UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Damjan Šonc

SAMODEJNO UČEČI SE SISTEM ZA SINTEZO ČLOVEŠKEGA GOVORA

DOKTORSKA DISERTACIJA

Mentor: prof. dr. Dušan Kodek

Ljubljana, 2011

Povzetek

Že desetletja se raziskovalci po celem svetu trudijo, da bi naredili sistem, ki bi iz poljubnih besedil tvoril tak umetni govorni signal, ki bi bil hkrati in razumljiv in bi zvenel kot naravni govor. Zelo bi bilo tudi zaželeno, da bi sistem deloval na taki računalniški strojni opremi, ki je cenovno dostopna čim širšemu krogu uporabnikov. Skozi zgodovino so se kot rezultati raziskav pojavljali različni sistemi, ki so skušali navedene cilje doseči. Razdelimo jih lahko v tri generacije. Toda tudi sistemi zadnje - tretje generacije - še niso dosegli vseh naštetih ciljev. Najtežje dosegljiv cilj je visoka naravnost umetnega govora.

Pri izdelavi sistema računalniške sestave govora je potrebno združiti različna znanja, tako s področja procesiranja signalov, kot tudi jezikoslovja, glasoslovja in drugih znanosti.

Sistemi za sestavo govora, ki tvorijo najkakovostnejši govorni signal, delujejo tako, da s pomočjo raznih algoritmov obdelujejo in sestavljajo skupaj kratke posnetke naravnega govora. Za sestavo ustrezne zbirke posnetkov pa še vedno potrebujemo veliko ur strokovnega dela.

V disertaciji je skupaj s postopki za njegovo izgradnjo opisan sistem NGS (Nauči se Govoriti Sam), ki se iz posnetkov naravnega govora in pripadajočih besedil sam nauči tvoriti umetni govor. Sistem spada v skupino sistemov, ki govorni signal tvorijo z izbiro in lepljenjem glasovnih enot iz zbirke posnetkov naravnega govora (Unit Selection Synthesis). Ustrezno zbirko posnetkov si v procesu učenja sproti ustvarja sistem sam s pomočjo algoritma za samodejni razrez, ki je rezultat raziskav doktorske disertacije. Za sestavo zbirke posnetkov tako ne potrebujemo velike količine strokovnega dela, kar je še posebej primerno za slovenski jezik in tudi druge jezike, ki jih govori manjše število ljudi. Za jezike, ki jih govori veliko število ljudi, je namreč lažje dobiti potrebne strokovnjake.

Sistem sestavlja govorni signal s pomočjo sinusnega generatorja, za katerega smo s pomočjo statističnih testov dokazali, da lahko tvori umeten govorni signal, ki ga ne moremo ločiti od naravnega govora. S tem smo nakazali, da obstaja način tvorbe umetnega govora, ki zveni čisto naravno.

Prototipna različica NGS, ki je nastala kot rezultat doktorske disertacije, tvori razumljiv govor, končni cilj nadaljnjega dela, pa je doseči tudi visoko stopnjo naravnosti. Pri nadaljnjem delu želimo izboljšati tudi natančnost samodejnega razreza in najti metode za samodejno pridobivanje prozodičnih značilk iz naravnega govornega signala.

Ključne besede: sinteza govora, sestava govora, sestava z izbiro enot, samodejni razrez govora, učeči se sistem.

Abstract

The ultimate goal of speech synthesis is to build a system that could convert arbitrary written messages into intelligible and natural sounding speech. Such a system should also run on hardware platforms that we meet in everyday's life like a personal computer.

The solutions that appeared in the last five decades can be divided into three different generations. Unfortunately, even the latest systems from the third generation are far from generating perfectly natural sounding speech. Currently, the best quality of the synthetic speech is obtained from the systems that belong to the group of Unit Selection Synthesis Systems. To build an adequate database of speech units a lot of work from trained engineers is required.

The main objective of this Ph.D. thesis was to develop a system that could learn how to produce a high quality synthetic speech from the text and corresponding speech samples only, without requirements for skilled human labor or trained ASR (Automatic Speech Recognition) systems. The system should use statistical, machine learning techniques instead and algorithms for the automatic speech segmentation that do not require ASR.

For the purposes of the thesis a prototype of the speech synthesis system named Learn to Speak by Yourself (LSY) was constructed. LSY belongs to the group of Unit Selection Synthesis Systems. The core of the LSY is made of the newly developed algorithm for the automatic speech segmentation that does not require the usage of an ASR system. The algorithm exploits the spectral differences between different phonemes (allophones) of a language. This approach is particularly useful for the Slovene or some other language with a relatively small number of speakers where it is more difficult to find skilled engineers or well trained ASR systems for the speech database construction. The system can start from scratch – i.e. no speech unit database is required. The database is automatically built during learning process.

For generation of the speech samples the LSY uses a sinusoidal generator. The statistical results obtained from the listening tests show that synthetic speech produced by the generator in a synthesis by analysis process cannot be distinguished from a natural human speech. We may conclude that in theory a perfectly natural sounding synthetic speech can be produced by LSY.

At this time the speech produced by a prototype version of the LSY is highly intelligible but not yet natural sounding. The main reason is the fact that only a few minutes of speech samples were fed to the prototype system while research results found in the literature recommend at least one hour of speech samples and even systems with five hours or more of speech samples are not uncommon.

The future work will be concentrated on methods for the automatic extraction of prosody parameters from the speech samples. We would also like to improve the algorithm for the automatic speech segmentation.

Keywords: speech synthesis, unit selection synthesis, automatic speech segmentation, trainable speech synthesis.

I Z J A V A O A V T O R S T V U

doktorske disertacije

Spodaj podpisani/-a Damjan Šonc ,

z vpisno številko 24001162 ,

sem avtor/-ica doktorske disertacije z naslovom

Samodejno učeči se sistem za sintezo človeškega govora

S svojim podpisom zagotavljam, da:

- sem doktorsko disertacijo izdelal/-a samostojno pod vodstvom mentorja (naziv, ime in priimek)

  prof. dr. Dušan Kodek

  in somentorstvom (naziv, ime in priimek)

- so elektronska oblika doktorske disertacije, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko doktorske disertacije

- in soglašam z javno objavo elektronske oblike doktorske disertacije v zbirki »Dela FRI«.

V Ljubljani, dne 10.01.2011 Podpis avtorja/-ice: _____

Zahvala

Hvala vsem, ki ste tako ali drugače prispevali, da me je življenjska pot privedla do doktorske disertacije.

Dodatno se zahvaljujem svojemu mentorju prof. dr. Dušanu Kodeku za spodbudo, veliko koristnih nasvetov ter za pripombe med pisanjem doktorata. Prof. dr. Romanu Trobcu se zahvaljujem za pazljivo branje osnutka doktorata ter za vse pripombe in nasvete, ki mi jih je pri tem posredoval. Vsem sodelavcem v laboratoriju: Andreju Štrancarju, Robertu Rozmanu, Zvonetu Petkovšku, Igorju Škrabi in Andražu Božičku hvala za odlično vzdušje, ki da delu v laboratoriju poseben čar. Hvala tudi Miru Romihu in Simonu Rozmanu s katerima že leta sodelujem na področju sestave govora.

Prav posebna zahvala pa gre vsem mojim bližnjim, ki ste me spodbujali, prenašali, spremljali v času pisanja doktorata. Ste smisel in nevidno gonilo življenjske poti.

Kazalo vsebine

Slikovno kazalo

1 Uvod

Govor, kot eden najvažnejših načinov sporazumevanja med ljudmi, vedno bolj prodira tudi v komunikacijsko povezavo med človekom in strojem – računalnikom. Glavni razlog, da ta prehod ni hitrejši, je nezadostna kakovost razpoznave in sestave govornega signala, ki predvsem ne sme biti vsebinsko omejen.

Razpoznava govora pomeni spreminjanje govornega signala v informacijo v obliki besedila, operacija sestave (sinteze) govora pa poteka ravno v obratni smeri in sicer tvori govorni signal iz podanega besedila. V nadaljevanju se bomo omejili samo na sestavo govora.

Danes obstaja že kar nekaj sistemov [12], [13], [14], ki lahko iz besedila tvorijo govorni signal zelo visoke kakovosti. Ti sistemi pa so izdelani le za nekaj svetovno najbolj razširjenih jezikov kot so: angleščina, nemščina, francoščina, španščina, japonščina, kitajščina, itn. Glavni razlogi za tako stanje so predvsem v veliki zahtevnosti izdelave sistema, saj za njegovo izdelavo potrebujemo veliko človek-ur dela in znanja s področja procesiranja signalov ter jezikoslovja. Šele dosežki zadnjih nekaj let in velik porast računalniških zmogljivosti so omogočili izdelavo sistemov za sestavo govornega signala, ki se po kakovosti že lahko kosa z naravnim govorom.

1.1 Značilnosti sistemov za sestavo govora

Pri sistemih za umetno sestavo govora nas zanimajo predvsem naslednje značilnosti, ki hkrati določajo tudi uporabnost takih sistemov:

- Kakovost umetnega govornega signala.

  Sistem mora tvoriti govorni signal visoke kakovosti, pri čemer kakovost določata dve merili:
  1. razumljivost in
  2. naravnost.

- Zmožnosti pri tvorbi govornega signala.

  Sistem mora omogočati tvorbo govornega signala z neomejenim naborom besed, po možnosti z različnimi glasovi, z različno hitrostjo izgovorjave ter z različnimi načini naglaševanja in to za različne jezike.

- Računalniške zmogljivosti, ki so potrebne za delovanje sistema.

  Računalniške zmogljivosti merimo kot potrebno procesorsko moč ter potrebno količino pomnilnika. Zaželeno je, da sistem deluje kot običajna aplikacija na osebnem računalniku.

1.1.1 Razumljivost in naravnost umetnega govornega signala

Razumljivost govora je merilo kakovosti govornega signala in predstavlja razmerje med pravilno razpoznanimi glasovnimi enotami (glasovi, besede, stavki) in vsemi glasovnimi enotami, ki jih zajema slušni test, s katerim razumljivost govora merimo. Je subjektivno merilo, ker je odvisno od poslušalcev, ki na testu sodelujejo.

Visoko razumljivost govora so dosegali že sistemi v poznih 70-tih letih dvajsetega stoletja. Sistem MITalk-79 [19] je na testih, kjer so razumljivost ocenjevali s pomočjo spremenjenih rim (modified ryme test), dosegal 93% razumljivost, samo nekaj let novejši sistem DECtalk [20] pa kar 97% razumljivost. Za primerjavo povejmo, da so v istih testih z naravnim govorom dosegli 99% razumljivost. Ker človek govorni signal v veliki meri razpoznava tako, da slabše razumljene besede naveže na vsebino pogovora, 100% razumljivost posameznih glasovnih enot za normalno sporazumevanje sploh ni potrebna. Čeprav so poslušalci pri poslušanju signala sistema DECtalk na testu spremenjenih rim zabeležili trikrat večjo napako, kot pri poslušanju naravnega govora, lahko ravno zaradi človekovega načina razpoznave govora trdimo, da sistem DECtalk tvori govorni signal enakovredne razumljivosti, kot je naravni govorni signal. Če bi torej ocenjevali kakovost računalniških govornih sistemov samo z merilom razumljivosti, bi lahko zaključili, da je problem računalniške sestave govora že rešen, saj najboljši sodobni sistemi dosegajo na testih razumljivosti le malenkost boljše rezultate. Seveda je to daleč od resnice in problem računalniške sestave govora bo mučil raziskovalce po vsem svetu še precej časa. Če izpustimo težave, ki se pojavljajo pri prevedbi poljubno zapisanega besedila v obliko, ki verno odraža zaporedje glasov (npr. zamenjava okrajšav in raznih simbolnih zapisov z besedami), ostane še vedno ogromno težav, ki jih moramo rešiti, če želimo poleg razumljivosti doseči tudi visoko stopnjo naravnosti umetnega govora.

Pod **naravnost umetnega govora** razumemo lastnost, da umetni govor zveni kot naravni človeški govor in ga ljudje težko oziroma v idealnem primeru sploh ne morejo razlikovati od naravnega človeškega govora.

Podatek o tem, ali ljudje umetni govor razlikujejo od naravnega, lahko pridobimo hitro in na sorazmerno preprost način. Težje pa s slušnimi testi natančneje opredelimo stopnjo naravnosti umetnega govora, ker so rezultati takih testov še bolj odvisni od razpoloženja ocenjevalcev kot pri določanju razumljivosti. Na tem mestu si vsekakor lahko zastavimo vprašanje, ali ljudje želijo, da bi stroj (računalnik) proizvajal naravno zveneč govorni signal, ali pa so zadovoljni tudi z "robotsko" zvenečim govorom. Raziskave [18] so pokazale, da ljudje mnogo raje sprejemajo umetni govorni signal, ki zveni čim bolj naravno. Še več. Ljudje so celo precej nestrpni do "robotsko" zvenečih sistemov in večina ljudi zavrača vsakršno uporabo nenaravno zvenečih sistemov, ne glede na to, kakšne dodatne koristi jim tak sistem sicer lahko ponudi. Ta nestrpnost večine ljudi do nenaravno zvenečih glasov je verjetno eden od glavnih razlogov sorazmerno majhne razširjenosti uporabe sistemov za sestavo govora v primerjavi z ostalimi računalniškimi tehnologijami, saj sama razumljivost umetnih govornih signalov ni več ovira že nekaj desetletij.

1.1.2 Zmožnosti pri tvorbi govornega signala

Glede na nabor besed, ki jih sistemi za sestavo govora lahko tvorijo, ločimo sisteme z omejenim in neomejenim naborom besed. **Sistemi z omejenim naborom** hranijo posnetke vnaprej določenih besed ali celih sporočil, ki jih nato lahko še kombinirajo med seboj za sestavo daljših sporočil. Ti sistemi so preprosti, tvorijo govorni signal zelo visoke kakovosti (naravni govor), njihova glavna pomanjkljivost, ki zelo omejuje področja njihove uporabe, pa je ravno omejenost v naboru besed oziroma sporočil, ki jih lahko tvorijo. Kljub temu take sisteme srečamo pri samodejnih sistemih za napoved časa, vremena, voznih redov javnih prevoznih sredstev ipd.

Sistemi z neomejenim naborom lahko v govor pretvorijo poljubno besedilo in so zapleteni. Razumljivost govora je sicer lahko visoka, naravnost pa le nizka do srednja. Sposobnost tvorbe tekočega govora je zelo visoka, saj lahko poljubno besedilo izgovorijo z različnimi glasovi, naglasi in hitrostmi. Ko bomo uspeli izboljšati še naravnost govornega signala, se bo uporaba teh sistemov precej razmahnila.

Pri pretvorbi poljubnega besedila v umetni govorni signal moramo opraviti tri glavne korake:

1. Najprej moramo razbrati jezikovno predstavitev sporočila, ki ga vsebuje besedilo. Razne okrajšave, številke in oznake moramo v vhodnem besedilu nadomestiti z ustreznimi besedami. Običajne zapise besed nato zamenjamo z glasovnimi zapisi, ki verno ustrezajo dejanski izgovarjavi vsake besede.

2. Glasovne zapise iz prvega koraka dodatno dopolnimo z informacijami, ki določajo osnovno višino glasu, podajajo krivulje naglaševanja ter druge podatke o prozodiji[1] in so odvisne od modela, ki ga uporabimo pri tvorbi govornega signala.

3. Iz podrobnih glasovnih določil, ki smo jih dobili v drugem koraku, tvorimo časovno odvisne funkcije krmilnih parametrov akustičnega modela, s katerim nazadnje tvorimo vzorce umetnega govornega signala.

1.2 Razvoj sistemov za sestavo govora

Sisteme za sestavo govora lahko razporedimo v tri glavne generacije. Prva generacija, ki je prevladovala do poznih 80-tih let dvajsetega stoletja, je temeljila na modelu govorne cevi. Glavna predstavnika te generacije sta formantna sestava govora [19] in sestava s pomočjo klasičnega linearnega napovedovanja (linear prediction) [26] .

Druga generacija je predstavljala glavno smer razvoja sestave govora v 90-tih letih dvajsetega stoletja in še vsaj prvo polovico prvega desetletja 21-ega stoletja. Metode druge generacije so uporabljale že podatkovno voden pristop in so temeljile na lepljenju in preoblikovanju kratkih

[1] Prozodija - vzorec poudarkov in naglasov jezika, ki vključuje tudi ritem.

enot, običajno dvoglasnikov [2] (difonov) posnetega govora. Najbolj znane metode druge generacije so PSOLA (Pitch Synchronous OverLap and Add) [23] in različni sinusni modeli [21], [31].

V zadnjo - tretjo - generacijo sistemov za sestavo govora spadajo metode, ki modelirajo govorni signal s pomočjo prikritih modelov Markova (HMM – Hidden Markov Model) [2], [36] in pa predvsem metode, ki govorni signal tvorijo z izbiranjem in lepljenjem glasovnih enot naravnega govornega signala (Unit Selection Synthesis) [16], [35], [28].

1.3 Namen disertacije in pregled vsebine

Osnovni namen te doktorske disertacije je raziskati možnosti za izdelavo sistema, ki bi ga lahko naučil govoriti "skoraj" vsak uporabnik, pri tem pa bi uporabnik moral samo posneti govorni signal določenega nabora besedil, ki bi mu ga posredoval računalnik. Sistemi, ki se naučijo govoriti iz vzorcev besedil in posnetkov govornega signala sicer že obstajajo, vendar potrebujejo za zagon v večini primerov že naučen razpoznavalnik ali sestavljalnik govora. Govorni signal, ki ga tvorijo, ne dosega stopnje naravnosti in razumljivosti, ki jo imajo "ročno nastavljeni" sistemi.

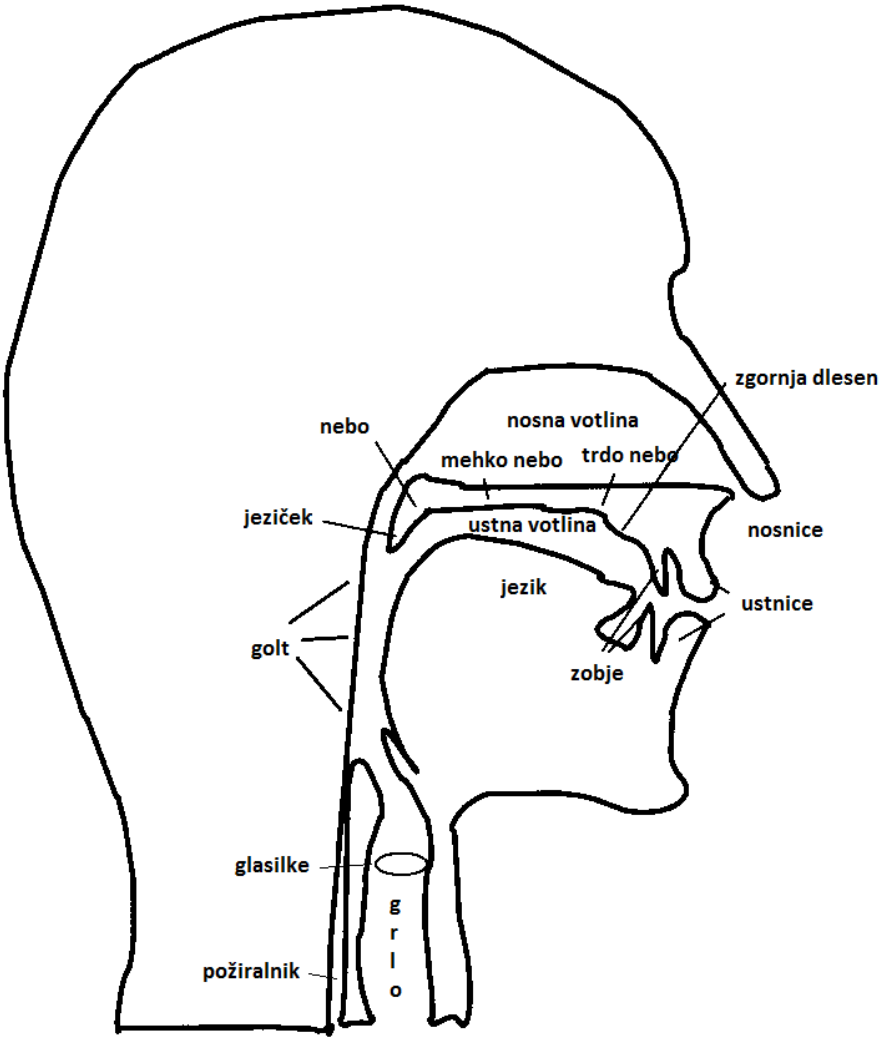
Področje sestave govora zahteva poleg znanj o procesiranju signalov tudi znanja s področja avtomatskega učenja, jezikoslovja in glasoslovja, zato je izgradnja sistema, ki se nauči govoriti sam in hkrati tvori govorni signal visoke stopnje razumljivosti in naravnosti zelo težka naloga.

V poglavjih 2 in 3 je najprej podan pregled obstoječega stanja na področju sestave govora. Akustični model, ki je podlaga za večino raziskav na tem področju, je opisan v poglavju 2 v poglavju 3 pa so nato dodane še metode in opisi sistemov za sestavo govornega signala, ki vsak za svoje obdobje predstavlja vrh dosežkov sestave govora. Zgradbo sistema NGS ali Nauči se Govoriti Sam, ki je plod raziskav te disertacije, opisuje poglavje 4, v poglavju 5 pa so podani rezultati raziskav, ki so vodili k izgradnji sistema NGS. Zaključki in smernice za nadaljnje delo so na koncu zapisani v poglavju 6.

[2] Dvoglasnik - določen je kot par sosednjih fonemov, ki se začne na sredini prvega fonema in konča na sredini drugega fonema.

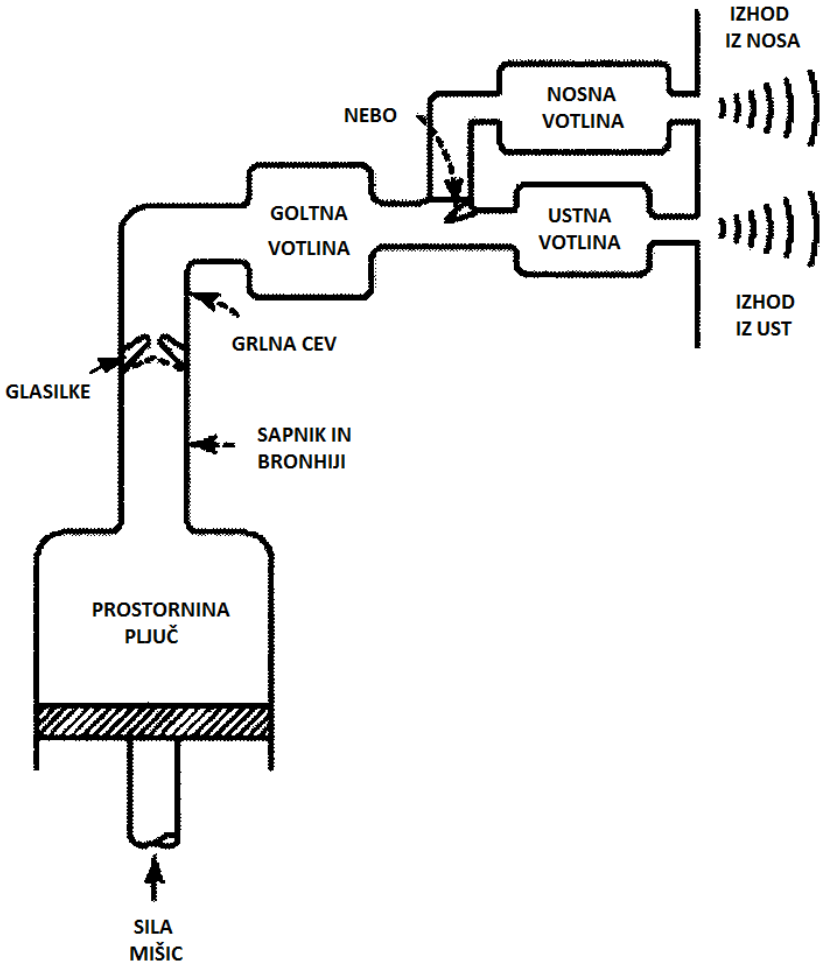
2 Akustični model sestave govora

Pri izdelavi modela, ki opisuje proces tvorbe govornega signala v človeškem govornem sistemu, smo prisiljeni narediti kompromis med natančnostjo in zahtevnostjo. Lahko naredimo tak model, ki zelo verno opisuje govorni proces, a je zaradi tega zelo zapleten in lahko tudi neobvladljiv. Lahko pa izdelamo model, ki je preprost in primeren za praktično uporabo, a ne zmore opisati govornega signala, ki bi bil enakovreden naravnemu govornemu signalu. Če bi želeli modelirati čisto vse značilnosti človekovega govornega sistema, bi morali upoštevati vsak mišični gib, poznati vse mogoče položaje in oblike organov človekovega govornega sistema, poznati bi morali zvočne absorpcijske značilnosti tkiv in tudi določiti, kako vse našteto vpliva na govorni signal. Tudi, če bi z meritvami lahko prišli do podatkov, ki bi dovolj natančno določali vse želene značilke govornega sistema, bi bil model, ki bi vse povezal v neko celoto, zelo verjetno preveč obsežen in zapleten, da bi bil tudi uporaben.



Slika 1: Človekov govorni sistem.

Zaradi vsega naštetega se je pri izdelavi akustičnega modela tvorbe umetnega govornega signala, ki je danes najbolj splošno sprejet in iz katerega so potem izšle tudi metode za sestavo govora prve generacije, izvedlo precej poenostavitev, ki so omogočile, da je model tudi obvladljiv in uporaben. Model temelji na človeškem govornem sistemu, ki ga prikazuje slika 1 (povzeto po [37]), poenostavljen mehanski model, ki ga modelira, pa slika 2 (povzeto po [7]) .



Slika 2: Mehanski model človekovega govornega sistema.

Zapleten človekov govorni sistem je v modelu porazdeljen na manjše število neodvisnih sestavnih delov. Pljuča skupaj z energijo mišic predstavljajo izvor zraka, ki iz pljuč teče mimo glasilk v goltno ter nato v ustno in/ali nosno votlino in na koncu skozi usta in nos v okoliški prostor. Če glasilke v tem procesu zanihajo, tvorijo osnovni ton govornega signala in zveneče glasove, sicer pa se v govorni cevi[3] (vocal-tract) tvorijo ostali nezveneči glasovi, ki so

[3] Govorna cev je votli prostor od grla navzven do ustnic oziroma nosnic. Zajema goltno, ustno in nosno votlino.

podobni različno obarvanemu šumu.

Pri matematičnem opisu modela [26], [35] zaradi lažje obvladljivosti upoštevamo dve predpostavki, ki sta bistveni za dovolj preprost opis, in sicer:

- linearnost in
- časovna nespremenljivost.

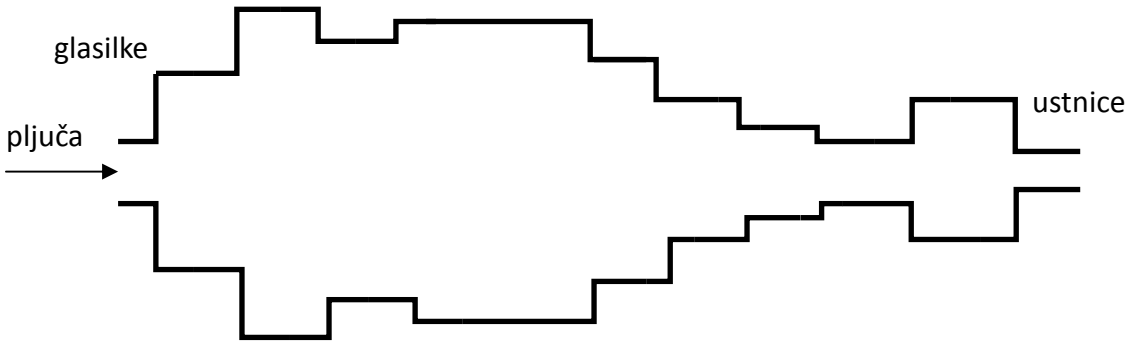
S tem ko predpostavimo linearnost modela, omogočimo, da odziv celotnega modela izračunamo kot preprosto linearno kombinacijo odzivov posameznih sestavnih delov. Dodatna zahteva po časovni nespremenljivosti nam skupaj z linearnostjo omogoča, da pri matematičnem modeliranju uporabimo teorijo linearnih časovno nespremenljivih sistemov in s tem posledično teorijo digitalnih filtrov. Predpostavka o časovni nespremenljivosti seveda drži le za dovolj kratke časovne intervale. Oblika govorne cevi se med govorjenjem neprestano spreminja, vendar so meritve pokazale, da lahko za časovne intervale, ki so krajši od 10 ms, privzamemo, da se znotraj tega intervala oblika govorne cevi ne spreminja in so posledično značilke, ki jo opisujejo, konstantne.

Govorno cev lahko torej opišemo kot linearen, časovno nespremenljiv sistem, sestavljen iz množice diskretnih, med seboj povezanih komponent. Posamezna komponenta lahko predstavlja izvor zvoka, ali pa filter, ki ta zvok preoblikuje. Model tudi predpostavlja, da je izvor zvoka popolnoma neodvisen od filtra, ki ga preoblikuje. To je ena od večjih pomanjkljivosti modela, saj je popolnoma jasno da, ko na primer govorimo z različnimi višinami glasu, z mišicami ne spreminjamo samo napetosti glasilk in s tem frekvence osnovnega tona, ampak spreminjamo tudi napetost in obliko drugih delov govornega sistema (slika 1), kar vpliva na značilnosti filtra, ki ga ti drugi deli govornega sistema predstavljajo.

2.1 Prenosna funkcija govorne cevi

Poglejmo si obnašanje človekovega govornega sistema pri izgovorjavi samoglasnikov. Pri izgovorjavi ne-nosnih samoglasnikov glasilke tvorijo zvočni val, ki potuje skozi grlo in ustno votlino in na koncu izzveni skozi ustnice. Nosna votlina je v tem primeru ločena od ostalih delov govorne cevi in bistveno ne vpliva na oblikovanje zvoka. Grlo, goltno ter ustno votlino lahko modeliramo kot eno cev. Presek cevi se od glasilk do ustnic spreminja in ravno zmožnost govorca, da spreminja obliko govorne cevi, mu omogoča izgovorjavo različnih samoglasnikov. Vsaka oblika drugače vpliva na zvočni val in pri določenih oblikah govorne cevi dobimo na izhodu iz ust zvok, ki ga razpoznamo kot enega od samoglasnikov.

Modeliranje cevi, ki se stalno spreminja, je zapleteno, vendar lahko dober približek izvedemo z zaporedno vezavo kratkih cevi različnih presekov, kot to prikazuje slika 3. Bolj podrobna izpeljava in razlaga tega postopka je opisana v več virih, med drugimi tudi v [26]. Z večanjem števila cevi lahko dosežemo poljubno natančnost modela, kar je eden od pomembnih rezultatov, ki so tam navedeni.



Slika 3: Model govorne cevi, sestavljen iz krajših odsekov cevi s konstantnim presekom.

Ko tvorbo samoglasnikov modeliramo z digitalnimi filtri, lahko enačbo govornega sistema po [26] zapišemo v obliki z-transformacije kot

$$Y(z) = \mathrm{U}(z)V(z)R(z), \tag{2.1}$$

kjer U(z) predstavlja prenosno funkcijo (z-transformacijo odziva na enotin impulz) izvora signala – v našem primeru glasilk, V(z) prenosno funkcijo govorne cevi, R(z) pa prenosno funkcijo (sevalno karakteristiko) ustnic. S pomočjo z-transformacije [24] smo konvolucijo zaporedij, ki predstavljajo impulzne odzive digitalnih filtrov, prevedli na zmnožek njihovih z-transformacij.

Prenosno funkcijo govorne cevi V(z) izpeljemo iz modela govorne cevi, ki je, kot smo prej omenili, sestavljen iz krajših odsekov cevi s konstantnim presekom. Širjenje zvoka v cevi lahko opišemo kot potujoče valovanje, ki se na robovih cevi odbija. Pri tem tudi predpostavimo, da ne prihaja do nobenih energijskih izgub valovanja skozi stene cevi. Stopnjo odboja valovanja podajajo odbojni koeficienti, ki so odvisni le od preseka cevi. Poenostavljen model govorne cevi, sestavljen le iz dveh cevi, prikazuje slika 4.

Iz večcevnega modela (N-cevni model) sta Rabiner in Schaffer v [26] izpeljala enačbo

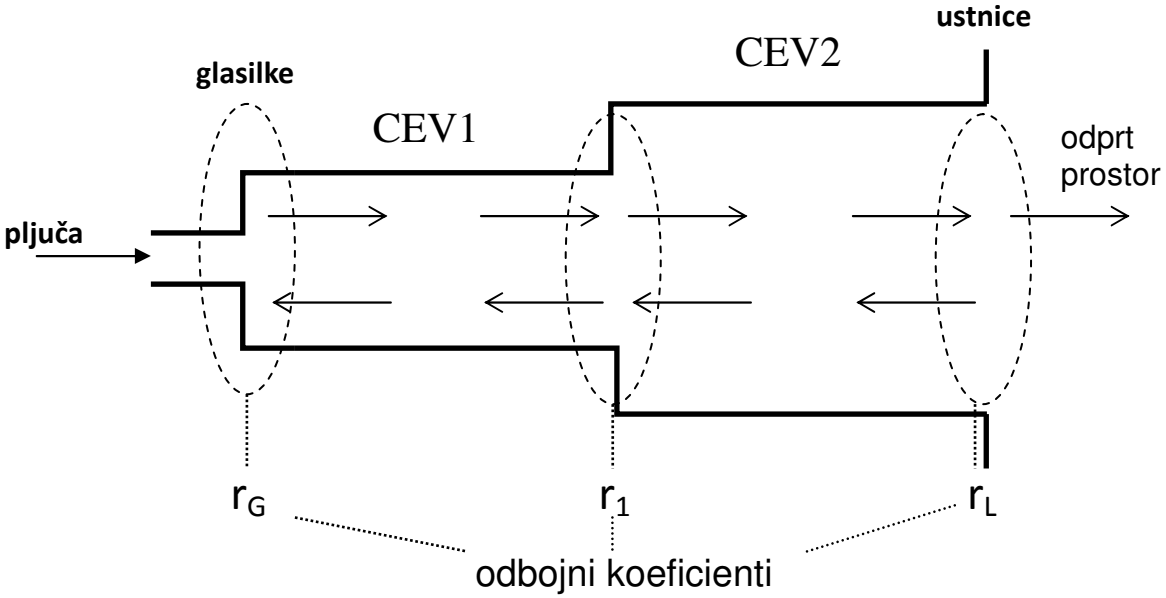
$$V(z) = \frac{1+r_G}{2} \frac{z^{-N/2} \prod_{k=1}^{N} (1+r_k)}{1 - \sum_{k=1}^{N} a_k z^{-k}}, \tag{2.2}$$

ki je zelo pomemben rezultat, saj nam pove, da lahko brez-izgubno govorno cev modeliramo

z digitalnim filtrom, ki ima samo pole v z-ravnini. Koeficienti r_k v enačbi (2.2) predstavljajo odbojne koeficiente med spoji večcevnega modela, koeficient r_G je odbojni koeficient na izvoru zvoka (glasilke), koeficient $r_N = r_L$ pa predstavlja odbojni koeficient na izhodu govorne cevi – ustnicah. Koeficiente a_k v imenovalcu ulomka izračunamo iz zmnožka matrik po enačbi

$$1-\sum_{k=1}^{N} a_k z^{-k} = [1,-r_G]\begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix}\cdots\begin{bmatrix} 1 & -r_N \\ -r_N z^{-1} & z^{-1} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{2.3}$$

Prenosno funkcijo (enačba 2.2) lahko nadalje razstavimo na filtre drugega reda, s katerimi verno modeliramo posamezne formante[4], ki sestavljajo samoglasnik.



Slika 4: Model govorne cevi, sestavljen iz dveh krajših cevi s konstantnim presekom.

2.2 Modeliranje sevanja ustnic

Govorni signal se pri tvorbi ne-nosnih samoglasnikov dokončno izoblikuje na izhodu iz ustnic. Natančno modeliranje sevalnih značilnosti ustnic R(z) je precej zahtevno, v večini primerov

[4]Formant je okrepljen snop harmoničnih tonov in je posledica resonanc govorne cevi.

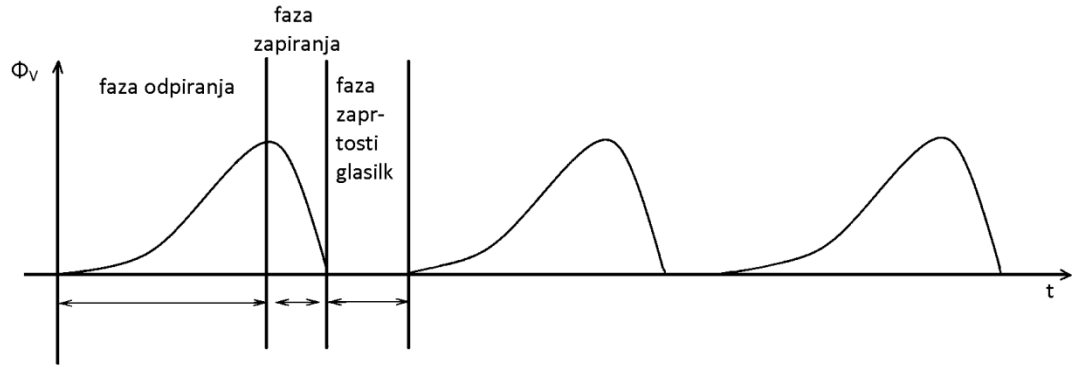
pa se je kot dovolj dober približek izkazal filter s končnim enotinim odzivom (KEO filter), ki ga podaja z-transformacija

$$R(z) = 1 - \alpha z^{-1}, \tag{2.4}$$

kjer je vrednost koeficienta α malo manjša od ena ($0{,}98 \leq \alpha < 1$). Učinek filtra je, da ojači visoke frekvence s 6 dB na oktavo.

2.3 Modeliranje izvora govornega signala

Osnovni izvor signala pri zvenečih glasovih predstavljajo glasilke, ki odpirajo in zapirajo pretok zraka iz pljuč na poti proti ostalim delom govorne cevi. Ko tlak zraka med pljuči in glasilkami naraste do določene meje, ki je odvisna od napetosti glasilk, se glasilke odprejo. Zrak steče skozi, kar povzroči padec tlaka in njihovo ponovno zapiranje. Proces se periodično ponavlja in hitrost ponavljanja predstavlja osnovni ton (pitch, F0) govornega signala. Pri moških se frekvenca osnovnega tona tipično giblje med 80 Hz in 250 Hz, pri ženskah in otrocih pa se tipične frekvence osnovnega tona nahajajo med 120 Hz in 400 Hz. Na sliki 5 lahko vidimo tri faze delovanja glasilk v procesu tvorbe osnovnega tona in sicer kot spreminjanje volumskega pretoka zraka ϕ_V v odvisnosti od časa. Faze delovanja so označene na eni periodi osnovnega tona.



Slika 5: Idealiziran prikaz delovanja glasilk (izvora zvoka), podan kot volumski pretok ϕ_V zraka v odvisnosti od časa.

Prva faza je faza odpiranja glasilk, ko je tlak zraka med pljuči in glasilkami dovolj velik, da se glasilke odprejo in začne zrak iz pljuč teči skoznje. Druga faza je faza zapiranja, ko se glasilke zaradi padca tlaka zapirajo, tretja faza pa je faza zaprtosti glasilk, ko so glasilke zaprte in tlak zraka med pljuči in glasilkami ponovno narašča. Časovni delež posamezne faze

znotraj periode osnovnega tona se spreminja, kot primer pa navedimo meritve iz [30], po katerih prva faza traja med 0,2 in 0,4 periode osnovnega tona, druga med 0,18 do 0,3 periode osnovnega tona, tretja pa med 0,3 in 0,5 periode osnovnega tona.

Medtem, ko je preprost model glasilk sorazmerno lahko narediti, je bolj natančno modeliranje izredno težavno, zato model izvora zvoka, ki bi natančno posnemal obnašanje glasilk v vseh mogočih primerih, sploh ne obstaja. Dober model mora upoštevati tudi sekundarne pojave pri delovanju glasilk kot so potresavanje (jitter), migljanje (shimmer) in valovanje (ripple).

Med obstoječimi modeli omenimo le dva, ki sta v raziskavah najpogosteje omenjena kot referenčna modela in sicer Rosenbergov model [27] in Liljencrants-Fantov model [4] ali tudi LF model. Rosenbergov model v vzorčeni obliki podaja enačba

$$g_R[n] = \begin{cases} 0{,}5\left(1 - \cos\left(\frac{\pi n}{N_1}\right)\right) & 0 \le n \le N_1 \\ \cos\left(\frac{\pi(n-N_1)}{2N_2}\right) & N_1 \le n \le N_1 + N_2 \\ 0 & sicer \end{cases}, \tag{2.5}$$

kjer je n indeks vzorca, N_1 število vzorcev v prvi fazi delovanja glasilk (faza odpiranja), N_2 pa število vzorcev v drugi fazi delovanja glasilk (faza zapiranja).

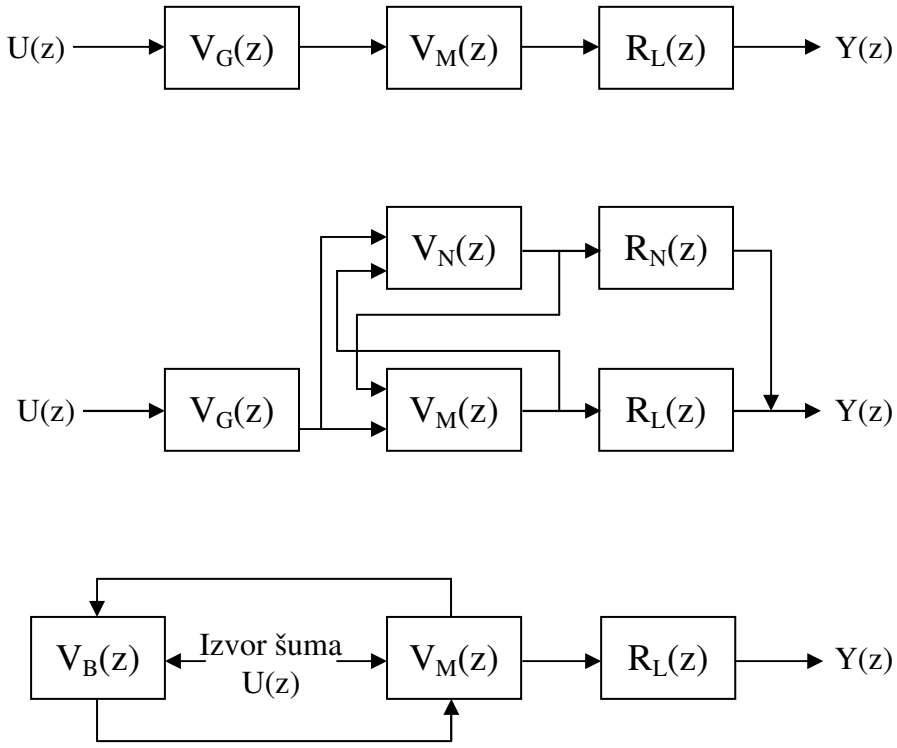
2.4 Izboljšave akustičnega govornega modela

Pri tvorbi osnovnega modela smo se naslonili predvsem na proces tvorbe ne-nosnih samoglasnikov in zanemarili ostale glasove. S tem smo se izognili vplivom nosne votline in upoštevali kot izvor zvoka le glasilke. Pri tvorbi soglasnikov pa moramo upoštevati tudi izvore šuma v govorni cevi. Izvori šuma so turbulence zraka v govorni cevi, ki nastanejo ob zožitvah. Pri nezvenečih soglasnikih so to edini izvori zvoka, saj pri tvorbi teh glasov glasilke ne sodelujejo. Morebitno delovanje glasilk med tvorbo teh glasov je le posledica vpliva zvenečih glasov, ki so nastali neposredno pred nezvenečim glasom, ali pa bodo nastali tik za njim. Za tvorbo vseh glasov je potrebno osnovni model nadgraditi oziroma uporabiti različne modele za različne skupine glasov.

Zato, da lahko dobimo jasnejšo predstavo o zahtevnosti modeliranja človeškega glasu, poleg osnovnega modela za tvorbo ne-nosnih samoglasnikov navajamo [35] še dva modela in sicer model za tvorbo nosnih samoglasnikov in model za tvorbo nezvenečih pripornikov[5]. Vse tri modele prikazuje slika 6. Na vrhu se nahaja že znani model za ne-nosne samoglasnike. Prenosna funkcija filtra govorne cevi V(z) je tu razdeljena na prenosno funkcijo goltne votline $V_G(z)$ in ustne votline $V_M(z)$, $R_L(z)$ pa predstavlja prenosno funkcijo filtra, ki posnema sevalne značilnosti na izhodu ustne votline oziroma ustnicah. Če temu modelu dodamo še

[5] Nezveneči priporniki slovenskega jezika so fonemi **f**, **s**, **z**, **š**, **ž** in **h**. Vsi fonemi slovenskega jezika so podani v prilogi B.

filtre, ki modelirajo vplive nosne votline, dobimo drugi model s slike 6, ki je primeren za tvorbo nosnih samoglasnikov. Z $V_N(z)$ in $R_N(z)$ smo označili prenosni funkciji filtra nosne votline in filtra, ki posnema sevalne značilnosti nosu. Za tvorbo nezvenečih glasov potrebujemo nekoliko drugačen model, ki se od osnovnega razlikuje predvsem v izvoru signala. Ker nezveneči glasovi v bistvu predstavljajo neke vrste obarvan šum, je najbolje, če za izvor signala U(z) uporabimo kar generator šuma. Tako pridemo do spodnjega modela, ki je prikazan na sliki 6. Ker se šum v govorni cevi pri človeku ne tvori v glasilkah, ampak v zožitvah ustne votline, ki jih oblikujemo predvsem z jezikom, moramo to upoštevati tudi pri tvorbi filtrov, ki šume oblikujejo v končne glasove. Tako imamo na zadnjem modelu poleg običajne prenosne funkcije filtra ustne votline $V_M(z)$ še prenosno funkcijo filtra $V_B(z)$, ki posebej modelira zadnji del ustne votline. Ta pri tvorbi nezvenečih glasov deluje kot stranski resonator, izvor šuma pa se nahaja med obema filtroma.



Slika 6: Modeli za tvorbo različnih glasov. Od vrha proti dnu se vrstijo: model za tvorbo ne-nosnih samoglasnikov, model za tvorbo nosnih samoglasnikov in model za tvorbo nezvenečih pripornikov.

3 Sistemi za sestavo govora

Skozi zgodovino lahko sisteme za sestavo govora razporedimo v tri generacije. Na razvoj vsake generacije so bistveno vplivale računalniške zmogljivosti, ki so bile v danem času na razpolago in sicer procesorska moč ter cenovno dostopna zadostna količina pomnilnika.

3.1 Prva generacija sistemov za sestavo govora

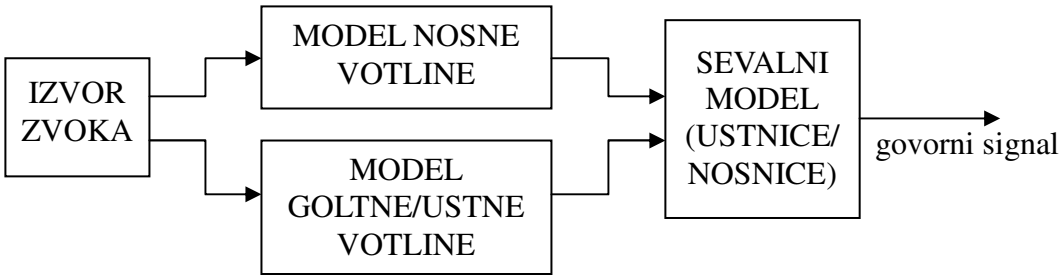
Prva generacija sistemov za sestavo govora, ki so prevladovali vse do konca 80-tih let dvajsetega stoletja, je temeljila na modelu govorne cevi, ki je opisan v poglavju 2. Danes so ti sistemi sicer že zastareli, vendar so v aplikacijah, kjer imamo na razpolago zelo omejene računalniške zmogljivosti, še vedno konkurenčni, saj sodobnejši sistemi potrebujejo večje procesorske in pa predvsem pomnilniške zmogljivosti, ki so na voljo šele v zadnjih letih. Preko sistemov prve generacije dobimo vpogled tudi v današnje stanje pri sestavi govora in lahko pojasnimo, zakaj so danes vodilni modeli sestave govora tako narejeni kot so.

Sistemi prve generacije običajno zahtevajo precej podroben, nizkonivojski zapis o tem, kar naj bi sistem izgovoril. Poleg samega glasovnega zapisa besedila potrebujejo še podatke o dolžinah posameznih glasov ter vsaj še podatke o poteku osnovnega tona za celoten stavek.

3.1.1 Formantna sestava govora

Formantna sestava govora je prva prava računalniška metoda sestave govora in je prevladovala do zgodnjih 80-tih let dvajsetega stoletja. Velikokrat se imenuje tudi sestava s pomočjo pravil. To pa zato, ker so s tem hoteli poudariti, da s to metodo tvorimo signal tako rekoč iz nič po določenih pravilih. Formantna sestava uporablja modularen, akustično-fonetični pristop, ki sloni na modelu govorne cevi. Metoda uporablja model na poseben način, tako da so značilke cevi preprosto povezljive z akustično-fonetičnimi lastnostmi, ki jih zlahka opazujemo, recimo s pomočjo spektrograma. Kot izvor zvoka služi periodični signal za zveneče in šum za nezveneče glasove. V praktično vseh formantnih sestavljalnikih sta ustna in nosna votlina modelirani posebej kot vzporedna sistema.

Blokovni prikaz tipične osnovne izvedbe formantnega sestavljalnika prikazuje slika 7.



Slika 7: Blokovni prikaz tipičnega formantnega sestavljalnika.

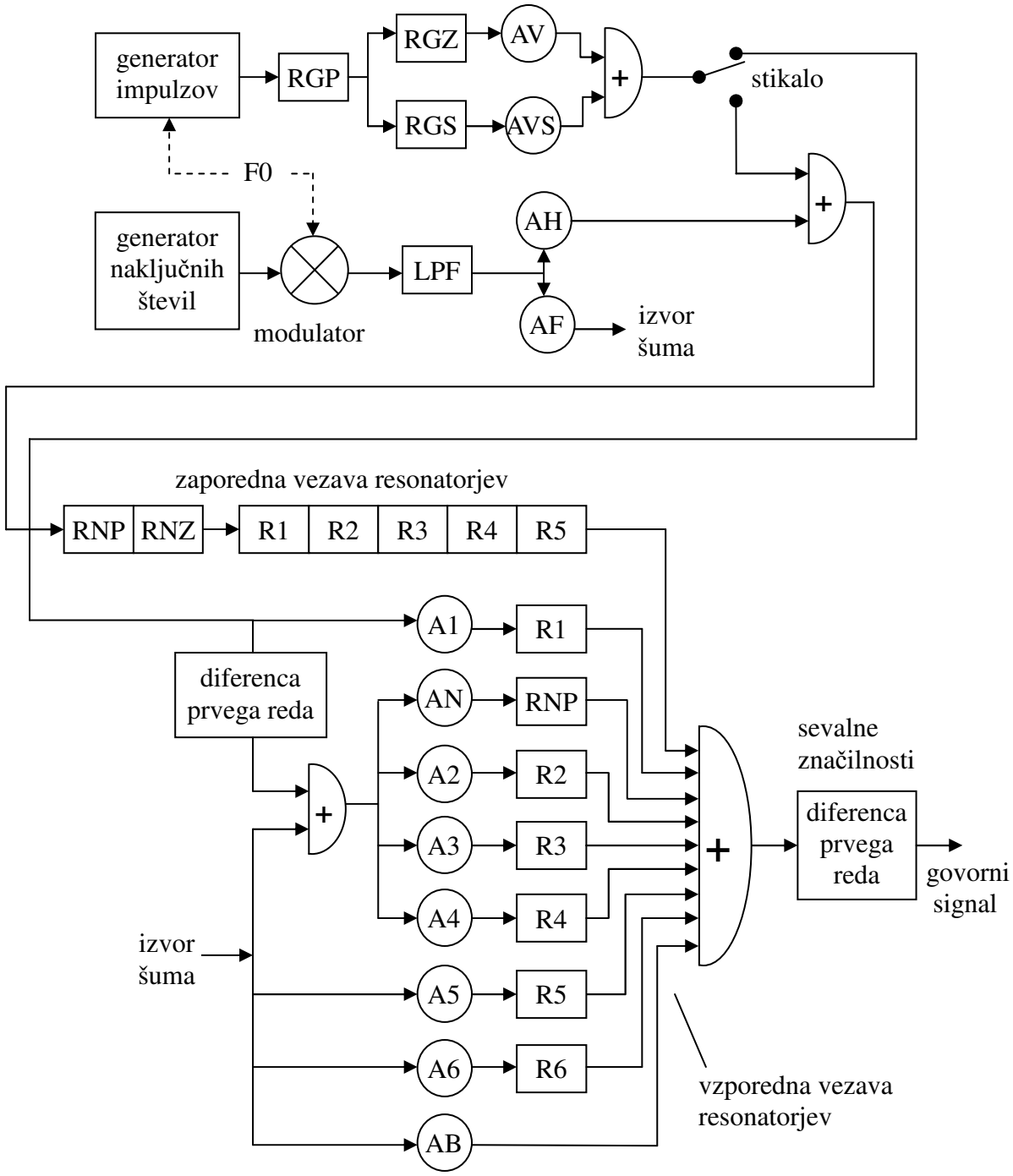
Enega najbolj popolnih formantnih sestavljalnikov je razvil Denis Klatt [19] in ga prikazuje slika 8, dodatno razlago oznak na sestavnih blokih sestavljalnika pa podaja tabela 1. Krmiliti ga je potrebno z množico značilk, ki se spreminjajo vsakih nekaj milisekund. Značilke tvorimo s pomočjo množice pravil in v prilogi C je kot primer podan seznam pravil, po katerih računamo dolžino trajanja posameznega fonema za angleški jezik. Ta seznam pravil je mogoče uporabiti v vseh sestavljalnikih umetnega govora, ki dolžino trajanja posameznih fonemov določajo s pomočjo pravil in ne samo pri formantnih sestavljalnikih.

Klattov sestavljalnik kot izvor periodičnega signala uporablja generator impulzov (simulacija glasilk), ki ga krmilimo s frekvenco osnovnega tona umetnega govornega signala. Izhodni signal iz generatorja se dodatno oblikuje s pomočjo dveh resonatorjev (RGP, RGS) in anti-resonatorja (RGZ), ki skupaj simulirajo učinke grla. Amplitudo signala določata dve značilki (AV, AVS). Drugi izvor signala v sestavljalniku predstavlja generator šuma, ki je sestavljen iz generatorja naključnih števil, iz modulatorja, ki ga krmilimo s frekvenco osnovnega tona, in iz nizko-propustnega filtra. Navedene izvore signala lahko vodimo v vejo zaporedno vezanih resonatorjev, kot tudi v vejo vzporedno vezanih resonatorjev. Vsak resonator določa en formant govornega signala. V veji vzporedno vezanih resonatorjev nastavljamo amplitudo vsakega izmed šestih formantov posebej (A1, A2, ..., A6), v veji zaporedno vezanih resonatorjev pa amplitudo določimo že z izvorom signala (AH, AV, AVS). Zaporedna vezava je najprimernejša za sestavo samoglasnikov, medtem ko je vzporedna vezava potrebna za sestavo pripornikov in zapornikov.

Če želimo z večcevnim modelom govorne cevi modelirati človeški govor s polno natančnostjo, ki jo model omogoča, potrebujemo en formant za vsakih 1000 Hz vzorčevalne frekvence. Ker je bil sestavljalnik v osnovi narejen za vzorčevalno frekvenco 10 kHz, bi torej moralo biti resonatorjev za formante več. Raziskave pa so pokazale [35], da za razlikovanje med različnimi glasovi zadostujejo samo trije formanti, medtem ko ostali le prispevajo k večji naravnosti umetnega govora.

Klattov formantni sestavljalnik torej ni natančen model govorne cevi. Bistveno se razlikuje tudi od večcevnega modela, saj omogoča ločen in neodvisen nadzor nad vsakim formantom posebej. Pri večcevnem modelu posamezna cev ne določa posameznega formanta, ampak tvori vse formante sistem kot celota. Ločen nadzor nad formanti pri formantnem

sestavljalniku je narejen zato, ker je iz spektrogramov mnogo lažje razbrati realne značilke formantov, ki jih potem uporabimo za krmiljenje sestavljalnika, kot pa določiti realne oblike govorne cevi, katere model bi potem uporabili za tvorbo formantov govornega signala. To velja še danes, kljub vsemu napredku tehnologije, ki omogoča natančno sledenje dogajanj v govorni cevi.



Slika 8: Klattov formantni sestavljalnik.

Oznaka	Opis
R1, R2, R3, R4, R5, R6	Resonatorji od 1 do 6 (določeni s frekvenco in pasovno širino)
RGP	Resonator grla
RGS	Resonator grla 2
RGZ	Anti-resonator grla
RNP	Nosni resonator
RNZ	Nosni anti-resonator
A1, A2, A3, A4, A5, A6	Amplitude formantov od 1 do 6
AN	Amplituda nosnega formanta
AB	Amplituda premostitve
AV	Amplituda periodičnega izvora signala (vlak impulzov)
AVS	Amplituda skoraj sinusnega periodičnega izvora signala
AF	Amplituda pri tvorbi pripornikov (f, s, z, š, ...)
AH	Amplituda pri tvorbi zapornikov (p, b, t, ...)
F0	Frekvenca osnovnega tona
LPF	Nizko prepustni filter (Low Pass Filter)

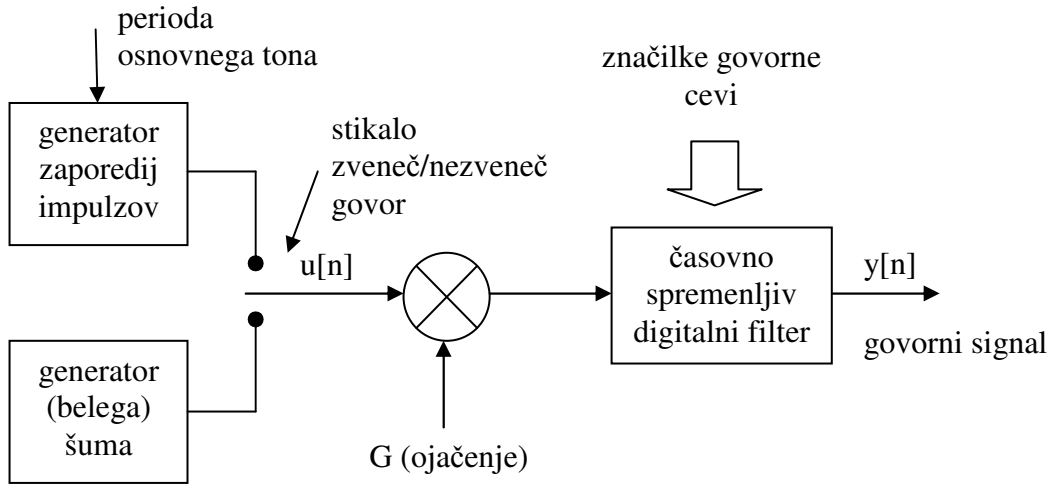
Tabela 1: Razlaga oznak na sestavnih delih Klattovega formantnega sestavljalnika.

Umetni govorni signal, ki ga tvori formantni sestavljalnik, je le malenkost slabše razumljiv od naravnega govora, zveni pa precej nenaravno, kar je posledica vseh pomanjkljivosti uporabljenega modela. Za visoko razumljivost umetnega govornega signala je namreč dovolj, če so posamezni glasovi dovolj jasno ločljivi med seboj in ni potrebno, da zvenijo naravno. Žal pa je nenaravnost umetnega govora glavna omejitev pri uporabi formantnega sestavljalnika, saj so raziskave pokazale, da ljudje nenaravnega govornega signala ne sprejemajo, oziroma jim "gre na živce".

3.1.2 Sestava govora s pomočjo klasične metode linearnega napovedovanja

Značilk za krmiljenje formantnega sestavljalnika ne moremo vedno zlahka in neposredno pridobiti iz govornega signala. Čeprav formante v večini primerov lahko določimo iz spektrogramov, je to časovno potratno in podvrženo človeškim napakam. Če uporabimo samodejni formantni sledilnik, lahko večino težav omilimo. Toda tudi najboljši samodejni sledilnik ne zmore vedno ločiti različnih formantov. Govorna cev ima takrat takšno obliko, da govorni signal nima izrazitih spektralnih vrhov. Lahko pa pri iskanju značilk za sestavo umetnega govornega signala uporabimo drugačen pristop. Namesto, da bi iskali ustrezne značilke za tvorbo različnih formantov, lahko uporabimo kar značilke prenosne funkcije govorne cevi. Če upoštevamo predpostavko o neodvisnosti izvora signala in predpostavko, da lahko govorno cev modeliramo s filtrom, ki ima samo pole v z-ravnini, potem lahko potrebne značilke vedno določimo samodejno s pomočjo linearnega napovedovanja (LP – linear prediction). Slika 9 prikazuje preprost model sestave govora, ki sloni na akustičnem modelu opisanem v poglavju 2 in upošteva obe navedeni predpostavki. Model, postopki analize

govornega signala, kodiranje ter postopki sestave govornega signala s pomočjo linearnega napovedovanja (Linear Predictive Analysis, Linear Predictive Coding, Linear Predictice Sytnhesis) so v literaturi obširno opisani. V nadaljevanju smo navedli samo pomembnejše rezultate iz [26].



Slika 9: Blokovni prikaz preprostega LP modela sestave govora.

Osnovni model LP sestave govora, ki ga prikazuje slika 9, uporablja dva izvora signala in sicer generator zaporedij impulzov, ki ga krmilimo z želeno periodo osnovnega tona govornega signala, ter generator šuma, ki ga običajno izvedemo z generatorjem psevdonaključnih števil. Stikalo, ki preklaplja med obema izvoroma, krmilimo z informacijo o tem, ali želimo tvoriti zveneč, ali nezveneč glas. Izbran izvor signala $u[n]$ ustrezno ojačimo (množenje s koeficientom ojačenja G) in vodimo v digitalni filter, ki ima samo pole v z-ravnini. Na izhodu iz filtra dobimo umetni govorni signal $y[n]$. Vse navedene značilke (perioda osnovnega tona, izbor zveneč/nezveneč glas, koeficient ojačenja G, koeficienti digitalnega filtra) se počasi spreminjajo, pridobimo pa jih z analizo naravnega govornega signala. Periodo osnovnega tona lahko izmerimo s pomočjo laringografa[6], ali pa uporabimo enega izmed algoritmov (na primer [1]), ki periodo osnovnega tona izračuna iz vzorcev naravnega govora. Informacija o tem, ali gre za zveneč ali nezveneč glas (prisotnost/odsotnost osnovnega tona), je največkrat dodaten stranski rezultat omenjenih algoritmov. Postopke določanja koeficientov digitalnega filtra in koeficienta ojačenja G pa si poglejmo malo podrobneje.

Umetni govorni signal $y[n]$ na izhodu modela je z izvorom signala povezan z enačbo

[6] Laringograf – merilna naprava za spremljanje pojavov v grlu (delovanje glasilk).

$$y[n] = \sum_{k=1}^{P} a_k y[n-k] + Gu[u], \tag{3.1}$$

oziroma v obliki z-transformacije

$$H(z) = \frac{Y(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}}, \tag{3.2}$$

kjer so a_k koeficienti digitalnega filtra, s $H(z)$ pa smo označili prenosno funkcijo modela. Približek trenutnega vzorca signala $\tilde{y}[n]$ lahko z metodo linearnega napovedovanja izračunamo kot linearno kombinacijo P prejšnjih vzorcev signala po enačbi

$$\tilde{y}[n] = \sum_{k=1}^{P} \alpha_k y[n-k]. \tag{3.3}$$

Število P imenujemo red linearne napovedi, z n pa smo označili indeks vzorca. Razliko med pravo vrednostjo vzorca $y[n]$ in napovedano vrednostjo vzorca $\tilde{y}[n]$ imenujemo napaka linearne napovedi za trenutni vzorec in jo označimo z $e[n]$. Z enačbo to zapišemo kot

$$e[n] = y[n] - \tilde{y}[n] = y[n] - \sum_{k=1}^{P} \alpha_k y[n-k], \tag{3.4}$$

oziroma v obliki z-transformacije

$$E(z) = A(z)Y(z), \tag{3.5}$$

kjer je

$$A(z) = 1 - \sum_{k=1}^{P} \alpha_k z^{-k}. \tag{3.6}$$

V primeru, ko model natančno modelira govorni signal, velja $a_k = \alpha_k$ in $e[n] = Gu[n]$. V tem primeru lahko prenosno funkcijo modela zapišemo kot

$$H(z) = \frac{G}{A(z)}. \tag{3.7}$$

Z $A(z)$ je določen tako imenovan analizni filter modela, katerega koeficiente izračunamo v postopku analize govornega signala, s $H(z)$ pa tako imenovani sestavni filter modela, ki ga uporabljamo v postopku sestave govornega signala. Osnovna naloga analize s pomočjo linearnega napovedovanja je določiti množico koeficientov α_k neposredno iz naravnega govornega signala tako, da kar najbolje zajamemo spektralne značilnosti signala. Ker se spektralne značilnosti govornega signala s časom spreminjajo, moramo analizo izvajati nad kratkimi odseki govornega signala in sicer tako, da zaporedje vzorcev naravnega govornega signala razrežemo na krajše odseke in potem na vsakem odseku izvršimo analizo.

Obstajata dve osnovni metodi za izračun koeficientov α_k: kovariančna in avtokorelacijska. Obe računata koeficiente s pomočjo minimizacije kvadrata napake linearne napovedi E po enačbi

$$E = \sum_n e^2[n]. \tag{3.8}$$

Pri obeh metodah rešujemo sistem P linearnih enačb s P neznankami. Metodi se razlikujeta v načinu določanja odseka signala nad katerim izvajamo analizo.

Pri avtokorelacijski metodi moramo zaporedje vzorcev najprej množiti z ustreznim N-točkovnim oknom, kjer je N število vzorcev v odseku signala na katerem želimo izvesti analizo. S tem dosežemo, da so vsi vzorci izven želenega intervala enaki 0. Uporabimo lahko trikotno, Hannovo, Hammingovo ali kakšno drugo okno. Največkrat uporabljamo kar Hammingovo okno, ki je podano z enačbo

$$w(n) = \begin{cases} 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right), & n = 0{,}1, \ldots, N-1 \\ 0, & sicer \end{cases}, \tag{3.9}$$

ker lahko njegove koeficiente sorazmerno preprosto izračunamo in ima od zgoraj navedenih oken ob isti širini glavnega vala amplitudnega odziva najmanjšo vršno amplitudo stranskih valov.

Pri kovariančni metodi preprosto vzamemo N zaporednih vzorcev signala, ki jim na začetku zaporedja dodamo še P predhodnih zaporednih vzorcev. Indeks vzorcev je tako med $-P$ in $N-1$.

Po vsaki izmed metod dobimo nekoliko drugačne koeficiente. Tako ene kot druge lahko uporabimo v LP modelu za sestavo govornega signala. V praksi se največkrat uporablja

avtokorelacijska metoda, ker je filter s koeficienti, ki jih izračunamo po tej metodi, vedno stabilen, medtem, ko to za filter s koeficienti, ki jih izračunamo po kovariančni metodi, ne velja vedno.

Pri računanju energijskega koeficienta G si pomagamo s predpostavko, da je energija signala napake enaka energiji izvora signala in G izračunamo s pomočjo enačbe

$$G^2 \sum_{m=0}^{N-1} u^2[m] = \sum_{m=0}^{N-1} e^2[m]. \tag{3.10}$$

Pri primerjanju prenosnih funkcij zaporednega formantnega sestavljalnika in modela, kjer prenosno funkcijo določimo s pomočjo linearnega napovedovanja (LP model), so v [35] navedene naslednje pomembne ugotovitve:

- Oba modela imata enak tip prenosne funkcije in tvorita enak obseg frekvenčnih odzivov. Prednost LP modela je, da lahko LP koeficiente izračunamo samodejno kar iz govornega signala in ne potrebujemo analize formantov, ki ne da vedno pravilnih rezultatov.

- Prenosno funkcijo sestavljeno iz samih polov lahko za poljuben govorni signal natančno ocenimo, ne moremo je pa vedno razstaviti na posamezne formante.

- Pri LP modelu za vse glasove uporabimo prenosno funkcijo sestavljeno iz samih polov, kar pri splošnem formantnem sestavljalniku, kot je na primer Klattov, ne velja.

- Pri splošnem formantnem sestavljalniku posebej modeliramo ustno in nosno votlino, česar pri LP modelu ne storimo.

- Pri LP modelu so značilnosti filtra izvora signala že vključene v filter govorne cevi.

Kako bi lahko LP model uporabili za sestavo poljubnega umetnega govornega signala? Najpreprostejša rešitev je, da posnamemo množico odsekov naravnega govora, ki predstavljajo en glas, ali pa kratka zaporedja glasov naravnega govora. Odseke nato analiziramo s pomočjo LP modela in tako dobljene značilke shranimo v bazo posnetkov. S tem, ko LP analizo izvajamo nad odseki naravnega govornega signala, zajamemo tudi dinamiko govorne cevi. Če odseki zajemajo več kot en glas, s tem hkrati razrešimo tudi težave, ki nastanejo pri modeliranju prehodov med različnimi glasovi. V posnetkih moramo zajeti vse tipe prehodov med glasovi, ki nas zanimajo in katerih število je pri normalnem človeškem govoru končno. Posnete odseke lahko nato sestavljamo v različna zaporedja, ki ustrezajo želenemu govornemu signalu. Ta način uporabe LP modela nas vodi k tako imenovani sestavi z lepljenjem, ki pa že predstavlja naslednjo generacijo sestave govora.

Najbolj pogosta enota pri sestavi z lepljenjem je dvoglasnik (diphone). Določen je kot par

sosednjih fonemov[7], ki se začne na sredini prvega fonema in konča na sredini drugega fonema. Glavni razlog, ki upravičuje uporabo dvoglasnikov je, da lahko uporabimo model tarča-prehod, kjer nam stabilno področje "tarče" predstavlja sredina glasu[8], ki ima zatem prehodno področje do sredine naslednjega glasu. Model tarča-prehod predpostavlja, da ima na sredini glasov govorna cev za kratek čas stabilno obliko, ostalo dogajanje v govorni cevi pa so prehodi med temi stanji. Model nadalje predpostavlja, da imajo isti glasovi enako konfiguracijo govorne cevi in bi se zato morali lepo lepiti med seboj. Lepljenje na robovih glasov bi bilo v nasprotju s tem manj uspešno, ker so tam spremembe oblik govorne cevi največje.

Ko izvajamo LP sestavo govornega signala, v bazi posnetkov najprej poiščemo značilke ustreznih dvoglasnikov. Značilke posameznih dvoglasnikov nato zložimo v enotno zaporedje. S tem smo načeloma zgradili zaporedje okvirjev z ustreznimi ciljnimi stanji in prehodi med njimi – za zaporedje glasov, ki ga želimo sestaviti. Hkrati moramo tudi preveriti, ali ima govor, ki ga tvorimo, ustrezen čas trajanja in ustrezno periodo osnovnega tona. Čas trajanja lahko prilagodimo na dva načina. Poglejmo si ju na primeru. Recimo, da imamo odsek govora v dolžini 150 ms. Če je dolžina posameznega okvirja 10 ms, to pomeni, da je odsek sestavljen iz 15 okvirjev. Recimo, da želimo dolžino skrajšati na 120 ms. Po prvem načinu lahko preprosto skrajšamo dolžino okvirjev iz 10 ms na 8 ms (15 x 8 ms = 120 ms). Če pa želimo dolžino odseka podaljšati, ustrezno podaljšamo dolžino posameznega sestavnega okvirja. Drugi pristop, ki ga lahko uporabimo, je ta, da pri skrajševanju ohranimo dolžino okvirjev nespremenjeno in preprosto izpustimo vsak k-ti okvir (v našem primeru vsak peti okvir). Pri daljšanju odsekov pa ustrezne okvirje preprosto podvojimo. Metoda krajšanja oziroma podaljševanja enot s spuščanjem oziroma podvajanjem okvirjev je bolj groba od prilagajanja dolžin okvirjev, vendar za manjše spremembe (do dvakratno podaljšanje ali skrajšanje) kljub temu zelo dobro deluje.

Ko smo določili vse potrebne dolžine okvirjev in ostale značilke, ki jih zahteva LP model, izberemo še ustrezen izvorni signal. Za zveneče glasove uporabimo zaporedje impulzov, za nezveneče pa generator šuma. Za sestavo vsakega okvirja govornega signala uporabimo svoj nabor značilk, ki smo ga pridobili v postopku analize. Pri zaporednih okvirjih, ki so na meji med dvema dvoglasnikoma ali kakima drugima enotama, ki jih uporabljamo pri sestavi govora, lahko med značilkami sosednjih okvirjev dodatno izvedemo interpolacijo, zato da so prehodi bolj gladki.

Čeprav lahko LP sestava verno posnema ciljne in prehodne dinamične pojave govorne cevi in s tem naravnega govora, vseeno zveni precej nenaravno. Govorni signal, ki ga dobimo s pomočjo klasičnega LP modela, ljudje opisujejo kot kovinski ali brneč. Razloga sta v preveč preprostem izvornem signalu in predpostavki o popolni neodvisnosti izvora in filtra. S pomočjo LP modela sicer lahko sestavimo tak signal, ki je popolnoma enak izvornemu, pod pogojem, da filter vzbujamo s signalom napake $e[n]$, ki smo ga pridobili pri analizi signala.

[7] Glasnik ali fonem je najmanjša glasovna enota, s katero govorci določenega jezika razlikujejo pomen besed. S spremembo enega glasnika v besedi vedno dobimo drugo besedo ali pa postane beseda nerazpoznavna.

[8] Glas ali alofon je uresničitev glasnika v govoru. Istemu glasniku lahko v posameznem jeziku ustreza več glasov (alofonov). V slovenščini na primer glasniku /v/ ustrezajo štirje glasovi: zobnoustnični v, dvoglasni u ter zveneči in nezveneči ustničnoustnični w.

Če ga analiziramo, vidimo večje špice (krajši odsek signala z veliko amplitudo) v intervalih, ki ustrezajo periodi osnovnega tona. V našem izvornem signalu smo ga nadomestili kar z zaporedjem impulzov. V signalu napake pa lahko opazimo tudi veliko manjših špic in ker jih v našem izvornem signalu čisto zanemarimo, nam klasični LP model da prej omenjeni, nenaravno zveneč zvok. Tu zelo dobro vidimo, da smo izvorni signal očitno preveč poenostavili.

3.2 Druga generacija sistemov za sestavo govora

Kot smo omenili že v poglavju 1.1.1, lahko s prvo generacijo sistemov za sestavo govora tvorimo umetni govorni signal tako visoke razumljivosti (97%), da je le malo manjša od razumljivosti naravnega govornega signala (99%). Naravnost umetnega govornega signala teh sistemov pa je prenizka, da bi se njih uporaba razširila med povprečne uporabnike, ki poleg visoke razumljivosti umetnega govora zahtevajo tudi visoko stopnjo naravnosti. Kje so glavne pomanjkljivosti oziroma omejitve metod prve generacije, da z njimi ne moremo tvoriti bolj naravno zvenečega umetnega govora?

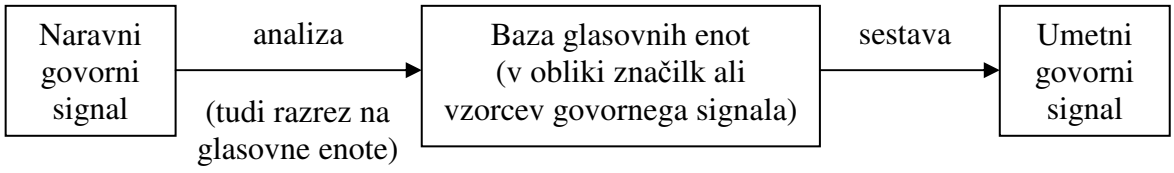
Omejitve formantne sestave govora so predvsem v tvorbi ustreznih značilk za krmiljenje sestavljalnika. Z ustreznim naborom značilk lahko tvorimo tudi govorni signal z zelo visoko stopnjo naravnosti, vendar mora biti nabor posebej prilagojen in ga ne moremo na preprost način pridobiti s pomočjo meritev naravnega govornega signala. Zaradi tega tudi z izboljšavami modela na podlagi podatkov, ki jih dobimo z natančnejšimi meritvami, ne pridobimo nič. Preslikovalne funkcije iz vhodnega glasovnega zapisa besedila v značilke, ki so potrebne za krmiljenje sestavljalnika, ostajajo zelo zapletene in jih ne znamo izraziti s pomočjo pravil, ki jih je sposoben sestaviti človek, ne glede na to kako dober strokovnjak je.

Pri LP sestavi smo se delom težav pri tvorbi značilk modela izognili tako, da značilk, ki opisujejo stanje govorne cevi, ne tvorimo umetno s pomočjo neke funkcije, ampak jih neposredno izmerimo iz govornega signala. Izvor signala pa je pri klasični LP sestavi še vedno preveč preprost (zaporedje impulzov oziroma generator šuma), da bi z njo dobili naravno zveneč govor.

Metode druge generacije sistemov za sestavo govora skušajo odpraviti slabosti, ki izvirajo iz preveč preprostega modeliranja izvora signala. Namesto zaporedja impulzov in/ali generatorja šuma uporabljajo kot izvor kompleksnejši signal, ki je lahko sestavljen kar iz vzorcev naravnega govora, ali pa ga tvori nek model. Skupni značilnosti metod druge generacije sta, da značilke za krmiljenje modela izvora govornega signala pridobivajo z analizo vzorcev naravnega govora, oziroma da uporabijo kar vzorce naravnega govora in da poljuben govorni signal tvorijo z lepljenjem kratkih odsekov govornega signala. Vhodne podatke za krmiljenje izvora signala, ki zajemajo vrsto in trajanje posameznega glasu ter vsaj še periodo osnovnega tona, pa še vedno nadzoruje eksplicitni model. Na primer: za tvorbo poteka osnovnega tona potrebujemo model, ki izračuna frekvenco osnovnega tona na vsakih nekaj milisekund. Metode prve in druge generacije potem izvornemu signalu spreminjajo trajanje in periodo osnovnega tona v skladu z vrednosti, ki jih model izračuna.

Razlike med metodami druge generacije izvirajo v glavnem iz načina izračunavanja značilk, ki so potrebne za tvorbo govornega signala. Vse uporabljajo podatkovno voden pristop, se pravi, da potrebne informacije pridobijo z analizo naravnega govornega signala. Pri tem pa nekatere izračunavajo značilke modela, ki ga uporabljajo (na primer koeficiente filtra za modeliranje govorne cevi), medtem ko druge uporabljajo kar neposredno vzorce naravnega govornega signala. Tipičen proces analize oziroma sestave govornega signala prikazuje slika 10.

Na metode druge generacije sistemov za sestavo govora lahko pogledamo tudi z drugačnega vidika. Gre za metode, kjer posnamemo vzorce naravnega govora in nad njimi izvedemo razne signalno-procesno-sestavne operacije, s katerimi tvorimo tudi glasovne in besedne zveze, ki niso prisotne v izvornih odsekih govora.



Slika 10: Tipičen proces analize/sestave govora v sistemih druge generacije.

Standardni sistem druge generacije krmilimo z zaporedjem glasovnih določil, kjer vsako določilo zajema znakovni zapis odseka govora, ki ga imenujemo glasovna enota, višino osnovnega tona in čas trajanja. V govorni bazi za vsak znakovni zapis glasovne enote hranimo vzorce naravnega govornega signala v surovi obliki ali v obliki značilk. Za vsako glasovno enoto običajno hranimo le en zapis. Med sestavo govora v govorni bazi samo poiščemo značilke oziroma vzorce govornega signala, ki ustrezajo znakovnim zapisom vhodnega zaporedja glasovnih določil in jih zlepimo skupaj. Hkrati skladno z glasovnimi določili tudi spremenimo trajanje in periodo osnovnega tona posamezne glasovne enote. Sprememba periode osnovnega tona in trajanja glasovnih enot pa je lahko kar precej zapletena operacija, če pri tem ne želimo imeti nezaželenih stranskih učinkov. Veliko naporov pri razvoju sistemov druge generacije je bilo vloženih ravno v signalno procesne algoritme, ki bi omogočali spreminjanje periode osnovnega tona in spreminjanje trajanja glasovnih enot, ne da bi v umetni govorni signal vnesli nezaželena popačenja.

Najpogostejša glasovna enota v sistemih za sestavo govora druge generacije je dvoglasnik (diphone). Ključna naloga pri izgradnji sistema je pazljiva izbira polnega nabora mogočih dvoglasnikov. V naravnem govoru se vse možne glasovne kombinacije seveda ne pojavijo, zato lahko že tu izbiro omejimo in izločimo vse nemogoče (nerealne) kombinacije, ki se v jeziku, za katerega gradimo sistem, ne pojavijo. Ko določimo poln nabor potrebnih dvoglasnikov, moramo najti tudi ustrezna zaporedja vzorcev naravnega govora, ki jim pripadajo. V sistemih druge generacije običajno potrebujemo samo po eno zaporedje vzorcev naravnega govora za vsak dvoglasnik iz nabora. Zaradi tega morajo biti ti vzorci čim bolj

ustrezno izbrani. Posnetki dvoglasnikov morajo zadostiti trem osnovnim merilom:

1. Izbrati moramo taka zaporedja vzorcev za posamezni dvoglasnik, da se dobro zlivajo skupaj. In kašno merilo naj določa dobro zlivanje? Ustrezno merilo so čim manjše akustične razlike do sosednjih dvoglasnikov. Sosednji dvoglasniki pa so tisti, ki imajo enak levi oziroma desni glas in jih lahko prilepimo z leve oziroma z desne strani.

2. Želimo imeti čim bolj tipične predstavnike dvoglasnikov. Izmed mnogih različic izberemo akustično najbolj nevtralne.

3. Zaradi spreminjanja trajanja in periode osnovnega tona glasovne enote je smiselno, da izmed več primerkov izberemo tiste, pri katerih navedene spremembe povzročijo kar najmanj popačenj. Izberemo torej tiste, ki imajo povprečno periodo osnovnega tona in nekoliko nadpovprečen čas trajanja. Nekoliko nadpovprečen čas trajanja pa izberemo zato, ker glasovno enoto kasneje lažje skrajšamo brez opaznejših posledic na kakovosti govornega signala, kot podaljšamo krajšo. Zakaj? Kratke enote običajno nimajo jasno določenih ciljnih (osrednjih) delov glasov, kot jih imajo daljše enote. Če te kratke enote podaljšujemo, zvenijo precej nenaravno. Če pa skrajšujemo daljše enote, sicer zaznamo glasovna popačenja (over-articulation), so pa slušni rezultati največkrat bolj sprejemljivi.

Spreminjanje periode osnovnega tona in trajanja posameznih glasov sta edina parametra prozodije, ki ju sistemi za sestavo govora druge generacije upoštevajo. To pa zato, ker sta to lastnosti signala, ki ju je najlažje spreminjati. Druge parametre, ki določajo prozodijo, kot je na primer izdihovanje na koncu stavkov, je težje modelirati, zato jih v sistemih za sestavo govora pogosto zanemarimo.

3.2.1 PSOLA

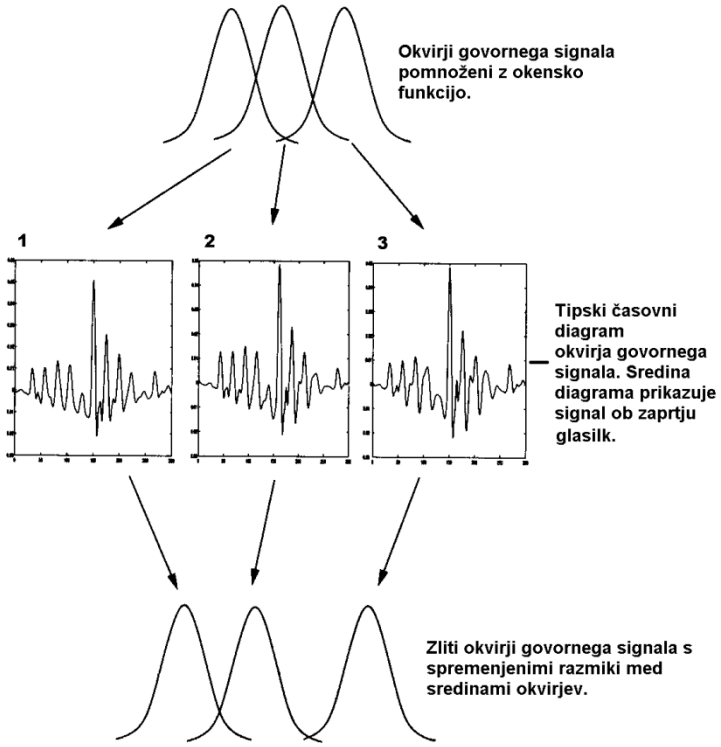
Najbolj znana družina signalno-procesnih metod druge generacije sistemov za sestavo govora je verjetno PSOLA (Pitch-Synchronous OverLap and Add) [9],[23]. Metode iz te družine ne spadajo neposredno med metode za sestavo govora, saj omogočajo le "gladko" lepljenje vnaprej posnetih glasovnih enot govora.

Osnovni princip vseh različic PSOLA algoritma je enak. Posnetek naravnega govora razbijemo na okvirje, ki jih množimo s Hannovim oknom časovno usklajeno z osnovnim tonom. Koeficiente Hannovega okna podaja enačba

$$w(n) = \begin{cases} 0{,}5 - 0{,}5\cos\left(\frac{2\pi n}{N-1}\right), & n = 0{,}1, \dots, N-1 \\ 0, & sicer \end{cases}. \qquad (3.6)$$

Namesto Hannovega okna bi lahko uporabili tudi katero izmed ostalih oken, pri katerih se koeficienti na robovih okna spustijo do vrednosti nič, ker s tem dosežemo najbolj gladko lepljenje. Ima pa Hannovo okno izmed vseh oken, katerih koeficiente lahko izračunamo na podobno preprost način pri prej omenjenem pogoju, ob isti širini glavnega vala amplitudnega odziva najmanjšo vršno amplitudo stranskih valov.

Po množenju s Hannovim oknom okvirje zlijemo skupaj v umetni govorni signal. Periodo osnovnega tona pri PSOLA metodah spreminjamo tako, da pred zlivanjem okvirje postavimo bliže skupaj (skrajšamo periodo), ali pa jih nekoliko razmaknemo (podaljšamo periodo). Trajanje umetno sestavljenega govora pa spreminjamo z brisanjem ali dodajanjem okvirjev. Samo zlivanje, po izvršenih popravkih osnovnega tona in trajanja, nazadnje izvedemo s pomočjo ene izmed metod s seštevanjem prekrivkov. Osnovni prikaz PSOLA metod prikazuje slika 11.



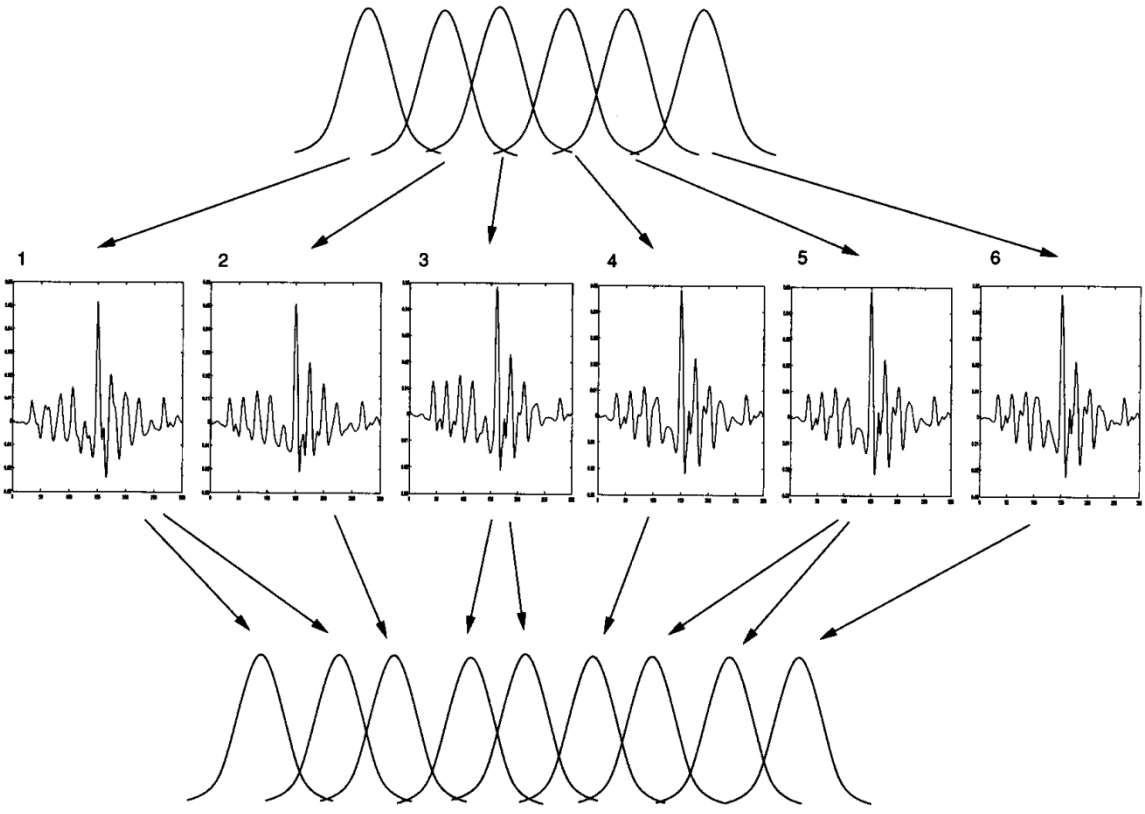
Slika 11: Osnovni prikaz delovanja PSOLA metod.

TD-PSOLA

Je najbolj priljubljena PSOLA metoda in jo lahko celo štejemo kot najbolj priljubljen algoritem za spreminjanje periode osnovnega tona in trajanja govornega signala.

TD-PSOLA deluje usklajeno (sinhrono) s periodo osnovnega tona, kar pomeni, da imamo en analizni okvir na periodo osnovnega tona govora. Predpogoj za uspešno delovanje metode je, da lahko dovolj natančno določimo periode osnovnega tona in pa okvirje vzorcev, kjer se osnovni ton sploh pojavi (ločitev zvenečih od nezvenečih glasov). Sredino analiznega okvirja običajno postavimo tako, da ustreza trenutku, ko so glasilke zaprte. Ta trenutek je mogoče zelo natančno določiti s pomočjo laringografa, metoda pa dobro deluje tudi v drugih položajih glasilk, če jih na enak način upoštevamo v vseh okvirjih. Okvir vzorcev govora množimo s Hannovim oknom, pri čemer je dolžina okna dve periodi osnovnega tona. Pri sestavi govora okvirje postavimo tako, da se njihove sredine ujemajo s ciljnimi periodami osnovnega tona in dele, ki se prekrivajo, preprosto seštejemo. Tako dobljen govorni signal je skoraj neločljiv od naravnega.

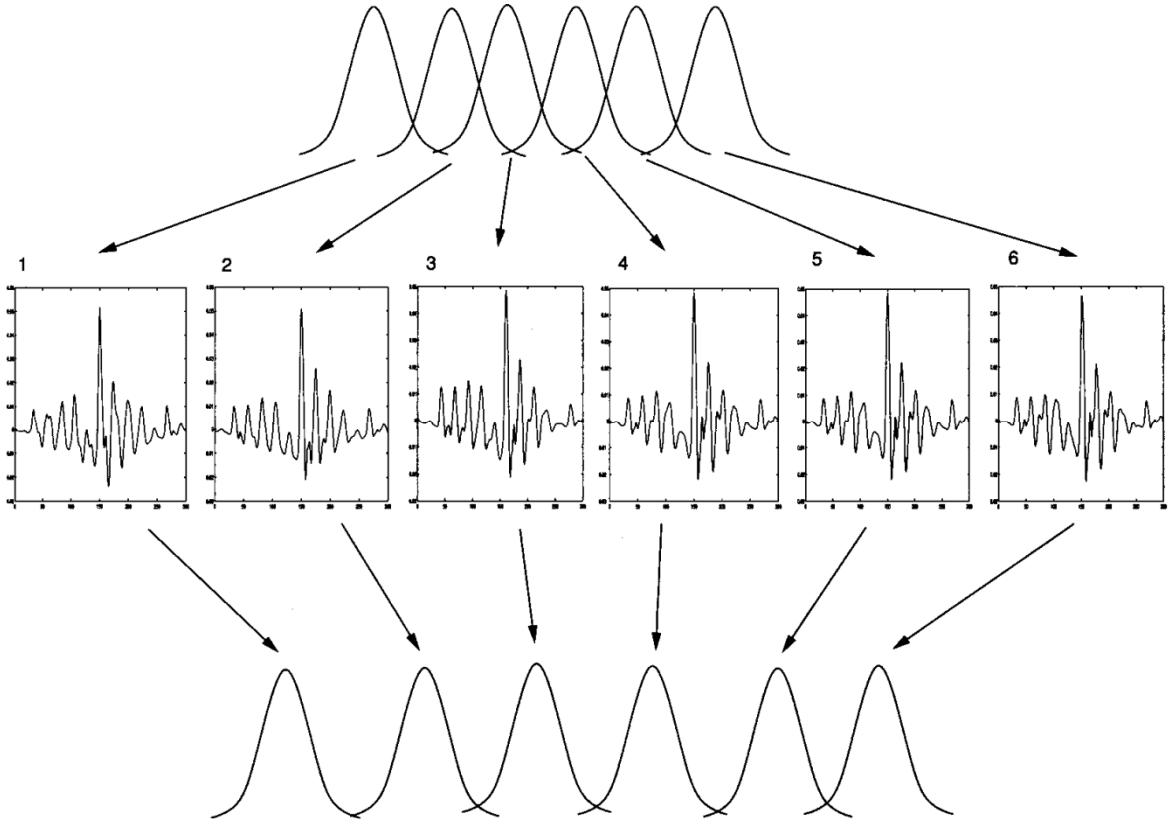
Raztegnitev po časovni osi (sprememba dolžine trajanja govora) dosežemo z izločanjem ali podvajanjem okvirjev (slika 12). Če so spremembe dovolj majhne, jih poslušalec komaj zazna. Izkustveno pravilo pravi, da so spremembe dovolj majhne (komaj zaznavne), če so manjše od velikostnega faktorja 2. Bolj pravilno pa je, če navedemo, da je sprememba tem bolj opazna, čim večja je.



Slika 12: Časovni razteg govornega signala s PSOLA algoritmom.
Slika prikazuje podaljševanje.

Drugi tip spremembe govornega signala, ki jo PSOLA algoritem omogoča, je sprememba periode osnovnega tona. Izvedemo jo, če okvirje sestavimo nekoliko bolj narazen, ali skupaj, kot pri analizi. Postopek si lahko ogledamo na sliki 13.

Poglejmo si delovanje algoritma še bolj podrobno na naslednjem primeru. Vzemimo za primer govorni signal, ki ima povprečno frekvenco osnovnega tona enako 100 Hz. Sredine analiznih okvirjev bodo v tem primeru 10 ms vsaksebi. Na koncu bi radi proizvedli govor z višjo frekvenco osnovnega tona, zato sredine okvirjev postavimo 8 ms narazen. Če okvirje zlepimo skupaj s seštevanjem prekrivkov, dobimo signal s frekvenco osnovnega tona, ki je enaka $100\ Hz * 10/8 = 125\ Hz$. Če po drugi strani okvirje postavimo bolj narazen, dobimo signal z nižjo frekvenco osnovnega tona. Tu lahko pojasnimo zakaj potrebujemo okvirje v dvojni dolžini periode osnovnega tona. S tem zagotovimo, da bomo tudi pri dvakratnem znižanju frekvence osnovnega tona še vedno imeli vsaj minimalno prekrivanje vzorcev na robovih okvirjev.



Slika 13: Postopek spreminjanja periode osnovnega tona s PSOLA algoritmom. Slika prikazuje podaljševanje periode osnovnega tona.

3.2.2 Sinusni modeli

Pri klasični sestavi govora z LP modelom smo videli, da je glavni vzrok sorazmerno slabe kakovosti govornega signala v preveč preprostem izvornem signalu, saj model kot izvor signala uporablja zaporedje impulzov in/ali generator belega šuma. Sinusni modeli skušajo to pomanjkljivost odpraviti. Izvorni signal, ali vsaj njegove periodične komponente, sestavijo kot vsoto sinusov, v splošnem poljubnih amplitud, frekvenc in faz.

Sinusni govorni model tipa izvor-filter

Sinusni govorni model [21],[25] je predstavnik govornih modelov izvor-filter in spada v drugo generacijo sistemov za sestavo govora. Govorni signal dobimo s konvolucijo vzbujevalnega signala $e(t)$ in impulznega odziva filtra govorne cevi $h(t)$ in ga opisuje enačba

$$s(t) = \int_0^t h(t-\tau)e(\tau)d\tau. \tag{3.7}$$

Posebnost sinusnega govornega modela je predstavitev vzbujevalnega signala in impulznega odziva filtra govorne cevi. Namesto klasičnega vzbujevalnega modela tipa zveneč/nezveneč in namesto bolj splošnega več impulznega vzbujevalnega signala kot ga poznamo pri LP modelu, uporabimo vzbujevalni signal sestavljen iz vsote sinusov poljubnih amplitud, frekvenc in faz

$$\begin{aligned} e(t) &= \sum_{l=1}^{L(t)} a_l(t)\cos[\Omega_l(t)], \\ \Omega_l(t) &= V_l(t) + \phi_l, \\ V_l(t) &= \int_{t_l}^{t} \omega_l(\sigma)d\sigma, \end{aligned} \tag{3.8}$$

kjer je $L(t)$ je število sinusnih komponent v času t, t_l pa predstavlja čas pojavitve l-te sinusne komponente. $a_l(t)$ in $\omega_l(t)$ predstavljata časovno odvisno amplitudo in frekvenco l-te sinusne komponente, $V_l(t)$ pa je časovno odvisni fazni prispevek vzbujevalnega signala. ϕ_l predstavlja konstantni fazni zamik l-te sinusne komponente, kar pomeni, da sinusne komponente niso fazno usklajene.

Impulzni odziv govorne cevi si poglejmo s pomočjo Fourierjeve transformacije

$$H(\omega, t) = M(\omega, t)e^{j\Phi(\omega,t)}, \tag{3.9}$$

kjer $M(\omega, t)$ označuje amplitudo, $\Phi(\omega, t)$ pa fazo. Naj

$$M_l(t) = M[\omega_l(t), t],$$
$$\Phi_l(t) = \Phi[\omega_l(t), t] \tag{3.10}$$

označujeta amplitudo in fazo impulznega odziva govorne cevi v odvisnosti od posamezne frekvenčne komponente $\omega_l(t)$. Če nadalje predpostavimo, da so vzbujevalni parametri v (3.8) nespremenljivi med trajanjem impulznega odziva filtra govorne cevi, dobimo po izračunu konvolucije (3.7) sinusno predstavitev govornega modela

$$s(t) = \sum_{l=1}^{L(t)} A_l(t)\cos[\theta_l(t)]\,, \tag{3.11}$$

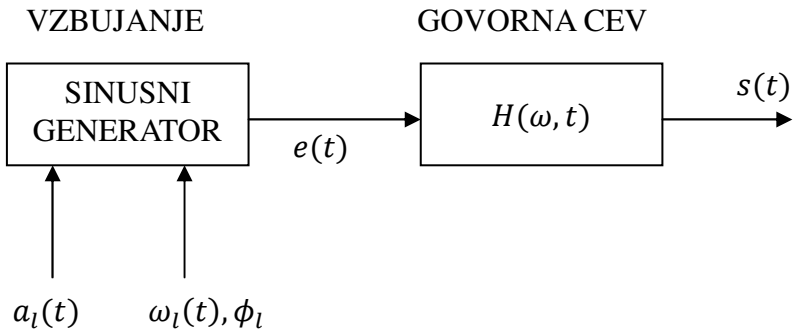
kjer

$$A_l(t) = a_l(t)M_l(t),$$
$$\theta_l(t) = \Omega_l(t) + \Phi_l(t) = V_l(t) + \phi_l + \Phi_l(t) \tag{3.12}$$

predstavljata amplitudo in fazo *l*-te sinusne komponente po frekvenčni stezi $\omega_l(t)$. Konvolucijo smo izračunali kot obratno Fourierjevo transformacijo zmnožka Fourierjevih transformacij vzbujevalnega signala (3.8) in impulznega odziva govorne cevi, ki je v (3.9) in (3.10) že zapisan v obliki Fourierjeve transformacije. Na izhodu modela dobimo vsoto sinusnih komponent, kjer se amplituda posamezne sinusne komponente vzbujevalnega signala $a_l(t)$ množi z amplitudnim odzivom filtra govorne cevi $M_l(t)$, faza pa se spremeni za fazni odziv filtra govorne cevi $\Phi_l(t)$.

Amplitude $A_l(t)$ in faze $\theta_l(t)$ v predstavitvi sinusnega govornega modela (3.11) lahko ocenimo na primer s pomočjo kratko-časovne Fourierjeve transformacije. Mnogo večji problem predstavlja ločitev tako dobljenih amplitud in faz na prispevke vzbujevalnega signala (3.8) in na učinke filtra, ki ponazarja značilnosti govorne cevi (3.10). Postopek ločitve je opisan v [25], ločeno predstavitev amplitud, frekvenc in faz za vzbujevalni signal in za filter govorne cevi pa potrebujemo, če želimo spreminjati periodo osnovnega tona oziroma čas trajanja govornega signala. Če želimo na primer simulirati učinek hitrejšega govorjenja, moramo pri filtru govorne cevi hitreje spreminjati amplitudni $M(\omega, t)$ in fazni odziv $\Phi(\omega, t)$, pri vzbujevalnem signalu pa ustrezno skrajšati potek frekvenčnih stez $\omega_l(t)$ in pohitriti spreminjanje amplitud $a_l(t)$, da ustrezajo novemu časovnemu intervalu.

Delovanje sinusnega govornega modela tipa izvor-filter prikazuje slika 14.

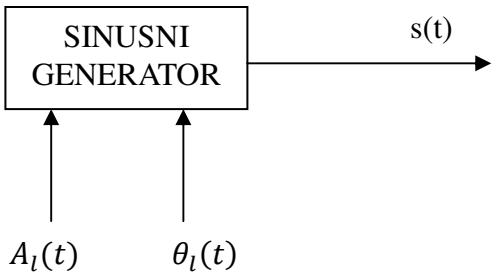


Slika 14: Sestava govornega signala s pomočjo sinusnega modela tipa izvor-filter.

Čisti sinusni govorni model

V sinusnem govornem modelu tipa izvor-filter smo vzbujevalni signal predstavili kot vsoto sinusov poljubnih amplitud, frekvenc in faz. Sestavljene amplitude in faze frekvenčnih komponent v (3.11) smo v tem modelu razstavili na prispevke vzbujevalnega signala in na učinke filtra, ki ponazarja značilnosti govorne cevi. Če ta korak izpustimo, dobimo nov preprostejši tip modela, ki ga imenujemo čisti sinusni govorni model. Model ne ločuje več vzbujevalnega signala in vplivov filtra govorne cevi, zato ni več tipa izvor-filter. Govorni signal tvorimo ravno tako s (3.11), le da amplitud $A_l(t)$ ter frekvenc in faz, zajetih v $\theta_l(t)$ in dobljenih z analizo naravnega govornega signala, ne ločimo na del, ki pripada vzbujevalnemu signalu in na del, ki je prispevek filtra govorne cevi. Model lahko uporabljamo pri sestavi kot dober pripomoček za lepljenje odsekov govora. Govorni signal spada namreč v skupino signalov (končna energija v poljubnem končnem intervalu), ki jih lahko bolj ali manj natančno razstavimo na vsoto sinusnih komponent in s tem predstavimo v frekvenčnem prostoru. Glajenje nezveznosti med različnimi odseki govora je v frekvenčnem prostoru lažje.

Delovanje čistega sinusnega modela nam podaja slika 15, bolj podrobno pa smo sestavo govornega signala s tem modelom opisali v podpoglavju 4.3.2



Slika 15: Sestava govornega signala s pomočjo čistega sinusnega modela.

HNM model

HNM (Harmonic plus Noise Model) model je predstavnik sinusnih govornih modelov, ki ga je avtor [31] razvil prav za potrebe sestave govora. Z njim je skušal odpraviti eno izmed glavnih pomanjkljivosti sinusnih modelov in to je velika količina značilk (sinusnih komponent), ki so potrebne za krmiljenje sinusnega generatorja. Veliko število sinusnih komponent potrebujemo predvsem za modeliranje delov govornega signala, ki imajo značilnosti šuma. Iz šuma so sestavljeni nezveneči glasovi, prisoten pa je tudi v zgornjem delu spektra pri zvenečih glasovih. Frekvenčni spekter govornega signala po tem modelu razstavimo na harmonični del, ki ga predstavimo kot vsoto osnovnega tona in njegovih harmonskih komponent do vključno neke maksimalne zgornje harmonske komponente. Preostali del spektra model obravnava kot šumni preostanek in ga modelira z generatorjem šuma. Potrebne vrednosti amplitud, frekvenc in faz za sestavo umetnega govora dobimo z analizo naravnega govora kot pri vseh sinusnih modelih. HNM model uporablja kompleksne amplitude frekvenčnih komponent govornega signala, ki vsebujejo tudi fazno informacijo in jih izračunamo z

$$C_k = \frac{\sum_{n=n_a{}^{i-N_0}}^{n_a{}^{i+N_0}} w^2(n)s(n)e^{-j2\pi k f_0 n}}{\sum_{n=n_a{}^{i-\mathrm{N}_0}}^{n_a{}^{i+N_0}} w^2(n)}, \tag{3.13}$$

kjer so $w(n)$ koeficienti Hammingovega okna, $s(n)$ so vzorci naravnega govora, f_0 je osnovna frekvenca, N_0 je perioda osnovne frekvence v vzorcih, $n_a{}^i$ pa označuje sredinski vzorec i-tega analiznega okvirja. Analiza poteka sinhrono s periodo osnovne frekvence. Zaradi tega je pri HNM modelu prvi korak pri analizi ravno določitev osnovne frekvence f_0. Ta je za kakovost signala zelo pomembna, saj se napaka pri višjih harmonskih komponentah množi. Avtor HNM modela uporablja algoritem opisan v [8], namesto njega pa lahko uporabimo tudi kakšen drug algoritem za določitev osnovne frekvence. Začetni približek za f_0 izboljšamo tako, da za $\hat{f_0}$ uporabimo vrednost, ki minimizira napako

$$E(\hat{f_0}) = \sum_{i=1}^{L_n} | f_i - i\hat{f_0} |^2, \tag{3.14}$$

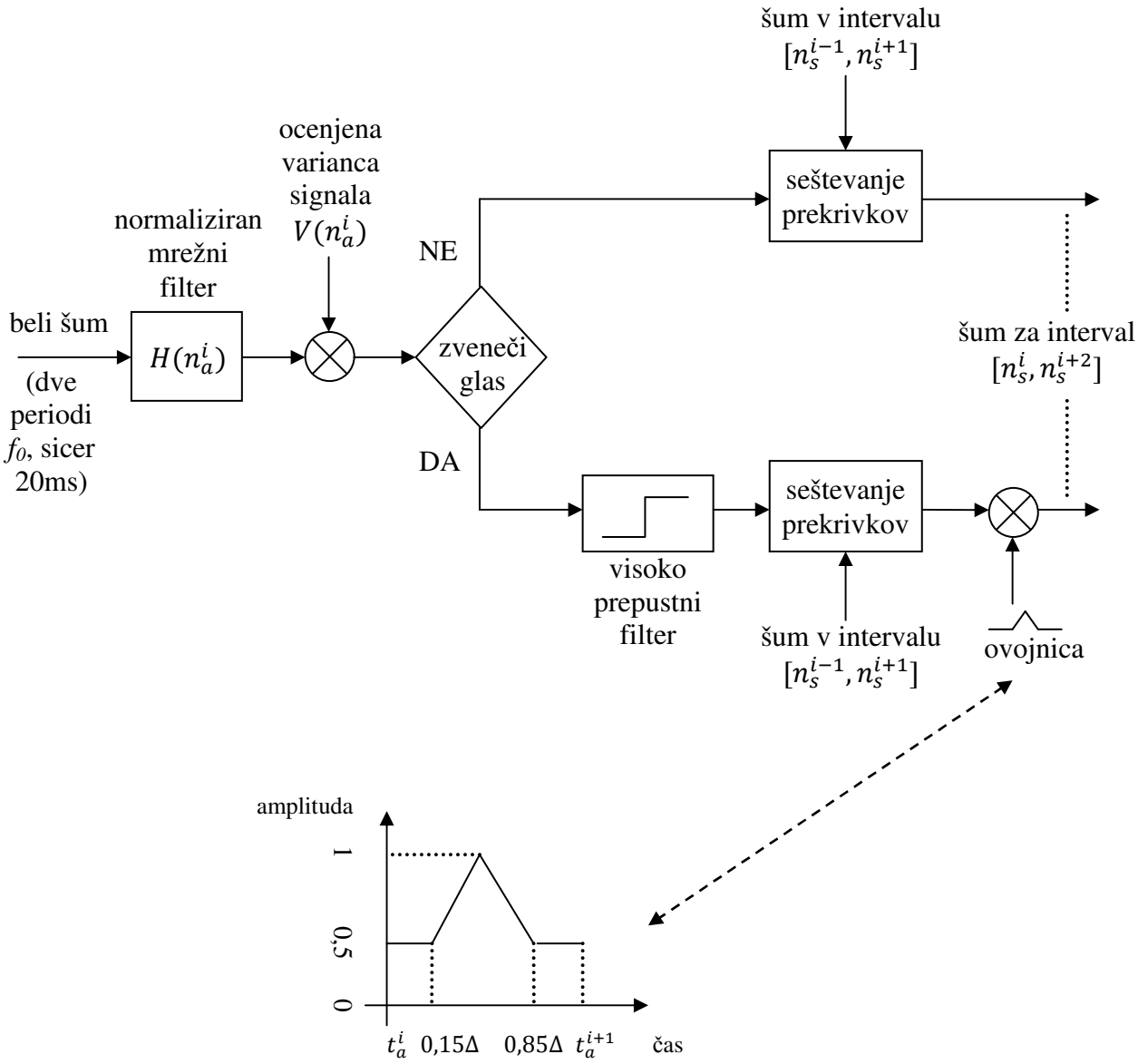
kjer L_n število harmonskih komponent osnovne frekvence. Frekvenca f_i predstavlja najvišji vrh amplitudnega spektra v i-tem intervalu podanem z $[f_c + \hat{f_0}(2\mathrm{i}-1)/2, f_c + \hat{f_0}(2\mathrm{i}+1)/2]$. V poštev pridejo vse frekvence f_i, ki zadostijo dodatnim "harmonskim" kriterijem podanim v [31]. Frekvenca f_c predstavlja približek za $i\hat{f_0}$ in velja $f_c = f_{i-1}$.

Šumni del govornega spektra modeliramo s pomočjo mrežnega filtra p-tega reda, ki je sestavljen le iz polov v z-ravnini.

Pri sestavi signala harmonični del govornega signala izračunamo kot vsoto sinusov po enačbi

$$s(t) = \sum_{k=0}^{L_n} C_k e^{jk\omega t}, \tag{3.15}$$

ki ji prištejemo še šumni del signala. Tvorbo šumnega dela signala prikazuje slika 16.



Slika 16: Blokovni prikaz tvorbe šumnega signala pri HNM modelu.

Osnovni vir šumnega dela signala je generator Gaussovega belega šuma z enotsko varianco.

Signal iz generatorja potuje skozi mrežni filter p-tega reda, katerega koeficiente izračunamo v postopku analize naravnega govornega signala, nato ga pa še množimo z varianco, ki smo jo tudi izračunali pri analizi naravnega govornega signala. Za vsak okvir vzorcev končnega umetnega govornega signala tvorimo dva okvirja vzorcev šuma. Z $n_s{}^i$ je označen sredinski vzorec *i*-tega sestavnega okvirja. Pri zvenečih glasovih je okvir dolg eno periodo osnovnega tona, pri nezvenečih pa je dolžina postavljena na 10 ms. Vzorce v delih okvirjev šumnega signala, ki se v tem postopku med seboj prekrivajo, seštejemo. Pri zvenečih glasovih signal pred seštevanjem še dodatno filtriramo z visoko prepustnim filtrom, ki iz šumnega signala odstrani vse frekvence do meje, ki ustreza frekvenci najvišje harmonske komponente iz harmoničnega dela govornega signala. Šum v tej veji postopka še dodatno oblikujemo s poenostavljeno ovojnico govornega signala trikotne oblike, ki je prikazana v grafu na spodnjem delu slike 16. S $t_a{}^i$ je na grafu označen čas sredinskega vzorca $n_a{}^i$ *i*-tega analiznega okvirja, časovna razlika do sredinskega vzorca *i+1*-ega analiznega okvirja pa je označena z $\Delta = t_a^{i+1} - t_a^i$.

3.3 Tretja generacija sistemov za sestavo govora

Tretjo generacijo sistemov za sestavo govora označujeta dva pristopa.

- Procese, ki so prisotni pri sestavi govora, modeliramo s pomočjo statističnih metod, v glavnem s pomočjo prikritih modelov Markova. Preko teh modelov se sestava govora navezuje na proces razpoznave govora. Pri sestavi poskušamo izkoristiti statistične informacije o govornem signalu, ki jih zberemo v procesu učenja razpoznavalnika.

- Govorni signal sestavljamo s pomočjo izbire glasovnih enot. Ta pristop nadgrajuje pristop druge generacije, ki temelji na sestavi s pomočjo lepljenja glasovnih enot.

V nadaljevanju si bomo pogledali samo metode sestave govora s pomočjo izbire glasovnih enot (Unit Selection Synthesis), ki prevladujejo med metodami tretje generacije sistemov za sestavo govora in danes tvorijo tudi najkakovostnejši umetni govorni signal.

Metode tretje generacije sistemov za sestavo govora uporabljajo signalno procesne algoritme druge generacije (npr. TD-PSOLA ali algoritmi sinusnih modelov) le za gladko lepljenje pri sestavi, ne pa tudi za spreminjanje prozodičnih lastnosti govora. Prozodične lastnosti skušajo zajeti z velikim številom posnetkov glasovnih enot, izmed katerih potem v procesu sestave izberejo tako zaporedje glasovnih enot, ki najbolj ustreza specifikacijam želenega govornega signala.

3.3.1 Sestava z izbiro glasovnih enot

Sistemi druge generacije, ki so uporabljali metode lepljenja glasovnih enot (prevladujoča enota dvoglasnik), so temeljili na dveh predpostavkah:

1. Vse variacije posamezne glasovne sestavne enote, ki jo uporabljamo v procesu lepljenja, lahko obvladamo samo s spremembo trajanja in s spremembo periode osnovnega tona enote.

2. Z različnimi signalno-procesnimi algoritmi lahko izvedemo vse potrebne spremembe periode osnovnega tona in časa trajanja posamezne sestavne enote, ne da bi pri tem poslabšali naravnost govornega signala.

Ti dve predpostavki sta se uveljavili predvsem zaradi inženirskih razlogov, saj ju je izmed vseh značilnosti govornega signala, ki vplivajo na naravnost, najlaže določiti in spreminjati, hkrati pa predstavljata glavno omejitev za doseganje višje kakovosti umetnega govornega signala.

Glavna ideja sestave s pomočjo izbire glasovnih enot je v tem, da skušamo v posnetih glasovnih enotah zajeti čim več informacij o spremenljivosti govornega signala, ki vključuje tudi naravno prozodijo govornega signala in se manj naslanjati na signalno-procesne metode. Namesto, da v zbirki hranimo samo en primerek dvoglasnika ali kakšne druge glasovne enote, ki mora zadostovati za sestavo vseh primerov govornega signala, ne glede na poudarke in prozodične lastnosti govora, damo v zbirko več primerkov glasovnih enot. Z več primerki bolje pokrijemo variacije naravnega govora in zato je potrebno glasovne enote v procesu sestave tudi manj preoblikovati. V procesu sestave skušamo s pomočjo ustreznega algoritma med več posnetki posamezne glasovne enote izbrati tistega, ki tako dopolni zaporedje glasovnih enot, da je ujemanje med posameznimi enotami zaporedja najboljše. Kakovost ujemanja pa seveda ocenjujemo z izbrano kriterijsko funkcijo. Ker imamo pri sestavi govora z izbiro glasovnih enot v zbirki podatkov zajetih več informacij o prozodiji, kot pri sestavi s pomočjo lepljenja, se morajo razlikovati tudi opisi enot v zbirki podatkov. Tipičen opis določil dvoglasnika "na" v zbirki za sestavo s pomočjo lepljenja glasovnih enot je podan z

$$d_t = \begin{bmatrix} S_1 \begin{bmatrix} glas & n \\ F0 & 122 \\ trajanje & 45 \end{bmatrix} \\ S_2 \begin{bmatrix} glas & a \\ F0 & 125 \\ trajanje & 75 \end{bmatrix} \end{bmatrix}, \tag{3.16}$$

v zbirki za sestavo s pomočjo izbire glasovne enote pa z

$$d_t = \begin{bmatrix} S_1 \begin{bmatrix} glas & n \\ F0 & 122 \\ trajanje & 45 \\ naglašen & da \\ konec\ izraza & ne \end{bmatrix} \\ S_2 \begin{bmatrix} glas & a \\ F0 & 125 \\ trajanje & 75 \\ naglašen & da \\ konec\ izraza & ne \end{bmatrix} \end{bmatrix}. \tag{3.17}$$

Z dodatki v zapisu želimo pokriti večji del prozodije, kot to lahko storimo, če spreminjamo le hitrost izgovorjave in periodo osnovnega tona. V primeru podanem z enačbo (3.17) sta tako tudi podatka o tem, ali je dvoglasnik del naglašenega zloga ali ne in ali je bil dvoglasnik del končnega zloga v izrazu oziroma stavku.

Sistemi z izbiro enot lahko uporabljajo zelo različne glasovne enote. Med najbolj znanimi tipi glasovnih enot so: polglasovi, glasovi, dvoglasniki, triglasniki, polzlogi, zlogi, dvozlogi, besede, stavki, ali tudi okvirji vzorcev govornega signala konstantne dolžine.

Algoritem, ki danes predstavlja osnovo standardne sestave s pomočjo izbiranja glasovnih enot, sta avtorja Hunt in Black prvič opisala v [16]. Algoritem definira postopek izbire enot kot iskanje skozi vsa mogoča zaporedja enot in skuša najti "najboljše" zaporedje, ki je določeno z najmanjšo ceno. Splošna funkcija za določanje cene je podana kot

$$C(U,S) = \sum_{t=1}^{N_B} T(u_t, s_t) + \sum_{t=1}^{N_B-1} J(u_t, u_{t+1}) \tag{3.18}$$

in je sestavljena iz cene zadetka $T(u_t, s_t)$, ki predstavlja razdaljo med naborom določil s_t in vrednostjo enote u_t v zbirki (kako dobro enota ustreza določilom), ter cene lepljenja $J(u_t, u_{t+1})$, ki podaja ceno (napako) lepljenja. N_B predstavlja število enot v zbirki. Nizka vrednost cene pomeni, da enoti sestavljata (slušno) dober spoj. Cilj iskanja je najti tako zaporedje enot v zbirki, ki nam da najmanjšo možno ceno oziroma minimalno vrednost funkcije

$$\widehat{U} = \min_{u} \left\{ \sum_{t=1}^{N_B} T(u_t, s_t) + \sum_{t=1}^{N_B-1} J(u_t, u_{t+1}) \right\}. \tag{3.19}$$

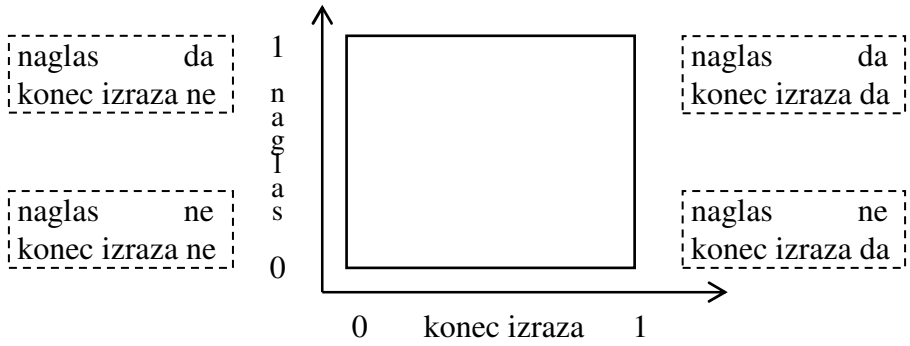
Funkcija cene zadetka mora podati ceno tudi v primeru, če v zbirki ni enote, ki bi natančno

ustrezala podanim določilom, tako recimo tudi če posnetka določenega dvoglasnika sploh ni v zbirki, ali pa obstaja, ampak z drugačno vrednostjo naglaševanja.

Pri natančnejšem določanju funkcije cene zadetka sta se uveljavila dva različna pristopa in sicer:

- izračun cene pri predpostavki o neodvisnosti značilk in
- izračun cene kot razdalje v akustičnem prostoru.

Po prvem pristopu predstavlja vsaka kombinacija značilk eno oglišče hiperkocke, razdalja med ogliščem, kateremu ustreza nabor določil, in med najbližjim ogliščem, v katerem imamo enoto iz zbirke, pa predstavlja ceno zadetka. Če nabor določil ustreza kakemu oglišču z enoto iz zbirke, je cena zadetka enaka 0. Na sliki 17 lahko vidimo primer dvorazsežnega prostora značilk, ki ga uporabimo pri računanju cene zadetka. Cena zadetka je seveda odvisna tudi od metrike, ki jo uporabimo pri računanju.



Slika 17: Prikaz prostora značilk za izračun funkcije cene zadetka razsežnosti 2.

Poleg Evklidske metrike, ki jo podaja enačba

$$D = \sqrt{\sum_i (x_i - y_i)^2}, \tag{3.21}$$

lahko uporabimo tudi metriko Manhattan, ki je podana z enačbo

$$D = \sum_i abs(x_i - y_i), \tag{3.22}$$

ali pa tudi Mahalanobisovo metriko, ki razdaljo med posameznimi komponentami dodatno uteži z obratno vrednostjo standardnega odstopanja komponente x_i v vzorcu in jo podaja enačba

$$D = \sqrt{\sum_i \left(\frac{x_i - y_i}{\sigma_i}\right)^2}. \tag{3.23}$$

Poglejmo si primer kako izračunamo ceno zadetka v primeru, ko iščemo glasovno enoto, ki je naglašena in sestavlja konec izraza, v bazi pa imamo samo glasovno enoto, ki je sicer sestavljena iz enakih glasov, ni pa naglašena in smo jo pri analizi izrezali iz zaporedja glasov na sredini, ali pa na začetku izraza. Prostor značilk v bazi naj bo enak kot na sliki 17. Izračunati moramo razdaljo med točko (0, 0), ki je v bazi in točko (1,1), ki jo iščemo. Če uporabimo metriko Manhattan je razdalja in s tem cena enaka 2, če pa uporabimo Evklidsko metriko, je razdalja in s tem cena enaka $\sqrt{2} = 1{,}41$.

V splošnem lahko posamezno razsežnost hiperkocke dodatno obtežimo in ceno zadetka izračunamo kot

$$T(u_i, s_i) = \sum_{q=1}^{Q} w_q \left[T_q\left(u_i(q), s_i(q)\right)\right], \tag{3.24}$$

kjer je:
s_i – i-ti nabor določil s Q različnimi značilnimi lastnostmi,
u_i – enota v zbirki s Q različnimi značilnimi lastnostmi,
$s_i(q)$ – vrednost značilne lastnosti q v naboru določil s_i,
$u_i(q)$ – vrednost značilne lastnosti q enote v zbirki u_i,
$T_q(x,y)$ – funkcija, ki vrne razdaljo med x in y za značilno lastnost q,
$T(u_i,s_i)$ – skupna vrednost cene med enoto v zbirki in naborom določil,
w_q – utež za funkcijo T_q.

Za učenje uteži lahko uporabimo odločitvena drevesa, nevronske mreže, prikrite modele Markova, ali kakšno drugo metodo avtomatskega učenja [22].

Po drugem pristopu izračunamo ceno zadetka kot razdaljo v akustičnem prostoru. Iz obstoječih glasovnih enot skušamo sestaviti porazdelitveno funkcijo v akustičnem prostoru (npr. prostor kepstralnih [9] koeficientov) v odvisnosti od značilnih lastnosti (npr. naglašen/nenaglašen). Sedaj se lahko zgodi primer, ko dve precej različni kombinaciji

[9] Kepstrum – (angl. Cepstrum) rezultat Fourierjeve transformacije logaritma spektra signala.

značilnih lastnosti ležita zelo blizu skupaj v akustičnem prostoru, kar po prvem pristopu ni bilo mogoče. Ko računamo ceno zadetka po drugem pristopu, s pomočjo porazdelitvene funkcije najprej postavimo nabor določil v neko predvideno točko v akustičnem prostoru, nato pa računamo ceno zadetka kot Evklidsko razdaljo do najbližjih točk v akustičnem prostoru, ki jih predstavljajo enote iz zbirke.

Poleg cene zadetka moramo pri računanju cene sestave govornega signala upoštevati in izračunati tudi ceno lepljenja glasovnih enot. Zanimiva ugotovitev pri lepljenju glasovnih enot iz [35] je ta, da je spoj med enotami lahko zelo dobro viden recimo na spektrogramu, na slušnih testih pa ga poslušalci sploh ne zaznajo. Rezultati raziskav [16] so pokazali, da samo velikost razlike med zadnjim vektorjem značilk glasovne enote *i* in prvim vektorjem značilk glasovne enote *i+1*, ni dovolj zanesljivo merilo o slušni kakovosti spoja. Ta rezultat je precej neodvisen tako od vrste značilk kot tudi od tipa metrike (Manhattan, Evklidska, Mahalanobisova) [32].
Funkcija za računanje cene spajanja mora upoštevati tudi opažanja, da se nekateri glasovi bolj neopazno lepijo kot drugi in temu primerno obtežiti ceno lepljenja. Ista vrednost akustične razlike je namreč pri lepljenju določenih glasov lahko neslišna, medtem ko je pri lepljenju drugih lahko zelo moteča.

Algoritem za računanje optimalne cene v splošnem zahteva računanje cene za vse mogoče kombinacije enot iz zbirke, kar je že pri majhnem številu enot neizvedljivo. Zato se pri iskanju uporablja algoritme za bolj usmerjeno iskanje, med drugimi pogosto tudi znan Viterbijev algoritem [17].

4 NGS ali Nauči se Govoriti Sam

Že v uvodu smo postavili merila, ki jim mora zadostiti kakovosten sestavljalnik umetnega govornega signala. Tvoriti mora govorni signal visoke razumljivosti in predvsem tudi visoke stopnje naravnosti, saj ga sicer uporabniki nočejo uporabljati. Poleg tega mora omogočati tvorbo govornega signala za neomejen nabor besed in za poljuben jezik. Hkrati je zaželeno, da so potrebne računalniške zmogljivosti, ki jih sestavljalnik potrebuje za svoje delovanje in jih izrazimo kot potrebno procesorsko moč in potrebno količino pomnilnika, dovolj majhne, da sestavljalnik deluje kot aplikacija na povprečnem osebnem računalniku. Te zahteve, ki določajo tudi merila, narekujejo uporabniki. Če pogledamo še na merila, ki jih želimo kot snovalci izpolniti pri snovanju takega sistema, se lahko omejimo kar na stroške izdelave, ki morajo biti čim nižji. To merilo zajema tako potrebno količino človeškega dela različne zahtevnosti, kot tudi vrsto in količino opreme, ki je potrebna pri izdelavi in kasneje pri uporabi sistema.

Sistem, ki bi zadovoljil vse navedene zahteve in izpolnil vsa merila, seveda še zdaleč ne obstaja, med drugim tudi zato, ker si določena merila in zahteve med seboj nasprotujejo. Prav zato moramo merila in zahteve ustrezno obtežiti in poudariti tiste, katerih izpolnitev je v danih okoliščinah najpomembnejša.

Cilj raziskav doktorske disertacije je predvsem proučitev tehničnih možnosti za izdelavo takega sistema, ki bi ga lahko "naučil govoriti" skoraj vsak uporabnik kar sam in sicer tako, da bi v mikrofon prebral razumno količino vnaprej pripravljenih besedil, vse ostalo pa bi opravil sistem sam. Tako naučen sistem bi moral seveda proizvajati umetni govorni signal visoke stopnje naravnosti ter razumljivosti in to za poljubno besedilo. Glede na cilj, da bi se "naučil govoriti sam", sta s tem izpolnjena tudi cilja o poljubnem jeziku ter o minimalni potrebni količini vloženega dela. Danes namreč še ne obstaja sistem za sestavo govora, ki ne bi zahteval precejšnje količine strokovnega človeškega dela in sicer predvsem pri izgradnji ustrezne zbirke govornih enot iz katerih lahko potem sistem tvori govorni signal iz poljubnega besedila.

V tem poglavju je opisana izvedba sistema Nauči se Govoriti Sam (NGS), ki je glavni dosežek doktorske disertacije.

4.1 Izbira metode za sestavo govora

Eno prvih vprašanj pri izgradnji želenega sistema za sestavo govora je, ali je katera izmed metod prve, druge ali tretje generacije, ki so opisane v poglavju 3, še posebej primerna za to, da bi predstavljala osnovo za nadaljnje raziskave. Večina metod sloni na zbirki posnetkov naravnega govornega signala, bodisi v surovi obliki ali v obliki značilk. Izjema so metode, ki uporabljajo eksplicitni model za tvorbo vzorcev govora. V to skupino spadata formantna sestava govora, ki spada v prvo generacijo sistemov, in sistemi, ki tvorijo govorni signal s pomočjo prikritih modelov Markova (HMM), ki pa spadajo že v tretjo generacijo sistemov.

Posebno slednji so že več kot desetletje v glavnem toku raziskav na področju sestave govora. So posebej primerni za uporabo v aplikacijah, kjer želimo sistem naučiti govoriti samo iz vzorcev besedil in pripadajočih posnetkov naravnega govora, saj pridobivajo značilke modela samodejno in neposredno iz posnetkov govora s pomočjo statističnih metod učenja. Sistemi so zelo privlačni tudi zaradi skromnih pomnilniških zahtev, pri tvorbi novega glasu pa je potrebno spremeniti le značilke modela. Metode, ki tvorijo govorni signal s pomočjo lepljenja, pa po drugi strani zahtevajo precej prostora v pomnilniku za hranjenje posnetkov. Pri tvorbi novega glasu moramo zato zamenjati cel nabor posnetkov.

Glavno pomanjkljivost metod, ki uporabljajo eksplicitni model za tvorbo vzorcev govora, predstavlja premajhna stopnja naravnosti govornega signala, ki ga tvorijo. Vse dosedanje raziskave kažejo na to, da je modeliranje vseh podrobnosti govorne cevi, ki so potrebne, da dosežemo visoko stopnjo naravnosti, zelo težak problem. Noben obstoječ sistem ne dosega take stopnje naravnosti govora, kot jo sistemi, ki tvorijo govorni signal z lepljenjem glasovnih enot. Osrednji del glavnega toka raziskav je zato danes usmerjen v metode, ki v taki ali drugačni obliki uporabljajo značilke naravnega govornega signala ali kar neposredno vzorce krajših odsekov naravnega govornega signala.

Izbira modela, ki tvori govorni signal s pomočjo prikritih modelov Markova, bi bila po eni strani dobra izbira, saj ravno na področju statističnih metod učenja HMM metode dosegajo najboljše rezultate. Pomnilniške zahteve so nizke, procesorske pa primerljive z drugimi metodami sestave govora, ki obdelujejo značilke govornega signala. Toda na koncu je pri izbiri naše metode prevladala zahteva po večji kakovosti (predvsem naravnosti) govornega signala. Osnovo sistema NGS tako predstavlja metoda tretje generacije sistemov za sestavo govora, ki temelji na izbiri enot.

4.2 Določitev glasovne enote

Glasovna enota, ki bi bila primerna za izgradnjo zbirke posnetkov NGS sistema, mora nujno zadostiti vsaj dvema meriloma:

- v povprečju mora biti sestavljena iz dovolj majhnega števila glasov, da je število različnih realnih kombinacij, ki se pojavijo v živem jeziku, obvladljivo, se pravi, da lahko zbirko vseh potrebnih enot shranimo vsaj na diskovnem pomnilniku povprečnega računalnika;

- zbirko enot je mogoče pridobiti iz posnetkov naravnega govora s pomočjo algoritma za samodejni razrez.

Dvoglasniki, kot enota lepljenja, ki se najbolj pogosto uporablja v sistemih za sestavo govora, niso najbolj primerni gradniki zbirke glasovnih enot NGS sistema, predvsem zaradi tega, ker ne obstaja algoritem za dovolj zanesljiv samodejni razrez oziroma samodejno označevanje dvoglasnikov. Položaj določenih vrst glasov v dvoglasnikih je zelo težko določiti in bo ustrezen algoritem zaradi tega razloga tudi zelo težavno razviti.

Pristopi, ki jih danes večinoma srečujemo pri samodejnem označevanju vzorcev oziroma samodejnem razrezu govornega signala, uporabljajo že naučene razpoznavalnike govora [18], ali pa že izdelane sestavljalnike govora. Oba pristopa povzročata težave pri izdelavi NGS sistema, ker mora imeti snovalec na voljo dovolj kakovosten razpoznavalnik ali sestavljalnik. Zaradi tega smo želeli v sistemu NGS uporabiti algoritem, ki bi razrez oziroma označevanje vršil samo na podlagi spektralnih značilnosti govornega signala in ne bi uporabljal že naučenih razpoznavalnikov ali sestavljalnikov govora.

Natančna in zanesljiva določitev položaja poljubnega glasu v govornem signalu je v splošnem izredno težak problem, ki mogoče sploh ni rešljiv. Najboljši razpoznavalniki govora na nivoju glasov namreč dosegajo 75% natančnost razpoznave [5]. To ni nič presenetljivega, saj so določeni glasovi, kot so na primer zaporniki (p,b,t,d,k,g), posledica kratkotrajnih prehodnih pojavov v govorni cevi, ki jim je težko pripisati enolično določljive spektralne značilnost, saj se le te spreminjajo tudi zaradi vplivov sosednjih glasov. Lažje rešljiv problem je zanesljivo določiti položaj skupin glasov, ki jih imenujmo kar značilni glasovi, saj imajo take spektralne značilnosti, da se dovolj razlikujejo od ostalih glasov. Pri takih glasovih je oblika govorne cevi stabilna dalj časa in niso posledica prehodnih pojavov. Izkaže se, da so med soglasniki slovenskega jezika dovolj spektralno značilni nezveneči soglasniki **s**, **š**, **c**, **č** ter zveneča **z** in **ž**. Spekter ali vsaj del spektra navedenih soglasnikov ima obliko šuma in je pomaknjen proti višjim frekvencam, glasovi pa v povprečju trajajo tudi dovolj časa, da jih sorazmerno preprosto določimo iz spektrograma. Prav tako so spektralno sorazmerno preprosto ločljivi vsi samoglasniki skupaj s polglasnikom **ə**. Tudi skupna energija zvočnega signala je pri samoglasnikih običajno višja od ostalih glasov. Vse navedene glasove (samoglasniki, polglasnik ə in soglasniki c, č, s, š, z in ž) označimo torej s skupno oznako **značilni glasovi**. Ker imamo v sistemu NGS kot vhodni podatek že na razpolago glasovni zapis govornega signala, moramo z dovolj visoko natančnostjo samo določiti, ali določen del zaporedja vzorcev predstavlja samoglasnik, polglasnik ali enega izmed soglasnikov iz prej omenjene skupine. Iz zaporedja glasov v vhodnem besedilu pa potem lahko določimo katere glasove dejansko predstavljajo vzorci signala.

Poglejmo si postopek prirejanja še na primeru. Recimo, da vhodno besedilo predstavlja zapis "čezcesto". Algoritem za določanje značilnih glasov nam vrne naslednje zaporedje **zs-sa-zs-zs-sa-zs-sa** (zs - značilni soglasnik, sa – samoglasnik) skupaj s časi pojavitve posameznega značilnega glasu. Iz teh podatkov lahko vzorce od začetka posnetka do pojavitve prvega značilnega glasu priredimo glasu **č**, vzorce do pojavitve naslednjega značilnega glasu priredimo glasu **e** itn. Na koncu imamo vzorce govornega signala razdeljene na zaporedje glasov **č-e-z-c-e-st-o**.

Iz navedenih spoznanj in predpostavk lahko določimo ustrezne glasovne enote. Ker določamo le položaje značilnih glasov in ne vseh glasov, potem so najkrajše glasovne enote, ki jih lahko dobimo, vedno v eni izmed oblik:

- ZNAČILNI GLAS -ZAPOREDJE NEZNAČILNIH GLASOV-ZNAČILNI GLAS,
- ZAČETNI GLAS-ZNAČILNI GLAS,
- ZNAČILNI GLAS-KONČNI GLAS.

Kot začetni glas v glasovni enoti se lahko pojavi poljuben glas jezika, ki je prvi v

neprekinjenem zaporedju glasov od začetka govorjenja, ali po premoru v izgovorjavi. Kot končni glas lahko tudi nastopa poljuben glas, ki je zadnji v neprekinjenem zaporedju glasov od začetka govorjenja, ali pa zadnji glas pred premorom v izgovorjavi. Premori v izgovorjavi se pri nevtralnem načinu izgovorjave pojavijo ob koncu stavkov in ne med besedami. V zapisih besedil jih predstavljajo ločila (pika, vejica, podpičje, klicaj, vprašaj). Če je začetni, ali končni glas že značilni glas, so v glasovno enoto vključeni še glasovi do naslednjega oziroma od prejšnjega značilnega glasu.

Ko poznamo položaje značilnih glasov, je mogoča tudi drugačna sestava glasovnih enot, ki ustreza delitvi na zloge oziroma na dvozloge (od sredine samoglasnika prvega zloga do sredine samoglasnika drugega zloga). V tem primeru so glasovne enote v eni izmed naslednjih oblik:

- SAMOGLASNIK -ZAPOREDJE SOGLASNIKOV-SAMOGLASNIK,
- ZAČETNI GLAS-SAMOGLASNIK,
- SAMOGLASNIK-KONČNI GLAS.

Kot začetni glas v glasovni enoti se lahko pojavi poljuben glas jezika, ki je prvi v neprekinjenem zaporedju glasov od začetka govorjenja, ali po premoru v izgovorjavi. Kot končni glas lahko tudi nastopa poljuben glas, ki je zadnji v neprekinjenem zaporedju glasov od začetka govorjenja, ali pa zadnji glas pred premorom v izgovorjavi. Če je začetni ali končni glas že samoglasnik, so v govorno enoto vključeni še glasovi do naslednjega oziroma od prejšnjega samoglasnika.

Tako določene glasovne enote ustrezajo obema prej določenima meriloma. Z rezanjem govora na značilnih glasovih lahko dosežemo dovolj visoko zanesljivost samodejnega rezanja s pomočjo računalnika, kar je razvidno iz rezultatov v poglavju 5. Tudi dolžina enot je po predvidevanjih obvladljiva, saj spada v skupino značilnih 14 od 29 fonemov slovenskega jezika. Statistike o številu tako določenih enot v slovenskem jeziku sicer ni, jo bo pa mogoče sčasoma pridobiti s pomočjo rezultatov projekta Fida oziroma FidaPLUS [6]. Ker spadajo med značilne glasove vsi samoglasniki in šest soglasnikov, predpostavljamo, da je najkrajša dolžina glasovne enote, ki smo jo določili, primerljiva vsaj s triglasniki, ki se pojavljajo v slovenskem jeziku. To sklepamo iz podatka, da se zaporedja soglasnikov med samoglasniki skupaj s polglasnikom, ki so daljša od treh glasov v slovenskem jeziku ne pojavljajo. Podatek ni strogo preverjen, vendar na podlagi splošnega znanja slovenskega jezika, lahko sklepamo, da če taka zaporedja le obstajajo, so zelo redka. Za boljšo predstavo o velikostnem redu števila glasovnih enot navedimo, da lahko iz fonemov slovenskega jezika sestavimo $29^3 = 24389$ mogočih različnih zaporedij dolžine tri, pri čemer pa se jih v živem jeziku pojavlja precej manj.

V prototipni različici sestavljalnika smo uporabili posnetke, ki so bili skupaj sestavljeni iz 139 glasovnih enot. Podrobnejši opis preizkusa prototipne različice smo podali v nadaljevanju v poglavju 5.2.

4.3 Zgradba sistema NGS

Zgradba sistema NGS je podana kar v obliki diagrama poteka operacij v sistemu in jo prikazuje slika 18. Vhodno besedilo se najprej obdela v modulu, ki znakovni zapis pretvori v glasovni zapis (grapheme-to-phoneme conversion). Ta del se nanaša na jezikovna znanja in ni obdelan v doktorski disertaciji. V sistem lahko vgradimo obstoječe module, ki so že razviti za posamezne svetovne jezike in tudi za slovenski jezik [10], [11].

Glasovni zapis se v naslednjem modulu razreže na zapise sestavnih glasovnih enot, katerih obliko smo podrobneje opisali v poglavju 4.2. V bazi posnetkov hranimo vzorce od začetka prvega do konca zadnjega glasu glasovne enote. V procesu lepljenja, ko med seboj lepimo različne glasovne enote, približno polovico vzorcev prvega in zadnjega glasu zavržemo. Te dodatne vzorce hranimo zato, da lažje določimo točko lepljenja, kjer je akustična razdalja med sosednjimi enotami v končnem zaporedju najmanjša.

V naslednji operaciji v procesu pretvorbe vhodnega besedila v umetni govorni signal moramo poiskati posnetke v bazi posnetkov glasovnih enot, ki dajo optimalno, to je najnižjo ceno pri končni operaciji lepljenja. Za vsako glasovno enoto vhodnega besedila poiščemo v bazi tako enoto, ki da najnižjo ceno lepljenja. Če je takih enot več, izberemo prvo. Zapisi določil glasovnih enot v bazi prototipne različice sistema NGS vsebujejo naslednje podatke:

- število glasov v enoti,
- niz znakov kot glasovni zapis vseh glasov govorne enote,
- frekvenco osnovnega tona za prvi in zadnji glas,
- dolžino trajanja za prvi in zadnji glas,
- dvojiško označbo, ali je bila posneta govorna enota začetna enota stavka, ali ne,
- dvojiško označbo, ali je bila posneta govorna enota končna enota stavka, ali ne,
- kazalec na besedilo celega stavka iz katerega je bila enota izrezana in
- zaporedno številko enote v stavku iz katerega je bila izrezana.

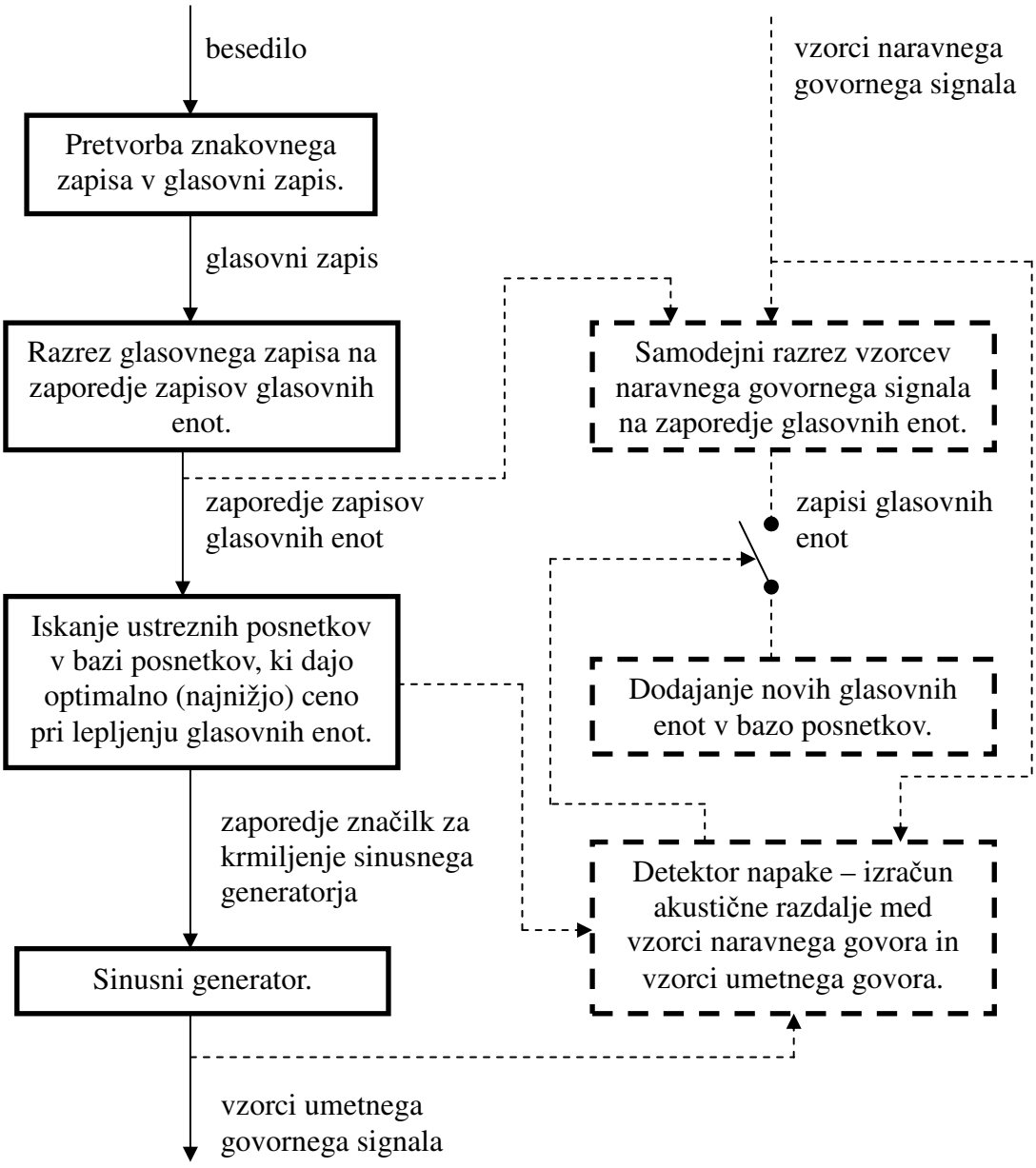
Pri izračunu cene lepljenja (enačba 3.18) smo uporabili prvi pristop, ki je opisan v poglavju 3.3.1. Kot funkcijo za izračun cene zadetka T smo uporabili enačbo (3.24). Utež za glasovni zapis in utež za število glasov v enoti smo postavili na visoko konstantno vrednost, tako da funkcija že v primeru, da ni ujemanja na enem glasu, vrne vrednost, ki je za velikostni razred večja od maksimalne možne celotne cene lepljenja v primeru popolnega ujemanja glasovnega zapisa. Pri frekvenci osnovnega tona, dolžini glasu ter pri oznakah za začetno oziroma končno enoto so uteži v prototipni različici postavljene na 1, v splošnem pa so učljive.

Zadnja dva podatka v zapisu sta pomembna pri računanju cene spoja. Funkcija za izračun cene spoja J (v enačbi 3.18) računa ceno spoja kot Evklidsko razdaljo med vektorjem amplitudnega spektra na sredi zadnjega glasu govorne enote i in vektorjem amplitudnega spektra na sredi prvega glasu govorne enote $i+1$. V primeru, da sta zaporedni govorni enoti hkrati tudi zaporedni govorni enoti v učnem posnetku naravnega govornega signala, kar lahko razberemo iz zadnjih dveh podatkov v zapisu določil, se cena spoja J postavi na 0.

Za lepljenje posnetih glasovnih enot in tvorbo govornega signala smo uporabili sinusni generator, ki temelji na čistem sinusnem modelu, predstavljenem v podpoglavju 3.2.2. Točko

lepljenja določimo kot tisto točko v bližini sredine glasu, na katerem izvajamo lepljenje, kjer je Evklidska razdalja med amplitudnim spektrom sosednjih okvirjev značilk govornega signala najmanjša. Lepimo pa na prvih oziroma zadnjih glasovih glasovnih enot.

Del sistema, ki je na desni strani slike 18 prikazan s črtkanimi črtami, je aktiven le v procesu učenja, ko v sistem dodajamo nove posnetke glasovnih enot. Bistven del sistema predstavlja modul za samodejni razrez vzorcev naravnega govornega signala na glasovne enote, ki jih potem uporabljamo v procesu sestave govora. V bazo posnetkov v procesu učenja dodamo samo tiste glasovne enote, brez katerih bi bila akustična razdalja med naravnim in umetnim govorom večja od maksimalne napake, ki jo v sistem vnese sinusni generator, oziroma primerke glasovnih enot za zaporedja glasov, za katera v bazi še ni nobenega posnetka.



Slika 18: Zgradba in prikaz delovanja sistema NGS.

V nadaljevanju tega podpoglavja je najprej podrobneje opisan algoritem za samodejni razrez, zatem algoritem učenja sistema, nato pa še postopek tvorjenja vzorcev umetnega govornega signala s pomočjo sinusnega generatorja, ki ga uporabljamo tudi za lepljenje glasovnih enot.

4.3.1 Algoritem za samodejni razrez

Algoritem je sestavljen iz dveh glavnih delov. V prvem delu iz značilk govornega signala najprej pridobimo nabor točk na časovni osi, ki služijo kot potencialne sredine značilnih glasov vhodnega besedila. Kot značilke smo uporabili izhode 393 za $\pi/2$ fazno zamaknjenih parov KEO filtrov, enakomerno porazdeljenih med 70 Hz in 7930 Hz z razmikom srednjih frekvenc 20 Hz in s pasovno širino prav tako 20 Hz. Na izhodu posameznega para filtrov, ki se osveži vsakih 5 ms, dobimo energijo govornega signala v frekvenčnem pasu para filtrov kot amplitudo osrednje frekvence para filtrov. Tak nabor značilk narekuje sinusni generator, ki ga uporabljamo za tvorbo vzorcev umetnega govornega signala.

V drugem delu algoritma značilnim glasovom priredimo ustrezne točke na časovni osi in jih tako poravnamo z vzorci govornega signala. Potencialnih točk na časovni osi, ki jih dobimo v prvem delu algoritma, je v splošnem več kot je značilnih glasov v vhodnem besedilu, zato v drugem delu algoritma značilnim glasovom priredimo tiste točke na časovni osi, ki padejo v časovni interval, ki ustreza mestu glasu v besedilu.

Primer rezultatov algoritma prikazuje slika 19, ki prikazuje samodejni razrez za besedi "čez cesto".

DOLOČANJE TOČK NA ČASOVNI OSI, KI USTREZAJO ZNAČILNIM GLASOVOM

Korak 1:
Določi raven šuma posnetka kot *R*-to najšibkejšo energijo vektorja značilk (izhodi iz nabora KEO filtrov) pomnoženo s pragovnim koeficientom *E*. Vrednosti za *R* in *E* sta odvisni od uporabljene snemalne opreme in pogojev snemanja, v praksi pa dajejo dobre rezultate vrednosti za *R*, ki so med 3 in 5, za *E* pa med 1,1 in 1,3. V nadaljnji analizi uporabimo samo zaporedja značilk, ki presežejo navedeni energijski prag.

Korak 2:
Za vsak nabor značilk *i* izračunaj vrednost funkcije W po enačbi

$$W(i) = \frac{1}{S-1} \sum_{k=-S/2}^{S/2} \sqrt{\sum_{j=0}^{F-1} (a_i[j] - a_k[j])^2}, \tag{4.1}$$

kjer je *S* dolžina okvirja, ki predstavlja zaporedno število vektorjev značilk, ki jih uporabimo

za izračun funkcije W in mora biti liho število. S F smo označili število komponent vektorja značilk, vektor značilk, ki predstavlja približek amplitudnega spektra dela govornega signala, pa z a_i oziroma a_k, z j pa je označen indeks posamezne komponente vektorjev. Vrednost za F je v našem primeru 393, ker imamo toliko filtrov v naboru, za dolžino okvirja S pa smo vzeli 51, tako da pokriva 250 ms govornega signala.
$W(i)$ je povprečna Evklidska razdalja med i-tim in ostalimi zaporednimi vektorji značilk znotraj okna dolžine S, kjer je i-ti vektor značilk srednji vektor znotraj okvirja.

Izračunaj vrednost funkcije $E_v(i)$ kot energijo vektorja značilk a_i v zgornji polovici amplitudnega spektra

$$E_v(i) = \sum_{k=F/2}^{F-1} a_i[k]. \tag{4.2}$$

Korak 3:

Poišči vrhove funkcij $W(i)$ in $E_v(i)$.

Korak 4 (odstranitev neuporabnih vrhov) :

Zavrzi vse vrhove funkcije $W(i)$, ki imajo vrednost manjšo od $p * \max_i W(i)$. Za pragovno vrednost p smo v naših poskusih vzeli 0,05 (5% maksimalne vrednosti $W(i)$).

Od vrhov funkcije $W(i)$, ki so bliže skupaj kot 30 ms, obdrži samo vrhove z večjo vrednostjo, ostale zavrzi.

PRIREJANJE TOČK NA ČASOVNI OSI ZNAČILNIM GLASOVOM

V nadaljnjih korakih algoritma nato izvajamo prirejanje med samoglasniki oziroma polglasnikom in vrhovi funkcije $W(i)$ ter soglasniki c, č, s, š, z,ž in vrhovi funkcije $E_v(i)$. Vrhovi funkcije $W(i)$ namreč označujejo potencialne položaje samoglasnikov oziroma polglasnika, vrhovi funkcije $E_v(i)$ pa potencialne položaje značilnih soglasnikov.

Korak 1:

Izračunaj grobo oceno povprečnega časa trajanja glasu $\boldsymbol{\tau}$ po enačbi

$$\tau = \frac{t_{nmš}}{2\mathrm{N}_{sa} + N_{so}}, \tag{4.3}$$

kjer je $t_{nmš}$ trajanje posnetka nad mejo šuma, N_{sa} število samoglasnikov skupaj s polglasnikom e, N_{so} pa število soglasnikov v zapisu besedila, ki ustreza posnetku govornega signala nad mejo šuma. Zapis besedila je skupaj z vzorci govora vhodni podatek algoritma. Časovno mejo T postavi na začetek signala, ki je nad mejo šuma, dolžino iskalnega intervala δ pa postavi na 0.

Korak 2:

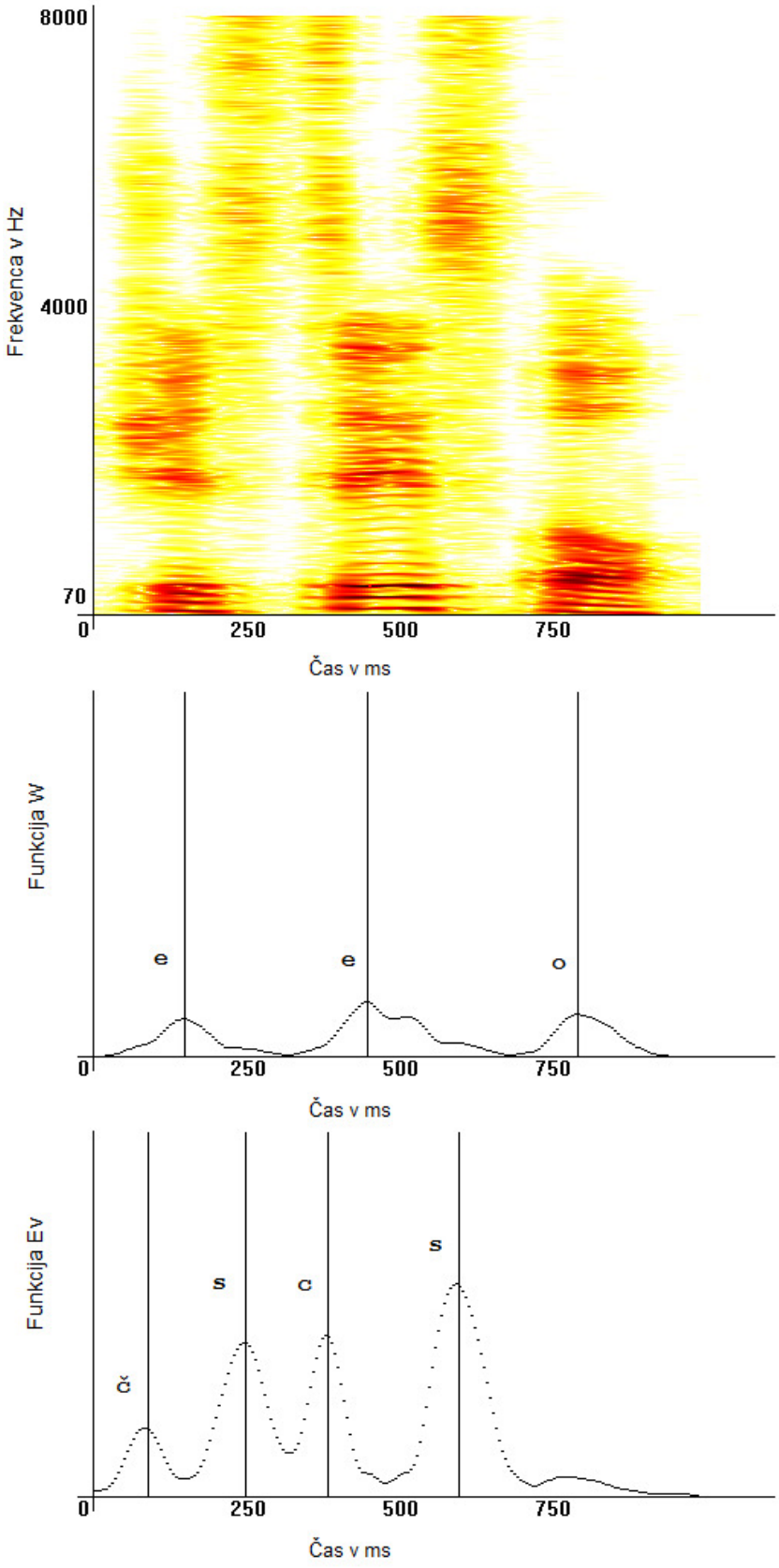
Če je trenutni glas eden izmed značilnih glasov, potem povečaj dolžino iskalnega intervala δ za 2τ, sicer povečaj dolžino iskalnega intervala δ za $\boldsymbol{\tau}$.

Če je trenutni glas samoglasnik ali polglasnik, mu priredi tisti vrh funkcije $W(i)$, ki ima največjo vrednost v intervalu $[T, T+\delta]$, če pa je trenutni glas soglasnik iz skupine značilnih glasov, mu priredi tisti vrh funkcije $E_v(i)$, ki ima največjo vrednost v intervalu $[T, T+\delta]$.

Trenutno časovno mejo postavi na čas prirejenega vrha ustrezne funkcije (W ali E_v) plus $\delta/2$ (predviden konec označenega glasu). Izračunaj nov povprečen čas trajanja glasu po enačbi iz koraka 1. Pri računu upoštevaj preostali čas od trenutne časovne meje do konca dela posnetka nad nivojem šuma in preostale glasove besedila. Postavi dolžino iskalnega intervala δ na 0.

Korak 3:

Pomakni se na naslednji glas. Če je v vhodnem besedilu še kakšen glas, pojdi nazaj na korak 2, sicer končaj.



Slika 19: Spektrogram in vrhovi funkcij W ter E_v za besedi "čez cesto".

4.3.2 Sinusni generator

Pri tvorbi vzorcev govornega signala sistem NGS uporablja sinusni generator, ki je primer čistega sinusnega modela (poglavje 3.2.2). Glede na izvorni model smo v sinusnem generatorju predelali ujemalni algoritem frekvenc in algoritem za izbiro sinusnih komponent, ki vstopajo v postopek sestave. Navedene predelave smo podrobneje opisali že v [32], [33].

Za sinusni generator smo se odločili, ker umetnega govornega signala, ki ga z njim tvorimo z značilkami pridobljenimi iz naravnega govornega signala (analizno-sestavni sistem), ne moremo ločiti od naravnega govora. Z izbiro sinusnega generatorja smo zmanjšali vpliv nizkonivojske sestave govornega signala pod prag človeške zaznave, tako da lahko vzroke za morebitni nenaravno zveneč ali celo nerazumljiv umetni govor, ki bi ga tvoril sistem NGS, iščemo v drugih gradnikih sistema.

Pri tvorbi vzorcev govornega signala s pomočjo sinusnega generatorja je potrebno rešiti več problemov. Prvi problem, ki ga je potrebno rešiti, je glajenje nezveznosti med okvirji. Z rešitvijo tega problema rešimo tudi problem zlivanja okvirjev. Pri glajenju nezveznosti med okvirji rešimo najprej ujemanje frekvenc med okvirji. Ker se število in položaj frekvenčnih komponent spreminja od okvirja do okvirja, je potrebno najti način, kako frekvenci iz predhodnega okvirja prirediti frekvenco iz naslednjega okvirja. Rešiti je potrebno tudi "rojevanje" in "umiranje" frekvenc, to sta primera, ko je število frekvenc v predhodnem okvirju manjše oziroma večje kot v naslednjem okvirju. Največ sprememb nastane v hitro spreminjajočih se delih govornega signala, kot so na primer prehodi med zvenečimi in nezvenečimi glasovi.

Pri ujemanju frekvenc smo uporabili naslednjo preprosto metodo, ki je podrobneje obdelana v [32]. Predpostavimo, da smo že uspeli prirediti frekvence do okvirja k in da imamo v okvirju $k+1$ novo množico frekvenc, ki bi jim radi priredili frekvence iz okvirja$k$, in s tem dosegli frekvenčno ujemanje med okvirjema. Frekvence iz okvirja k označimo z

$$\omega_0^k, \omega_1^k, \dots, \omega_{N-1}^k,$$

frekvence iz okvirja $k+1$ pa z

$$\omega_0^{k+1}, \omega_1^{k+1}, \dots, \omega_{M-1}^{k+1},$$

kjer v splošnem $M \neq N$. Postopek prirejanja frekvenc lahko zapišemo v naslednjih treh korakih

Korak1:

Predpostavimo, da smo frekvence $\omega_0^k, \omega_1^k, \dots, \omega_{n-1}^k$ že priredili in da skušamo zdaj prirediti ustrezno frekvenco frekvenci ω_n^k. Zgodi se lahko več različnih primerov.

a) Vse frekvence ω_m^{k+1} v okvirju $k+1$ ležijo izven ujemalnega intervala Δ za frekvenco ω_n^k

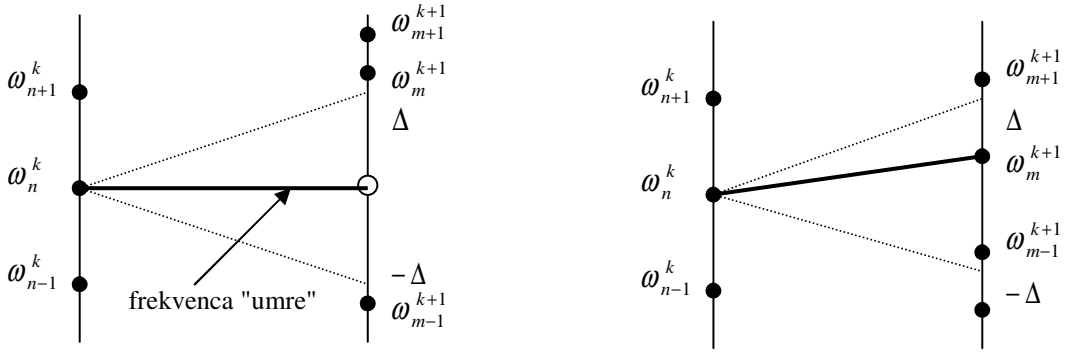
$$|\omega_n^k - \omega_m^{k+1}| \geq \Delta \text{ za } \forall m. \tag{4.4}$$

V tem primeru frekvenca "umre" in ima v okvirju $k+1$ amplitudo 0. S tem je tudi izločena iz nadaljnje obravnave. Korak 1 ponovimo za naslednjo frekvenco. Širina ujemalnega intervala Δ je običajno manjša od dvakratne frekvenčne ločljivosti pri analizi signala.

b) Posreči se nam najti frekvenco ω_m^{k+1} v okvirju $k+1$, ki leži v ujemalnem intervalu Δ in je najbližja taka frekvenca

$$|\omega_n^k - \omega_m^{k+1}| < \left|\omega_n^k - \omega_i^{k+1}\right| < \Delta \text{ za } \forall i \neq m. \tag{4.5}$$

Frekvenco ω_m^{k+1} označimo za možnega kandidata za frekvenčno ujemanje, vendar se o tem odločimo šele v koraku 2, ker je lahko frekvenca ω_m^{k+1} boljši par kaki drugi frekvenci iz okvirja k. Korak 1 ponovimo za naslednjo frekvenco.



Slika 20: Možnosti a in b prvega koraka ujemalnega algoritma za sledenje frekvenčnih stez.

Korak2:

V tem koraku dokončno potrdimo vse izbire, narejene v prvem koraku. Če smo frekvenci ω_m^{k+1} iz okvirja $k+1$ priredili le frekvenco ω_n^k iz okvirja k, potem to ostane tudi končna prireditev, sicer pa izmed frekvenc ω_i^k iz okvirja k, ki jih hočemo prirediti frekvenci ω_m^{k+1} iz okvirja $k+1$, poiščemo najbližjo frekvenco ω_n^k, tako da velja

$$|\omega_m^{k+1} - \omega_n^k| < \left|\omega_m^{k+1} - \omega_i^k\right| \text{ za } \forall i \neq n. \tag{4.6}$$

Frekvenci ω_n^k in ω_m^{k+1} damo na seznam končnih odločitev, za preostale frekvence iz okvirja k pa ponovimo korak 1, vendar pri tem ne upoštevamo več frekvenc, ki so na seznamu končnih odločitev.

Korak 3:

Ko smo tako obdelali vse frekvence iz okvirja k, nam lahko ostane še nekaj neprirejenih frekvenc v okvirju $k+1$. Za vse take frekvence zaključimo, da so "rojene" v okvirju k z amplitudo 0, in jih priredimo samim sebi v okvirju $k+1$.

Po končanem postopku ujemanja frekvenc dobimo množico tako imenovanih frekvenčnih stez, ki prikazujejo potek spreminjanja posamezne frekvence govornega signala.

INTERPOLACIJA AMPLITUD

Poleg ujemanja frekvenc moramo rešiti tudi problem glajenja amplitud med okvirji. Naj bosta $(A_l^k, \omega_l^k, \theta_l^k)$ in $(A_l^{k+1}, \omega_l^{k+1}, \theta_l^{k+1})$ množici značilk za k in $k+1$ okvir za frekvenčno stezo l. Amplitudo preprosto interpoliramo z

$$B_l(n) = A_l^k + \frac{(A_l^{k+1} - A_l^k)}{N_f} n \text{ za } n = 0, 1, \dots, N_f - 1, \tag{4.7}$$

kjer N_f predstavlja število vzorcev v okvirju.

INTERPOLACIJA FAZ IN FREKVENC

Pri interpolaciji frekvenc in faz ne moremo uporabiti tako enostavnega pristopa kot pri amplitudah, ker so meritve faz pridobljene po modulu 2π. Zagotoviti moramo, da bodo imele frekvenčne steze pri faznem odvijanju (unwraping) maksimalno gladke prehode med okvirji. Zaradi tega bomo fazo interpolirali s kubičnim polinomom

$$\theta(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3. \tag{4.8}$$

Priročno je, če privzamemo, da je fazna funkcija θ zvezna funkcija časa t, kjer $t = 0$ ustreza okvirju k, $t = T$ pa okvirju $k+1$. Ker je frekvenca odvod faze po času, mora vrednost kubične fazne funkcije in njenega odvoda ustrezati merjenim vrednostim faz in frekvenc na robovih okvirjev. Te omejitve moramo upoštevati pri odvijanju faz, ker dobimo pri meritvi le

glavno vrednost (vrednost po modulu 2π).

Odvod kubične fazne funkcije je enak

$$\dot{\theta}(t) = \gamma + 2\alpha t + 3\beta t^2. \tag{4.9}$$

V kubično fazno funkcijo in njen odvod vstavimo robne pogoje:

začetna točka, $t = 0$,

$$\begin{aligned} \theta(0) &= \zeta = \theta^k, \\ \dot{\theta}(0) &= \gamma = \omega^k \end{aligned} \tag{4.10}$$

in končna točka $t = T$,

$$\begin{aligned} \theta(T) &= \theta^k + \omega^k T + \alpha T^2 + \beta T^3 = \theta^{k+1} + 2\pi X, \\ \dot{\theta}(T) &= \omega^k + 2\alpha T + 3\beta T^2 = \omega^{k+1}. \end{aligned} \tag{4.11}$$

Dobimo dve enačbi s tremi neznankami α, β in X, kjer α in β predstavljata koeficienta interpolacijskega polinoma, X pa celoštevilčni mnogokratnik faznega zasuka za 2π, ki ga moramo upoštevati pri interpolaciji. V matrični obliki lahko vse skupaj zapišemo kot

$$\begin{bmatrix} \alpha(X) \\ \beta(X) \end{bmatrix} = \begin{bmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{bmatrix} \begin{bmatrix} \theta^{k+1} - \theta^k - \omega^k T + 2\pi X \\ \omega^{k+1} - \omega^k \end{bmatrix}. \tag{4.12}$$

Ker je faza θ^{k+1} izmerjena po modulu 2π, jo je potrebno raztegniti s členom $2\pi X$ in to tako, da bo dobljena odvita kubična fazna funkcija "maksimalno gladka". Ta dodatni pogoj "maksimalne gladkosti" nam da potrebno tretjo enačbo. Kot "maksimalno gladko" funkcijo razumemo tisto fazno funkcijo, ki ima najmanjšo spremenljivost (variation). Taka funkcija nakazuje na enakomerne fazne spremembe brez velikih skokov, kar posredno pogojuje majhne in enakomerne spremembe frekvence, ki je prvi odvod faze po času. Tako za določitev vrednosti X poiščemo minimum funkcije

$$f(X) = \int_0^T [\ddot{\theta}(t, X)]^2 dt = 4\alpha(X)^2 T + 12\alpha(X)\beta(X)T^2 + 12\beta(X)^2 T^3, \tag{4.13}$$

kjer je $\ddot{\theta}(t,X)$ drugi odvod $\theta(t,X)$ po času *t*. Ker je *X* celoštevilčen, vzamemo za njegovo vrednost tisto celo število, ki je najbližje minimumu funkcije *f(x),* kjer je *x* zvezna spremenljivka.

Rešitev za *x* je desna stran enačbe

$$x = \frac{1}{2\pi}\left[(\theta^k + \omega^k T - \theta^{k+1}) + (\omega^{k+1} - \omega^k)\frac{T}{2}\right], \tag{4.14}$$

iz katere potem dobimo ustrezno celoštevilsko vrednost za *X*. To vrednost vstavimo v interpolacijsko funkcijo

$$\theta(t) = \theta^k + \omega^k t + \alpha(X)t^2 + \beta(X)t^3. \tag{4.15}$$

Tako dobljena fazna funkcija zadovolji postavljene začetne pogoje in da zahtevano maksimalno gladko odvito fazo. Določiti moramo le še začetno fazo pri tistih frekvenčnih stezah, ki se v okvirju k šele rodijo (amplituda 0) in za njih neka izmerjena vrednost obstaja šele v okvirju $k+1$. Začetno fazo v tem primeru določimo kot

$$\theta^k = \theta^{k+1} - \omega^{k+1}Y, \tag{4.16}$$

kjer je *Y* število vzorcev od okvirja $k+1$ na začetek okvirja k. S tem imamo določeno trenutno fazo za vsako frekvenčno stezo, ki sledi hitrejšim spremembam, ki so posledica spreminjanja frekvence in počasnim spremembam, ki so posledica faznih sprememb. Izhod iz sinusnega generatorja lahko končno zapišemo kot

$$s(n) = \sum_{i=1}^{L^k} B_i(n)\cos[\theta_i(n)], \tag{4.17}$$

ki predstavlja govorni signal sestavljen z NGS.

4.3.3 Učni algoritem

Postopek učenja sistema lahko poteka v sprotnem sodelovanju z uporabnikom. Vsaj v začetnem delu pa je proces hitrejši, če uporabnik vnaprej posname človeški glas pri branju določene količine besedil. Raziskave [35] so pokazale, da je za visoko kakovost umetnega

govornega signala potrebno posneti vsaj za eno uro naravnega govornega signala iz besedil, ki pokrijejo čim večji del glasovnih značilnosti jezika. Za bolj fino nastavitev sistema pa je nujno sprotno sodelovanje človeškega uporabnika, ki je lahko hkrati ocenjevalec kakovosti sestave govora in izvor dodatnih posnetkov govornega signala. Uporabnik v procesu učenja vnese v sistem besedilo in zatem posluša umetno sestavljen govorni signal. Če je s kakovostjo zadovoljen, je proces učenja končan, sicer posname novo zaporedje vzorcev govornega signala za tiste stavke besedila, kjer kakovost umetnega govornega signala ni zadovoljiva. Podrobneje je algoritem opisan v naslednjih 8 korakih.

Korak 1:

Posnemi govorni signal za en stavek besedila. V postopek učenja vstopajo znakovni zapis besedila in vzorci govornega signala.

Korak 2:

Vhodni znakovni zapis besedila pretvori v glasovni zapis in ga razreži na zapise glasovnih enot sistema s pomočjo algoritma za samodejni razrez, ki je podrobno opisan v poglavju 4.3.1.

Korak 3:

V bazi posnetkov s pomočjo funkcije cene zadetka C (enačba 3.18) za vsako glasovno enoto poišči najboljši (najcenejši) zadetek.

Korak4:

Zaporedje glasovnih enot pridobljenih v koraku 3 zlepi s pomočjo sinusnega generatorja, ki hkrati tvori vzorce umetnega signala.

Korak 5:

Izračunaj razliko med posnetkom naravnega govora in v koraku 4 pridobljenim zapisom umetnega govornega signala kot Evklidsko razdaljo med amplitudnim spektrom naravnega in umetnega govornega signala. Amplitudni spekter izračunaj na vsakih 5 ms s frekvenčno ločljivostjo 20 Hz.

Korak 6:

Ta korak ima dve možni izvedbi glede na to, ali človeški uporabnik sproti sodeluje v procesu učenja, ali ne. Če uporabnik ne sodeluje, se v bazo posnetkov dodajo tiste glasovne enote, kjer razlika med naravnim in umetnim govornim signalom presega maksimalno napako sinusnega

generatorja. Maksimalno napako sinusnega generatorja določimo kot maksimalno Evklidsko razdaljo med amplitudnim spektrom naravnega in umetnega govornega signala vzorčnega besedila, le da za razliko od koraka 5 pri sestavi uporabimo tiste značilke za krmiljenje sinusnega generatorja, ki smo jih pridobili pri analizi naravnega govornega signala vzorčnega besedila (sestava neposredno iz analize).

V primeru, če uporabnik sodeluje, je dopustna napaka lahko večja, če uporabnik s poslušanjem oceni, da je umetni govorni signal kljub napaki, ki je večja od napake sinusnega generatorja, dovolj kakovosten.

Korak 7:

Če so v bazo posnetkov dodani novi posnetki glasovnih enot, izvedi korak učenja odločitvenega drevesa, ki ustrezno popravi uteži pri funkciji za izračun cene. Cena sestave posnetka, ki smo ga ravnokar obdelali, mora biti po tem koraku učenja najnižja, če uporabimo na novo dodane posnetke govornih enot.

Korak 8:

Če želimo sistem še naprej učiti z dodatnimi zapisi besedil in posnetki ustreznih govornih signalov, potem ponovimo korak 1, sicer se učni algoritem konča.

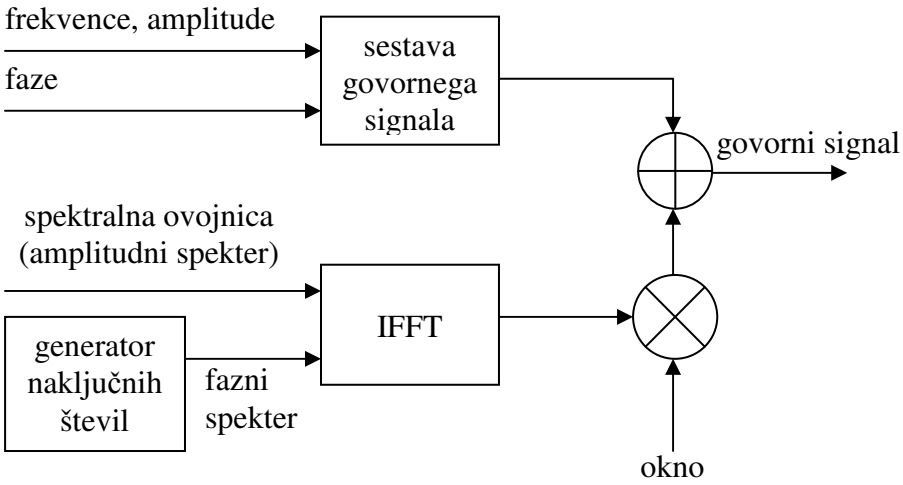
4.4 Zmanjšanje števila značilk za krmiljenje sinusnega generatorja

Del raziskav je bil namenjen tudi zmanjšanju števila potrebnih značilk za krmiljenje sinusnega generatorja. Značilke, ki jih uporablja sinusni generator, niso primerne za zapis govornega signala v zbirki glasovnih enot, ker bi bili zapisi skoraj dvajsetkrat daljši od običajnega PCM zapisa govornega signala. Tudi z drugimi sinusnimi modeli bi brez izgube akustične informacije govornega signala težko dosegli skrajšanje zapisa. Glavni razlog za zmanjšanje števila značilk ob ohranitvi enake kakovosti umetnega govornega signala je predvsem v tem, da bi se s tem zmanjšala tudi računska zahtevnost postopka zlivanja.

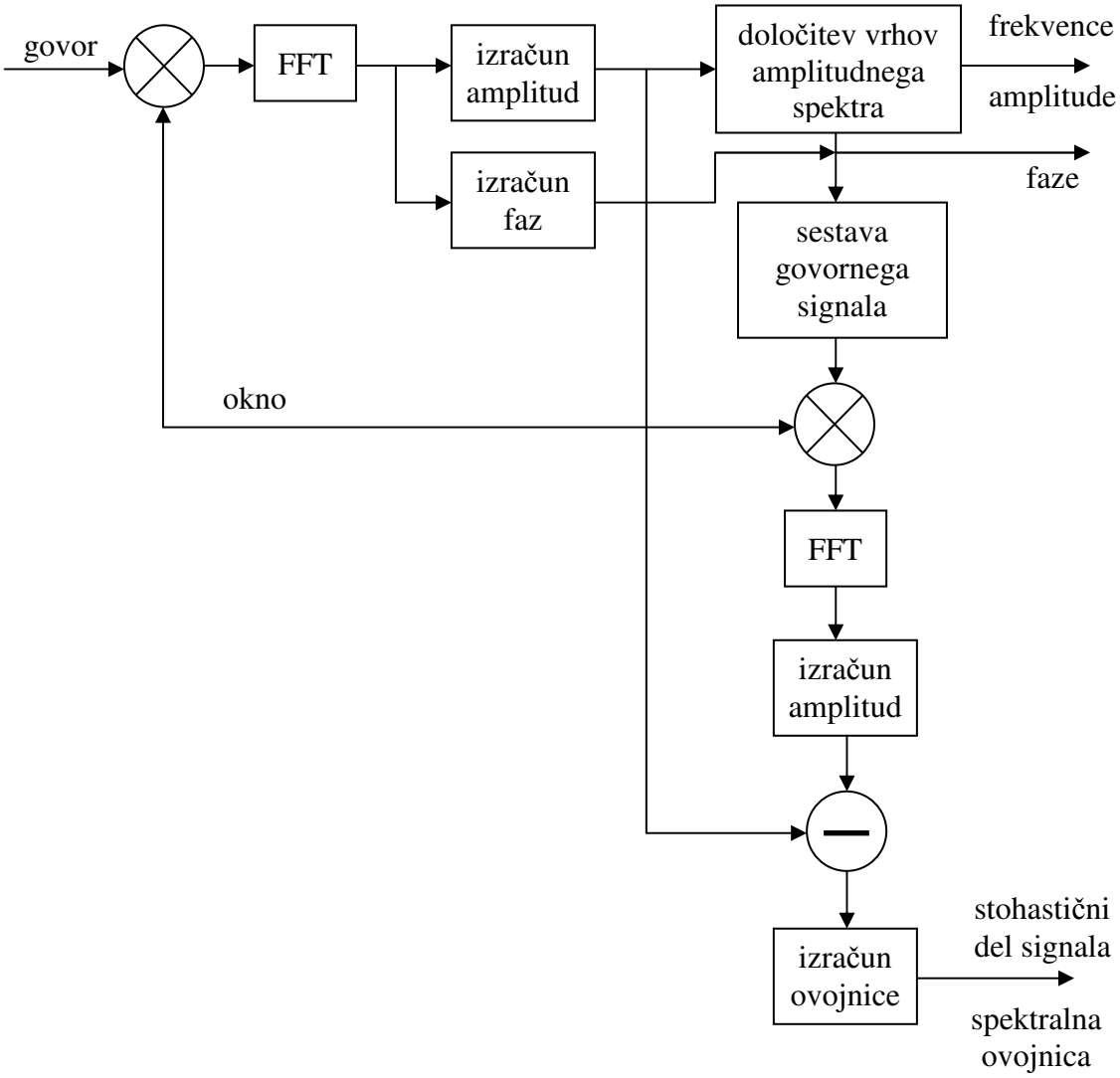
Za znižanje števila značilk predlagamo nov model sinusnega generatorja, ki deloma sloni na modelu, ki je opisan v [29] kot deterministično plus stohastično model (DSM). Prvi del postopka do pridobitve podatkov o naboru frekvenc, amplitud in faz govornega signala je pri obeh modelih enak – uporabimo kratko časovno Fourierjevo transformacijo oziroma nabor filtrov. Od tu se postopek nekoliko razlikuje. Pri izvornem sinusnem generatorju smo upoštevali vse frekvenčne komponente z dovolj velikimi amplitudami, da so skupaj lahko prispevale vnaprej določen energijski delež η okvirja govornega signala. Pri na novo predlaganem sinusnem generatorju pa kot pri DSM modelu upoštevamo le tiste frekvenčne komponente, ki predstavljajo vrhove amplitudnega spektra in jih označimo kot deterministični del signala. Iz teh komponent sestavimo umetni govorni signal, katerega amplitudni spekter odštejemo od amplitudnega spektra naravnega signala in dobljeni preostanek potem

obravnavamo kot šum oziroma stohastični del signala. S pomočjo LPC analize, ali pa s pomočjo zlepkov dobimo spektralno ovojnico šumnega signala, ki jo pri sestavi uporabimo za krmiljenje generatorja šuma. Generator za deterministični del spektra ostane enak kot pri izvornem sinusnem generatorju.

Cel postopek analize in sestave govora sinusnega generatorja po zgledu na DSM model prikazujeta sliki 21 in 22. Algoritma za ujemanje frekvenčnih stez ter interpolacijo amplitud in faz na meji med okvirji ostaneta enaka kot v izvorni izvedbi sinusnega generatorja.



Slika 21: Sestavljalnik sinusnega generatorja, ki je predelan po vzoru na DSM model.

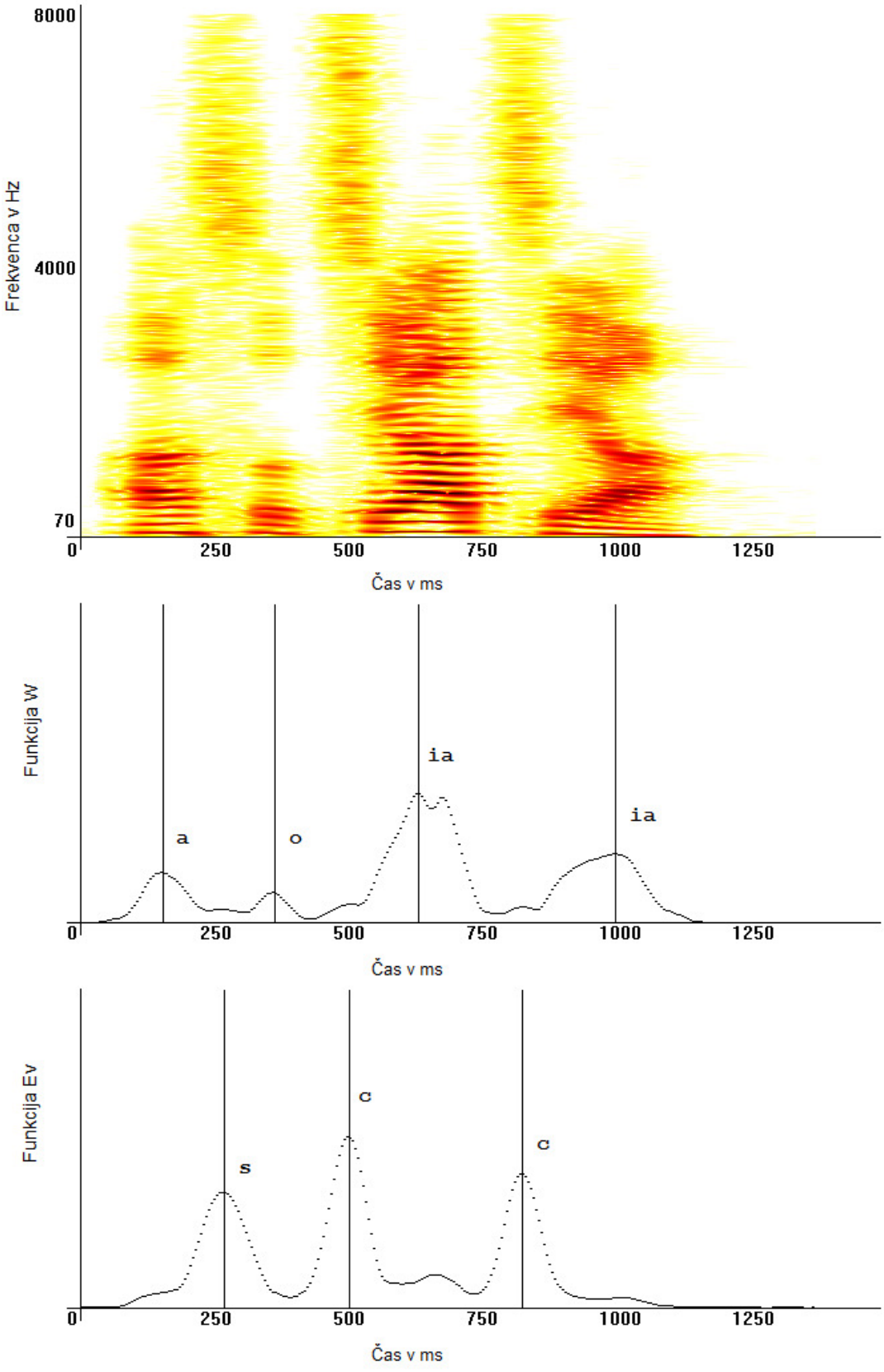


Slika 22: Analizni del sinusnega generatorja, ki je predelan po vzoru na DSM model.

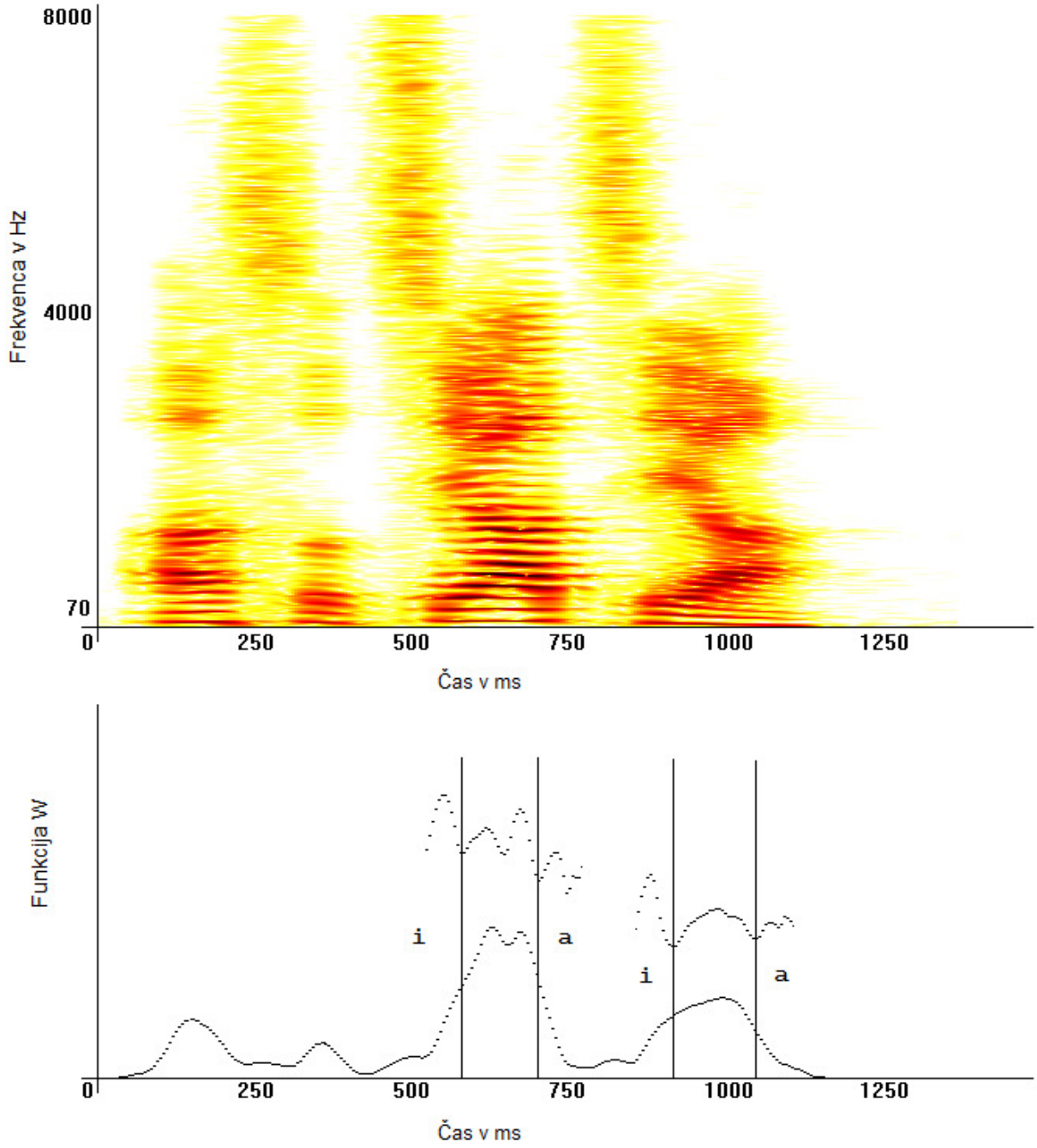
5 Rezultati

Pri razvoju algoritma za avtomatsko označevanje posnetkov govora smo se omejili na določevanje položaja lažje določljivih glasov. Sredine soglasnikov **c**, **č**, **s**, **š**, **z** in **ž** določajo vrhovi funkcije E_v, določene z enačbo (4.2), sredine samoglasnikov in polglasnika pa vrhovi funkcije W, določene z enačbo (4.1). Na sliki 23 lahko tako vidimo spektrogram besede "asociacija" z označenimi vrhovi obeh funkcij, ki ustrezajo glasovom iz navedenih skupin. Iz vrhov funkcije E_v so položaji glasov **s** in **c** v besedi jasno določljivi, ravno tako tudi položaji samoglasnikov **a**, **o** in samoglasniškega para **ia**. Glas **j** na koncu besede se pri izgovorjavi zlije z **i**, tako da imamo v besedi v bistvu dva samoglasniška para **ia**. Posamezne samoglasnike znotraj samoglasniškega para določimo kot lokalne minimume Evklidske razdalje med amplitudnimi spektri dveh vektorjev značilk, ki sta razmaknjena za 40 ms, kar lahko vidimo na sliki 24. Tak pristop smo uporabili na podlagi hipoteze, da so spektralne razlike znotraj posameznega samoglasnika manjše od spektralnih razlik med različnimi samoglasniki. Razmik 40 ms je v povprečju najbolj ustrezen, ker je skupna dolžina trajanja samoglasniškega para gotovo daljša od 40 ms, v točkah lokalnega maksimuma spektralne razlike pa gotovo zajamemo vektorje iz različnih samoglasnikov v samoglasniškem paru.

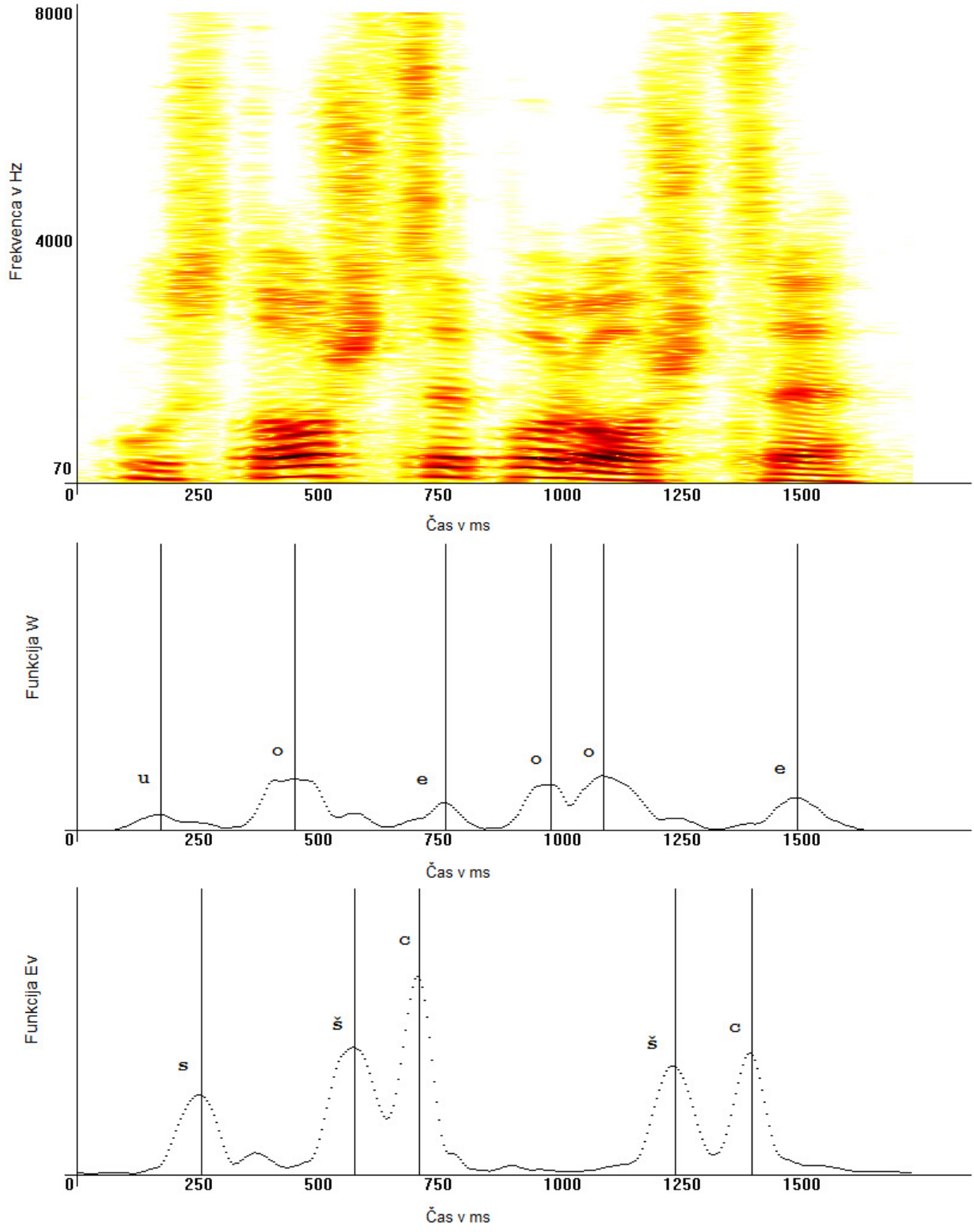
Slika 25 prikazuje spektrograme in vrhove funkcij W in E_v za skupino besed "v Stožce po rožce". Tudi tu ni nobenih težav z ločevanjem značilnih glasov. Na določene težave naletimo v primerih, ki ju prikazujeta sliki 27 in 28. V besedni zvezi "z zažigalnico" se glas v predlogu zlije skupaj s prvim glasom besede, kar je seveda običajno pravilo pri izgovorjavi besed in to moramo pri našem iskalnem algoritmu značilnih glasov tudi upoštevati. Enak pojav se pojavi pri besedni zvezi "s seštevanjem" (slika 28). Večjo težavo povzroča zelo neizrazit maksimum funkcije W pri zadnjem samoglasniku **e**. Pri dovolj nerazločni izgovorjavi se lahko zgodi, da za določen samoglasnik v zadnjih zlogih besed v stavku sploh ni ustreznega vrha funkcije W. V tem primeru mora aplikacija, ki uporablja algoritem, od uporabnika preprosto zahtevati, da stavek še enkrat izgovori.



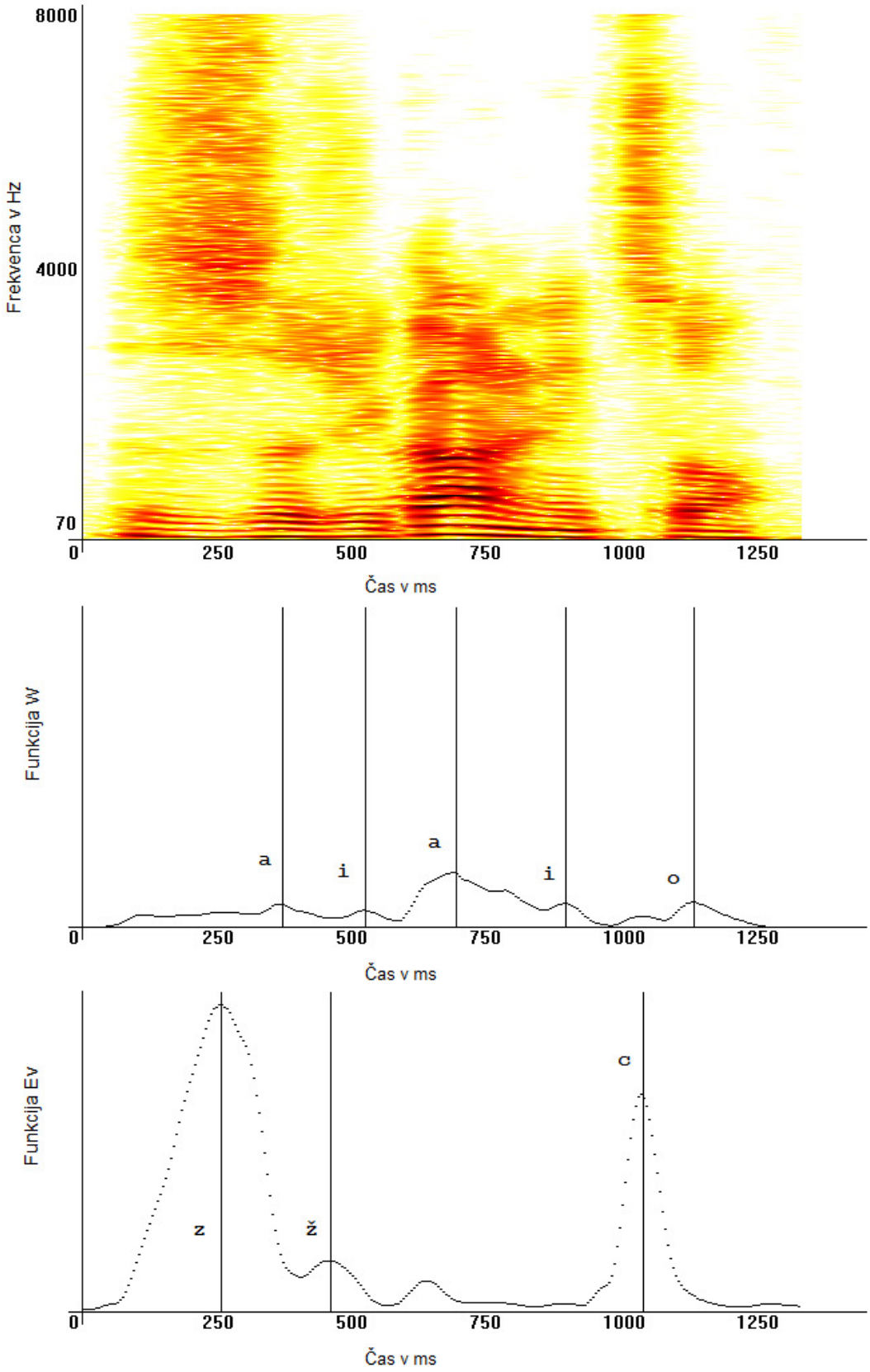
Slika 23: Spektrogram in vrhovi funkcij W ter E_v za besedo "asociacija".



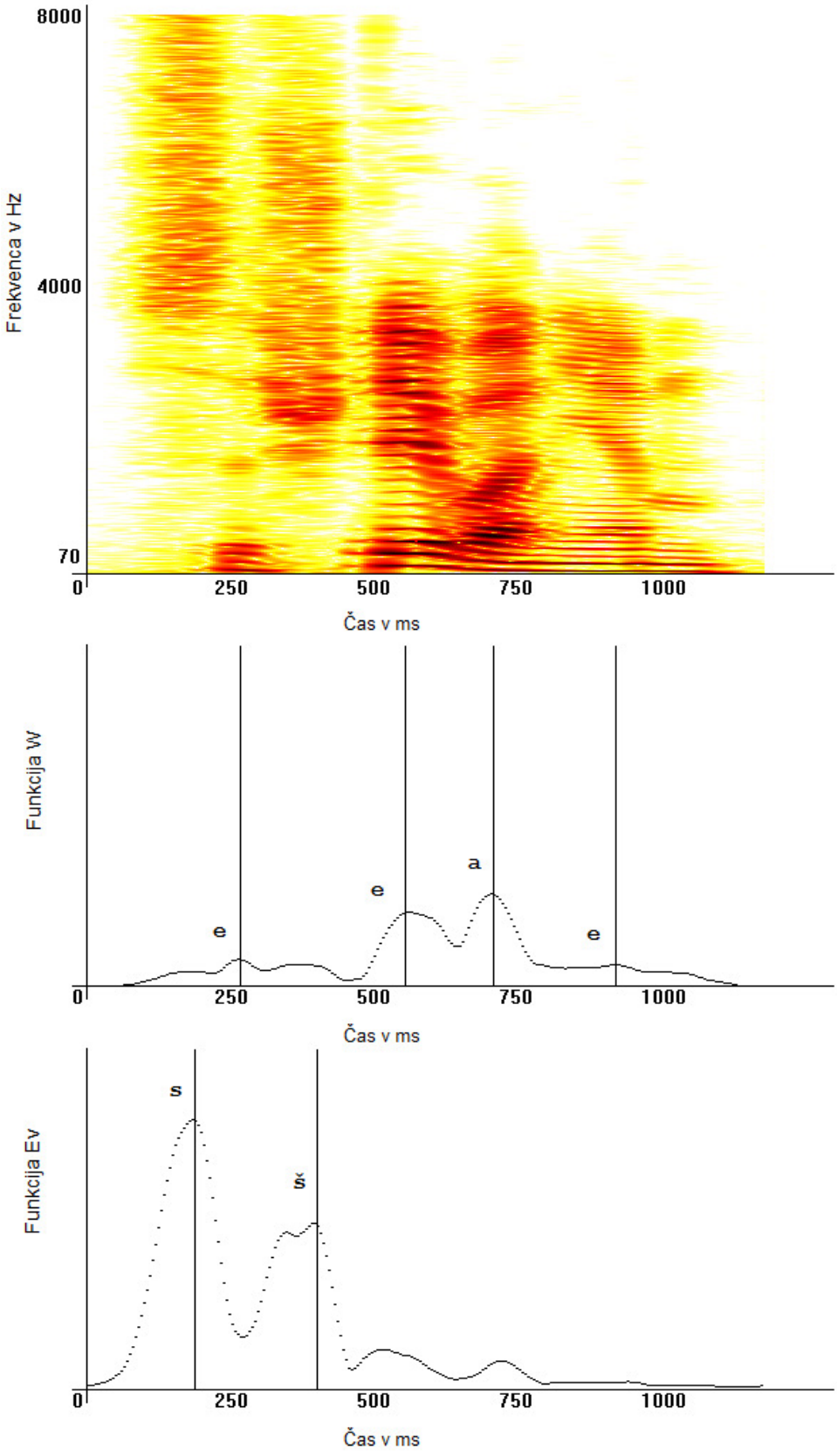
Slika 24: Spektrogram in graf funkcije W za besedo "asociacija" z nadrisanimi odseki Evklidske razdalje med amplitudnimi spektri dveh vektorjev značilk govornega signala, ki sta razmaknjena za 40 ms. Z navpičnimi črtami v grafu funkcije W so označene sredine samoglasnikov v samoglasniškem paru ***ia***.



Slika 25: Spektrogram in vrhovi funkcij W ter E_v za besede "v Stožce po rožce".



Slika 26: Spektrogram in vrhovi funkcij W ter E_v za besedo s predlogom"z zažigalnico".



Slika 27: Spektrogram in vrhovi funkcij W ter E_v za besedo s predlogom " s seštevanjem".

Kakovost algoritma za samodejni razrez smo preverili na ročno označenih oziroma razrezanih primerih posnetkov besedil. Pri testu smo naključno izbrali stran iz slovenskega prevoda romana "Odiseja 2001" pisca Arthurja C. Clarka in posneli govorni signal 10 zaporednih stavkov iz besedila. Proces smo ponovili 5 krat in primerjali razrez, ki ga je izvedel algoritem z ročnim razrezom. Rezultati so podani v tabeli 2.

Zaporedje stavkov	Število glasov	Število napak
1	288	3
2	218	5
3	461	9
4	247	3
5	313	6

Tabela 2: Napaka algoritma za samodejni razrez glede na ročno označevanje značilnih glasov govornega signala.

Pri napakah gre za napačno prireditev vrha funkcij W oziroma E_v, kar ima za posledico, da se del glasu, ali tudi cel glas pri razrezu premakne v skupino v katero ne spada. V redkih primerih, ko se vrh funkcij ne pojavi, lahko zahtevamo tudi ponovno snemanje določenega stavka. Napake napačnega samodejnega razreza lahko izločimo s poslušanjem posnetkov razrezanih delov govornega signala. Če na posnetku slišimo glas premalo ali preveč, je razrez napačen.

5.1 Slušni testi metode za lepljenje glasovnih enot

Glavno merilo pri izboru ustrezne metode za lepljenje glasovnih enot je vpliv metode na kakovost govornega signala, ki ga dobimo. V postopku sestave govora se vsa popačenja, ki jih v tem postopku vnašamo, seštevajo in če metoda, s katero na koncu tvorimo vzorce govornega signala tudi pri optimalnem naboru vhodnih vrednosti ne more proizvesti vzorcev kakovostnega govornega signala, je rezultat, kjer so prisotne tudi ne-optimalne vrednosti, še slabši. Pri velikem številu metod so avtorji zapisali, da tvorijo umetni govorni signal, ki je skoraj neločljiv, ali povsem neločljiv od naravnega. Ne navajajo pa nobenega statističnega testa, ki bi to trditev tudi podpiral. Za sinusni generator (SG), ki je primer čistega sinusnega modela pa opravljeni slušni testi kažejo, da lahko tvori govorni signal, ki je neločljiv od naravnega.

Slušni preizkus kakovosti govornega signala SG modela je potekal na naslednji način. Vsak poslušalec se je usedel pred računalnik opremljen s parom zvočnikov. Najprej je poslušal nekaj stavkov umetnega govora različne kakovosti ter nekaj stavkov naravnega govora, da je dobil slušni vtis, kakšne so razlike med umetnim in naravnim govorom. Računalnik je nato v naključnem zaporedju predvajal deset različnih stavkov in sicer vsak stavek dvakrat. Vrstni

red predvajanja stavka z umetno sestavljenim govorom je bil naključno izbran. Računalnik je lahko najprej predvajal naravni govor, lahko umetno sestavljen govor, lahko pa je obakrat predvajal naravni govor. Poslušalec je nato na listu papirja z vnaprej pripravljenimi možnimi odgovori obkrožil eno izmed treh možnosti za vsak stavek:

- prvo predvajanje je umetno sestavljen govor,
- drugo predvajanje je umetno sestavljen govor,
- obe predvajanji predstavljata naravni govor.

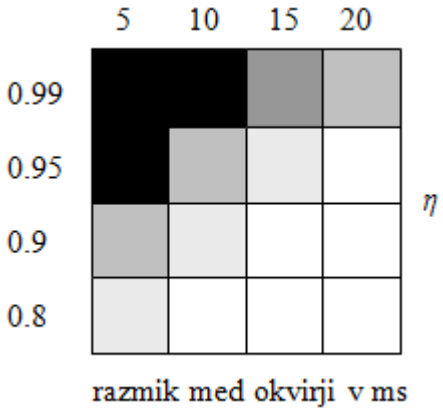
Vsako dvojno predvajanje stavka je poslušalec lahko slišal le enkrat – brez ponavljanj. Stavki so bili različnih dolžin in so trajali od dveh do petih sekund. Umetni govorni signal je bil sestavljen iz zaporedja značilk, ki smo jih pridobili iz posnetka naravnega govora in sicer smo uporabili izhode 393 parov KEO filtrov s faznim zamikom $\pi/2$, ki so bili enakomerno porazdeljeni med 70 Hz in 7930 Hz z razmikom srednjih frekvenc 20 Hz in s pasovno širino prav tako 20 Hz. Pri sestavi govora smo upoštevali samo izhode KEO filtrov z dovolj velikimi amplitudami, ki jih je določal skupni energijski prispevek η, ki je bil določen kot

$$\eta = \frac{\sum izhodi\ filtrov\ z\ največjimi\ amplitudami}{\sum izhodi\ vseh\ filtrov}. \tag{5.1}$$

Med testi smo spreminjali dva parametra: čas osveževanja filtrov ter vrednost energijskega prispevka η. Zaporedja značilk, ki smo jih pridobili pri analizi, smo v enakem vrstnem redu takoj uporabili v postopku sestave. Slušni preizkus je vsak poslušalec ponovil za štiri različne čase osveževanja filtrov (5ms, 10ms, 15ms in 20ms) in štiri različne vrednosti energijskega prispevka η (0.8, 0.9, 0.95 in 0.99).

Govorni signal smo vzorčili s frekvenco 16 kHz, kvantiziran pa je bil na 16 bitov natančno. Pri snemanju smo uporabili samostoječ kondenzatorski mikrofon, za predvajanje posnetkov pa multimedijski komplet zvočnikov za domačo uporabo.

Na sliki 28 so prikazane stopnje ločevanja naravnega in umetno sestavljenega govora, ki so jih dosegli trije poslušalci. S črno barvo so označena tista področja parametrov, pri katerih so vsi poslušalci pri vseh stavkih označili, da sta oba naravna. Bela barva označuje področja parametrov, pri katerih so poslušalci pri vseh stavkih pravilno ločili umetno sestavljen in naravni govor. Področja različnih odtenkov sivin pa označujejo bolj ali manj zanesljivo ločevanje umetno sestavljenega in naravnega govora. Pri vrednostih parametrov za katere so področja na sliki 28 označena z najtemnejšim sivim odtenkom, so imeli vsi poslušalci precej težav pri razločevanju, je pa vsak pravilno razločil kakšen stavek. S srednje sivo barvo so označeni tiste vrednosti parametrov, kjer so poslušalci pri razločevanju še delali napake, večina razvrstitev je bila pravilnih, nekateri pa so že pravilno razločili vse stavke. Z najsvetlejšim sivim odtenkom so označene vrednosti parametrov, kjer je le posamezen poslušalec napačno ocenil kakšen primer.



Slika 28: Ločljivost naravnega in umetno sestavljenega govora pri modelu s 393 filtri v odvisnosti od časovnega razmika med okvirji in energijskega prispevka η.

Za tiste vrednosti parametrov, pri katerih so poslušalci označili, da so vsi stavki naravni, smo preizkus ponovili na dvajsetih stavkih. Poslušalca, ki je dosegel najboljše rezultate, smo prosili naj še natančneje posluša in če se mu kakorkoli zazdi, da je en posnetek zveni bolj umetno od drugega, naj to označi. Vsi ostali pogoji preizkusa so bili enaki kot v prvem primeru. Pri razlagi rezultatov smo uporabili statistični preizkus za srednjo vrednost pravilnih razvrstitev. Če je poslušalec razločeval naravni in umetno sestavljeni govor, bi se moralo število pravilnih odgovorov bistveno razlikovati od pričakovane srednje vrednosti pri naključnem ugibanju. Za verjetnost pravilnega zadetka smo vzeli p = 1/3. Ker je vzorec n = 20 manjši od 30, smo pri preizkusu uporabili Studentovo porazdelitev. Za srednjo vrednost smo vzeli vrednost $\mu = np = 20/3$. Pri stopnji značilnosti $\alpha = 0{,}05$, za katero je $t_\alpha = 2{,}093$ in pri izpolnjeni relaciji $P(t > t_\alpha) = \alpha$, smo dobili rezultate, ki so prikazani v tabeli 3.

Značilke	**Število napak**	**Ocena testa**
393 filtrov, *t* = 5ms, η = 0,99	13	0,69 < 2,093
393 filtrov, *t* = 5ms, η = 0,95	5	17,23 > 2,093
393 filtrov, *t* = 10ms, η = 0,99	9	8,96 > 2,093

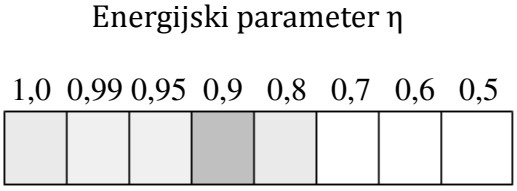
Tabela 3: Rezultati preizkušanja statistične predpostavke, da je število pravilnih razlikovanj umetnega in naravnega govora enako srednji vrednosti pri naključnem ugibanju.

Rezultati kažejo, da je pri naboru 393 filtrov, časovnem razmiku med okvirji $t = 5\ ms$, vrednostjo $\eta = 0{,}95$ ter pri časovnem razmiku med okvirji $t = 10\ ms$ in vrednostjo $\eta = 0{,}95$,

poslušalec še razločeval med naravnim in umetno sestavljenim govorom, pri istem naboru filtrov, časovnem razmiku med okvirji $t = 5\ ms$ in vrednostjo $\eta = 0{,}99$ pa ne več.

Prikazani rezultati raziskav dodatno potrjujejo hipotezo modela govorne cevi, da lahko obliko govorne cevi pri človeku obravnavamo kot nespremenljivo za čase, ki so krajši od 10 ms.

Enak slušni test [34] smo opravili tudi s HNM modelom, ki je bil razvit prav posebej za sisteme za sestavo govora z lepljenjem. Model smo želeli uporabiti namesto SG modela, ker je za njegovo krmiljenje potrebno mnogo manjše število značilk. Že pri prvih poskusih smo s poslušanjem ugotovili, da je umetni govor, ki smo ga dobili s pomočjo HNM sicer zelo visoke kakovosti, vendar se ga zaradi šibko zaznavnih motenj v nezvenečih delih govora da ločiti od naravnega govora. Zaradi tega smo se odločili, da izmerimo s kakšno zanesljivostjo lahko ločimo naravni in umetni govor in kako vpliva spreminjanje števila uporabljenih harmonskih komponent osnovne frekvence na kakovost govora. Za slušni test smo sestavili vzorce umetnega govora, ki so vsebovali različen energijski prispevek najmočnejših harmonskih komponent osnovne frekvence. Energijski prispevek η se je gibal v razponu $0{,}5 \leq \eta \leq 1$.

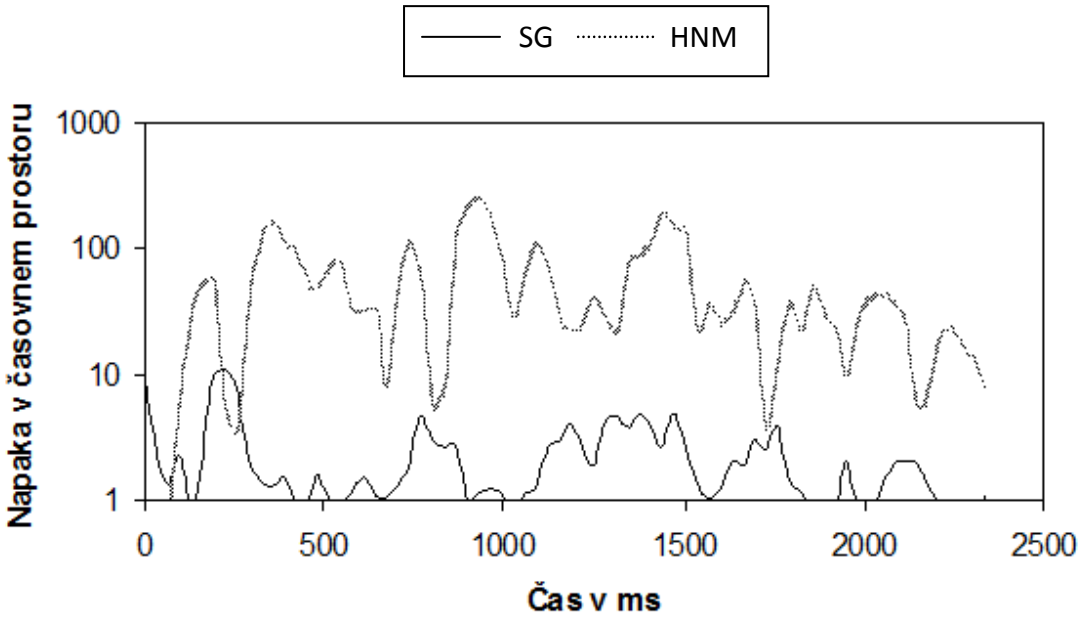


Slika 29: Ločljivost naravnega in umetno sestavljenega govora s pomočjo HNM modela v odvisnosti od energijskega prispevka η.

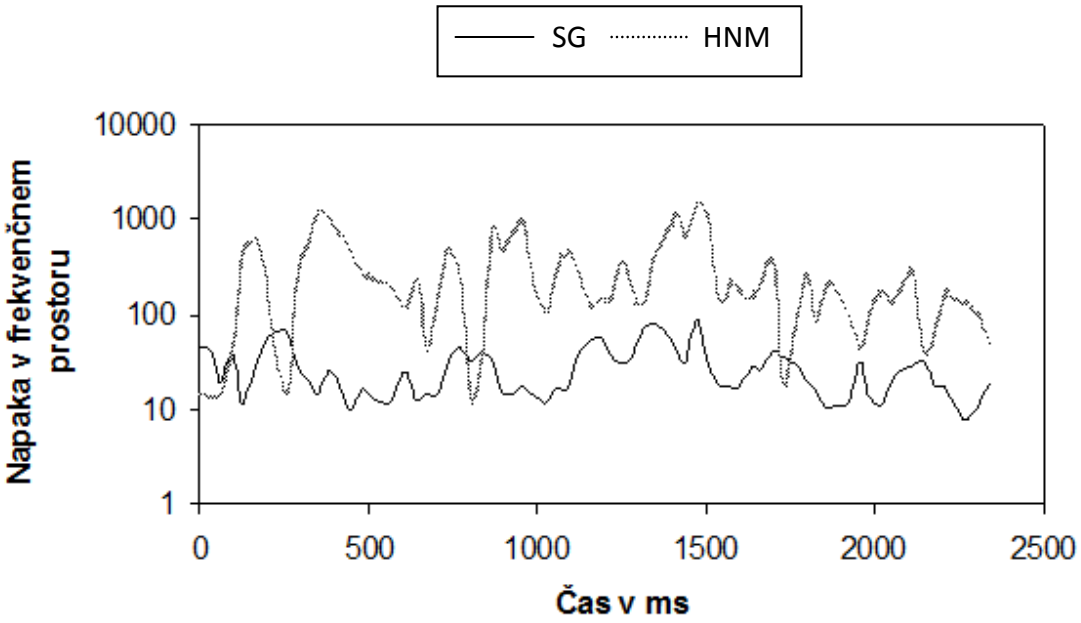
Slika 29 prikazuje stopnje ločevanja naravnega in umetno sestavljenega govora s pomočjo HNM modela. Bela barva označuje področja energijskega prispevka η, pri katerih so poslušalci pri vseh stavkih pravilno ločili umetno sestavljen in naravni govor. Področja različnih odtenkov sivin pa označujejo bolj ali manj zanesljivo ločevanje umetno sestavljenega in naravnega govora. Verjetnost napake je bila povsod manjša od 0,1.

Primerjali smo tudi napake obeh modelov. Napako smo izračunali kot kvadrat razlike med amplitudo vzorcev umetnega in naravnega govora (v časovnem prostoru) in kot kvadrat razlike amplitudnih spektrov (v frekvenčnem prostoru).

Slika 30 prikazuje napako SG in HNM modela po posameznih okvirjih govora v časovnem prostoru, slika 31 pa še napaki za oba modela v frekvenčnem prostoru. V obeh primerih je napaka pri SG modelu za velikostni razred manjša od napake HNM modela, kar seveda pojasni večjo kakovost govornega signala SG modela.



Slika 30: Napaka SG in HNM modelov po posameznih okvirih v časovnem prostoru; pri obeh modelih smo uporabili značilke, ki dajo najkakovostnejši govor. Napaka je podana kot kvadrat razlike med amplitudo vzorcev umetnega in naravnega govora.

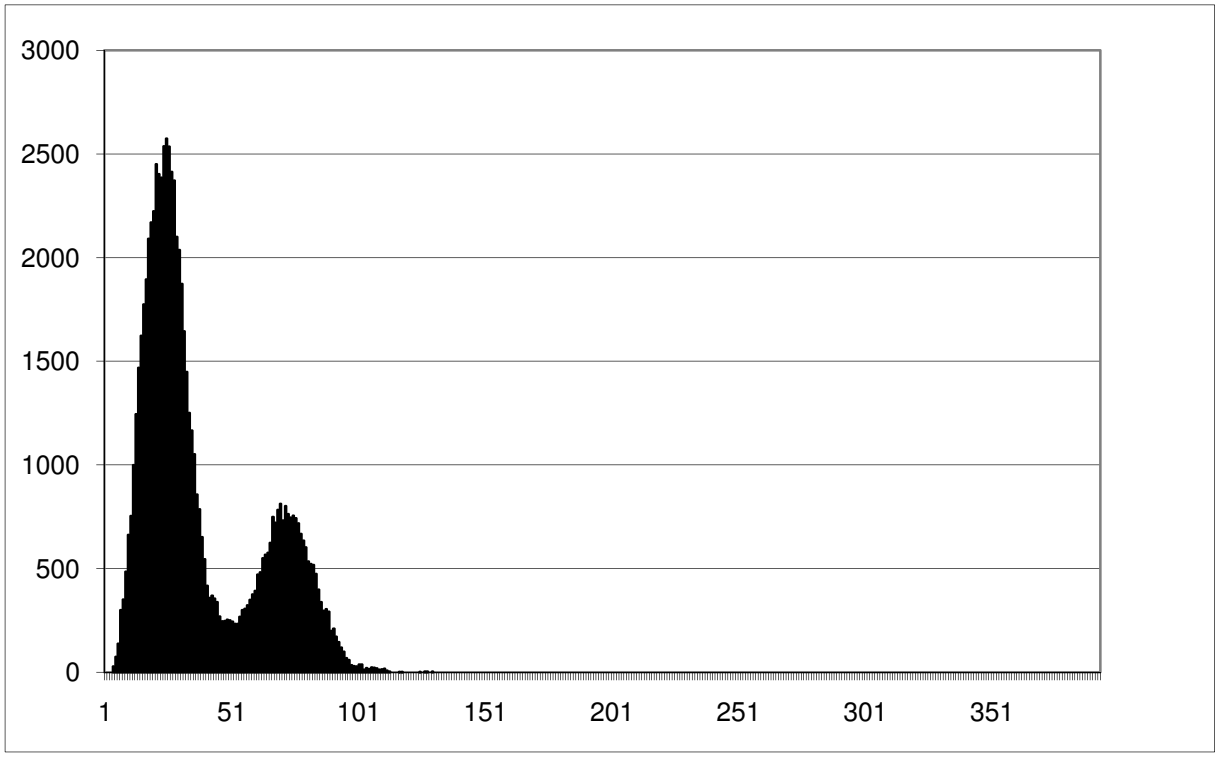


Slika 31: Napaka SG in HNM modelov po posameznih okvirjih v frekvenčnem prostoru; pri obeh modelih smo uporabili parametre, ki dajo najkakovostnejši govor. Napaka je podana kot kvadrat razlike amplitudnih spektrov

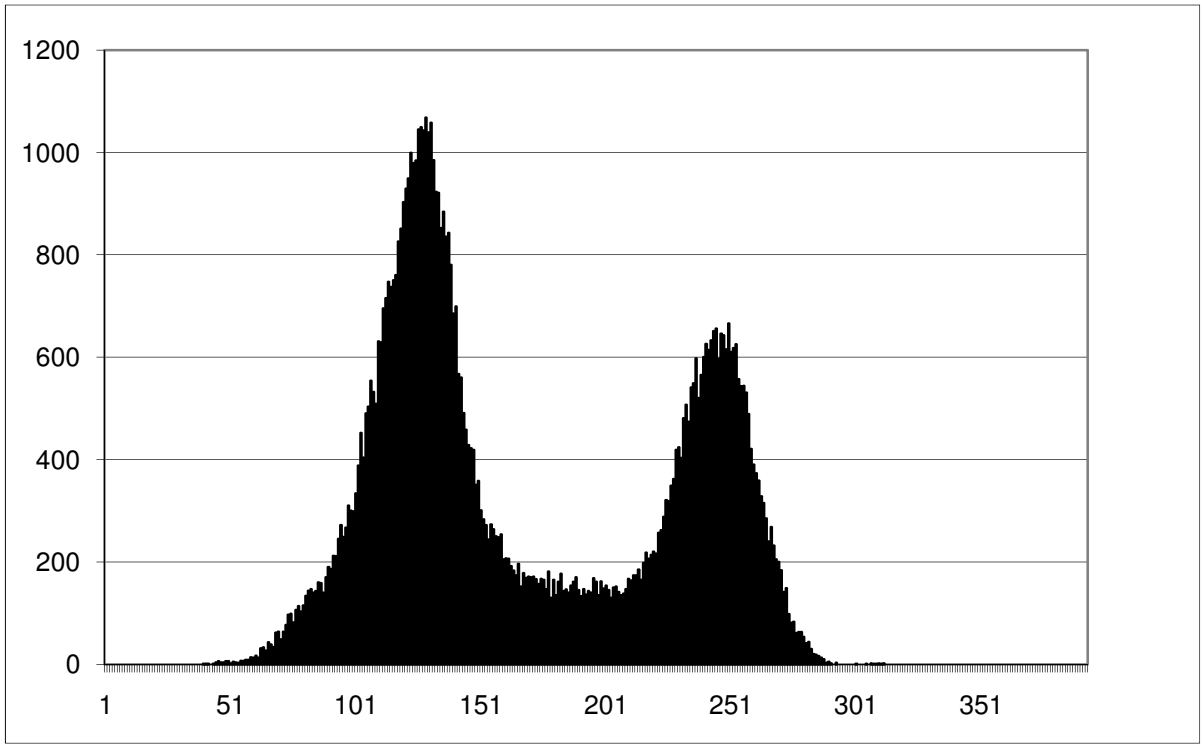
Glede na rezultate raziskav je glavna pomanjkljivost SG modela veliko število značilk, ki so potrebne za doseganje naravne kakovosti govornega signala. Na slikah od 32 do 34 lahko vidimo porazdelitve števila značilk, ki so potrebne, da zajamemo različne energijske prispevke (η=0,5, η=0,9, η=0,99) v okvirjih govornega signala. Analizirali smo 390 sekund govornega signala (78000 okvirjev govornega signala po 5 ms), grafi na slikah pa prikazujejo števila okvirjev, kjer smo posamezen energijski prispevek η dosegli s številom značilk, ki se je gibalo med 1 in 392.

Na prvih dveh slikah lepo vidimo dva vrhova v porazdelitvi, ki ju povzročajo zveneči in nezveneči glasovi. Večji del energije signala pri zvenečih glasovih je namreč porazdeljen na manjše število frekvenčnih komponent, ki sestavljajo formante, pri nezvenečih glasovih, ki so podobni obarvanemu šumu, pa je energija bolj enakomerno porazdeljena čez celo frekvenčno področje posnetkov govornega signala (od 70 Hz do 7930 Hz), ki smo jih uporabili pri meritvah.

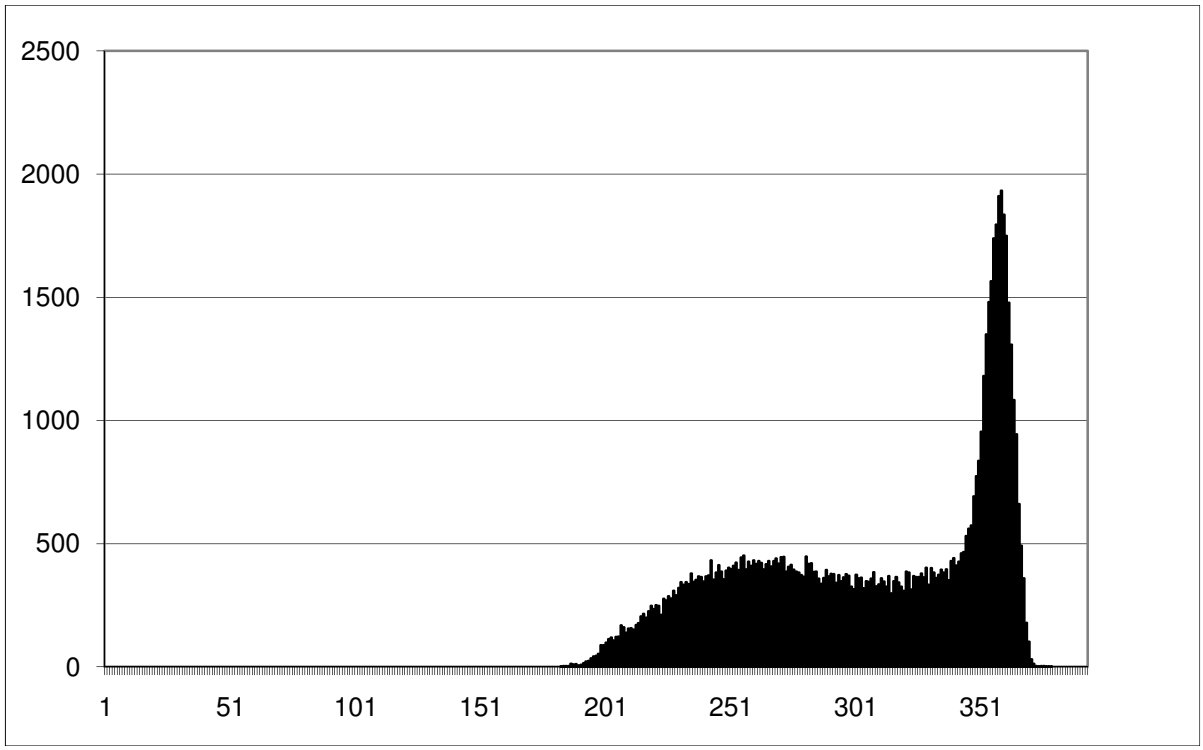
Če želimo pri sestavi govora uporabiti toliko frekvenčnih komponent, da skupaj dosegajo 99% energije signala, s čimer tudi dosežemo takšno kakovost umetnega govora, da ga ne moremo več ločiti od naravnega, lahko na podlagi grafa na sliki 34 sklepamo, da mora umetni govorni signal, ki dosega visoko stopnjo naravnosti, vsebovati tudi zaznaven delež različno obarvanega šuma. Za tvorbo šuma potrebujemo namreč veliko število sinusnih komponent.



Slika 32: Porazdelitev števila sinusnih komponent SG modela v okvirjih govornega signala dolžine 5ms za η=0,5.



Slika 33: Porazdelitev števila sinusnih komponent SG modela v okvirjih govornega signala dolžine 5ms za η=0,9.



Slika 34: Porazdelitev števila sinusnih komponent SG modela v okvirjih govornega signala dolžine 5ms za η=0,99.

5.2 Rezultati sestave s sistemom NGS

Delovanje sistema NGS smo preizkusili na manjši bazi stavkov oziroma zaporedij besed, ki je podana v prilogi E. Pri sestavi učne baze smo pazili, da v njej ni besed iz testnega primera, in za vsak stavek oziroma zaporedje besed smo posneli ustrezni govorni signal. Osnovna glasovna enota baze je bilo glasovno zaporedje samoglasnik – neznačilni glasovi – samoglasnik. Učna baza je skupno vsebovala 22 različnih posnetkov v skupni dolžini 34 sekund, ki jih je sistem razrezal v 139 glasovnih enot. Po učenju sistema z zapisi besedil in posnetki govornega signala smo sestavili umetni govorni signal za testno zloženo poved: "Nekaj časa je naporno premišljal, nato pa se je domislil sijajne razlage." Navedena poved je sestavljena iz 25 različnih glasovnih enot, izgovorjava povedi pa traja 7 sekund.

Umetni govor, ki ga je sistem NGS tvoril, je bil dobro razumljiv. Ker pa je bil za večino glasovnih enot na razpolago samo en vzorec in pri sestavi nismo izvajali nobenih dodatnih glajenj na spojih enot, se pozna nekoliko robotski zven in zavijanje glasov. WAV datoteka z vzorci umetnega govora se nahaja v datoteki 'NGS.WAV' na CD-ju v prilogi. Za primerjavo so na CD-ju tudi govorni signali treh drugih sistemov, ki sodijo v svetovni vrh sistemov za sestavo umetnega govora in imajo na spletni strani javno dostopne demo različice [12], [13], [14]. Za sistem Festival (datoteka 'FESTIVAL_NEKAJ_CASA.WAV') smo uporabili angleškega govorca (Mike), ostala dva sistema (ATT in Cepstral) pa sta imela na voljo glas italijanske govorke in smo ga tudi uporabili, ker je italijanščina glasovno mnogo bližje slovenščini kot angleščina. Za sistema ATT in Cepstral smo za primerjavo kakovosti sestave na CD poleg posnetkov stavka v slovenščini (datoteki 'ATT_NEKAJ_CASA.WAV' in 'CEPSTRAL_NEKAJ_CASA.WAV') priložili tudi posnetek stavka "Buon giorno a tutti." v italijanščini (datoteki 'ATT_BUON_GIORNO.WAV' in 'CEPSTRAL_BUON_GIORNO.WAV').

6 Zaključki in nadaljnje delo

V doktorski disertaciji smo predstavili sistem za tvorbo umetnega govornega signala NGS (Nauči se Govoriti Sam), ki se nauči govoriti sam iz vzorcev besedil in pripadajočih posnetkov naravnega govornega signala. Bistvena prednost predstavljenega sistema pred ostalimi učljivimi sistemi za sestavo govora je, da za njegovo učenje ne potrebujemo že naučenih razpoznavalnikov ali drugih delujočih sestavljalnikov govornega signala. Za samodejni razrez govornega signala smo razvili nov algoritem, ki ne skuša določiti položaja vseh glasov v govornem signalu, ampak le položaj značilnih glasov jezika. Kot značilne glasove smo definirali tiste glasove jezika, ki jih tvorijo govorila v stabilnih stanjih in so spektralno dovolj lahko ločljivi med seboj in ostalih neznačilnih glasov.

Kot nizkonivojski generator vzorcev govornega signala sistem uporablja sinusni generator, ki sicer glede procesorskih in pomnilniških zahtev ni optimalen, smo pa s slušnimi testi uspeli pokazati, da lahko tvori govorni signal, ki ga ni mogoče ločiti od naravnega govornega signala. S tem smo zagotovili, da se lahko v nadaljnjih raziskavah osredotočimo na ostale dele sistema in na njihov vpliv na kakovost govornega signala, saj sama nizko nivojska sestava signala ne vnaša slušno zaznavnih napak v sistem NGS.

Predvidevamo, da bomo pri nadaljnjih raziskavah lahko uporabljali bazo 4000 posnetih stavkov slovenskega jezika, ki naj bi bila predvidoma dokončana v letu 2011 in ki pokriva večino glasovnih značilnosti slovenskega jezika. Vsebuje vse dvoglasnike in večino triglasnikov, posneti stavki pa so izbrani na podlagi rezultatov projekta FidaPLUS [6].

Nadaljnje raziskave bomo posvetili iskanju prozodičnih značilk, ki najbolj vplivajo na kakovost umetnega govornega signala. Pri tem bomo skušali najti tudi metode za njihovo samodejno pridobivanje iz naravnega govornega signala. Veliko je še tudi možnosti izboljšav pri algoritmu za samodejno rezanje. Med drugim predvidevamo, da bomo še izboljšali natančnost razreza s tem, ko bomo vanj vključili sprotno merjenje napake na testnih zaporedjih glasov oziroma besed. Algoritem bo takoj po razrezu iz razrezanih glasovnih enot sestavil krajše zaporedje glasov umetnega govora in ga primerjal s posnetki naravnega govora. Če bo v razrezanih glasovnih enotah kakšen glas premalo ali preveč, glede na znakovni zapis glasovne enote, bo od neke točke naprej razlika med naravnim in umetnim govorom skokovito porasla in na podlagi te informacije bo mogoče ustrezno popraviti meje razreza.

Eden od pomembnih ciljev nadaljnjega dela je tudi nadgradnja prototipne različice sistema NGS, ki je bila narejena med izdelavo doktorske disertacije, v polno zmogljiv sistem.

6.1 Prispevki k znanosti

1. Razvili smo nov algoritem za samodejni razrez govornega signala v zaporedje glasovnih enot. Algoritem predstavlja osnovo idejo doktorske disertacije, na kateri je potem zgrajen sistem za sestavo govora NGS.

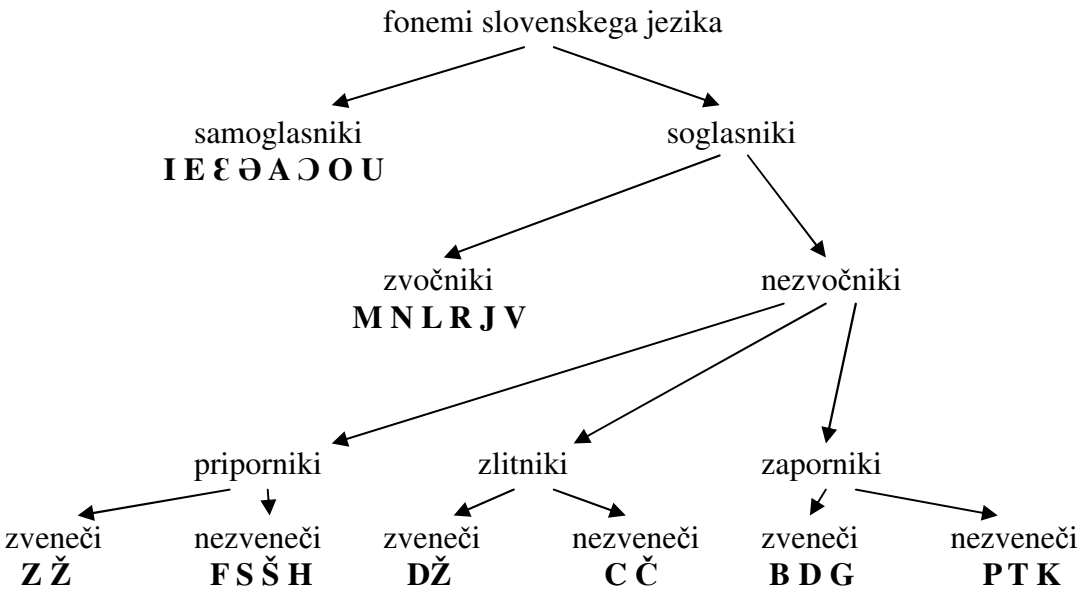
2. Razdelali smo postopke za izgradnjo sistema NGS, ki se samodejno nauči govoriti iz besedil in pripadajočih posnetkov govora, ter pri tem kot bistveni del uporablja algoritem iz točke 1. Postopki izgradnje sistema ne potrebujejo že naučenih razpoznavalnikov, ali že zgrajenih sestavljalnikov govora in omogočajo izgradnjo sistema tako rekoč iz nič, kar pomeni novost na področju računalniške sestave govora.

3. Kot stranski rezultat izgradnje sistema NGS smo s pomočjo statističnih testov pokazali, da lahko s čistim sinusnim modelom tvorimo umeten govorni signal, ki ga ne moremo ločiti od naravnega govora. S tem smo nakazali, da obstaja način tvorbe umetnega govora, ki zveni čisto naravno.

A. Razlage nekaterih pojmov

Dvoglasnik	Dvoglasnik (diphone) je določen kot par sosednjih fonemov, ki se začne na sredini prvega fonema in konča na sredini drugega fonema.
Fonem	Fonem ali glasnik je najmanjša glasovna enota, s katero govorci določenega jezika razlikujejo pomen besed. S spremembo enega glasnika v besedi vedno dobimo drugo besedo ali pa postane beseda nerazpoznavna.
Formant	Formant je okrepljen snop harmoničnih tonov in je posledica resonanc govorne cevi.
Frekvenčna steza	S frekvenčno stezo označujemo počasno spreminjanje frekvence sinusne komponente v sinusnem modelu za sestavo govora.
Glas	Glas ali alofon je uresničitev fonema v govoru. Istemu fonemu lahko v posameznem jeziku ustreza več glasov (alofonov). V slovenščini na primer fonemu /v/ ustrezajo štirje glasovi: zobnoustnični v, dvoglasni u ter zveneči in nezveneči ustničnoustnični w.
Govorna cev	Govorna cev je votli prostor od grla navzven do ustnic oziroma nosnic. Zajema goltno, ustno in nosno votlino.
Značilka	Značilka (atribut) je opis neke značilne lastnosti sistema, ki je lahko podan kot številska vrednost ali tudi kot simbolni zapis.

B. Fonemi slovenskega jezika



Dodatna pojasnila:

Ɛ – široki e,
Ə – polglasnik,
Ɔ – široki o.

C. Seznam pravil za računanje dolžin glasov angleškega jezika

Kot zanimivost si poglejmo množico deterministčnih pravil za izračun dolžine trajanja glasov angleškega jezika, ki jih je razvil Dennis Klatt za uporabo na MITalk formantnem sestavljalniku.

Trajanje je podano z dvema vrednostma in sicer z minimalnim in inherentnim trajanjem. Inherentno trajanje je trajanje glasu izgovorjenega v nevtralnem načinu govora. Enačba za izračun dolžine trajanja je podana kot:

trajanje = (obvezno trajanje – minimalno trajanje) * A + minimalno trajanje.

Količnik A izračunamo z zaporedno uporabo množice pravil, kjer vsako pravilo količnik A pomnoži z nekim dodatnim količnikom. Pravila so pa naslednja:

Raztegovanje na koncu stavka	Če je odsek, ki ga tvorimo, samoglasnik, ali pa je v repu zloga, ki je na koncu stavka, potem A=1,4*A.
Skrajševanje – ne na koncu izraza	Če odsek ni v zadnjem slogu izraza, potem A=0,6*A. Če je odsek na koncu izraza jezičnik po samoglasniku, ali nosnik, potem A=1,4*A.
Skrajševanje – ne na koncu besede	Če odsek ni v končnem zlogu besede, potem A=0,85*A.
Večzložno skrajševanje	Če je samoglasnik v večzložni besedi, potem A=0,80*A.
Soglasniško skrajševanje – ne na začetku besede	Če samoglasnik ni na začetku besede, potem A=0,85*A.
Nenaglašeno skrajševanje	Nenaglašeni odseki so dvakrat bolj stisljivi, zato velja minimalno_trajanje = minimalno_trajanje/2. Če je odsek nenaglašen ali skrajšan, nastavi A kot: • samoglasnik v zlogu na sredi besede A=0,5*A, • ostali samoglasniki A=0,7*A, • jezičnik ali drsnik pred samoglasnikom A=0,1*A, • vsi ostali A=0,7*A.
Naglašenost zlogov	Če je odsek v naglašenem zlogu, potem A=1,4*A.
Vpliv soglasnikov, ki sledijo samoglasnikom	Soglasnik spremeni predhodni samoglasnik na naslednje načine: • če ni nobenega soglasnika več - konec besede - potem A=1,2*A, • če je zveneč pripornik, potem A=1,6*A,

	• če je zveneč zapornik, potem A=1,2*A, • če je nosnik, potem A=0,85*A, • če je nezveneč zapornik, potem A=0,7*A, • vsi ostali A=A. Ti učinki so manjši v položajih, ki se ne nahajajo na koncu izraza, kjer velja A=0,7+0,3*A.
Skrajševanje v gručah	• Samoglasnik pred samoglasnikom A=1,2*A. • Samoglasnik za samoglasnikom A=0,7*A. • Soglasnik obkrožen s soglasniki A=0,5*A. • Soglasnik pred soglasnikom A=0,7*A. • Soglasnik za soglasnikom A=0,7*A.

D. Osnovni tehnični podatki uporabljene opreme za snemanje in predvajanje govornega signala

Za zajem govornega signala smo povsod, razen tam, kjer je označeno drugače, uporabljali zunanji, na USB vmesnik priključen vzorčevalnik z naslednjimi tehničnimi značilnostmi:
- podprte vzorčevalne frekvence: 44.1, 48, 88.2, 96, 176.4, 192kHz (urin signal se tvori neposredno s pomočjo kvarčnega kristala - ni pretvarjanja vzorčevalne frekvence),
- 32-bitno procesiranje signalov,
- vmesnik USB 2.0 Hi-Speed,
- 24-bitna kvantizacija signala pri vseh frekvencah vzorčenja,
- stereo vhod in izhod pri vseh frekvencah vzorčenja,
- zelo nizko potresavanje (jitter) urinega signala < 100ps RMS.

Tehnične značilnosti vgrajenega mikrofonskega predojačevalnika:
- ultra nizkošumni combo predojačevalnik z uravnoteženim (balanced) in neuravnoteženim (unbalanced) vhodom,
- A/D pretvornik AK5385,
- maksimalni dovoljen nivo vhodnega signala: +6,5dBV (+8,7 dBU),
- frekvenčni odziv: (pri minimalnem ojačenju, 20Hz-20kHz) +0.0/-0.16dB,
- dinamični obseg: (A-utežen, 1kHz, pri minimalnem ojačenju) 113dB,
- razmerje signal/šum (S/N): (A-uteženo, minimalno ojačenje) 113dB,
- skupna harmonska popačenja plus šum (THD+N): (1kHz pri -1dBFS, minimalno ojačenje): -103dB (0,0007%),
- vhodna impedanca: 1,5 Kohm,
- presluh med kanali: (1 kHz, minimalno ojačenje, -1dBFS) < -110dB
- maksimalno ojačenje signala: +60 dB.

Tehnične značilnosti visoko-impedančnega linijskega vhoda:
- vhodna impedanca: 1 Mohm,
- maksimalni dovoljen nivo vhodnega signala: +12dBV (+14,2 dBU),
- dinamični obseg: (A-utežen, 1kHz, pri minimalnem ojačenju) 113dB,
- razmerje signal/šum (S/N): (A-uteženo, minimalno ojačenje) 113dB,
- skupna harmonska popačenja plus šum (THD+N): (1kHz pri -1dBFS, minimalno ojačenje): -101dB (0,0009%).

Tehnične značilnosti linijskega izhoda:
- uravnotežen, AC-sklopljen, dvopolni nizkoprepustni diferenčni izhodni filter,
- D/A pretvornik: CS4392,
- maksimalni izhodni nivo signala: +6,7dBV,
- frekvenčni odziv : (20Hz - 20kHz) 0,00/-0,01dB,
- dinamični obseg: (1kHz, A-utežen) 111dB,
- razmerje signal/šum (S/N): (A-uteženo) 112dB,
- skupna harmonska popačenja plus šum (THD+N): (1kHz pri -1dBFS) -98dB (0,0013%)

– presluh med kanali: (1kHz pri -1dBFS) < -120 dB.

Tehnične značilnost ojačevalnika za slušalke:
– tip: ojačevalnik v razredu A,
– D/A pretvornik: CS4392 (iz linijskega izhoda),
– maksimalno ojačenje: 60dB,
– maksimalna izhodna moč: 16mW,
– izhodna impedanca: 22 ohmov,
– frekvenčni odziv: (20Hz–20kHz) +0,02/-0,08dB,
– dinamični obseg: (A-utežen) 110dB,
– razmerje signal/šum (S/N): (A-uteženo) 108dB,
– skupna harmonska popačenja plus šum (THD+N): (1kHz, maksimalno ojačenje, impedanca bremena 300 ohmov) -98dB (0.0013%),
– presluh med kanaloma: (1kHz pri -1dBFS, impedanca bremena 300 ohmov) < -91dB.

Pri snemanju in poslušanju signala smo uporabljali naglavni komplet s slušalkami in z na pomični ročici vgrajenim usmerjenim kondenzatorskim mikrofonom. Osnovne tehnične značilnosti kompleta:
– frekvenčni obseg (slušalke): 18 Hz – 22000 Hz,
– ustvarjen zvočni tlak: 114 dB,
– frekvenčni obseg (mikrofon): 80 Hz – 15000 Hz,
– občutljivost (mikrofon): -38dB.

Mikrofon je bil med snemanjem od ust oddaljen 5 cm in poravnan na sredino ust.

E. Seznam stavkov oziroma zaporedij besed v testni bazi sistema NGS

	Običajni zapis	Zapis razreza
1.	Neravne podlage.	Ne-ra-vne-po-dla-ge.
2.	Nekateri nesmisli so.	Ne-ka-te-ri-ne-smi-sli-so.
3.	Pastir in sejalec.	Pa-sti-ri-nse-ja-le-c.
4.	Domači kruh.	Do-ma-či-kru-h.
5.	Barvna niansa.	Ba-rvna-ni-a-nsa.
6.	On je utopil sivo miš.	On-je-uto-pi-lsi-vo-mi-š.
7.	Zakaj čakaš.	Za-ka-jča-ka-š.
8.	Nasadil je sekiro.	Na-sa-di-lje-se-ki-ro.
9.	Melasa in kakav.	Me-la-sa-inka-ka-v.
10.	Razvajena deklica.	Ra-zva-je-na-de-kli-ca.
11.	Vremenska napoved.	Vre-me-nska-na-po-ve-d.
12.	Vzorno vedenje.	Vzo-rno-ve-de-nje.
13.	Leno premikanje.	Le-no-pre-mi-ka-nje.
14.	Le kaj si domišljamo.	Le-ka-jsi-do-mi-šlja-mo.
15.	In je postal.	Inje-po-sta-l.
16.	Nakladal je cel dan.	Na-kla-da-lje-ce-lda-n.
17.	Na lopato je dal pesek.	Na-lo-pa-to-je-da-lpe-se-k.
18.	Mast in zaseka.	Ma-sti-nza-se-ka.
19.	Vejevje in deževje.	Ve-je-vje-inde-že-vje.
20.	Ledolomilec Arktika.	Le-do-lo-mi-le-cA-rkti-ka.
21.	Lapajne Mitja.	La-pa-jne-Mi-tja.
22.	Cijazla na kolesu.	Ci-ja-zla-na-ko-le-su.

F. Literatura

[1] CHEVEIGNE Alain de, Kawahara Hideki;YIN, a fundamental frequency estimator for speech and musica, J. Acoust. Soc. Am. 111 (4), April 2002.

[2] DONOVAN Robert E.; Trainable Speech Synthesis, Ph.D. Thesis, University of Cambridge, England, 1996.

[3] EIDE, E. et al; Recent Improvements to the IBM Trainable Speech Synthesis System, Proc. ICASSP 2003, Hong Kong, Volume 1, pp. 708-711.

[4] FANT, G., Liljencrants, J., and Lin, Q.; A four parameter model of vocal flow, STL-QPSR 4/1985, pp. 1-13.

[5] FERNANDEZ Santiago, Graves Alex, Schmidhuber Jürgen; Phoneme recognition in TIMIT with BLSTM-CTC, Technical Report No. IDSIA-04-08, April 15, 2008.

[6] FidaPlus – korpus slovenskega jezika; http://www.fidaplus.net/Info/Info_index.html.

[7] FLANAGAN James L., Allen Jont B., Hasegawa-Johnson Mark A.; Speech Analysis Synthesis and Perception, Third Edition, 2008

[8] GRIFFIN D.W., Lim J.S.; Multiband excitation vocoder, IEEE Transactions on Acoustics, Speech and Signal Processing, Aug. 1988.

[9] HAMON Christian, Moulines Eric, Charpentier Francis, "A Diphone Synthesis System Based on Time-Domain Prosodic Modifications of Speech", Proc. Int. Conf. ASSP, 238-241, 1989.

[10] http://ai.ijs.si/govorec/

[11] http://www.alpineon.com/proteus/test/

[12] http://www.cepstral.com/demos/.

[13] http://www.cstr.ed.ac.uk/projects/festival/morevoices.html.

[14] http://www2.research.att.com/~ttsweb/tts/demo.php.

[15] HUANG Xuedong, Acero A., Hon H., Ju Y., Liu J., Meredith S., Plumpe M.; Recent improvements on Microsoft's trainable text-to-speech system-Whistler, Proc. ICASSP, Vol. 22, pp. 959 - 962, April 1997.

[16] HUNT, A.J., Black A.; Unit selection in a concatenative speech synthesis system using a large speech database, Proceedings of the International Conference On Speech and

Language Processing 1996

[17] JELINEK Frederick; Statistical Methods for Speech Recognition, The MIT Press, 2nd Edition, 1999.

[18] KELLER E., Bailly G., Monaghan A., Terken J., Huckvale M. – editors; Improvements in Speech Synthesis, COST258: The naturalness of Sythetic Speech, Wiley, 2002.

[19] KLATT D.H.; Software for a Cascade/Parallel Formant Synthesiser, Journal of the Acoustical Society of America", Vol. 67, No.3, pp.971-995, 1980.

[20] KLATT D.H.; Review of Text-to-Speech Conversion for English, Journal of the Acoustical Society of America, Vol. 82, No.3, pp.737-793, 1987.

[21] MCAULAY Robert J., Quatieri Thomas F., Speech Analysis/Synthesis Based on a Sinusoidal Representation, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34,No. 4, August 1986, pp. 744-754.

[22] MITCHELL Tom M.; Machine Learning, McGraw Hill, 1997, ISBN:0071154671.

[23] MOULINES Eric, Charpentier Francis, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Speech Communications, No. 9, 1990, pp. 453-467.

[24] OPPENHEIM Alan V., Schafer Ronald W., Buck John R.; Discrete-Time Signal Processing (2nd Edition), Prentice-Hall, 1999, ISBN:0137549202.

[25] QUATIERI Thomas F., McAulay Robert J., Speech Transformations Based on a Sinusoidal Representation, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No. 6, December 1986, pp. 1449-1464.

[26] RABINER Lawrence R., Schafer Ronald W.; Digital Processing of Speech Signals, Prentice-Hall, 1975.

[27] ROSENBERG, S.; Glottal pulse shape and vowel quality, J. Acoust. Soc. Am., 49, pp. 583-590, (1970).

[28] SHRIKANTH Narayanan, Abeer Alwan; Text to Speech Synthesis: New Paradigms and Advances, Prentice Hall, 2004, ISBN: 013145661X.

[29] SERRA Xavier, Smith Julius; A Sound Analysis/Synthesis System Based on a Deterministic Plus Stohastic Decomposition, Computer Music Journal, Vol . 14, No. 4, 1990, pp. 12-24.

[30] SULTER Arend M., Wit Hero P.; Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity,

fundamental frequency and age; J. Acoust. Soc. Am., Vol. 105, Issue 3, pp. 1965-1971, March 1999.

[31] STYLIANOU Ioannis; Harmonic plus Noise Models for Speech combined with Statistical Methods for Speech and Speaker Modification, Ph.D. Thesis, Ecole Nationale Superieure des Telecommunications, Paris, 1996.

[32] ŠONC Damjan; Generator za sintezo govornega signala : magistrsko delo, Ljubljana: [D. Šonc], 2000.

[33] ŠONC Damjan, Kodek Dušan. Generator govornega signala visoke kakovosti. V: ZAJC, Baldomir (ur.). Zbornik enajste mednarodne Elektrotehniške in računalniške konference ERK 2002, 23.-25. september 2002, Portorož, Slovenija. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, [2002], zv. B, str. 243-246.

[34] ŠONC Damjan, Rozman Robert, Štrancar Andrej. Kvaliteta umetnega govornega signala pri čistem sinusnem ter pri hibridnem modelu sestave govora. V: ZAJC, Baldomir (ur.), TROST, Andrej (ur.). Zbornik štirinajste mednarodne Elektrotehniške in računalniške konference ERK 2005, 26. - 28. september 2005, Portorož, Slovenija. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, 2005, zv. B, str. 225-228.

[35] TAYLOR Paul; Text-to-Speech Sythesis, Cambridge University Press, 2009, ISBN:9780521899277.

[36] TOKUDA, K. , Yoshimura, T. , Masuko, T. , Kobayashi, T. , Kitamura, T. ; Speech parameter generation algorithms for HMM-based speech synthesis, ICASSP 2000.

[37] TOPORIŠIČ Jože, "Slovenska slovnica", Založba Obzorja Maribor, 1976.

Stvarno kazalo