju značilk

Relativne spremembe jakosti ΔL so opisane s številom z intervala [-1, 1], kar omogoča mehko spreminjanje frekvenčno-časovne ločljivosti spektra. Ostri prehodi na mejah fonemov pri poskusih iz prejšnjega poglavja so namreč problematični. Tudi če je fonemska zgradba določena pravilno, običajno pride do odstopanja pri časovni določitvi mej med fonemi. Zaradi odstopanja, se pri parametrizaciji nekaterih okvirjev na meji fonemov uporabi napačna dolžina okvirja. Z mehčanjem prehodov je ta problem mogoče omiliti.

Mehčanje prehodov pri prilagajanju frekvenčno-časovne ločljivosti spektra na prehodih med fonemi je bilo doseženo z uporabo obtežene vsote dveh spektrov. Spekter je bilo potrebno izračunati dvakrat, z uporabo kratkega ($\mathbf{X_s}$(l)) in z uporabo dolgega okvirja ($\mathbf{X_l}$(l)). Obtežena vsota spektrov je bila določena z enačbo

$$\mathbf{X}(l) = u\mathbf{X}_s(l) + (1-u)\mathbf{X}_l(l). \qquad (37)$$

Na mestih, kjer so spremembe jakosti velike, se poudari s kratkim okvirjem izračunani spekter. Utež *u* je sorazmerna izračunani relativni spremembi jakosti zvoka.

$$u = 0{,}8\left|\Delta L(l)\right| \tag{38}$$

V enačbi (37) je **X**(l) je izračunana obtežena vsota, $\mathbf{X_s}$(l) s kratkim okvirjem izračunani spekter in $\mathbf{X_l}$(l) z dolgim oknom izračunani spekter. Utež *u* je določena kot absolutna vrednost sprememb jakosti zvoka, pomnožena s faktorjem 0,8. Ta faktor je bil v predhodnih poskusih določen s poskušanjem [97]. Večji *u* torej pomeni, da se nahajamo na odseku, kjer so spremembe jakosti velike in se v nadaljnjem postopku parametrizacije bolj upošteva s kratkim okvirjem izračunani spekter.

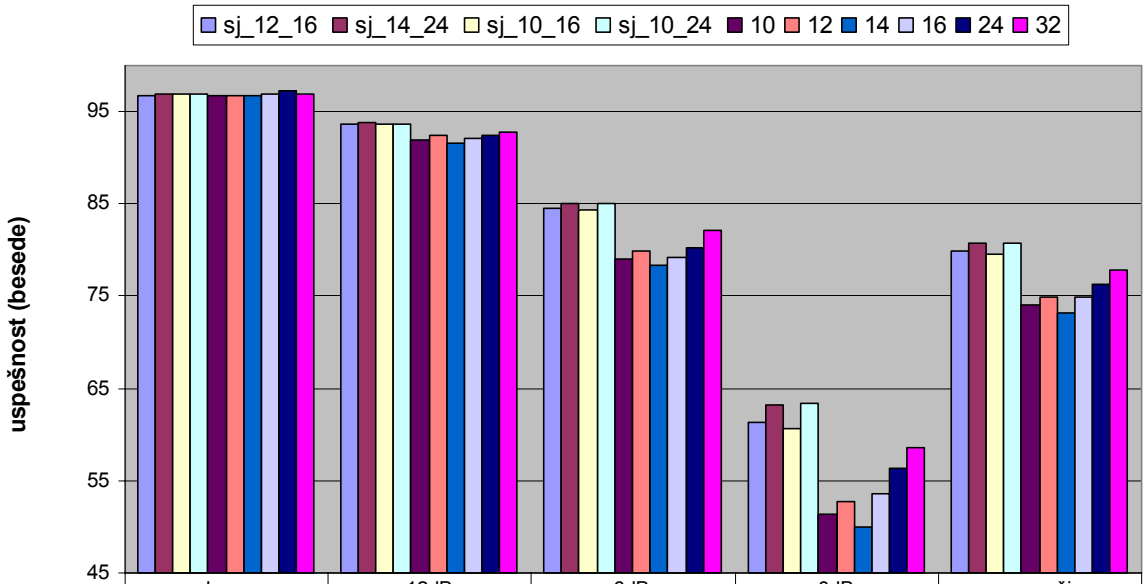
6.3.2 Poskusi z zbirko ŠTEVKE

Postopek prilagajanja frekvenčno-časovne ločljivosti spremembam jakosti zvoka je bil preizkušen z uporabo različno dolgih okvirjev pri računanju obeh spektrov. Za računanje spektra z boljšo časovno ločljivostjo so bili uporabljeni okvirji dolžine 10, 12 in 14 ms, za računanje spektra z boljšo frekvenčno ločljivostjo pa okvirji dolžine 16, 24 in 32 ms.

poskus	brez	12dB	6dB	0dB	povprečje
sj_10_16	96,9	93,6	84,3	60,7	79,5
sj_10_24	97,0	93,7	85,1	63,5	80,8
sj_10_32	96,3	93,5	83,4	62,0	79,6
sj_12_16	96,8	93,6	84,6	61,4	79,9
sj_12_24	96,7	93,3	84,3	62,6	80,1
sj_12_32	96,3	93,5	83,9	63,4	80,2
sj_14_16	96,4	93,8	85,1	62,7	80,5
sj_14_24	96,8	93,8	85,0	63,3	80,7
sj_14_32	96,0	93,8	84,0	62,4	80,0
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Tabela 15: Rezultati prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Dodani so rezultati, ki so bili doseženi z uporabo okvirjev fiksne dolžine 24 in 32 ms. Nanašajo se na govorno zbirko ŠTEVKE.

V tabeli so prikazani rezultati poskusov, pri katerih je bila pri računanju značilk uporabljena opisana obtežena vsota dveh spektrov. SRG so označeni podobno kot v prejšnjem poglavju - z dolžinama okvirjev, ki sta bila uporabljena pri računanju spektra z boljšo časovno ločljivostjo in spektra z boljšo frekvenčno ločljivostjo. Uspešnost razpoznavanja testne množice je bila v večini primerov primerljiva z uspešnostjo SRG s fiksno dolžino okvirja 32 ms. Opazimo lahko, da pri uspešnosti razpoznavanja testne množice brez dodanih motenj navzdol odstopajo poskusi, v katerih je bil pri računanju spektra z boljšo frekvenčno ločljivostjo uporabljen 32 ms dolg okvir. Očitno je bil okvir predolg. Pri robustnosti večjih razlik ni. V vseh primerih je bila občutno večja kot pri obeh SRG s fiksno dolžino okvirja.



	brez	12dB	6dB	0dB	povprečje
sj_12_16	96,8	93,6	84,6	61,4	79,9
sj_14_24	96,8	93,8	85,0	63,3	80,7
sj_10_16	96,9	93,6	84,3	60,7	79,5
sj_10_24	97,0	93,7	85,1	63,5	80,8
10	96,8	91,9	79,0	51,3	74,1
12	96,7	92,4	79,8	52,7	75,0
14	96,8	91,6	78,4	49,9	73,3
16	96,9	92,1	79,3	53,5	74,9
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Slika 24: Primerjava prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Rezultati se nanašajo na govorno zbirko ŠTEVKE.

Primerjava SRG iz štirih poskusov, pri katerih je bila uspešnost razpoznavanja testne množice brez dodanih motenj največja (*sj_12_16*, *sj_14_24*, *sj_10_16* in *sj_10_24*), je prikazana na sliki 24. Razvidno je, da je prilagajanje frekvenčno-časovne ločljivosti spremembam jakosti govornega signala v vseh primerih povečalo robustnost razpoznavanja. Na uspešnost razpoznavanja brez dodanih motenj je bil vpliv zanemarljiv. Primerjava *sj_10_24* in *32* kaže, da se je število napak pri razpoznavanju testnih množic z dodanimi motnjami v povprečju zmanjšalo za 13%[24]. Uspešnost razpoznavanja testne množice brez dodanih motenj je pri obeh SRG skoraj enaka.

6.3.3 Poskusi z zbirko NUMBERS

Za preizkus prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti je bila uporabljena tudi govorna zbirka NUMBERS. Rezultati so prikazani v tabeli 16. Pri računanju spektrov so bili uporabljeni enako dolgi okvirji kot pri poskusih z zbirko ŠTEVKE.

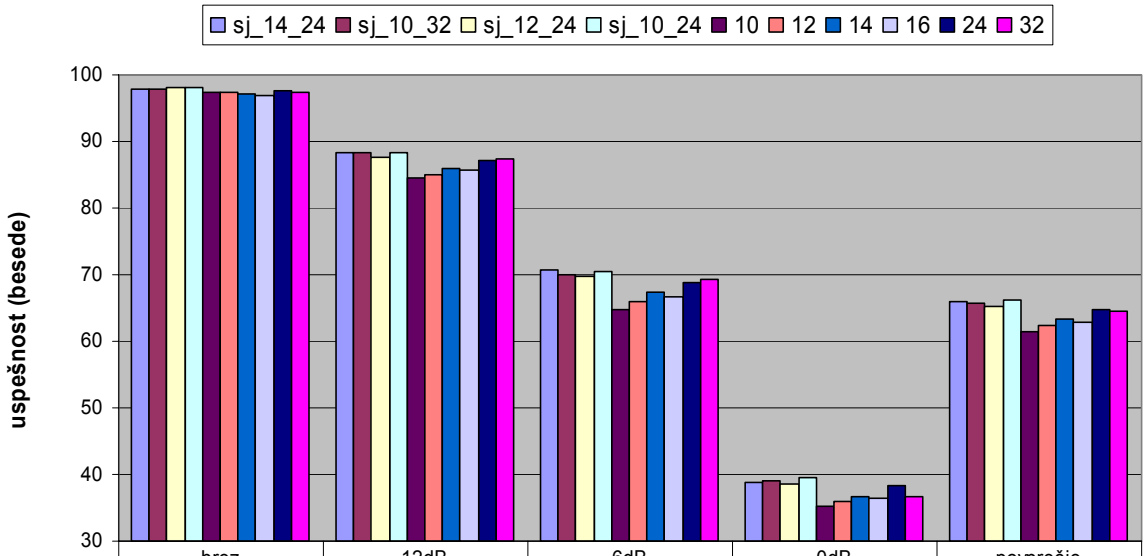
poskus	brez	12dB	6dB	0dB	povprečje
sj_10_24	98,17	88,28	70,48	39,52	66,09
sj_10_32	97,95	88,22	70,00	39,05	65,76
sj_12_24	97,98	87,71	69,75	38,47	65,31
sj_12_32	97,57	88,21	69,91	38,30	65,47
sj_14_24	97,91	88,27	70,68	38,76	65,90
sj_14_32	97,68	88,39	70,55	39,23	66,06
sj_16_24	97,68	87,49	69,89	38,33	65,24
sj_16_32	97,83	88,90	72,12	40,66	67,23
24	97,57	87,03	68,78	38,22	64,68
32	97,41	87,50	69,22	36,64	64,45

Tabela 16: Rezultati prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Dodani so rezultati, ki so bili doseženi z uporabo okvirjev fiksne dolžine 24 in 32 ms. Nanašajo se na govorno zbirko NUMBERS.

Razvidne so majhne razlike pri uspešnosti in robustnosti posameznih SRG. Tako glede uspešnosti pri razpoznavanju testne množice brez dodanih motenj kot pri robustnosti navzgor nekoliko odstopa SRG označen s *sj_10_24*.

[24] *podano je relativno zmanjšanje števila napak; angl. Word Error Rate (WER)*

Na sliki 25 je prikazana primerjava SRG označenih s *sj_14_24*, *sj_10_32*, *sj_12_24* in *sj 10_24* s SRG s fiksno dolžino okvirja. Postopek je v vseh primerih izboljšal tako uspešnost razpoznavanja brez dodanih motenj kot robustnost razpoznavanja. Pri razpoznavanju testne množice brez dodanih motenj se je število napak glede na najuspešnejši SRG, pri katerem je spekter izračunan z okvirjem fiksne dolžine 24 ms, zmanjšalo za 24% (primerjava *sj_10_24* in *32*). Pri razpoznavanju testnih množic z dodanimi aditivnimi motnjami se je število napak v povprečju zmanjšalo za 4% (primerjava istih SRG).



	brez	12dB	6dB	0dB	povprečje
sj_14_24	97,91	88,27	70,68	38,76	65,90
sj_10_32	97,95	88,22	70,00	39,05	65,76
sj_12_24	97,98	87,71	69,75	38,47	65,31
sj_10_24	98,17	88,28	70,48	39,52	66,09
10	97,49	84,51	64,85	35,26	61,54
12	97,30	85,08	65,93	36,05	62,35
14	97,15	85,99	67,48	36,68	63,38
16	96,86	85,60	66,70	36,31	62,87
24	97,57	87,03	68,78	38,23	64,68
32	97,41	87,50	69,22	36,64	64,45

Slika 25: Primerjava prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Rezultati se nanašajo na govorno zbirko NUMBERS.

6.3.4 *Ugotovitve*

Prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka omogoča izboljšanje uspešnosti in robustnosti SRG. Glede na to, da je postopek računsko nezahteven, ga je priporočljivo uporabljati. Izboljšanje robustnosti je bilo pri poskusih z govorno zbirko NUMBERS v primerjavi z rezultati iz poskusov z zbirko ŠTEVKE razmeroma majhno (če primerjamo robustnost *sj_10_24* in *24*, se je število napak zmanjšalo le za 4%, v primerjavi s 13% pri zbirki ŠTEVKE) . Uspešnost razpoznavanja testne množice brez dodanih motenj se je pri govorni zbirki NUMBERS povečala. Razlogi za razlike pri poskusih z obema govornima zbirkama so najverjetneje v fonemski zgradbi besed v zbirkah.

6.4 Prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora

Rezultati prilagajanja dolžine okvirja na osnovi fonemske zgradbe besed, ki je bilo opisano v poglavju 6.2 kažejo, da je s prilagajanjem dolžine okvirja zvenečim in nezvenečim fonemom mogoče izboljšati robustnost razpoznavanja govora. Ker pa je potrebno določiti fonemsko zgradbo, sta potrebna dva prehoda razpoznavanja, kar je glavna slabost pristopa. V tem poglavju je predstavljeno prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora, ki se določijo s pomočjo ustrezne akustične značilke, za kar dodatni prehod ni potreben.

Za določanje zvenečih odsekov govora je bil uporabljen novejši postopek, ki ga je predstavil Hosom [82]. Ta postopek upošteva spoznanja o človekovem sluhu. V preteklosti smo ga že uspešno uporabili za računanje akustičnih značilk [45]. Opisan je v nadaljevanju poglavja.

6.4.1 Določanje zvenečih odsekov govora

Postopek za določanje zvenečih in nezvenečih odsekov govora temelji na zaznavanju periodičnih sprememb moči govornega signala, ki so povezane z osnovno frekvenco. Zaznavanje sprememb moči je omejeno na območje prvega formanta oziroma vrha z najnižjo frekvenco v spektru govornega signala.

Za določanje sprememb jakosti zvoka je uporabljen pristop, ki je bil podrobneje opisan v poglavju 6.3. Prilagojen je zaznavanju periodičnih sprememb moči signala v območju prvega formanta. Ker želimo zaznati spremembe moči, ki so povezane z osnovno frekvenco, sta bila za računanje jakosti in spremembe jakosti v enačbi (35) uporabljena zelo kratka okvirja. Okvir za računanje jakosti $I_l(l)$ mora zajeti nekaj period signala. Okvir za računanje $I_s(l)$ mora biti dovolj kratek, da z njim zaznamo spremembe energije signala znotraj posamezne periode. Tako dobimo spremembe jakosti znotraj periode, kot jih zazna človek v odvisnosti od absolutne jakosti zvoka, ki je izračunana za nekaj period signala. Ali je odsek govora zveneč ali ne, se določi glede na prisotnost periodičnih relativnih sprememb jakosti. Če so periodične spremembe jakosti prisotne, je odsek govora zveneč,

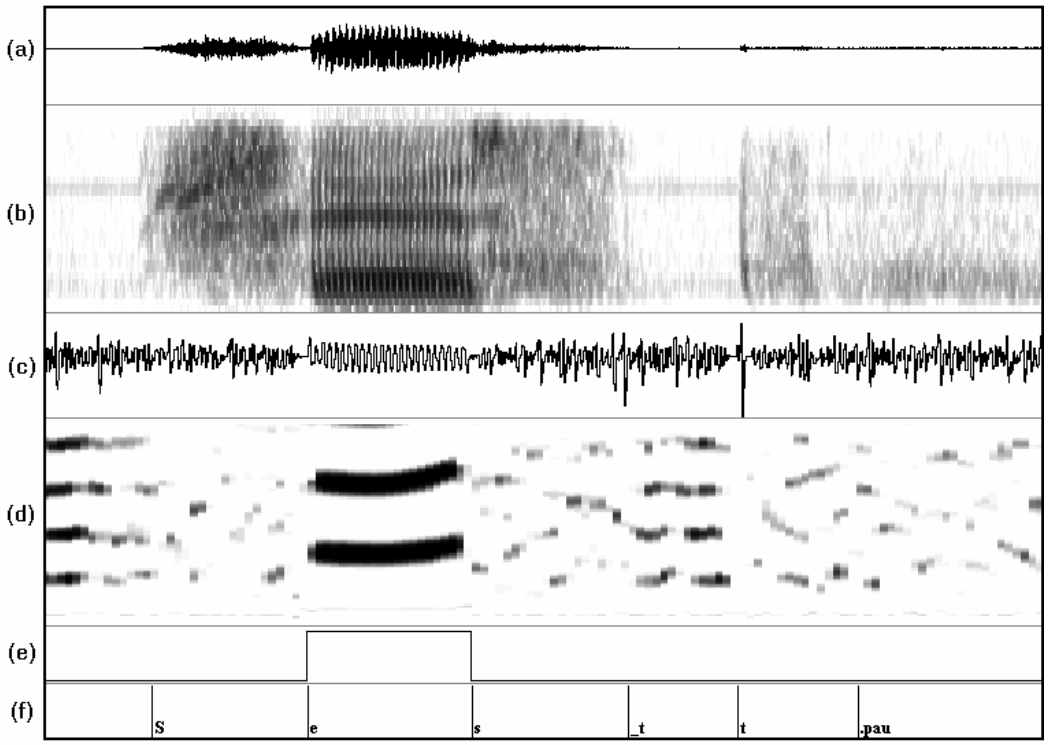
sicer pa ne. Postopek sestoji iz naslednjih korakov:

1. Filtriranje govornega signala s pasovno prepustnim filtrom 160 - 700 Hz. S tem se iz signala odstrani nizkofrekvenčni šum in zvoke povezane z dihanjem. Odstranijo se tudi visokofrekvenčne komponente signala.

2. V dobljenem signalu se z uporabo enačbe (35) izračuna relativne spremembe jakosti zvoka. Pri tem se uporabi zelo kratka okvirja, za računanje $I_l(l)$ okvir dolžine 45 ms in za $\Delta I(l)$ 4 ms dolg okvir. Okvir za računanje absolutne jakosti $I_l(l)$ zajema nekaj period osnovne frekvence. Okvir za računanje $I_s(l)$ oziroma $\Delta I(l)$ mora biti dovolj kratek, da omogoča zaznavanje sprememb energije signala, ki so povezane z osnovno frekvenco. Rezultat analize je pri zvenečih odsekih govora zaporedje enakomerno razmaknjenih impulzov (razmik je perioda osnovne frekvence). Če je odsek govora nezveneč, je rezultat zaporedje neenakomerno razmaknjenih impulzov, povezanih z neperiodičnimi spremembami moči signala v območju prvega formanta.

3. Za dobljeno zaporedje impulzov se izračuna avtokorelacija. Nato se z različnimi pragovi glede na vrhove v avtokorelaciji in relativne spremembe moči signala določi zveneče in nezveneče odseke govora. Metoda upošteva skupno 17 parametrov, ki so bili določeni s poskušanjem. Podrobnosti so v [82], stran 84.

Na sliki 26 je prikazano določanje periodičnih odsekov govora za besedo *"šest"*. V spektru signala (b) so pri fonemu /e/ opazne enakomerno razmaknjene vertikalne temne lise, ki so povezane z osnovno frekvenco. Posledično je rezultat računanja sprememb jakosti zvoka zaporedje enakomerno razmaknjenih impulzov (c). V (d) je prikazana avtokorelacija zaporedja impulzov z razločno vidnima vrhovoma pri zvenečem fonemu /e/. Določitev zvenečih odsekov (e) se zelo dobro ujema s fonemsko označitvijo (f).



Slika 26: Določanje zvenečih in nezvenečih odsekov govora pri besedi "šest". Na sliki so od zgoraj navzdol prikazani: signal (a), spekter signala (b), izračunano zaporedje impulzov (c), avtokorelacija (d), določitev periodičnosti (e) in ročno določena fonemska zgradba besede (f).

6.4.2 Prilagajanje dolžine okvirja

Kot je bilo že omenjeno, temelji prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora na predpostavki, da se pri zvenečih odsekih govora spekter signala spreminja razmeroma počasi. Zato se lahko zadovoljimo s slabšo časovno ločljivostjo spektra in na takih odsekih uporabimo daljši okvir. Tudi če dobre frekvenčne ločljivosti ne bi potrebovali, je pri zvenečih odsekih potrebno uporabiti daljši okvir - če se uporabi prekratek okvir, se v spektru, podobno kot na sliki 26 pri fonemu /e/, pojavijo z osnovno frekvenco povezana nihanja. Pri klasifikaciji v akustične modele taka nihanja predstavljajo motnjo in so nezaželena. Pri nezvenečih fonemih je zaradi boljše časovne ločljivosti smiselno uporabiti kratek okvir.

6.4.3 *Poskusi z zbirko ŠTEVKE*

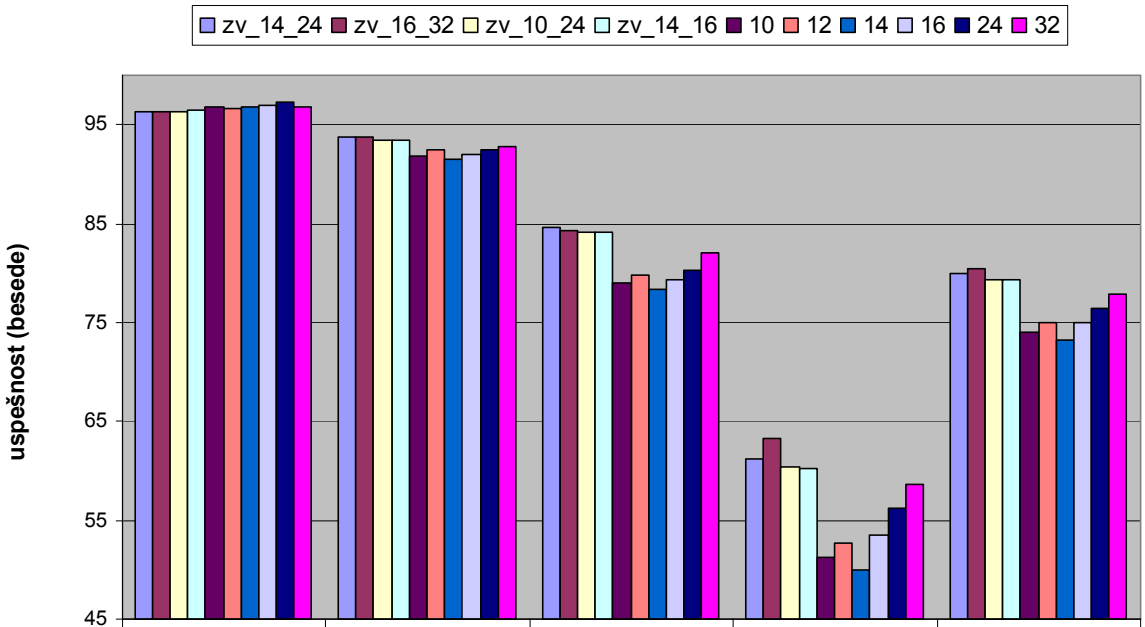
Rezultati poskusov z govorno zbirko ŠTEVKE so podani v tabeli 17. Razvidno je, da se je z uporabo opisane metode za določanje zvenečih odsekov govora izboljšala robustnost razpoznavanja. Glede robustnosti so rezultati precej boljši od rezultatov, ko se zveneče odseke določi na osnovi fonemske zgradbe besed (stran 67). Uspešnost razpoznavanja testne množice brez dodanih motenj se je poslabšala.

poskus	brez	12dB	6dB	0dB	povprečje
zv_10_16	96,3	93,2	82,6	59,1	78,3
zv_10_24	96,4	93,5	84,1	60,3	79,3
zv_10_32	96,1	93,4	84,1	60,4	79,3
zv_12_16	96,2	92,8	82,5	58,9	78,1
zv_12_24	96,0	93,4	85,2	62,9	80,5
zv_12_32	96,0	93,5	83,3	61,0	79,3
zv_14_16	96,6	93,5	84,2	60,3	79,3
zv_14_24	96,3	93,8	84,7	61,2	79,9
zv_14_32	96,2	93,6	84,5	63,5	80,5
zv_16_24	96,2	93,6	83,9	62,9	80,1
zv_16_32	96,3	93,7	84,3	63,3	80,5
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Tabela 17: Rezultati prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora. Dodani so rezultati, ki so bili doseženi z uporabo okvirjev fiksne dolžine 24 in 32 ms. Nanašajo se na zbirko ŠTEVKE.

Primerjava prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora in SRG s fiksno dolžino okvirja je prikazana na sliki 27.

V primerjavo so vključeni SRG, označeni z *zv_14_24*, *zv_16_32*, *zv_10_24* in *zv_14_16*. Če te SRG primerjamo z rezultati, ki so bili doseženi z uporabo okvirja fiksne dolžine, lahko opazimo, da se je uspešnost razpoznavanja testne množice brez dodanih motenj zmanjšala. Robustnost je bila v vseh primerih boljša, kot če bi pri parametrizaciji uporabili okvir fiksne dolžine.



	brez	12dB	6dB	0dB	povprečje
zv_14_24	96,3	93,8	84,7	61,2	79,9
zv_16_32	96,3	93,7	84,3	63,3	80,5
zv_10_24	96,4	93,5	84,1	60,3	79,3
zv_14_16	96,6	93,5	84,2	60,3	79,3
10	96,8	91,9	79,0	51,3	74,1
12	96,7	92,4	79,8	52,7	75,0
14	96,8	91,6	78,4	49,9	73,3
16	96,9	92,1	79,3	53,5	74,9
24	97,3	92,5	80,3	56,3	76,4
32	96,8	92,8	82,1	58,6	77,8

Slika 27 : Primerjava prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora. Rezultati se nanašajo na govorno zbirko ŠTEVKE.

Zanimiva je tudi primerjava z rezultati, ki so bili doseženi z določanjem zvenečih in nezvenečih odsekov na osnovi fonemske zgradbe (slika 22 na strani 68). Uspešnost razpoznavanja testne množice brez dodanih motenj se je nekoliko poslabšala. Sklepamo lahko, da je v odsotnosti motenj s pomočjo fonemske zgradbe zveneče in nezveneče odseke mogoče določiti natančneje. Prednost postopka, ki je predstavljen v tem poglavju, je torej razpoznavanje v enem prehodu ter razmeroma veliko izboljšanje robustnosti. Pri *zv_16_32* je bilo število napak v primerjavi s *32* manjše za 12%.

6.4.4 *Poskusi z zbirko NUMBERS*

Postopek je bil na podoben način kot z zbirko ŠTEVKE preizkušen tudi z govorno zbirko NUMBERS. Rezultati so prikazani v tabeli:

poskus	brez	12dB	6dB	0dB	povprečje
zv_10_24	97,38	85,85	66,04	34,16	62,02
zv_10_32	97,41	86,83	67,19	33,41	62,48
zv_12_24	97,45	87,39	68,00	36,53	63,97
zv_12_32	97,30	86,20	67,06	35,68	62,98
zv_14_24	97,41	88,55	69,94	38,91	65,80
zv_14_32	97,00	85,81	67,52	37,00	63,44
zv_16_24	97,53	88,52	70,47	38,98	65,99
zv_16_32	97,45	87,83	69,96	38,51	65,43
24	97,57	87,03	68,78	38,22	64,68
32	97,41	87,50	69,22	36,64	64,45

Tabela 18 : Rezultati prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora. Dodani so rezultati, ki so bili doseženi z uporabo okvirjev fiksne dolžine 24 in 32 ms. Nanašajo se na govorno zbirko NUMBERS.

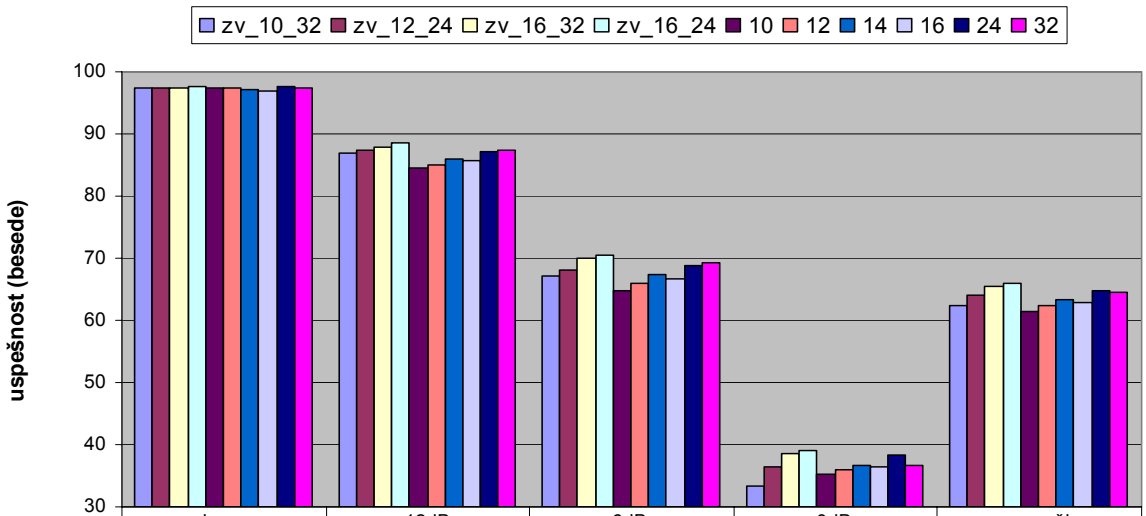
Uspešnost razpoznavanja testne množice brez dodanih motenj se v primerjavi s SRG z okvirji fiksne dolžine skoraj ni spremenila. Robustnost se je pri večini poskusov v primerjavi z uporabo daljših okvirjev fiksne dolžine poslabšala. Uporaba kratkih okvirjev na nezvenečih odsekih je nanjo vplivala negativno.

S SRG s fiksno dolžino okvirja so bili primerjani naslednji SRG: *zv_10_32*, *zv_12_24*, *zv_16_32* in *zv_16_24*. Primerjava je prikazana na sliki 28. Tudi ta slika kaže na nespremenjeno uspešnost razpoznavanja v odsotnosti motenj. Razvidno je tudi, da je robustnost odvisna od dolžine krajšega izmed obeh okvirjev.

6.4.5 *Ugotovitve*

Poskusi z govorno zbirko ŠTEVKE kažejo, da je s prilagajanjem dolžine okvirja zvenečim in nezvenečim odsekom govora mogoče izboljšati robustnost razpoznavanja. Določanje zvenečih odsekov z metodo, opisano v tem poglavju, se odrazi v nekoliko boljših rezultatih kot če se zveneče

odseke določi s pomočjo fonemske zgradbe besed. To velja predvsem za robustnost. Zaradi večje hitrosti parametrizacije je v tem poglavju opisan pristop boljši od določanja zvenečih odsekov s pomočjo fonemske zgradbe, ki zahteva razpoznavaje v dveh prehodih.



	brez	12dB	6dB	0dB	povprečje
zv_10_32	97,41	86,83	67,19	33,41	62,48
zv_12_24	97,45	87,39	68,00	36,53	63,97
zv_16_32	97,45	87,83	69,96	38,51	65,43
zv_16_24	97,53	88,52	70,47	38,98	65,99
10	97,49	84,51	64,85	35,26	61,54
12	97,30	85,08	65,93	36,05	62,35
14	97,15	85,99	67,48	36,68	63,38
16	96,86	85,60	66,70	36,31	62,87
24	97,57	87,03	68,78	38,23	64,68
32	97,41	87,50	69,22	36,64	64,45

Slika 28: Primerjava prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora. Rezultati se nanašajo na govorno zbirko NUMBERS.

Na rezultate poskusov z govorno zbirko NUMBERS je bil vpliv prilagajanja dolžine okvirja manjši. Ne glede na to, da gre po velikosti slovarja za podobni zbirki, med njima obstajajo precejšnje razlike. Med fonemi besed iz govorne zbirke ŠTEVKE je 11 zapornikov. V govorni zbirki NUMBERS najdemo le 3. Opazimo lahko tudi, da so v zbirki NUMBERS skoraj vsi fonemi zveneči. Zato prilagajanje dolžine okvirja glede na zveneče in nezveneče odseke govora ni smiselno, saj prostora za izboljšavo uspešnosti skoraj ni. Vpliv na uspešnost razpoznavanja je bil pri govorni zbirki NUMBERS zanemarljiv. Tudi vpliv na robustnost je bil zelo majhen. Pri SRG označenih z *zv_16_32* in *zv_16_24* je opazno manjše izboljšanje glede na najrobustnejša SRG s

fiksno dolžino okvirja (označena sta s *24* in *32*). Robustnost se je torej izboljšala, če smo kot krajši okvir uporabili okvir dolžine 16 ms, ki je bil najdaljši izmed vseh preizkušenih. To potrjuje prej omenjeno tezo; ker je večina fonemov z zbirki NUMBERS zvenečih, uporaba zelo kratkih okvirjev ni smiselna.

7. PRIMERJAVA REZULTATOV IN DODATNO TESTIRANJE ROBUSTNOSTI

V prejšnjem poglavju predstavljeni rezultati kažejo, da je s prilagajanjem dolžine okvirja, odvisno od uporabljenega postopka, mogoče izboljšati tako uspešnost kot robustnost razpoznavanja. Vpliv na robustnost je bil v tem poglavju še dodatno preverjen z več motnjami. Za vsak metodo iz prejšnjega poglavja sta bila po dva poskusa, pri katerih je bila uspešnost razpoznavanja testne množice brez dodanih motenj največja, razširjena z dodatnimi aditivnimi motnjami in dvema konvolutivnima motnjama. Pri poskusih, ki so opisani v tem poglavju, je bila pri poskusih z govorno zbirko NUMBERS uporabljena celotna zbirka.

7.1 Motnje za testiranje robustnosti

Za testiranje robustnosti so bile dodatno uporabljene naslednje aditivne motnje iz zbirke NOISEX [25]:

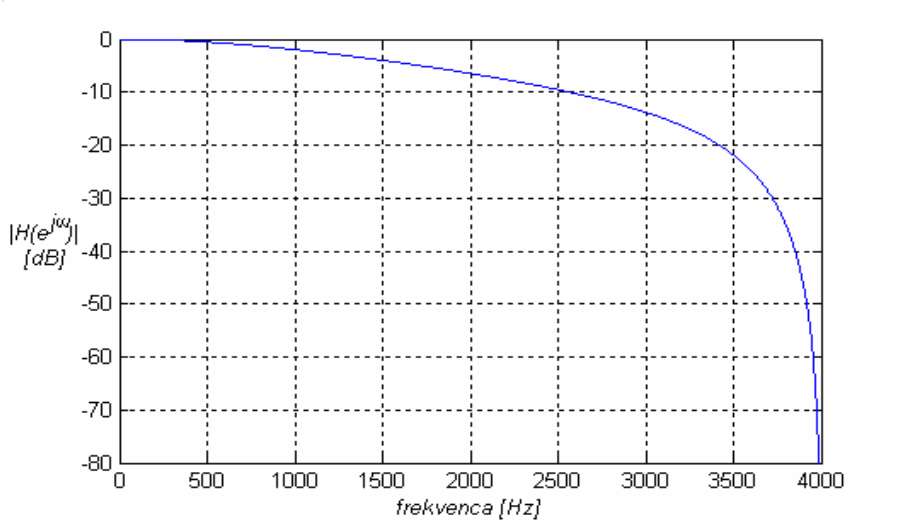
- hrup v vojaškem letalu F16 (angl. oznaka *F16*);
- pasovno omejen šum v frekvenčnem pasu 800 - 1000 Hz (angl. oznaka *Pass900hz*);
- hrup v tovarni (angl. oznaka *Factory*).

Pasovno omejen šum je dobljen iz belega šuma. Ta je bil s pasovno prepustnim filtrom omejen na frekvenčni pas 800 - 1000 Hz. Motnje so bile, enako kot pri poskusih v prejšnjem poglavju, dodane treh razmerjih signal/šum: 0dB, 6dB in 12dB. Poleg aditivnih motenj sta bili samostojno in v kombinaciji z aditivnimi motnjami uporabljeni še dve konvolutivni motnji.

7.1.1 *Konvolutivne motnje*

Konvolutivne motnje se v govornem signalu najpogosteje pojavijo zaradi vpliva prenosnega kanala (npr. telefonska linija) in značilnosti naprav, s katerimi govor zajemamo. Mednje spadajo tudi akustični odmevi, ki nastanejo pri govorjenju v zaprtih prostorih, kjer se zvok odbija od sten in drugih predmetov. Za preizkušanje robustnosti sta bili izbrani dve konvolutivni motnji, ki se pojavljata zelo pogosto: slabljenje visokih frekvenc in odjek[25].

Slabljenje visokih frekvenc je bilo simulirano z uporabo nizkoprepustnega filtra z neskončnim enotinim odzivom. Njegov amplitudni odziv je prikazan na sliki 29.



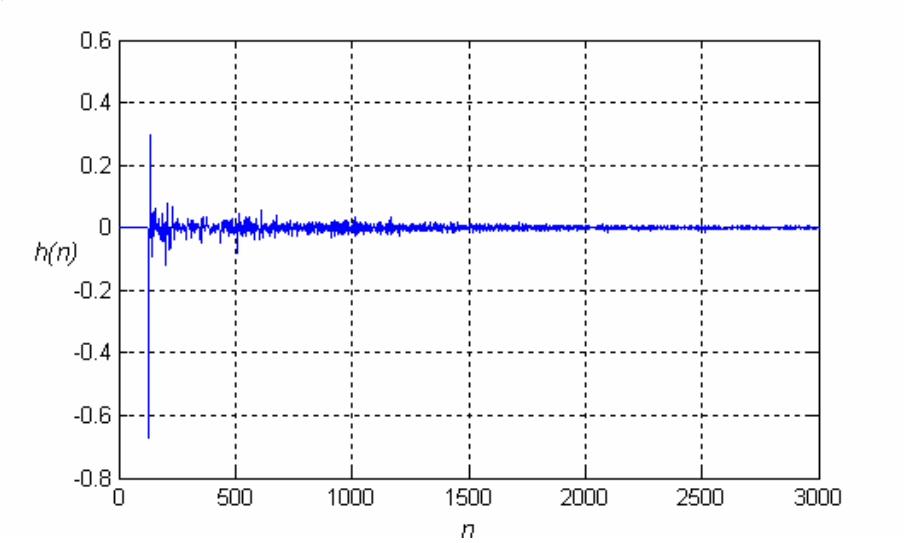
Slika 29: Amplitudni odziv nizkoprepustnega filtra za simulacijo slabljenja visokih frekvenc.

Do podobnih popačitev signala med drugim pride v primeru, ko je med govorcem in poslušalcem oziroma mikrofonom ovira.

Za simulacijo odjeka je bil uporabljen filter s končnim odzivom na enotin impulz dolžine N=8192. Ker so koeficienti z višjimi indeksi zelo majhni, je na sliki 30 prikazanih samo prvih 3000 koeficientov. Enotin odziv na sliki je bil izmerjen v realnem akustičnem prostoru.

---

[25] *angl. reverb – motnja je značilna za zaprte prostore, v katerih se zvok odbija od sten*



Slika 30: Odziv na enotin impulz filtra za simulacijo odjeka.

Obe konvolutivni motnji sta bili uporabljeni za popačitev testne množice brez dodanih motenj in v kombinaciji z vsemi aditivnimi motnjami. Robustnost razpoznavanja je bila testirana s skupno 7 različnimi aditivnimi motnjami v treh razmerjih signal/šum v kombinaciji z dvema konvolutivnima motnjama, učno množico brez dodanih motenj, ter učno množico brez dodanih aditivnih motenj, popačeno s konvolutivnima motnjama. Skupaj je bil vsak SRG torej preizkušen s 7*3*3+1*3=66 testnimi množicami.

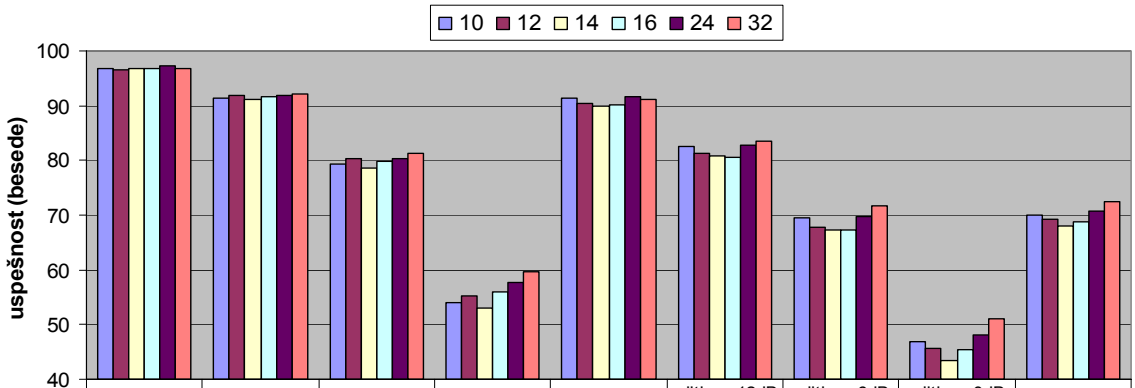
7.2 Poskusi z govorno zbirko ŠTEVKE

V nadaljevanju so prikazani rezultati podrobnejšega testiranja vpliva prilagajanja frekvenčno-časovne ločljivosti spektra na robustnost razpoznavanja. V rezultatih je naveden odstotek pravilno razpoznanih besed. Zaradi velikega števila različnih testnih množic so, podobno kot v prejšnjem poglavju, navedeni povprečni rezultati razpoznavanja skupin testnih množic, v katerih so bile prisotne motnje podobne vrste. V prikazu rezultatov je v skrajnem levem stolpcu podana uspešnost razpoznavanja testne množice brez dodanih motenj. Sledijo stolpci, ki prikazujejo povprečno uspešnost razpoznavanja pod vplivom različnih aditivnih motenj. Gre za povprečno uspešnost razpoznavanja sedmih testnih množic, ki so dobljene z dodajanjem govora v ozadju, hrupa v avtomobilu, hrupa v vojaškem letalu, hrupa v tovarni, pasovno omejenega šuma ter belega in roza šuma v navedenem razmerju signal/šum. Sledi povprečna uspešnost razpoznavanja testnih množic,

ki sta iz testne množice dobljeni z uporabo konvolutivnih motenj. V naslednjih treh stolpcih so rezultati kombiniranja aditivnih in konvolutivnih motenj. V najbolj desnem stolpcu je navedena povprečna uspešnost razpoznavanja vseh množic z dodanimi motnjami.

Z obema govornima zbirkama so bili izvedeni tudi poskusi z različno dolgimi okvirji fiksne dolžine. Robustnost pri obeh zbirkah je bila večja v primeru, ko je bil uporabljen razmeroma dolg okvir. Zato sta bila v primerjavo s SRG, pri katerih frekvenčno-časovno ločljivost prilagaja po eni izmed metod iz poglavja 6, vedno vključena tudi SRG s fiksno dolžino okvirja 24 in 32 ms.

Na sliki 31 je prikazana primerjava uspešnosti SRG, pri katerih v postopku parametrizacije uporabimo okvir fiksne dolžine. Enako kot v poglavju 6, so bili uporabljeni 10 - 32 ms dolgi okvirji.

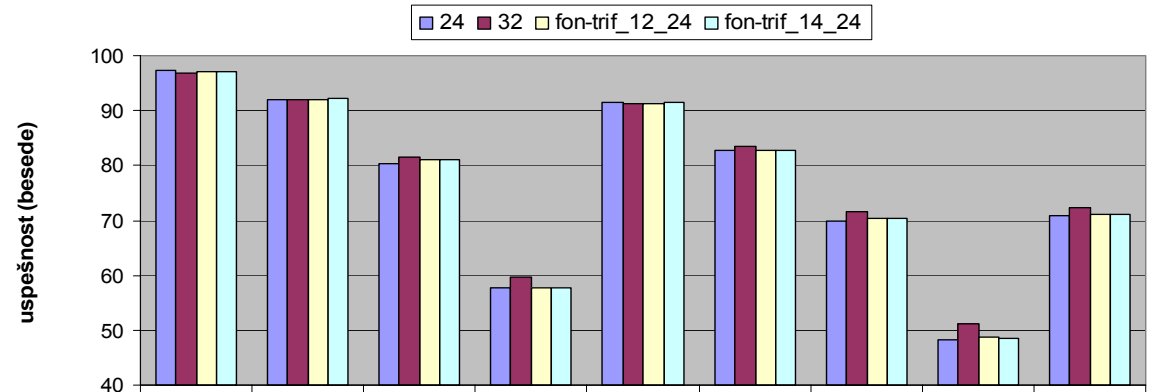


	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
10	96,8	91,4	79,3	54,0	91,4	82,6	69,6	46,8	69,9
12	96,7	92,0	80,3	55,3	90,4	81,2	67,8	45,6	69,2
14	96,8	91,1	78,7	53,0	90,0	80,8	67,2	43,5	68,0
16	96,9	91,7	79,8	55,9	90,1	80,6	67,3	45,4	68,9
24	97,3	92,0	80,4	57,7	91,6	82,7	69,8	48,2	70,8
32	96,8	92,1	81,4	59,7	91,3	83,5	71,6	51,1	72,3

Slika 31: Vpliv dolžine okvirja na uspešnost in robustnost razpoznavanja. Rezultati se nanašajo na govorno zbirko ŠTEVKE.

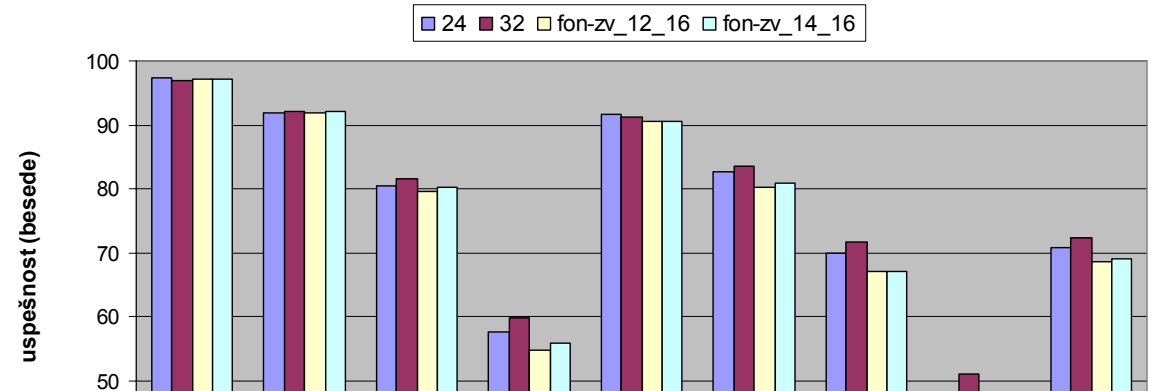
Sledijo rezultati, ki so bili doseženi s prilagajanjem dolžine okvirja na osnovi fonemske zgradbe besed, torej z razpoznavanjem v dveh prehodih. Na sliki 32 so prikazani rezultati, doseženi z uporabo daljšega okvirja pri parametrizaciji srednjih delov trifonov (podpoglavje 6.2.2), na sliki 33 pa rezultati uporabe daljšega okvirja pri parametrizaciji zvenečih fonemov (podpoglavje 6.2.3). Fonemska zgradba razpoznavanih vzorcev je bila, enako kot v poglavju 6.2, določena v

predhodnem prehodu razpoznavanja, v katerem je bil pri računanju spektra uporabljen 32 ms dolg okvir.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	97,3	92,0	80,4	57,7	91,6	82,7	69,8	48,2	70,8
32	96,8	92,1	81,4	59,7	91,3	83,5	71,6	51,1	72,3
fon-trif_12_24	97,1	92,0	81,1	57,7	91,3	82,8	70,3	48,8	71,2
fon-trif_14_24	97,1	92,2	81,0	57,7	91,5	82,6	70,3	48,4	71,1

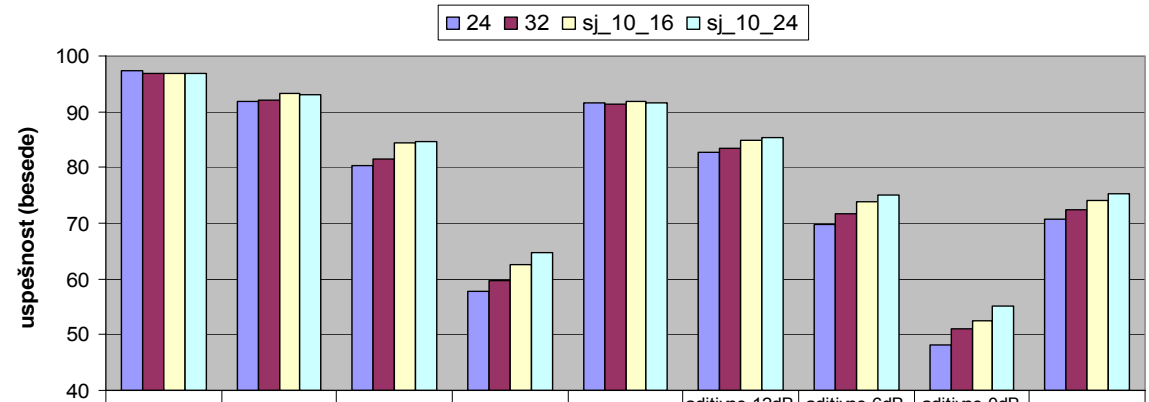
Slika 32: Primerjava uporabe daljšega okvirja pri parametrizaciji srednjih delov trifonov. Rezultati se nanašajo na govorno zbirko ŠTEVKE.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	97,3	92,0	80,4	57,7	91,6	82,7	69,8	48,2	70,8
32	96,8	92,1	81,4	59,7	91,3	83,5	71,6	51,1	72,3
fon-zv_12_16	97,2	91,9	79,7	54,8	90,6	80,2	67,0	44,9	68,5
fon-zv_14_16	97,0	92,0	80,2	55,9	90,5	81,0	67,1	45,4	69,0

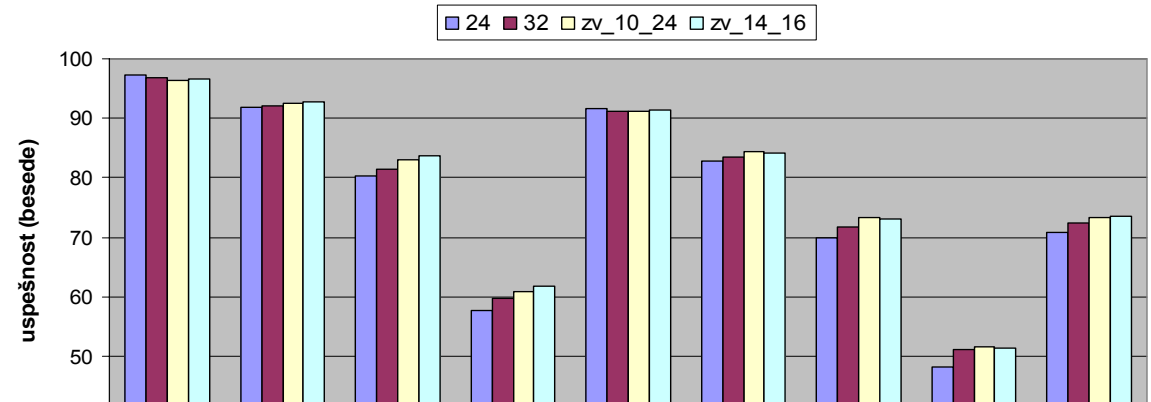
Slika 33: Primerjava uporabe daljšega okvirja pri parametrizaciji zvenečih fonemov. Rezultati se nanašajo na govorno zbirko ŠTEVKE.

Sledijo rezultati poskusov, pri katerih je bilo uporabljeno prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka (poglavje 6.3). Prikazani so na sliki 34.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	97,3	92,0	80,4	57,7	91,6	82,7	69,8	48,2	70,8
32	96,8	92,1	81,4	59,7	91,3	83,5	71,6	51,1	72,3
sj_10_16	96,9	93,3	84,3	62,6	91,9	84,8	74,0	52,4	74,2
sj_10_24	97,0	93,2	84,7	64,7	91,6	85,3	75,1	55,1	75,4

Slika 34: Primerjava prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Rezultati se nanašajo na govorno zbirko ŠTEVKE.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	97,3	92,0	80,4	57,7	91,6	82,7	69,8	48,2	70,8
32	96,8	92,1	81,4	59,7	91,3	83,5	71,6	51,1	72,3
zv_10_24	96,4	92,4	83,1	60,9	91,1	84,3	73,2	51,5	73,3
zv_14_16	96,6	92,8	83,6	61,7	91,5	84,1	73,1	51,4	73,4

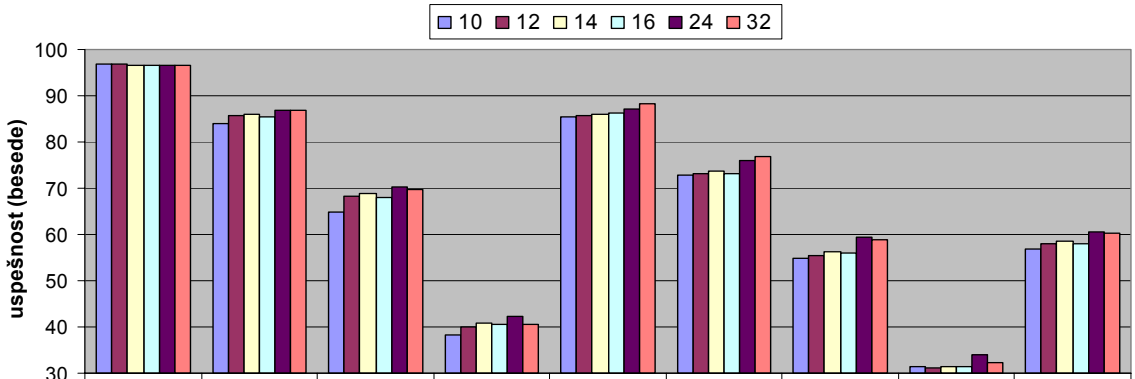
Slika 35: Primerjava prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora. Rezultati se nanašajo na govorno zbirko ŠTEVKE.

Na sliki 35 so rezultati, doseženi s prilagajanjem dolžine okvirja zvenečim in nezvenečim odsekom govora (poglavje 6.4).

Dodatno testiranje robustnosti v vseh primerih potrjuje rezultate iz poglavja 6. Tudi ob dodanih konvolutivnih motnjah je uspešnost pri vseh poskusih padla približno enako kot ob uporabi okvirja fiksne dolžine, torej lahko sklepamo, da dinamično prilagajanje frekvenčno-časovne ločljivosti spektra nima pomembnejšega vpliva na uspešnost razpoznavanja v prisotnosti konvolutivnih motenj.

7.3 Poskusi z govorno zbirko NUMBERS

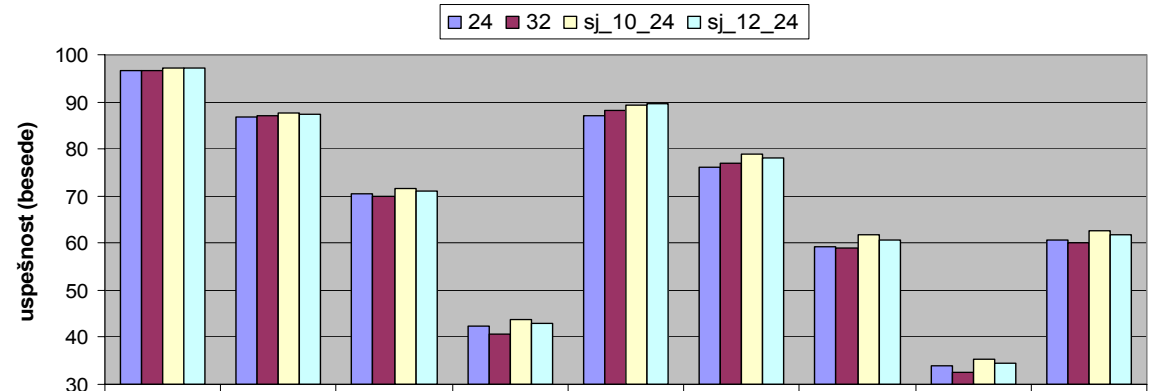
V tem podpoglavju so predstavljeni rezultati podrobnejših testov robustnosti z govorno zbirko NUMBERS. Oblika prikaza je enaka kot v prejšnjem poglavju. Pri poskusih je glede na poglavje 6 več sprememb: poskusi so izvedeni z večjo razvojno in večjimi testnimi množicami. Enako kot pri poskusih v prejšnjem poglavju, je bilo uporabljenih tudi več oblik aditivnih motenj in dve konvolutivni motnji. Na sliki 36 so rezultati poskusov, pri katerih je bil spekter izračunan z okvirji fiksne dolžine 10 - 32 ms.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
10	96,76	83,95	64,99	38,16	85,51	72,72	54,82	31,35	57,00
12	96,77	85,71	68,23	40,12	85,70	73,22	55,56	31,18	57,99
14	96,66	85,91	68,82	40,93	86,14	73,75	56,29	31,44	58,50
16	96,69	85,36	67,95	40,63	86,27	73,13	56,00	31,48	58,13
24	96,69	86,88	70,38	42,36	87,20	75,97	59,31	33,91	60,62
32	96,70	86,95	69,85	40,64	88,23	76,83	58,82	32,39	60,17

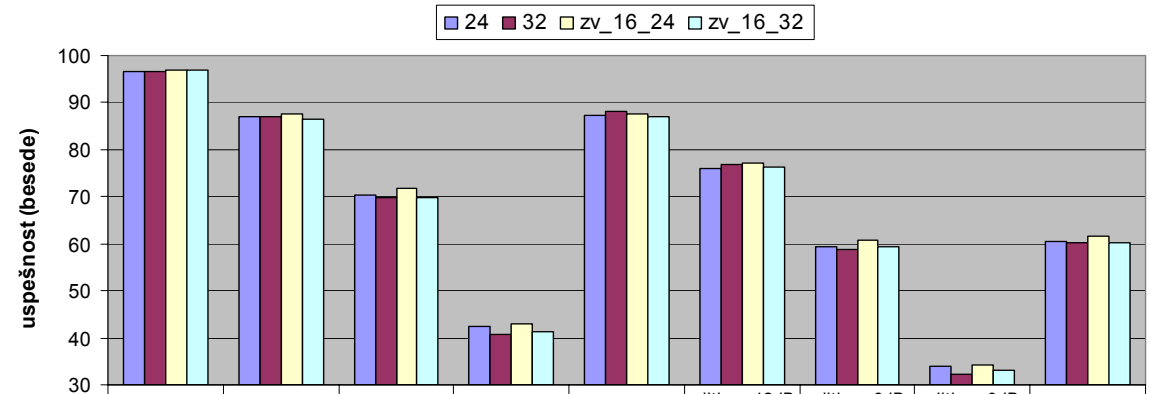
Slika 36: Vpliv dolžine okvirja na uspešnost in robustnost razpoznavanja. Rezultati se nanašajo na govorno zbirko NUMBERS.

Sledijo rezultati, ki so bili doseženi s prilagajanjem frekvenčno-časovne ločljivosti spektra signala spremembam jakosti zvoka (slika 37) in prilagajanjem dolžine okvirja zvenečim in nezvenečim odsekom govora (slika 38).



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	96,69	86,88	70,38	42,36	87,20	75,97	59,31	33,91	60,62
32	96,70	86,95	69,85	40,64	88,23	76,83	58,82	32,39	60,17
sj_10_24	97,13	87,75	71,74	43,78	89,41	78,78	61,67	35,33	62,50
sj_12_24	97,24	87,22	70,94	42,81	89,70	77,95	60,71	34,53	61,70

Slika 37: Primerjava prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Rezultati se nanašajo na govorno zbirko NUMBERS.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	96,69	86,88	70,38	42,36	87,20	75,97	59,31	33,91	60,62
32	96,70	86,95	69,85	40,64	88,23	76,83	58,82	32,39	60,17
zv_16_24	96,81	87,67	71,74	43,06	87,49	77,01	60,67	34,33	61,54
zv_16_32	96,88	86,47	69,82	41,18	87,14	76,26	59,43	33,16	60,32

Slika 38: Primerjava prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora. Rezultati se nanašajo na govorno zbirko NUMBERS.

Rezultati so skladni z rezultati poskusov iz poglavja 6, pri katerih je bil uporabljen samo del govorne zbirke NUMBERS. Dodatne aditivne motnje torej po pričakovanju na rezultate vplivajo podobno kot aditivne motnje, ki so bile uporabljene v poglavju 6. Pri testiranju vpliva aditivnih motenj v kombinaciji s konvolutivnima motnjama se je uspešnost zmanjšala enako kot pri SRG s fiksno dolžino okvirja. Edina izjema je rezultat na sliki 38 pri čisti testni množici, ki je bila popačena s konvolutivnima motnjama. Tu je dinamično prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora na robustnost vplivalo negativno.

7.4 Ugotovitve

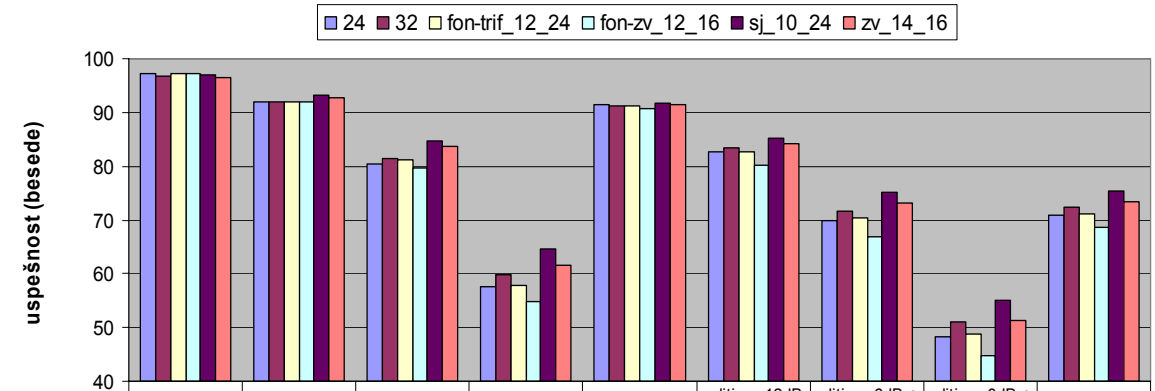
V tem poglavju je bil s pomočjo dveh govornih zbirk najprej preizkušen vpliv dolžine okvirja na uspešnost in robustnost razpoznavanja. Preizkušeni so bili različno dolgi okvirji fiksne dolžine 10 - 32 ms. Rezultati so bili podobni pri obeh govornih zbirkah. Na uspešnost razpoznavanja testne množice brez dodanih motenj izbira dolžine okvirja ni imela večjega vpliva. Izstopa dober rezultat pri poskusu s 24 ms dolgim okvirjem z govorno zbirko ŠTEVKE. Odstopanje je nenavadno veliko in si ga lahko razlagamo le s premajhno velikostjo testne množice. Pri vseh ostalih poskusih z obema govornima zbirkama se rezultati razlikujejo zelo malo. S stališča robustnosti ima primerna izbira dolžine okvirja večji pomen. Pri poskusih so bili, neodvisno od govorne zbirke, doseženi boljši rezultati, ko je bil uporabljen razmeroma dolg okvir dolžine 24 oziroma 32 ms. Rezultati dodatnega testiranja robustnosti potrjujejo, da je z dinamičnim prilagajanjem frekvenčno-časovne ločljivosti pri spektralni analizi govornega signala robustnost mogoče izboljšati. Izboljšanje je mogoče doseči s prilagajanjem dolžine okvirja zvenečim in nezvenečim odsekom govora, za še boljšo pa se je izkazala metoda prilagajanja frekvenčno-časovne ločljivosti spremembam jakosti zvoka. Prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed je dalo slabše rezultate. Ker je prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed izmed preizkušenih pristopov tudi računsko najbolj zahtevno, ga z večjo govorno zbirko NUMBERS nismo preizkusili.

Prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed je odvisno od natančnosti razmejitve fonemov, ki se jo dobi v prvem prehodu s pomočjo klasičnega SRG s fiksno dolgim okvirjem. Predvsem v prisotnosti motenj je ta razmejitev premalo natančna. To je najverjetnejši razlog, da se je v drugem prehodu razpoznavanja robustnost v primerjavi z rezultati iz prvega prehoda celo nekoliko poslabšala. Težnja po čim boljši razmejitvi v prvem prehodu tudi v

prisotnosti motenj je razlog, da je bil v prvem prehodu za določanje razmejitve izbran SRG s fiksno dolžino okvirja 32 ms - ob uporabi okvirja take dolžine je bila robustnost najboljša (slika 31). Pri razpoznavanju testne množice brez dodanih motenj v prvem prehodu dobimo boljšo razmejitev, kar se v nekaterih primerih odrazi tudi v izboljšanju uspešnosti glede na rezultate iz prvega prehoda.

Prilagajanje frekvenčno-časovne ločljivosti spremembam jakosti zvoka se je s stališča robustnosti izkazalo za najuspešnejši pristop. To je še posebej razveseljivo, saj je ta pristop izmed vseh preizkušenih tudi računsko najmanj zahteven. Robustnost se je izboljšala pri vseh poskusih, neodvisno od uporabljene govorne zbirke. Pri govorni zbirki NUMBERS je opazno tudi razmeroma veliko (število napak se je zmanjšalo za približno 16% - primerjava *sj_16_32* in *32*) izboljšanje uspešnosti razpoznavanja testne množice brez dodanih motenj.

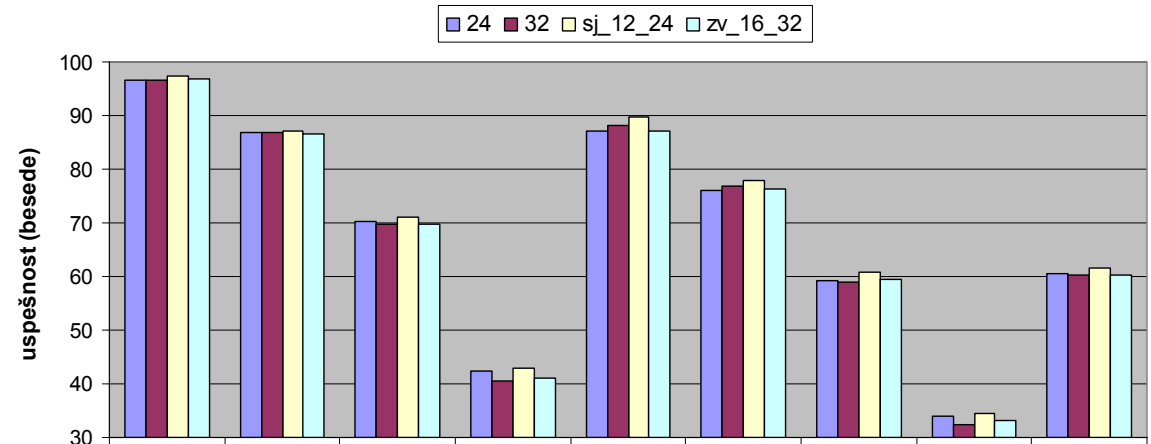
Tudi s prilagajanjem dolžine okvirja zvenečim in nezvenečim odsekom govora z uporabo ustrezne akustične značilke je robustnost mogoče nekoliko izboljšati. Izboljšanje je opazno pri obeh govornih zbirkah, vendar je manjše kot pri prilagajanju frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	97,3	92,0	80,4	57,7	91,6	82,7	69,8	48,2	70,8
32	96,8	92,1	81,4	59,7	91,3	83,5	71,6	51,1	72,3
fon-trif_12_24	97,1	92,0	81,1	57,7	91,3	82,8	70,3	48,8	71,2
fon-zv_12_16	97,2	91,9	79,7	54,8	90,6	80,2	67,0	44,9	68,5
sj_10_24	97,0	93,2	84,7	64,7	91,6	85,3	75,1	55,1	75,4
zv_14_16	96,6	92,8	83,6	61,7	91,5	84,1	73,1	51,4	73,4

Slika 39: Primerjava prilagajanja dolžine okvirja fonemski zgradbi besed, prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka in prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora z zbirko ŠTEVKE.

Primerjava najuspešnejših poskusov prilagajanja dolžine okvirja fonemski zgradbi besed, prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka in prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora je za zbirko ŠTEVKE prikazana na sliki 39. Rezultati podobnega preizkusa z govorno zbirko NUMBERS so prikazani na sliki 40.



	brez	aditivne 12dB	aditivne 6dB	aditivne 0dB	konvolutivne	aditivne 12dB + konvolutivne	aditivne 6dB + konvolutivne	aditivne 0dB + konvolutivne	povprečje
24	96,69	86,88	70,38	42,36	87,20	75,97	59,31	33,91	60,62
32	96,70	86,95	69,85	40,64	88,23	76,83	58,82	32,39	60,17
sj_12_24	97,24	87,22	70,94	42,81	89,70	77,95	60,71	34,53	61,70
zv_16_32	96,88	86,47	69,82	41,18	87,14	76,26	59,43	33,16	60,32

Slika 40: Primerjava prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka in prilagajanja dolžine okvirja zvenečim in nezvenečim odsekom govora z zbirko NUMBERS.

Pri razpoznavanju v prisotnosti motenj je za uspešno razpoznavanje bistveno, da postopek za dinamično prilagajanje frekvenčno-časovne ločljivosti deluje čim bolj podobno kot v učni množici, v kateri teh motenj ni. Gre torej za robustnost samega postopka za prilagajanje frekvenčno-časovne ločljivosti spektra. Če je postopek premalo robusten, se razmejitev, s katero je določena frekvenčno-časovna ločljivost spektra na določenem odseku, v testni množici preveč razlikuje od razmejitve v učni množici. To pomeni, da se pri vseh preizkušenih pristopih prilagajanja dolžine okvirja, pri parametrizaciji določenih odsekov govora v testni množici uporabi okvir napačne dolžine. Ker se dolžini obeh uporabljenih okvirjev lahko precej razlikujeta, ima uporaba napačnega okvirja razmeroma velik vpliv. Pri prilagajanju frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka je ta problem manj izrazit. To je najverjetneje tudi pomemben razlog, da je bila (tudi) s stališča robustnosti ta metoda najboljša. Pri prilagajanju dolžine okvirja na osnovi

fonemske zgradbe besed in prilagajanju dolžine okvirja zvenečim in nezvenečim odsekom govora je razmejitev, s katero je določena dolžina okvirja "trda"; uporabi se ena izmed dveh možnih dolžin okvirja. To pomeni, da se tudi ob časovno nekoliko zgrešeni razmejitvi, do katere zaradi dodanih motenj zelo verjetno pride, uporabi drugačna dolžina okvirja kot bi se uporabila, če teh motenj ne bi bilo.

Mera za spremembe jakosti zvoka je zvezna. To je mogoče izkoristiti tako, da se tudi frekvenčno-časovno ločljivost spektra prilagaja zvezno. Uporabljena je bila obtežena vsota spektrov, ki sta izračunana z različno dolgima okvirjema. Če so spremembe jakosti zaradi dodanih motenj drugačne kot če motenj ne bi bilo, ima to na obteženo vsoto manjši vpliv kot pri prilagajanju dolžine okvirja, kjer napačna razmejitev pomeni uporabo okvirja napačne dolžine. Zato je boljša robustnost prilagajanja frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka najverjetneje tudi posledica same uporabe obtežene vsote dveh spektrov in ne zgolj boljše robustnosti določanja sprememb jakosti zvoka napram določanju zvenečih in nezvenečih odsekov oziroma fonemske zgradbe.

Tudi pri prilagajanju dolžine okvirja je prehode med različnima dolžinama okvirjev mogoče omehčati. To bi lahko storili tako, da bi prehod med dvema različno dolgima okvirjema izvedli postopoma, lahko s postopnim spreminjanjem dolžine okvirja v bližini prehoda ali pa bi na prehodu uporabili obteženo vsoto dveh spektrov, ki bi ju izračunali z vsakim izmed obeh okvirjev na obeh straneh prehoda. Na ta način bi se pri (manjših) napakah pri razmejitvi na zveneče in nezveneče odseke oziroma foneme, uporabil podobno dolg okvir kot v učni množici, v kateri dodanih motenj ni, tako kot to velja za obteženo vsoto dveh spektrov, ki je bila uporabljena v povezavi s spremembami jakosti zvoka. Verjetno bi tako mehčanje prehodov izboljšalo robustnost razpoznavanja pri prilagajanju dolžine okvirja na osnovi fonemske zgradbe oziroma prilagajanju dolžine okvirja zvenečim in nezvenečim odsekom govora.

8. ZAKLJUČEK

V disertaciji so bile predstavljene tri različne metode za dinamično prilagajanje frekvenčno-časovne ločljivosti spektra govornega signala. Pri vseh se na različne načine poskuša oceniti dinamiko spreminjanja spektra govornega signala. Na odsekih, kjer se spekter spreminja hitreje, se poudari časovno ločljivost, na ostalih pa frekvenčno ločljivost spektra. Izvedeno je bilo večje število poskusov, v katerih je bilo dinamično prilagajanje frekvenčno-časovne ločljivosti spektra s pomočjo hibridnega SRG primerjano s klasičnim postopkom parametrizacije, ki temelji na spektru z nespremenljivo frekvenčno-časovno ločljivostjo. Ob tem je bila, razen vplivu na uspešnost razpoznavanja, pozornost namenjena tudi vplivu na robustnost. Ocenjen je bil vpliv na inherentno robustnost, torej na razpoznavanje v prisotnosti motenj, ki jih v učni množici ni bilo.

Rezultati poskusov so bili podani z odstotkom pravilno razpoznanih besed v vzorcih, v katerih so bila naključna zaporedja besed. Izmerjen je bil torej končni rezultat razpoznavanja, čeprav bi bilo vpliv sprememb v postopku računanja značilk mogoče oceniti že na nižjih stopnjah, npr. s številom pravilno klasificiranih okvirjev. Glavna razloga za tako odločitev sta dva. Prvi je dejstvo, da se povečanje odstotka pravilno klasificiranih okvirjev pogosto ne odrazi tudi v večjem odstotku pravilno razpoznanih besed. Drugi razlog je v tem, da se je zelo težko opredeliti, ali je določen okvir klasificiran pravilno ali ne. Če testno množico označimo avtomatsko, se kategorije določijo s SRG, ki je bil uporabljen za označevanje. Na nivoju okvirjev bi zato pravzaprav merili skladnost s tem SRG. Na nivoju pravilno razpoznanih besed teh težav ni.

8.1 Komentar rezultatov

Rezultati poskusov kažejo, da je z dinamičnim prilagajanjem frekvenčno-časovne ločljivosti spektra v postopku parametrizacije mogoče izboljšati uspešnost razpoznavanja. To je še posebej opazno pri rezultatih poskusov z zbirko NUMBERS. S to govorno zbirko sta bili preizkušeni dve metodi: prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka in

prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora. Uporaba obeh je prinesla večjo robustnost kot če se pri parametrizaciji uporabi katerokoli izmed fiksnih dolžin okvirja. Pri zbirki ŠTEVKE po uspešnosti razpoznavanja zelo odstopa SRG, pri katerem se pri računanju spektra uporablja 24 ms dolg okvir. Ta rezultat z dinamičnim prilagajanjem frekvenčno-časovne ločljivosti spektra ni bi presežen. Ker je govorna zbirka NUMBERS za nekaj razredov večja od zbirke ŠTEVKE, lahko sklepamo, da je do velikega odstopanja pri 24 ms dolgem okvirju najverjetneje prišlo zaradi premajhne testne množice, v vsakem primeru pa lahko, zaradi nekajkrat večje obsežnosti zbirke, večji pomen pripišemo rezultatom, ki so bili doseženi z zbirko NUMBERS.

Za izboljšanje robustnosti razpoznavanja je bistvenega pomena že robustnost samega postopka za določanje frekvenčno-časovne ločljivosti spektra. Če imajo motnje na postopek prevelik vpliv, se določitev v testni množici od učne preveč razlikuje. Problem je bil najbolj izrazit pri prilagajanju dolžine okvirja na osnovi fonemske zgradbe besed, kjer je dinamično določanje dolžine okvirja robustnost celo poslabšalo. Prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka in prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora sta robustnost pri večini poskusov izboljšala. Tudi pri preverjanju robustnosti so bili rezultati, ki so bili doseženi z uporabo zbirke NUMBERS, bolj konsistentni od rezultatov doseženih z uporabo govorne zbirke ŠTEVKE.

Sklepamo torej lahko, da je z dinamičnim prilagajanjem frekvenčno-časovne ločljivosti spektra govornega signala mogoče izboljšati tako uspešnost kot robustnost (razen s prilagajanjem dolžine okvirja fonemski zgradbi besed) razpoznavanja, zato bi bilo to področje v prihodnosti smiselno podrobneje raziskati. Ker je za oceno vpliva sprememb postopka parametrizacije potrebno uporabiti razmeroma veliko govorno zbirko, so poskusi časovno zahtevni. Zato smo se pri svojem delu omejili na majhno število različnih pristopov, ki pa so dali razmeroma dobre rezultate. Seveda je parametrizacija, pri kateri se frekvenčno-časovno ločljivost spektra dinamično prilagaja, počasnejša od klasične. Podaljšanje časa razpoznavanja bi se predvsem pri SRG, ki tečejo na manj zmogljivih vgrajenih sistemih, lahko pokazalo kot problematično. Zato je potrebno poudariti, da se je v naših poskusih za najuspešnejše izkazalo prilagajanje frekvenčno-časovne ločljivosti spremembam jakosti zvoka. Ustrezna predlagana metoda je namreč izmed vseh preizkušenih računsko najmanj zahtevna. Zato nameravamo delo na tem področju nadaljevati predvsem v smereh, ki so nakazane v nadaljevanju.

8.2 Smernice za nadaljnje raziskave

Glede na rezultate poskusov, se nameravamo v prihodnosti omejiti na uporabo tistih postopkov za dinamično prilagajanje frekvenčno-časovne ločljivosti spektra, ki so računsko nezahtevni. Ocenjujemo, da je izboljšanje, ki ga s tem dosežemo, razmeroma majhno in v večini primerov ne odtehta prevelikega povečanja računske zahtevnosti parametrizacije. Izmed preizkušenih postopkov je s tega stališča najprimernejše prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka. Glavna slabost predlagane metode je, da je spekter potrebno izračunati dvakrat, s krajšim in z daljšim okvirjem. Dvakratno računanje spektra je mogoče odpraviti s spreminjanjem dolžine okvirja v majhnih korakih. Dolžino okvirja je mogoče določiti glede na absolutno vrednost relativne spremembe jakosti zvoka. Spreminjanje frekvenčno-časovne ločljivosti spektra bi bilo tudi na ta način, podobno kot pri obteženi vsoti dveh spektrov, skoraj zvezno. Prav tako bi bilo za določanje sprememb jakosti mogoče uporabiti parametre, ki so potrebni že za računanje MFCC koeficientov. MFCC koeficient z indeksom 0 namreč predstavlja energijo signala v okvirju, na osnovi katere bi bilo mogoče določiti spremembe jakosti.

A. PARAMETRIZACIJA GOVORNEGA SIGNALA Z UPORABO VALČNE TRANSFORMACIJE

V tem dodatku sta podani funkciji za računanje močnostnega spektra signala s pomočjo diskretne in zvezne valčne transformacije, ki sta bili uporabljeni v poglavju 5.3.

Ostale funkcije za računanje MFCC koeficientov z orodjem Matlab so bile napisane po zgledu parametrizacije v orodju CSLU Toolkit. Pri tem nam je bila v veliko pomoč knjižnica *Rastamat* za računanje MFCC koeficientov. Dosegljiva je na naslovu: *http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/*.

A.1 Parametrizacija z uporabo diskretne valčne transformacije

V nadaljevanju je podana funkcija za računanje spektra signala z uporabo paketne valčne transformacije. Vhodna parametra sta okvir signala *s*, ki je predhodno pomnožen s Hammingovim oknom ustrezne dolžine in vrsta valčka, določena s parametrom *wavelet*. V poglavju 5 sta bila uporabljena dva različna valčka: Haarov valček in valček iz družine Daubechies D^4. Parameter *s* je torej vektor, v katerem je 128 vzorcev (16 ms) signala, ki so predhodno pomnoženi s Hammingovim oknom enake dolžine. Parameter *wavelet* (v njem je oznaka želenega valčka) je imel vrednost *'db1'* oziroma *'db4'*.

V programu je implementirano iterativno filtriranje signala v frekvenčne pasove, ki se razmeroma dobro ujemajo z Mel-lestvico (glej sliko 20 na strani 54):

```
function [pspectrum,frqs] = sgram(s, wavelet)
% wavelet='db4' / wavelet='db1';

[sl,sh]=dwt(s,wavelet,'mode','zpd');
```

```
[shl,shh]=dwt(sh,wavelet,'mode','zpd');

[shhl,shhh]=dwt(shh,wavelet,'mode','zpd');
[shll,shlh]=dwt(shl,wavelet,'mode','zpd');
[shlll,shllh]=dwt(shll,wavelet,'mode','zpd');


[sll,slh]=dwt(sl,wavelet,'mode','zpd');

[slhl,slhh]=dwt(slh,wavelet,'mode','zpd');
[slll,sllh]=dwt(sll,wavelet,'mode','zpd');

[slhhl,slhhh]=dwt(slhh,wavelet,'mode','zpd');
[slhll,slhlh]=dwt(slhl,wavelet,'mode','zpd');
[sllhl,sllhh]=dwt(sllh,wavelet,'mode','zpd');
[sllll,sllll]=dwt(slll,wavelet,'mode','zpd');

[slhhll,slhhlh]=dwt(slhhl,wavelet,'mode','zpd');
[slhlhl,slhlhh]=dwt(slhlh,wavelet,'mode','zpd');
[slhlll,slhllh]=dwt(slhll,wavelet,'mode','zpd');
[sllhhl,sllhhh]=dwt(sllhh,wavelet,'mode','zpd');
[sllhll,sllhlh]=dwt(sllhl,wavelet,'mode','zpd');
[slllhl,slllhh]=dwt(sllll,wavelet,'mode','zpd');
[sllllll,slllllh]=dwt(slllll,wavelet,'mode','zpd');


[slllhll,slllhlh]=dwt(slllhl,wavelet,'mode','zpd');
[slllhhl,slllhhh]=dwt(slllhh,wavelet,'mode','zpd');
[slllllhl,slllllhh]=dwt(slllllh,wavelet,'mode','zpd');
[sllllllh,slllllhh]=dwt(slllllh,wavelet,'mode','zpd');

pspectrum=[S(sllllll), S(slllllh), S(slllllhl), S(slllllhh), S(slllhll),
S(slllhlh), S(slllhhl), S(slllhhh), S(sllhll), S(sllhlh), S(sllhhl),
S(sllhhh), S(slhlll), S(slhllh), S(slhlhl), S(slhlhh), S(slhhll),
S(slhhlh), S(slhhh), S(shlll), S(shllh), S(shlh), S(shhl), S(shhh)];

frqs=[31,94,156,219,281,344,406,469,563,688,813,938,1063,1188,1313,1438,1
563,1688,1875,2125,2375,2750,3250,3750];

%%%%%%%%%%%%%%%%%%%%%%%%%%
function r=S(v)
[n,m]=size(v);
r=dot(v,v)/n;
```

Izpis 1: Izpis programa za parametrizacijo s pomočjo paketne valčne transformacije. Program je napisan za orodje Matlab.

Rezultat je vektor *pspectrum* s 24 elementi, v katerem je močnostni spekter signala, podan za frekvenčne pasove različnih širin. Centralne frekvence frekvenčnih pasov funkcija vrne v *frqs*. Močnostni spekter je predstavljen kot povprečna vrednost vsote kvadratov koeficientov diskretne

valčne transformacije za posamezni frekvenčni pas.

A.2 Parametrizacija z uporabo zvezne valčne transformacije

V nadaljevanju je izpis funkcije za orodje Matlab, ki je bila v poglavju 5.3.1 uporabljena za računanje zvezne valčne transformacije govornega signala. Parametra sta vektor vzorcev celotnega signala *s* in želeni tip zvezne valčne transformacije *wavelet*.

```
function [pspectrum,frqs] = sgram(s, wavelet)

% Compute CWT power spectrum
% like in 'THE USE OF WAVELET TRANSFORMS IN PHONEME RECOGNITION
% Beng t. TAN, Minyue Fu, Andrew Spray, Philip Dermody

% wavelet = 'cmor0.46-16';
% wavelet = 'db1';
% wavelet = 'db4';


a=[];
for m=0:51
    a=[a,2^(m/8)];
end

%16 ms window
wlen=128;

%100 Hz frame rate
step=80;

c=abs(cwt(s,a,wavelet));
r=c;

%filtering
fc=filter(hamming(wlen),[1],c');
c=fc';

[cy,cx]=size(c);
pspectrum=c(:,[wlen/2:step:cx]);

frqs=scal2frq(a,wavelet,1/8000);
```

Izpis 2: Izpis programa za parametrizacijo s pomočjo zvezne valčne transformacije. Program je napisan za orodje Matlab.

Funkcija je pisana po zgledu [74]. Uporabljeni so bili naslednji valčki: kompleksni Morletov valček označen s *"cmor0.46-16"*, Haarov valček *"db1"* in funkcija D^4 iz skupine Daubechies *"db4"*. Najprej se izračunajo koeficienti zvezne valčne transformacije pri različnih vrednostih *a*. Koeficiente se nato filtrira z nizkoprepustnim filtrom, za katerega je bilo uporabljeno Hammingovo okno dolžine 16 ms. Rezultat se vzorči s frekvenco vzorčenja 100 Hz, ki ustreza frekvenci okvirjev. V *frqs* funkcija vrne frekvence glede na vrednosti *a*, pri katerih je podan močnostni spekter *pspectrum*.

B. SPECIFIKACIJA POSKUSOV ZA IZVEDBO Z ORODJEM CSLU TOOLKIT

V tem dodatku so navedene podrobnosti poskusov, ki so opisani v disertaciji. Vsi poskusi so bili opravljeni z orodjem CSLU Toolkit, ki je bilo razširjeno s paketom *Mfeture* tako, da je mogoče določati dolžino okvirja, ki se uporabi pri parametrizaciji posameznih odsekov govora. Mogoče je uporabljati tudi obtežene vsote dveh spektrov, ki sta izračunana z različno dolgima okvirjema (tak pristop je bil uporabljen v povezavi s spremembami jakosti zvoka). Parametrizacijo je mogoče izvesti izven orodja in namesto govornih vzorcev uporabiti datoteke z že izračunanimi značilkami. Tak pristop je bil uporabljen pri poskusih, pri katerih je bila parametrizacija na osnovi valčne transformacije izračunana z orodjem Matlab.

V obeh govornih zbirkah je na voljo manjši del učne množice, ki je bil ročno fonemsko označen. V prvem koraku je potrebno izdelati SRG, s katerim se nato avtomatsko fonemsko označi celotni učni množici, ki se v drugem koraku uporabita za učenje končnega SRG. V vektorju značilk je bilo vedno prvih 13 MFCC koeficientov in njihovi delta koeficienti. Podatki o velikosti učnih, razvojnih in testnih množic so podani v tabeli:

Govorna zbirka	Učna množica 1. korak	Učna množica 2. korak	Razvojna množica	Testna množica
ŠTEVKE	180, 2340	468, 6048	156, 2028	156, 2028
NUMBERS manjši	2705, 11420	6428, 35635	600, 3568	600, 2629
NUMBERS večji	2705, 11420	6428, 35635	1111, 6417	2287, 13130

Tabela 19: Podatki o velikosti množic za učenje in testiranje SRG. Vedno je najprej navedeno število stavkov, nato pa število besed v posamezni množici. Zaradi velikosti govorne zbirke NUMBERS, je bil v poglavju 6 uporabljen samo del zbirke, celotna zbirka pa je bila uporabljena samo za dokončno primerjavo in testiranje robustnosti v poglavju 7.

V nadaljevanju je navedena vsebina najpomembnejših konfiguracijskih datotek, ki določajo oba SRG, ki sta bila uporabljena za razpoznavanje govornih zbirk ŠTEVKE in NUMBERS.

B.1 Konfiguracija SRG za orodje CSLU Toolkit

Navedeni sta vsebini datotek *stevke.vocab* in *numbers.vocab*, s katerima je določen slovar besed, njihova fonemska zgradba ter gramatika oziroma veljavna zaporedja besed. V orodju CSLU ti dve datoteki, skupaj s časovnimi modeli in ustrezno nevronsko mrežo za klasifikacijo v kontekstno odvisne kategorije, določata SRG.

```
nic          {n i _tS tS}     ;
ena          {E n a}          ;
dve          {_d d v e}       ;
tri          {_t t r i}       ;
stiri        {S _t t i r i}   ;
pet          {_p p e _t t}    ;
sest         {S e s _t t}     ;
sedem        {s e _d d E m}   ;
sedem        {s e _d d @ m}   ;
osem         {o s E m}        ;
osem         {o s @ m}        ;
devet        {_d d E v e _t t};
devet        {_d d @ v e _t t};
ja           {j a}            ;
ne           {n E}            ;
stop         {s _t t O _p p}  ;
premor      { .pau [ .garbage .pau ] };

$digit   = nic | ena | dve | tri | stiri | pet | sest | sedem |
           osem | devet | ja | ne | stop;

$grammar = (premor%%  < $digit  premor%% > [premor%%]) ;
```

Izpis 3: Datoteka stevke .vocab za določitev slovarja besed, njihove fonemske zgradbe ter gramatike za razpoznavanje govorne zbirke STEVKE.

```
zero        {z I 9r oU}             ;
oh          {oU}                    ;
one         {w ^ n [^3]}            ;
two         {uc th u}               ;
three       {T 9r i:}               ;
four        {f >r}                  ;
five        {f aI v}                ;
six         {s I uc ks}             ;
seven       {s E v ^2 n [^3]}       ;
eight       {ei uc [th]}            ;
nine        {n aI n [^3]}           ;

separator   {.pau [.garbage] .pau}  ;

$digit   = zero | oh | one | two | three | four | five | six |
           seven | eight | nine;

$grammar = ([separator%%] < $digit [separator%%] > [separator%%]);
```

Izpis 4: Datoteka numbers .vocab za določitev slovarja besed, njihove fonemske zgradbe ter gramatike za razpoznavanje govorne zbirke NUMBERS.

B.2 Specifikacija poskusov z valčno transformacijo

Pri poskusih z valčno transformacijo se parametrizacija izvede s pomočjo funkcij, ki so opisane v dodatku A. Oblika konfiguracijske datoteke je prikazana v nadaljevanju:

```
test_name:        test_db4

vocab:            stevke.vocab
train_2p_info:    stevke.train_2p.info
dev_2p_info:      numbers.dev_2p.info
test_2p_info:     numbers.test_2p.info

train_2p_olddesc: numbers.train_2p_.olddesc
train_2p_pick:    stevke.train_2p_.pick

train_files:      big.train.numbers.files
dev_files:        big.dev.numbers.files
test_files:       big.test.numbers.files

feature_code:     ufeat_matlab.tcl
feature_files:    {pkgIndex.tcl}
```

```
vec_sample_size:  130
categories:       86
iter:             40

from:             20
garbage:          7
```

Izpis 5: Konfiguracijska datoteka s končnico .conf za poskuse z valčno transformacijo.

V konfiguracijski datoteki je podano ime poskusa, sledi razdelek z navedbami konfiguracijskih datotek orodja CSLU Toolkit. Sledita imeni datotek s časovnimi modeli in izborom okvirjev, katerih vektorji značilk se uporabijo za učenje nevronske mreže. Časovni modeli in izbor okvirjev sta neodvisna od načina parametrizacije, zato sta bili v vseh poskusih uporabljeni isti datoteki. Parameter *feature_code* mora vsebovati ime datoteke za parametrizacijo, *feature_files* pa seznam morebitnih dodatnih datotek, ki so potrebne pri parametrizaciji. Naslednja skupina parametrov podaja dolžino vektorja značilk, število kontekstno odvisnih kategorij in število iteracij učenja nevronske mreže. Parameter *from* določa, katere nevronske mreže bodo preizkušene z razvojno množico (od dvajsete dalje). Parameter orodja CSLU Toolkit *garbage* se uporablja za modeliranje motenj in besed, ki niso v slovarju.

V nadaljevanju je predstavljen skript za parametrizacijo vzorcev (njegovo ime podamo s parametrom *feature_code*):

```
proc UserComputeVector {wave up_feat sampling_rate wavefile}  {
      upvar $up_feat rfeat
      global info_file

      set info(cat_suffix) feat
      readInfoFile $info_file {cat_path cat_suffix} info
      set a [file split $wavefile]
      set f [lindex $a [expr [llength $a]-1]]
      set featfile [format "./feat/%s%s" [string trimright $f "wav"] \
          $info(cat_suffix)]

      mx fread -c 13 $featfile wmel

      set wcms  [mx zeromean $wmel]

      #--- add delta coefficients ---
      set fdel  [feature delta initialize 13 -order 2 -sigmaT2]
      set wdel  [feature delta $fdel $wcms -flush]

      set rfeat  [mx join col [list $wcms $wdel]]
```

```
      #--- nuke stuff ---
      nuke $fdel $wdel
      nuke $wmel $wcms
}
```

Izpis 6: Skript za računanje MFCC koeficientov za poskuse z valčno transformacijo.

Skript v imenu datoteke vzorca končnico *.wav* zamenja s končnico *.feat* in prebere ustrezno datoteko, v kateri morajo biti zaporedja vektorjev značilk s 13 MFCC koeficienti. To datoteko smo pri poskusih z valčno transformacijo tvorili z orodjem Matlab. MFCC koeficientom se dodajo delta koeficienti.

B.3 Specifikacija poskusov s prilagajanjem dolžine okvirja na osnovi fonemske zgradbe besed in zvenečim oziroma nezvenečim odsekom govora

Za izvedbo poskusov sta potrebni dve konfiguracijski datoteki. Prva je skoraj enaka kot konfiguracijska datoteka v poglavju B.2. Razširjena je s specifikacijami vrste šumov, s katerimi se preveri robustnost. Druga datoteka mora imeti enako ime kot ga je podano s parametrom *test_name*, njena končnica pa mora biti *.pht* (npr. *test10_32_32t.pht)*. V njej so določene dolžine okvirja za parametrizacijo posameznih kontekstno odvisnih kategorij. Oblika datoteke je podana na izpisu 12.

```
test_name:        test10_32_32t
vocab:            stevke.vocab
nnet_1p:          nnet1p32
train_2p_info:    stevke.train_2p.info
train_2p_olddesc: numbers.train_2p_.olddesc
train_2p_pick:    stevke.train_2p32.pick
dev_2p_info:      numbers.dev_2p.info
test_2p_info:     numbers.test_2p.info
train_files:      big.train.numbers.files
dev_files:        big.dev.numbers.files
test_files:       big.test.numbers.files

feature_files:    {pkgIndex.tcl feature.dat default.phl}
feature_code:     ufeat_wls32.tcl

vec_sample_size:  130
```

```
categories:       85
iter:             40

from:             20
garbage:          7

#noise tests configuration
name:             big
corpus:           numbers

noisedir:         /baza/telephone_noises_stevke
#noises:          muffle_lp
#noises:          muffle_reverb

noises:           white
noises:           pink
noises:           babble
noises:           volvo
#noises:          factory1
#noises:          f16
#noises:          pass900hz

snr:              12.0
snr:              6.0
snr:              0.0
```

Izpis 7: Konfiguracijska datoteka s končnico .conf za poskuse s prilagajanjem dolžine okvirja na osnovi fonemske zgradbe besed in zvenečim oziroma nezvenečim odsekom govora.

MFCC koeficienti so bili pri poskusih, v katerih je bila pri parametrizaciji različnih odsekov govora uporabljena različna dolžina okvirja, izračunani s pomočjo skripta:

```
proc UserComputeVector {wave up_feat sampling_rate wavefile} {
      upvar $up_feat rfeat
      global info_file

      # read additional feature parameters
      readInfoFile $info_file {cat_path cat_suffix} info
      set a [file split $wavefile]

      set f $info(cat_path)/[lindex $a [expr [llength $a]-1]]

      set phnfile [format "%s%s" [string trimright $f "wav"] \
            $info(cat_suffix)]

      #voicing $wave $phnfile 1 $sampling_rate

      #--- remove dc component from wave ---
      set nodc      [prep !dc initialize -rate $sampling_rate]
      set wave_nodc [prep !dc $nodc $wave]
```

```
      set fmel  [feature initialize mel \
            -samplerate 8000 -framesize 10.0 -windowsize 36.0 \
            -output 13 -filters 28 -ufreq 4000]

      set wmel  [feature $fmel $wave_nodc -mix $mix \
            -param_type wlswitch -phnfile $phnfile -phtfile ptimes.pht]

      set wcms  [mx zeromean $wmel]

      #--- add delta coefficients ---
      set fdel  [feature delta initialize 13 -order 2 -sigmaT2]
      set wdel  [feature delta $fdel $wcms -flush]
      set rfeat  [mx join col [list $wcms $wdel]]

      #--- nuke stuff ---
      nuke $wave_nodc
      nuke $fdel $wdel
      nuke $wmel $wcms
      nuke $fmel $nodc $line
}
```

Izpis 8: Skript za računanje MFCC koeficientov za poskuse s prilagajanjem dolžine okvirja na osnovi fonemske zgradbe besed in zvenečim oziroma nezvenečim odsekom govora.

Skript potrebuje datoteki z razdelitvijo vzorca v foneme *phnfile* in podatki o dolžini okvirja za parametrizacijo posameznega fonema *ptimes.pht*. Prvo datoteko dobimo ali v prvem prehodu razpoznavanja s SRG, pri katerem uporabimo okvir fiksne dolžine ali pa z razdelitvijo v zveneče in nezveneče foneme, če odkomentiramo vrstico *#voicing $wave $phnfile 1 8000*. Parametrizacija s fiksno dolžino okvirja se izvede s skriptom:

```
proc DefaultComputeVector {in_wave up_feat samplerate} {
    upvar $up_feat feat
    global PLPIIR PLPFIR UserComputeFeatures

     set nodc      [prep !dc initialize -rate $samplerate]
     set wave_nodc [prep !dc $nodc $in_wave]

     set fmel  [feature initialize mel \
               -samplerate $samplerate -framesize 10.0 -windowsize 32.0 \
               -output 13 -filters 28 -ufreq [expr $samplerate/2]]
     set fdel  [feature delta initialize 13 -order 2 -sigmaT2]

     set wmel  [feature $fmel $wave_nodc]
     set wcms  [mx zeromean $wmel]
     set wdel  [feature delta $fdel $wcms -flush]
     set feat  [mx join col [list $wcms $wdel]]
```

```
     nuke $nodc $wave_nodc $fmel $fdel
     nuke $wmel $wcms $wdel
     return 0
}
```

Izpis 9 : Skript za računanje MFCC koeficientov s fiksno dolžino okvirja.

Za razdelitev v zveneče in nezveneče odseke je bila uporabljena funkcija *voicing*:

```
proc voicing {wave voifile fsize samp_freq} {
      set pitch_obj  [pitch init -frame_size $fsize \
                              -reduce_method 2]
      set pitch      [pitch compute $pitch_obj $wave]
      set samp_per_frame [expr $samp_freq * $fsize / 1000.0]
      set rate           [expr $samp_freq/$samp_per_frame]

      set val      [mx cut $pitch :,6:6]
      set val_list [lindex [mx puts $val] 0]
      set label_fid [open $voifile "w"]
      puts $label_fid "MillisecondsPerFrame: 1.0000"
      puts $label_fid "END OF HEADER"
      for {set idx 0} {$idx < [llength $val_list]} {incr idx 3} {
            set time1 [expr round([lindex $val_list $idx])]
            set time2 [expr round([lindex $val_list [expr $idx+1]])]
            if {$time1 == 0.0 && $time2 == 0.0} {
                  break
                  }
            set value [lindex $val_list [expr $idx+2]]
            if {$value == 1.0} {
                  set vuv "<vo>"
                  } else {
                  set vuv "<uv>"
                  }
            puts $label_fid "$time1 $time2 $vuv"
            #puts "$time1 $time2 $vuv"
            }
      close $label_fid

      nuke $pitch_obj $pitch
      return 0
}
```

Izpis 10 : Skript za določanje zvenečih in nezvenečih odsekov govora.

Datoteka za razdelitev vzorca na foneme ima pri poskusih prilagajanja dolžine okvirja na osnovi fonemske zgradbe besed naslednjo obliko (prikazan je začetek datoteke):

```
MillisecondsPerFrame: 1.0
```

```
END OF HEADER
0 170 <.pau>
170 350 .pau<s
350 410 s>e
410 460 s<e
460 550 <e>
550 620 e>_d
620 640 <_d>
640 650 _d<d
650 660 d>E
660 680 d<E
680 710 <E>
...
```

Izpis 11 : Datoteka za razdelitvijo govora v kontekstno odvisne kategorij. Prikazan je začetek datoteke.

Dolžina okvirja, ki se uporabi pri parametrizaciji posameznih kontekstno odvisnih kategorij, je določena z datoteko oblike:

```
Phoneme wlengths:
END OF HEADER
<E>      24
<O>      24
<a>      24
<e>      24
<i>      24
<o>      24
default  10
```

Izpis 12 : Datoteka za določitev dolžine okvirja za parametrizacijo posameznik kontekstno odvisnih kategorij.

Prikazana je datoteka za poskuse z zbirko ŠTEVKE, pri katerih je bil pri parametrizaciji srednjih delov trifonov uporabljen daljši okvir kot pri parametrizaciji ostalih kategorij. Podobno je pri prilagajanju dolžine okvirja zvenečim in nezvenečim odsekom: zveneči odseki so predstavljeni z monofonom <*vo*>, nezveneči pa z monofonom <*uv*>.

B.4 Specifikacija poskusov s prilagajanjem frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka

Tudi parametrizacija tega tipa za delovanje potrebuje paket *Mfeature*. Konfiguracijski datoteki so dodani potrebni parametri, s katerimi sta določeni dolžini krajšega in daljšega okvirja ter faktor za računanje obtežene vsote obeh spektrov. Prikazan je primer konfiguracijske datoteke:

```
test_name:        test10_32
vocab:            stevke.vocab
train_2p_info:    stevke.train_2p.info
train_2p_olddesc: numbers.train_2p_.olddesc
train_2p_pick:    stevke.train_2p32.pick
dev_2p_info:      numbers.dev_2p.info
test_2p_info:     numbers.test_2p.info
train_files:      big.train.numbers.files
dev_files:        big.dev.numbers.files
test_files:       big.test.numbers.files

feature_files:    {pkgIndex.tcl}
feature_code:     ufeat_int.tcl
mix:              0.8
window_long:      32.0
window_short:     10.0

vec_sample_size:  130
categories:       217
iter:             40

from:             20
garbage:          7

name:             mfcc_delta_32
corpus:           stevke

noisedir:         /baza/telephone_noises_stevke
noises:           muffle_lp
noises:           muffle_reverb

noises:           white
noises:           pink
noises:           babble
noises:           volvo
noises:           factory1
noises:           f16
noises:           pass900hz

snr:              12.0
snr:              6.0
snr:              0.0
```

Izpis 13: Konfiguracijska datoteka s končnico .conf za poskuse s prilagajanjem frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka.

MFCC koeficienti so bili pri poskusih te vrste izračunani s skriptom:

```
proc UserComputeVector {wave up_feat sampling_rate wavefile}  {
      upvar $up_feat rfeat
      global info_file

      readInfoFile $info_file {cat_path cat_suffix} info
      set a [file split $wavefile]
      set f $info(cat_path)/[lindex $a [expr [llength $a]-1]]

      set intfile [format "%s%s" [string trimright $wavefile "wav"] int]

      readInfoFile feature.dat {mix window_long window_short} feat_info

      #--- remove dc component from wave ---
      set nodc      [prep !dc initialize -rate 8000]
      set wave_nodc [prep !dc $nodc $wave]

      set fmel  [feature initialize mel \
              -samplerate 8000 -framesize 10.0 -windowsize \
              $feat_info(window_long) \
              -output 13 -filters 28 -ufreq 4000]

      #set wmel  [feature $fmel $wave_nodc]

      set wmel  [feature $fmel $wave_nodc -mix $feat_info(mix) \
            -swsz  $feat_info(window_short) \
            -param_type intensity -intfile $intfile]

      #--- add delta coefficients ---
      set fdel  [feature delta initialize 13 -order 2 -sigmaT2]
      set wdel  [feature delta $fdel $wcms -flush]

      set rfeat  [mx join col [list $wcms $wdel]]

      #--- nuke stuff ---
      nuke $wave_nodc
      nuke $fdel $wdel
      nuke $wmel $wcms
      nuke $fmel $nodc
}
```

Izpis 14: Skript za računanje MFCC koeficientov za poskuse s prilagajanjem frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka.

C. SPECIFIKACIJE OZNAK REZULTATOV TESTIRANJA ROBUSTNOSTI

V tem dodatku je podan način podajanja rezultatov testiranja vpliva dinamičnega prilagajanja frekvenčno-časovne ločljivosti spektra govornega signala v postopku parametrizacije. Navedeno je, katere motnje so vključeno v skupine motenj, za katere je v rezultatih podana povprečna uspešnost razpoznavanja.

C.1 Motnje za testiranje robustnosti v poglavju 6

V poglavju 6 so v rezultatih podane povprečne vrednosti uspešnosti razpoznavanja v prisotnosti različnih oblik aditivnih motenj:

Oznaka v tabelah	V povprečju zajete množice
brez	testna množica
12dB	govor 12.0dB
	volvo 12.0dB
	roza šum 12.0dB
	beli šum 12.0dB
6dB	govor 6.0dB
	volvo 6.0dB
	roza šum 6.0dB
	beli šum 6.0dB
0dB	govor 0.0dB
	volvo 0.0dB
	roza šum 0.0dB
	beli šum 0.0dB
povprečje	vse navedene testne množice z dodanimi aditivnimi motnjami

Tabela 20 : Skupine testnih množic za podajanje rezultatov v poglavju 5.

C.2 Motnje za testiranje robustnosti v poglavju 7

Pri podrobnejšem testiranju robustnosti v poglavju 7 so bile uporabljene dodatne aditivne motnje in dve konvolutivni motnji. Katere testne množice so bile zajete v rezultatih, je podano v nadaljevanju.

V tabeli 21 so navedene zajete aditivne motnje:

Oznaka v tabelah	V povprečju zajete množice
brez	testna množica
12dB	govor 12.0dB
	tovarna 12.0dB
	f-16 12.0dB
	pas 900Hz 12.0dB
	volvo 12.0dB
	roza šum 12.0dB
	beli šum 12.0dB
6dB	govor 6.0dB
	tovarna 6.0dB
	f-16 6.0dB
	pas 900Hz 6.0dB
	volvo 6.0dB
	roza šum 6.0dB
	beli šum 6.0dB
0dB	govor 0.0dB
	tovarna 0.0dB
	f-16 0.0dB
	pas 900Hz 0.0dB
	volvo 0.0dB
	roza šum 0.0dB
	beli šum 0.0dB

Tabela 21: Skupine testnih množic za podajanje rezultatov testiranja robustnosti z aditivnimi motnjami v poglavju 7.

Zajete konvolutivne motnje v kombinaciji z aditivnimi so prikazane v tabeli:

Oznaka v tabelah	V povprečju zajete množice	
konvolutivne	testna množica + hp	testna množica + odjek
aditivne 12dB + konvolutivne	govor 12.0dB + hp	govor 12.0dB + odjek
	tovarna 12.0dB + hp	tovarna 12.0dB + odjek
	f-16 12.0dB + hp	f-16 12.0dB + odjek
	pas 900Hz 12.0dB + hp	pas 900Hz 12.0dB +
	volvo 12.0dB + hp	volvo 12.0dB + odjek
	roza šum 12.0dB + hp	roza šum 12.0dB + odjek
	beli šum 12.0dB + hp	beli šum 12.0dB + odjek
aditivne 6dB + konvolutivne	govor 6.0dB + hp	govor 6.0dB + odjek
	tovarna 6.0dB + hp	tovarna 6.0dB + odjek
	f-16 6.0dB + hp	f-16 6.0dB + odjek
	pas 900Hz 6.0dB + hp	pas 900Hz 6.0dB + odjek
	volvo 6.0dB + hp	volvo 6.0dB + odjek
	roza šum 6.0dB + hp	roza šum 6.0dB + odjek
	beli šum 6.0dB + hp	beli šum 6.0dB + odjek
aditivne 0dB + konvolutivne	govor 0.0dB + hp	govor 0.0dB + odjek
	tovarna 0.0dB + hp	tovarna 0.0dB + odjek
	f-16 0.0dB + hp	f-16 0.0dB + odjek
	pas 900Hz 0.0dB + hp	pas 900Hz 0.0dB + odjek
	volvo 0.0dB + hp	volvo 0.0dB + odjek
	roza šum 0.0dB + hp	roza šum 0.0dB + odjek
	beli šum 0.0dB + hp	beli šum 0.0dB + odjek
povprečje	vse navedene testne množice z dodanimi aditivnimi motnjami (tudi iz tabele 21)	

Tabela 22: Skupine testnih množic za podajanje rezultatov testiranja robustnosti s konvolutivnimi motnjami v povezavi z aditivnimi.

D. VITERBIJEV ALGORITEM

V tem dodatku je podan psevdo program za Viterbijevo iskanje, ki je uporabljen pri referenčnem SRG. Povzet je po [17]. Pri referenčnem sistemu se računa z logaritmom verjetnosti, s čimer se odpravi težave z zelo majhnimi števili, do katerih pride pri zaporednem množenju z majhnimi verjetnostmi kategorij.

```
/* initialization */
given N is the number of categories,
given T is the number of frames,
given matrix B[j][t]: category probabilities (neural network outputs),
    with 1 <= j <= N, 1 <= t <= T
given matrix A[i][j]: probability of transitioning from category i to
category j,
    with 1 <= i <= N, 1 <= j <= N

initialize delta[i][1] to B[i][1], with 1 <= i <= N
initialize psi[i][1] to 0, with 1 <= i <= N

/* main loop: compute delta (scores) and psi (category values
corresponding to best scores).  max_score is probability of being in
state i and transitioning from state i to state j */

for each frame t from 2 to T {
    for each category j from 1 to N {
        max_score = LOWEST_POSSIBLE_VALUE
        max_index = 0
        for each category i from 1 to N {
            if (delta[i][t-1] * A[i][j] > max_score) {
                max_score = delta[i][t-1] * A[i][j]
                max_index = i
                }
            }
        delta[j][t] = max_score * B[j][t]
        psi[j][t]   = max_index
        }
    }

/* backtracking to find state (category) sequence */

max_score = LOWEST_POSSIBLE_VALUE
max_index = 0
for each category i from 1 to N {
    if (delta[i][T] > max_score) {
```

```
        max_score = delta[i][T]
        max_index = i
        }
    }
state[T] = max_index

for each frame t from T-1 to 1 {
    state[t] = psi[state[t+1]][t+1]
    }
```

Izpis 15: Izpis psevdo programa za Viterbijevo iskanje

E. LITERATURA

D.1 Splošno področje razpoznavanja govora

[1] Bell, A. G., "The Mechanism of Speech," reprinted from the proceedings of the first summer meeting of the American association to promote the teaching of speech to the deaf, Funk & Wagnalls, New York, 1908.

[2] Boite, J. M., Bourlard, H., D'hoore, B., and Haesen, M., "A New Approach Towards Keyword Spotting," Proceedings of EUROSPEECH '93, vol. 2, pp. 1273-1276, September 1993.

[3] Cole, R., Noel, M., Lander, T., Durham, T., "New telephone speech corpora at CSLU," Proceedings of European Conference on Speech Communication and Technology, vol. 1, pp. 821-824, 1995.

[4] Cosi, P., Hosom, J.-P., Shalkwyk, J., Suttun, S. and Cole, R. A., "Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizers," Proceedings of 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA-ETWR98), pp. 135-140, Turin, Italy, September 1998.

[5] Darken C. and Moody J., "Note on Learning Rate Schedules for Stochastic Optimization," Proceedings of the Neural Information Processing Systems 3, pp. 832-838, San Mateo, 1991.

[6] Dudley, H., Riesz, R. R., and Watkins, S. A. (1939), "A synthetic speaker", J. Franklin Inst., 227, pp. 739-764.

[7] Ferguson, Ed. J., "Hidden Markov Analysis: An Introduction", Hidden Markov Models for Speech, Institute of Defense Analyses, Princeton, NJ, 1980.

[8] Flanagan, J. L., "Speech Synthesis, Analysis and Perception," 2nd edition, Springer-Verlag, 1972

[9] Greenberg, S., "Understanding Speech Understanding: Towards a Unified Theory of Speech Perception," Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, pp. 1-8, July 1996.

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

[11] Helmholtz, H. L. F., "On the Sensations of Tone as a Physiological Basis for the Theory of Music," 2nd edition, Dover Publications, New York, 1954, translated from the fourth German edition of 1877.

[12] Hosom, J.-P., Cosi, P., and Cole, R. A., "Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition," Proceedings of ICSLP '98, vol. 3, pp. 731-734, December 1998.

[13] Itakura, F., "Minimum Prediction Residual Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-23 (1), pp. 67-72, February 1975.

[14] Lippmann, R. P., "An introduction to Computing with Neural Nets," IEEE ASSP Mag., 4 (2), pp. 4-22, April 1987.

[15] Massaro, D. W., and Friedman, D., "Models of Integration Given Multiple Sources of Information," Psychological Review, 97, 2, pp. 225-252, 1990.

[16] Meyers, C. S. and Rabiner, L. R., "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-29, pp. 284-297, April 1981.

[17] Rabiner, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE 77(2):257–286, February 1989.

[18] Rabiner, L., and Juang, B., "Fundamentals of Speech Recognition," Prentice Hall, Englewood Cliffs, NJ, 1993.

[19] Robinson, D. J. M. and Hawksford, M. J., "Time-Domain Auditory Model for the Assessment of High-Quality coded Audio," Preprint 5017, presented at the 107th convention of Audio Engineering Society, New York, 1999.

[20] Sakoe, H. and Chiba, S., "Dynamic programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26 (1), pp. 43-49, February 1978.

[21] Sakoe, H., "Two Level DP Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27, pp. 588-595, December 1979.

[22] Sweet, H., "Handbook of Phonetics," Frowde, Oxford, 1877, reprinted by McGrath Publishing Co., 1970.

[23] Toporišič, J., "Slovenska slovnica," Založba Obzorja, Maribor, 1984.

[24] Tebelskis, J., "Speech Recognition using Neural Networks," Doctoral dissertation, School of Computer Science, Carnegie Mellon University Pittsburgh, Pennsylvania, 1995.

[25] Varga, A., Steeneken, H.J.M.,. Tomlinson, M.J and Jones, D., "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.

[26] Velichko, V. M., and Zagoruyko, N. G., "Automatic recognition of 200 words", Int. J. Man-Machine Studies 2: 223, June 1970.

[27] Viterbi, Andrew J., "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Transactions on Information Theory 13(2):260–269, April 1967.

[28] Wei, W. and van Vuuren, S., "Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition," ICASSP'98, pp. 497-500, May 1998.

[29] Werbos, P.J., "The roots of backpropagation," Wiley, 1994.

D.2 Parametrizacija govornega signala

[30] Bacchiani, M., Ostendorf, M., Sagisaka, Y., Paliwal, K., "Design of a speech recognition system based on acoustically derived segmental units," Proc. ICASSP'96, pp. 443 - 446, May 1996.

[31] Bou-Ghazale, S. E. and Hansen, J. H. L., "A comparative study of traditional and newly proposed features for recognition of speech under stress," IEEE Trans. Speech Audio Processing, vol. 8, pp. 429 - 442, July 2000.

[32] Hansen, John H. L., Clements, Mark A., "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," IEEE Trans. Speech Audio Processing, vol. 3, pp. 407 - 415, September 1995.

[33] Hariharan, R., Kiss, I., and Viikki, O., "Noise robust speech parameterization using multiresolution feature extraction," IEEE Trans. Speech Audio Processing, vol. 9, pp. 856 - 865, November 2001.

[34] Hernando, Javier, Nadeu, Climent, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," IEEE Trans. Speech Audio Processing, vol. 5, pp. 80 - 84, January 1997.

[35] Hosom, J. P., Cole, R. and Cosi, P., "Improvements in Neural-Network Training and Search Techniques for Continuous Digit Recognition," Australian Journal of Intelligent Information Processing Systems, 1999.

[36] Hunt, M. J. and Lefèbvre, C., "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," Proc. ICASSP'89, pp. 262 - 265, May 1989.

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

[38] Kramer, M. L. and Jones, D. L., "Improved Time-Frequency Filtering Using an STFT Analysis-Modification-Synthesis Method," Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, pp. 264—267, Philadelphia, October 1994.

[39] Rabiner, L. R., Schafer; R. W., "Digital Processing of speech signals," Prentice Hall, Englewood Clifs, NJ, 1978.

[40] Reynolds, D. A., "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Processing, vol. 2, pp. 639 - 643, October 1994.

[41] Rozman, R., Štrancar, A., Kodek, D., "Analiza vpliva oken na robustnost sistemov za razpoznavanje govora," Zbornik devete Elektrotehniške in računalniške konference ERK 2000, 21. - 23. september 2000, Portorož, Slovenija. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, 2000, zv. B, str. 177-180.

[42] Rozman, R., Štrancar, A., Kodek, D., "Povečevanje robustnosti sistemov za razpoznavanje govora in optimizacija procesa parametrizacije," Zbornik desete Elektrotehniške in računalniške konference ERK 2001, 24, str. 257-260, Portorož, september 2001.

[43] Schalkwyk, Johan, Hosom, Paul, Kaiser, Ed, Shobaki, Khaldoun, "CSLU-HMM: The CSLU Hidden Markov Modeling Environment," Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, March 2000.

[44] Smith, F. J., Ming, J., O'Boyle, P., Irvine, A. D., "A hidden Markov model with optimized inter-frame dependence," Proc. ICASSP'95, pp. 209 - 212, May 1995.

[45] Štrancar, A., "Analiza vpliva aktivacijskih funkcij, akustičnih značilk in dolžine oken na uspešnost avtomatskega razpoznavanja govora," magistrsko delo, Fakulteta za računalništvo in informatiko, Ljubljana 2001.

[46] Štrancar, A., "Homomorfna analiza govornega signala s kratkimi okvirji," Zbornik osme Elektrotehniške in računalniške konference ERK '99, 23. - 25. september 1999, Portorož, Slovenija. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, 1999.

[47] Štrancar, A., Rozman, R., Kodek, D., "Analiza vpliva akustičnih značilk na uspešnost razpoznavanja govora," Zbornik desete Elektrotehniške in računalniške konference ERK 2001, 24. - 26. september 2001, Portorož, Slovenija. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, 2001.

[48] Štrancar, A., Rozman, R., Kodek, D., "Analiza vpliva modificiranih postopkov parametrizacije na robustnost sistemov za razpoznavanje govora," Zbornik enajste mednarodne Elektrotehniške in računalniške konference ERK 2002, 23.-25. september 2002, Portorož, Slovenija. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, 2002.

[49] Sutton, S. and Cole, R. and de Villiers, J. and Schalkwyk, J and Vermeulen, P. and Macon, M and Yan, Y and Kaiser, E. and Rundle, B. and Shobaki, K and Hosom, P. and Kain, A. and Wouters, J and Massaro, M and Cohen, M, "Universal Speech Tools: The CSLU Toolkit," Proceedings of the International Conference on Spoken Language Processing (ICSLP), pp. 3221-3224, Sydney, Australia, Nov, 1998.

[50] Tebelskis, J., "Speech Recognition using Neural Networks," Doctoral dissertation, School of Computer Science, Carnegie Mellon University Pittsburgh, Pennsylvania, 1995.

[51] Zhao, Y., Wang, S., and Yen, K., "Recursive estimation of time-varying environments for robust speech recognition," Proceedings ICASSP'01, pp. 225 - 228, May 2001.

D.3 Parametrizacija z uporabo množice filtrov ali valčne transformacije

[52] Alessandro, C., "Auditory-based wavelet representation." Visual representations of speech signals, chapter 8, pp. 131-137. John Wiley & Sons Ltd., 1993.

[53] Allen, Jont B., "How do humans process and recognize speech?," IEEE Trans. Speech Audio Processing, vol. 2, pp. 567 - 577, October 1994.

[54] Biem, A. and Katagiri, S., "Cepstrum-based filter-bank design using discriminative feature extraction training at various levels," Proceedings of ICASSP'97, pp. 1503 - 1506, April 1997.

[55] Bourlard, Hervé, Dupont, Stéphane, "Subband-based speech recognition," Proceedings of ICASSP'97, pp. 1251 - 1254, April 1997.

[56] Cerisara, C., Haton, J.-P., Mari, J.-F., Fohr, D., "A recombination model for multi-band speech recognition," Proceedings of ICASSP'98, pp. 717 - 720, May 1998.

[57] Chengalvarayan, Rathinavelu, Deng, Li, "Use of generalized dynamic feature parameters for speech recognition," IEEE Trans. Speech Audio Processing, vol. 5, pp. 232 - 242, May 1997.

[58] Davis, Steven B., Mermelstein, Paul, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, pp. 357 - 366, August 1980.

[59] De Mori, R., D. Albesano, Gemello, R., and Mana, F., "Ear-model derived features for automatic speech recognition," Proceedings of ICASSP'00, pp. 1603 - 1606, June 2000.

[60] Dinei A. F. Florencio, Ronald W. Schafer, "Perfect reconstructing nonlinear filter banks," Proceedings of ICASSP'96, pp. 1815 - 1818, May 1996.

[61] Gemello, R., et al., "Integration of fixed and multiple resolution analysis in a speech recognition system," Proceedings of the International Conference on Acoustics, Speech, Signal Processing, Salt Lake City, UT, 2001.

[62] Gowdy, J. N. and Tufekci, Z., "MEL-Scaled discrete wavelet coefficients for speech recognition," Proceedings of ICASSP'00, pp. 1351 - 1354, June 2000.

[63] Kim, D., S. Lee, and Kil, R. M., "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] Kronland-Martinet R., "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," Computer Music J., 12(4), pp. 11- 20, 1988.

[65] Kronland-Martinet, R., Morlet , J., and Grossmann A., "Analysis of sound patterns through wavelet transforms," Int. J. Pattern Recog. Articial Intell., 1(2), pp. 273-302,1987.

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

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

[68] Okawa, Shigeki, Bocchieri, Enrico, Potamianos, Alexandros, "Multi-band speech recognition in noisy environments," Proceedings of ICASSP'98, pp. 641 - 644, May 1998.

[69] Sarikaya, R., Pellom, B., Hansen, J.H.L., "Wavelet Packet Transform Features with Application to Speaker Identification," NORSIG-98 IEEE Norsic Signal Processing Symposium, pp. 81-84, Vigso, Denmark, June 1998.

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

[71] Sreenivas, T. V. and Niederjohn, R. J., "Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise," IEEE Trans. Signal Processing, vol. 40, pp. 282 - 293, February 1992.

[72] Sreenivas, T. V., Singh, K., and Niederjohn, R. J., "Spectral resolution and noise robustness in auditory modeling," Proceedings of ICASSP'90, pp. 817 - 820, April 1990.

[73] Štrancar, A., "Uporaba multiresolucijske analize pri parametrizaciji govora," Zbornik sedme Elektrotehniške in računalniške konference ERK '98, 24. - 26. september 1998, Portorož, Slovenija. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, 1998.

[74] Tan, B. T., Fu, M., Spray, A., and Dermody, P., "The use of wavelet transforms in phoneme recognition," The Fourth International Conference on Spoken Language Processing (ICSLP), Philadelphia, PA, USA, October 3-6, 1996.

[75] Tibrewala, Sangita, Hermansky, Hynek, "Sub-band based recognition of noisy speech," Proceedings of ICASSP'97, pp. 1255 - 1258, April 1997.

[76] Tomlinson, M. J., Russell, M. J., Moore, R. K., Buckland, A. P., Fawley, M. A., "Modelling asynchrony in speech using elementary single-signal decomposition," Proceedings of ICASSP'97, pp. 1247 - 1250, April 1997.

[77] Vaseghi, Saeed, Harte, Naomi, Milner, Ben, "Multi-resolution phonetic/Segmental features and models for HMM-based speech recognition," Proceedings of ICASSP'97, pp. 1263 - 1266, April 1997.

D.4 Akustične značilke

[78] Ahn, R. and Holmes, W. H., "An improved harmonic-plus-noise decomposition method and its application in pitch determination," IEEE Speech Coding Workshop, pp. 41 - 42, September 1997.

[79] Atal, B. S. and Rabiner, L. R., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. 24, pp. 201 - 212, June 1976.

[80] Harris, J. D. and Nelson, D., "Glottal pulse alignment in voiced speech for pitch determination," Proceedings of ICASSP'93, pp. 519 - 522, April 1993.

[81] Hedelin, P. and Huber, D., "Pitch period determination of aperiodic speech signals," Proceedings of ICASSP'90, pp. 361 - 364, April 1990.

[82] Hosom, J. P., "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information," Doctoral dissertation, University of Massachusetts, 2000.

[83] Janer, L., "New pitch detection algorithm based on wavelet transform," IEEE-SP Int. Sym. Time-Fre. Time-Scale Anal., pp. 165 - 168, October 1998.

[84] Markel, John D., "The SIFT algorithm for fundamental frequency estimation, IEEE Transactions on Audio Electroacoust.," vol. 20, pp. 367 - 377, December 1972.

[85] McAulay, Robert J., Quatieri, Thomas F., "Pitch estimation and voicing detection based on a sinusoidal speech model," Proceedings of ICASSP'90, pp. 249 - 252, April 1990.

[86] Moore, B. C. J., "An Introduction to the Psychology of Hearing," Academic Press, San Diego, CA, 1997.

D.5 Dinamično prilagajanje frekvenčno-časovne ločljivosti spektra govornega signala

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

[88] George, E. Bryan, Smith, Mark J. T., "Speech analysis/Synthesis and modification using an analysis-by-synthesis/Overlap-add sinusoidal model," IEEE Trans. Speech Audio Processing, vol. 5, pp. 389 - 406, September 1997.

[89] Goodwin, M., "Multiresolution sinusoidal modeling using adaptive segmentation," Proceedings of ICASSP'98, pp. 1525 - 1528, May 1998.

[90] Goodwin, Michael, "Residual modeling in music analysis-synthesis," Proceedings of ICASSP'96, pp. 1005 - 1008, May 1996.

[91] Jones, Douglas L., Baraniuk, Richard G., "A simple scheme for adapting time-frequency representations," IEEE Trans. Signal Processing, vol. 42, pp. 3530 - 3535, December 1994.

[92] Kozek, W., "Optimally Karhunen-Loeve-like STFT expansion of nonstationary processes," Proceedings of ICASSP'93, pp. 428 - 431, April 1993.

[93] Levine, Scott N., Verma, Tony S., Smith III, Julius O., "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio," Proceedings 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 101 - 104, October 1997.

[94] Prandoni, Paolo, Goodwin, Michael, Vetterli, Martin, "Optimal time segmentation for signal modeling and compression," Proc. ICASSP'97, pp. 2029 - 2032, April 1997.

[95] Ramchandran, Kannan, Vetterli, Martin, "Best wavelet packet bases in a rate-distortion sense," IEEE Trans. Image Processing, vol. 2, pp. 160 - 175, April 1993.

[96] Stankovic, L. and Katkovnik, V., "Algorithm for the instantaneous frequency estimation using time-frequency distributions with adaptive window width," IEEE Signal Processing Lett., vol. 5, pp. 224 - 227, September 1998.

[97] Štrancar, A., Rozman, R., Kodek, D., "Parametrizacija govornega signala z dinamičnim določanjem dolžine okna," Zbornik dvanajste mednarodne Elektrotehniške in računalniške konference ERK 2003, str. 481-484, september 2003.

[98] Štrancar, A., Rozman, R., Kodek, D., "Dinamično prilagajanje dolžine okvirja fonemski zgradbi besed v postopku parametrizacije govornega signala," Zbornik trinajste mednarodne Elektrotehniške in računalniške konference ERK 2004, str. 167-170, september 2004.

[99] Sun, X. and Bao, Z., "Adaptive spectrogram for time-frequency signal analysis," Proceedings of 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing, pp. 460 - 463, June 1996.

D.6 Analiza lastnosti človekovega slušnega zaznavanja

[100] Bu, L. and Chiueh, T., "Perceptual speech processing and phonetic feature mapping for robust vowel recognition," IEEE Trans. Speech Audio Processing, vol. 8, pp. 105 - 114, March 2000.

[101] Ghitza, O. and Kroon, P., "Dichotic presentation of interleaving critical-band envelopes: An application to multi-descriptive coding," IEEE Speech Coding Workshop, pp. 74 - 76, September 2000.

[102] Ghitza, O., "Auditory models and human performance in tasks related to speech coding and speech recognition," IEEE Trans. Speech Audio Processing, vol. 2, pp. 115 - 132, January 1994.

[103] Hermansky, Hynek, Morgan, Nelson, "RASTA processing of speech," IEEE Trans. Speech Audio Processing, vol. 2, pp. 578 - 589, October 1994.

[104] Kates, James M., "A time-domain digital cochlear model," IEEE Trans. Signal Processing, vol. 39, pp. 2573 - 2592, December 1991.

[105] Kates, James M., "Accurate tuning curves in a cochlear model," IEEE Trans. Speech Audio Processing, vol. 1, pp. 453 - 462, October 1993.

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

[107] Philipos C. Loizou, "Mimicking the human ear," IEEE Signal Processing Magazine, Vol. 15, pp. 101 - 130, September 1998.

[108] Rozman, R., "Uporaba spoznanj o človekovi slušni percepciji v sistemu za razpoznavanje govora," magistrsko delo, Fakulteta za Računalništvo in Informatiko, Ljubljana, junij 1999.

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

[110] Skoglund, J. and Kleijn, W. B., "On time-frequency masking in voiced speech," IEEE Trans. Speech Audio Processing, vol. 8, pp. 361 - 369, July 2000.

[111] Virag, Nathalie, "Speech enhancement based on masking properties of the auditory system," Proceedings of ICASSP'95, pp. 796 - 799, May 1995.

ZAHVALA

Ob pisanju disertacije mi je stalo ob strani več ljudi. Na prvem mestu se zahvaljujem mentorju, prof. dr. Dušanu Kodeku. Ves čas me je podpiral pri raziskovanju in mi pomagal s prenekatero izkušnjo. Zahvaliti se želim tudi Robertu Rozmanu, sodelavcu, ki se ukvarja s področjem razpoznavanja govora. Pomagal je s svojimi izkušnjami in nasveti pri iskanju primernih člankov. Posebna zahvala gre tudi Johnu-Paulu Hosomu. Omogočil mi je vpogled v izvorno kodo njegovih metod, ki so bile uporabljene za prilagajanje frekvenčno-časovne ločljivosti spektra in mi s tem prihranil precej časa. Omeniti želim tudi avtorje orodja CSLU Toolkit, ki so mi pomagali z odgovori na vprašanja s tem v zvezi. Zahvaliti se želim še ostalim sodelavcem: Zvonetu Petkovšku, Igorju Škrabi in Damjanu Šoncu za podporo pri nastanku tega dela.

IZJAVA

Spodaj podpisani Andrej Štrancar 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. Andrej Štrancar, univ. dipl. inž.

IZVIRNI PRISPEVKI

Disertacija vsebuje naslednje izvirne prispevke k znanosti:

- *prilagajanje dolžine okvirja pri računanju MFCC koeficientov*

  Predlagan je postopek prilagajanja dolžine okvirja pri spektralni analizi govornega signala, s katerim se frekvenčno-časovno ločljivost prilagodi trenutni dinamiki spreminjanja spektra. Dinamika spreminjanja spektra govornega signala je bila določena na osnovi fonemske zgradbe ter z uporabo dveh akustičnih značilk s področja avtomatskega označevanja govora.

  - *prilagajanje dolžine okvirja na osnovi fonemske zgradbe besed*

    Znano je, da je pri nekaterih fonemih spekter skoraj stacionaren, pri drugih pa se spreminja hitro. Zato je pri spektralni analizi različnih fonemov dolžino okvirja smiselno prilagoditi posameznemu fonemu. To je mogoče ob znani fonemski zgradbi vzorca, ki se jo določi s SRG s fiksno dolžino okvirja. Dolžino okvirja se dobljeni fonemski zgradbi prilagodi v drugem prehodu.

  - *prilagajanje dolžine okvirja zvenečim in nezvenečim odsekom govora*

    Pri preizkušanju prilagajanja dolžine okvirja fonemski zgradbi besed se je pokazalo, da je pri parametrizaciji zvenečih fonemov smiselno uporabiti daljši okvir. Zato je bil predlagan postopek prilagajanja dolžine okvirja glede na to, ali je določen odsek govora zveneč ali ne. Za določanje zvenečih odsekov je bila uporabljena akustična značilka, s katero je mogoče zaznati periodične spremembe govornega signala, ki so povezane z vibriranjem glasilk. Ob uporabi takega pristopa dodaten prehod razpoznavanja ni potreben.

- *uporaba obtežene vsote dveh spektrov z različno frekvenčno-časovno ločljivostjo pri računanju MFCC koeficientov*

  Predlagan je bil postopek, pri katerem se pri računanju MFCC koeficientov uporabi obteženo vsoto dveh spektrov, pri čemer se en spekter izračuna z daljšim drugega pa s krajšim okvirjem. V obteženi vsoti spektrov se frekvenčno-časovno ločljivost prilagaja s spreminjanjem uteži, s katero je določeno, kateri izmed spektrov je v obteženi vsoti bolj poudarjen.

  - *prilagajanje frekvenčno-časovne ločljivosti spektra spremembam jakosti zvoka*

    Predlagan je postopek, ki sloni na spoznanju, da je pri govornem signalu večje spremembe jakosti mogoče povezovati s hitrimi spremembami spektra. Spremembe jakosti so posebej izrazite pri izgovorjavi zapornikov in na prehodih med fonemi. Za določanje uteži v obteženi vsoti dveh spektrov je bila uporabljena akustična značilka, ki spremembe jakosti opisuje skladno s človekovim slušnim zaznavanjem.

STVARNO KAZALO