UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA

PREVERJANJE KONCEPTA ZLOGOVNEGA SINTETIZATORJA GOVORA V SLOVENŠČINI

KAJA PRAPROTNIK

2022

ZAKLJUČNA NALOGA

PRAPROTNIK

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

Preverjanje koncepta zlogovnega sintetizatorja govora v slovenščini

(Concept testing of the syllable-based text to speech synthesis in the Slovenian language)

Ime in priimek: Kaja Praprotnik
Študijski program: Računalništvo in informatika
Mentor: izr. prof. dr. Jernej Vičič
Somentor: doc. dr. Branko Kavšek

Koper, september 2022

Ključna dokumentacijska informacija

Ime in PRIIMEK: Kaja PRAPROTNIK

Naslov zaključne naloge: Preverjanje koncepta zlogovnega sintetizatorja govora v slovenščini

Kraj: Koper

Leto: 2022

Število listov: 61 Število slik: 22 Število tabel: 9

Število prilog: 1 Število strani prilog: 3 Število referenc: 35

Mentor: izr. prof. dr. Jernej Vičič

Somentor: doc. dr. Branko Kavšek

Ključne besede: sinteza govora, sinteza govora s pomočjo združevanja zlogov, sinteza s pomočjo združevanja difonov

Izvleček:
V zaključni nalogi preverjamo koncept sintetizatorja govora s pomočjo združevanja difonov oziroma zlogov. V prvem delu je predstavljena struktura sintetizatorja, njegova uporaba in vrste sintetizatorjev. V drugem delu je predstavljena implementacija programa, s katerim preverjamo, če se izboljša kakovost govora z uporabo zlogov namesto difonov. V zaključku so podane morebitne izboljšave in predlogi.

Key document information

Name and SURNAME: Kaja PRAPROTNIK

Title of the final project paper: Concept testing of the syllable-based text to speech synthesis in the Slovenian language

Place: Koper

Year: 2022

Number of pages: 61 Number of figures: 22 Number of tables: 9

Number of appendices: 1 Number of appendix pages: 3 Number of references: 35

Mentor: Assoc. Prof. Jernej Vičič, PhD

Somentor: Assist. Prof. Branko Kavšek, PhD

Keywords: text-to-speech synthesis, concatenative synthesis, syllable-based concatenative synthesis, diphone-based concatenative synthesis

Abstract:
The thesis has tested the concept of a speech synthesizer by concatenation diphones or syllables. The first part of the thesis presents the structure of the speech synthesizer, its use and types of synthesizers. In the second part, the implementation of the program is presented, with which we test, whether the quality of speech is improved by using syllables concatenation instead of diphones concatenation. Possible improvements and suggestions are given in the conclusion.

Zahvala

Rada bi se zahvalila mentorju, izr. prof. dr. Jerneju Vičiču, in somentorju, doc. dr. Branko Kavšku, za vso podporo, strokovno pomoč in usmeritve pri zaključnem delu in v času izobraževanja. Poleg tega bi se rada zahvalila svoji družini za vso izkazano podporo. Posebna zahvala gre mojemu očetu Gorazdu, ki mi je svetoval in pomagal pri študiju. Zahvaljujem se tudi Maticu in Vidu, ki sta me skozi študijska leta inštruirala.

Kazalo vsebine

Kazalo tabel

Kazalo slik

Seznam kratic

TTS	Ang. Text-To-Speech (Slo. Tekst v govor)
NLP	Ang. Natural Language processing (Enota za procesiranje naravnega jezika)
DSP	Ang. Digital Signal Processing (Slo. Enota za procesiranje digitalnega signala)
LTS	Ang. Letter To Sound (Slo. Modul za pretvorbo teksta v foneme)
F0	Ang. Fundamental frequency (Slo. Osnovni ton)
LPC	Ang. Linear Predictive coding (Slo. Večpulzno linearno napovedovanje)
HMM	Ang. Hidden Markov model (Slo. Prikriti Markovi modeli)
HNM	Ang. Harmonic plus noise modulation (Slo. Neznana beseda)
PCM	Ang. Pulse Code Modulation (Slo. Neznana beseda)
TD–PSOLA	Ang. Time Domain Pitch Synchronous Overlap-Add synthesis (Slo. Neznana beseda)
ASCII	Ang. American Standard Code for Information Interchange (Slo. Neznana beseda)

1 Uvod

Namen sintetizatorja govora (Text To Speech – TTS) je samodejna pretvorba poljubnega besedila (korpus) v besedno izgovorjavo [9].

Tehnologija umetne izdelave govora s pomočjo TTS je v široko uporabljena na različnih področjih, poleg tega pa ima skoraj vsak človek pri sebi pametni telefon, ki je zmožen podajati glasovna sporočila in opozorila.

Med pomembnejšimi uporabami sintetizatorja govora na pametnih telefonih je nastavitev za slepe in slabovidne, ki jim pretvori dotik besedila na ekranu v glasovno sporočilo, tako da lahko uporabljajo pametne mobilne naprave. Seveda pa je to le ena izmed mnogih aplikacij sintetizatorja govora. Omenimo lahko še uporabo sintetizatorja govora za učenje tujih jezikov in raznih jezikovnih prevajalnikov, ki omogočajo ljudem, da slišijo pravilno izgovorjavo besed. Za primer lahko podamo google prevajalnik, ki ima možnost, da poda glasovno obliko prevedene besede. Tehnologija sinteze govora je lahko tudi na področju učenja otrok, obstaja namreč kar nekaj igrač, ki s pomočjo sintetizatorja govora otrokom predstavijo učno snov.

Čeprav je priložnosti za uporabo sintetizatorja govora veliko, je na žalost sintetizatorjev za slovenščino malo. Večina je zastarelih in nenaravno zvenečih. Izpostavimo lahko sintetizator govora za slovenščino eBralec, ki je trenutno najboljša možnost, na voljo pa je tudi v obliki mobilne aplikacije in njegov predhodnik, sintetizator govora Govorec [15, 35].

Zaključno delo predstavlja primerjavo dveh tehnik sinteze govora:

- sinteza s pomočjo združevanja zlogov,
- sinteza s pomočjo združevanja difonov.

Raziskava je namenjena predvsem primerjavi obeh tehnik lepljenja posameznih delov izgovorjenega besedila in smotrnosti uporabe (boljše) izbrane tehnike.

Zaključno delo je sestavljeno iz treh delov.

Prvi del se osredotoči na opis govora, naglas in slovnična pravila slovenskega knjižnega jezika. Namen tega dela je predvsem predstaviti osnove slovenskega jezika, ki jih je potrebno vedeti, če želimo izdelati sintetizator govora v slovenščini.

V drugem delu je opisana struktura sintetizatorja govora in različne tehnike, ki so uporabljene za sintezo umetnega govora. Bolj podrobno je opisana sinteza govora s

pomočjo združevanja, saj je to tehnika sinteze, ki smo jo uporabili za implementacijo našega programa.

Zaključni del se osredotoča na preverjanje koncepta sintetizatorja govora z metodo lepljenja zlogov, s pomočjo implementiranega poenostvaljenega sintetizatorja govora. Z namenom, da bi preverili katera tehnika sinteze (sinteza govora s pomočjo lepljenja difonov ali sinteza govora s pomočjo lepljenja zlogov) je bolj primerna za slovenski jezik, smo izvedli anketo, ki smo jo skupaj s programom razdelili testni skupini ljudi. Na podlagi rezultatov ankete smo preverili, ali je naša predpostavka pravilna. Na koncu smo podali še sklep zaključne naloge in možne izboljšave ter predloge.

2 Slovenski jezik in fonetika

Da lahko ustvarimo sintetizator govora, moramo poznati strukturo jezika. Potrebno je poznati izgovorjavo posameznih črk oziroma celotne besede, strukturo povedi, naglas in ostale značilnosti jezika. Ker je sintetizator govora, ki ga preučujemo v tem zaključnem delu, v slovenskem jeziku, bomo predstavili osnove slovenskega jezika, ki jih je potrebno upoštevati za generiranje funkcionalnega sintetizatorja govora.

2.1 Osnove fonetike

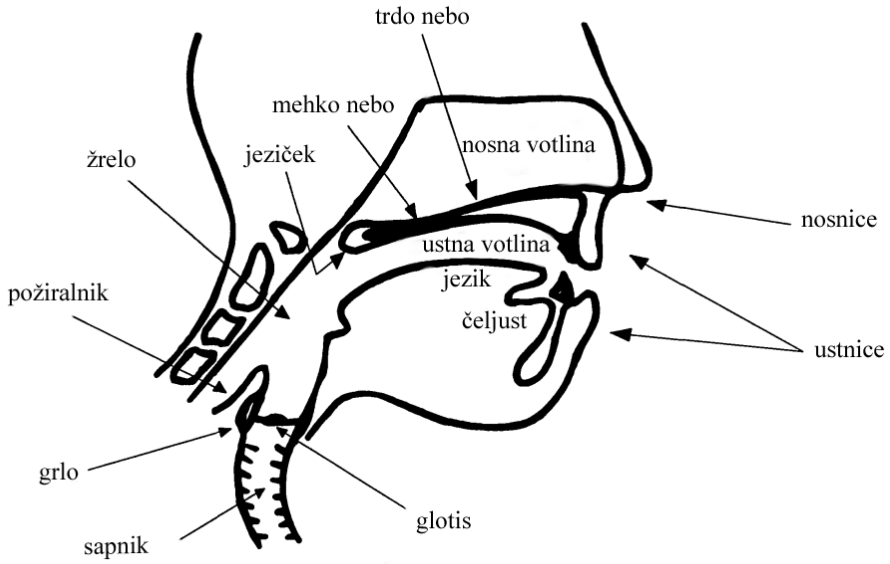
Ker delamo na sintezi govora, moramo najprej definirati, kaj je govor. Govor nastane zaradi vzbujanja človeškega govornega trakta (funkcionalno povezani govorni organi) in je odvisen od oblike goltne, ustne in nosne votline ter ostalih parametrov govornega trakta. Glasovni trakt lahko vzbujamo tako, da tvorimo zveneče ali nezveneče glasove. Pri vzbujanju zvenečih glasov glasilki nihata v enakomernih presledkih in pri tem nastane zven določene višine, poimenovan osnovni ton (F0). Ko tvorimo nezveneče glasove, sta glasilki razmaknjeni, kar pomeni, da ju zračni tok ne more vzbuditi v nihanje in je govorni trakt takrat vzbujan s turbolentnim pretokom zraka [19].

Izredno pomembno je, da sintetizator govora čim bolj pravilno in razločno pretvori besedilo v govorni jezik. Da pa to lahko stori, morajo razvijalci poznati osnove fonetike. Fonetika je izraz za glasoslovje, kar pomeni, da je predmet raziskovanja izrazna podoba glasov, besed, glasovnih zvez, besedil ... Najmanjše pomensko samostojne enote stavka so besede.

Fonem je najmanjši gradnik govora, ki loči pomen besed ali morfemov v nasprotju z drugimi glasovi. Kot primer lahko podamo glasova /p/ in /b/, ki pomensko razločujeta med besedami piti-biti, pik-bik, itd. V slovenskem knjižnem jeziku je devetindvajset fonemov [14].

Alofoni so množica fonemskih variant, saj imajo fonemi lahko več pojavnih oblik z določeno skupno značilnostjo na izgovrni in slušni ravni. Opišemo jih s pravilom glede na okolje ponavljanja. Za primer lahko podamo slovenski fonem /n/, ki se glede na to v kakšnem okolju je, izgovori drugače (primer: sani, sanke) [19].

Glasovi v besedah se načeloma razvrščajo po pravilu, da se okoli samoglasnikov zbirajo soglasniki in z njimi tvorijo zloge. V besedi je praviloma toliko zlogov, kot



Slika 1: Prerez glasovnega aparata. Glasovni trakt se začne pri glasilkah in konča pri ustnicah (Vir: [7]).

beseda vsebuje samoglasnikov. Če želimo govor razčleniti na zloge, moramo pogledati glasnost govora. Ta doseže lokalni maksimum nad jedrom zloga (samoglasnik), na začetku in na koncu zloga pa lokalni minimum [19].

Glasove našega knjižnega jezika označujemo s črkami. Slovenska abeceda je sestavljena iz petindvajsetih črk. Vsebuje pet črk za osem samoglasnikov in dvajset črk za 24 soglasnikov. Ker posebni znaki za izražanje sedmih dodatnih glasov niso del abecede, ta ni popolna. Tako s posebno črko e pišemo 3 različne glasove: široki ê, ozki é in polglasnik ə. Prav tako uporabljamo črko o za zapis širokega ô in ozkega ó; glasa i in u zapisujemo v nekaterih primerih s črkama j in v (primer: divje), za mehka glasova l in n pa pišemo črki lj in nj [3].

Kot že prej omenjeno, slovenske soglasnike delimo na zvočnike (m, n, r, l, j, v) in nezvočnike - ti so lahko zveneči (b, d, z, ž, dž, g) ali nezveneči (p, t, f, s, š, h, c, č, k). Razlika med pisnim in govorjenim jezikom se pokaže v primerih, ko zadnji različno zveneči nezvočnik prevlada (primer: odpreti), lahko tudi pride do zlitja glasov (primer: predsednik), vsak nezvočnik pred premorom pa je izgovorjen nezveneče (primer: Pojdite na odpad.) [3, 31].

V slovenščini se samoglasniki izgovarjajo jasno in navzven, tako da zvok ne ostaja v ustni/nosni votlini. Dolžino oziroma kračino samoglasnikov označujejo naglasna znamenja [31].

Za izdelavo sintetizatorja je prav tako potrebno vedeti, kako se kakšna beseda naglasi. Naglas je fonetično-fonološka lastnost zloga glede na druge zloge v govornem

Tabela 1: Samoglasniški trikotnik slovenskega knjižnega jezika (vir: [31]).

	sprednji	nesprednji	
		srednji	zadnji
visoki	i		u
sredinski	e ɛ	ə	o ɔ
nizki		a	

nizu. Velik problem pri izdelavi sintetizatorja govora je, da slovenski knjižni jezik nima stalno naglašenega zloga v besedi, ampak se je treba naglasa naučiti skupaj z besedo. Načeloma so nenaglašeni samoglasniki kratki, naglašeni pa dolgi. Za slovenščino je značilno, da ima dva tipa naglaševanja, tonemsko in jakostno. Beseda ima načeloma le en naglas – kratki ali dolgi naglašeni zlog (več kot enega imajo le nekatere zloženke). Naglas pa moramo obravnavati znotraj povedi ter v razmerju do drugih enot v stavku. V slovenskem knjižnem jeziku ločimo osem naglašenih fonemov. Za slovenski jezik je značilno, da imata e in o dva fonema. Zato imamo v slovenščini posebne znake, da označimo mesto in trajanje naglasa. V tabeli 2 in tabeli 3 so predstavljena naglasna znamenja [31].

Tabela 2: Jakostno naglaševanje (vir: [32]).

Znamenje	Ime	Zaznamuje		
´	ostrivec	mesto naglasa	dolžino	ozkost e-ja in o-ja
^	krativec	mesto naglasa	dolžino	širokost e-ja in o-ja
`	krativec	mesto naglasa	kračino	širokost e-ja in o-ja

Tabela 3: Tonemsko naglaševanje (vir: [32]).

Znamenje	Ime	Zaznamuje		
´	akut	mesto naglasa	dolžino	nizki ton
^	cirkumfleks	mesto naglasa	dolžino	visoki ton
`	gravis	mesto naglasa	kračino	nizki ton
¨	dvojni gravis	mesto naglasa	kračino	visoki ton

2.1.1 Stavčna fonetika

Dober sintetizator govora mora zveneti čim bolj naravno. To pomeni tudi, da upošteva stavčno intonatiko. Zaradi stavčne intonatike lahko razločimo poved od vprašanja,

razne govorčeve emocije itd. Stavčna intonatika se razlikuje glede na zapisano ločilo. Poznamo več ločil, ta so lahko postavljena znotraj povedi (vejica, pomišljaj ...) ali na koncu (pika, klicaj, vprašaj).

Pri piki, ki zaznamuje konec stavka, gremo z glasom navzdol (kadenčna intonacija). Če je na koncu ločilo vprašaj, gremo lahko z glasom navzdol, če vprašanje začne z vprašalnimi zaimki (kdo ..., kam ..., zakaj ..., čemu ..., čigav ...). Z glasom navzgor (antikadenčna intonacija) gremo običajno pri vprašanjih brez vprašalnice oziroma z vprašalnico ali. Prav tako gremo z glasom navzdol pri vprašanjih prve vrste in vprašanjih začudenja (primer: Kje si bil?! - Kje si bil?!). Pri končnemu ločilu pomišljaj in pri ločilu tri pike ponavadi glas pada. Pri dvopičju in podpičju se ponavadi pojavi nekončna (polkadenčna) intonacija. Vejica lahko v prirednih zvezah zaznamuje rastočo ali padajočo polkadenco, v podredjih rastočo polkadenco in pri zvalnikih, medmetih, ... padajočo polkadenco. Pri oklepajih, kjer kaj ponazarjamo, imajo pred seboj padajočo polkadenco, če pa kaj povemo, pa rastočo. Znotraj oklepaja se povedna intonacija ravna po ločilu te povedi [32].

2.1.2 Posebne glasovne zveze

V slovenščini določujemo pod posebne glasovne zveze tiste zveze, ki so na meji dveh morfemov ali besed [32].

Če sta dva enaka soglasnika sosednja fonema, potem se zlijeta eden v drugega (primer: sam misli, oddati). Če sta zobnik (soglasnik, tvorjen z vrhom jezika ob sekalcih) in za njim zlitnik v govorni verigi, potem izgovarjamo oba glasova ali pa samo ustrezni zlitnik (primer: po cekar/ pocekar). Če se v besedilu pojavijo zveze t, d ali c, č, dž z zvezami s, z, š, ž, potem izgovarjamo ali oba glasova ali le ustrezni dolgi zliti glas (primer: podse/ potse / poce). Če imamo sičnik pred šumnikom, ponavadi izgovarjamo šumevce š, ž oziroma č (primer: iz šole / is šole /išole). Namesto nenaglašenih ə + u, se knjižno izgovarja tudi u (primer: videl / vidu). Zveza i + samoglasnik (zlasti v grško-romanskih besedah) se izgovarja kot i + j + samoglasnik (primer: pacient / pacijent) [32].

2.1.3 Naglasni tipi v oblikoslovju

Naglas osnovnih oblik lahko uvrstimo v štiri naglasne tipe [32]:

- Nepremični naglasni tip na osnovi, kjer naglas ostane na istem zlogu, ne glede na obliko (primer: lip-a, lip-e).
- Premični naglasni tip na osnovi, ki ima naglas v ednini na predzadnjem zlogu osnove, če nima glasovne končnice. Če jo ima je naglas skoraj zmeraj (obstajajo

izjeme) na zadnjem zlogu osnove (primer: pleme, plemena).

- Končniški naglasni tip, kjer je naglas na kratkem zadnjem samoglasniku v osnovnih oblikah (primer: temen, temna, temno)
- Mešani naglasni tip, kjer je naglas v ednini v nekaterih oblikah na osnovi, nekje na končnici, itd. (primer: mož, moža, možu, možem)

3 Zgodovina sinteze govora in predstavitev raziskovalnega področja

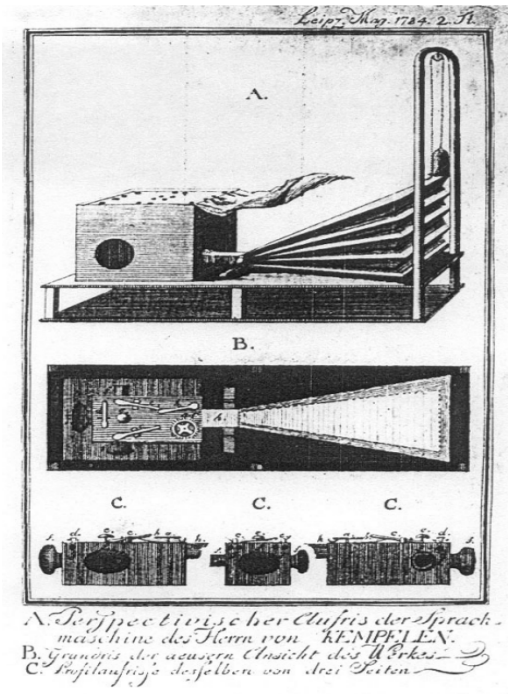
Prve zapise o poskusu izdelave govorečega stroja segajo v konec osemnajstega stoletja. Tako je leta 1779 Ch.T.Kratzenstein na Cesarski akademiji St. Petrsburg izdelal mehansko napravo, ki je proizvajala samoglasnike s pihanjem zraka skozi cev v spremenljivo resonančno komoro, ki je oponašala človeški vokalni trakt [6, 16].

Leta 1791 je W. Von Kempelen konstruiral napravo, ki je bila zmožna proizvesti številne soglasnike in samoglasnike. Tudi ta naprava je imela meh, ki je potiskal zrak skozi cev v resonančno komoro. Obliko resonančne komore pa je bilo možno spreminjati s prsti, tako da je lahko proizvajala samoglasnike. Druga komora pa je bila namenjena proizvajanju soglasnikov, zrak pa je namesto skozi tanko cev prehajal skozi vzporedno ozko tubo, kar je s pomočjo prstov, ki so zaprli "ustno votlino", povzročilo povečanje pritiska in tvorjenje samoglasnikov ob sprostitvi prstov. W. Von Kempelen velja za prvega raziskovalca fonetike. Svoja raziskovanja in opis govorne naprave je objavil v knjigi "Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine" [2, 6].

V devetnajstem stoletju so raziskovalci v glavnem izdelovali naprave, ki so delovale na podobni osnovi. Omenimo lahko Sira Charles Wheatstone, ki je zgradil izboljšano verzijo Kempelen-ga stroja [6].

Kljub temu, da proti koncu devetnajstega stoletja še ni bilo pravih pripomočkov kot npr. spektrogram, so začeli znanstveniki (H.L.F von Helmhotz) proučevati razmerja med zvokom in njegovim spektrom. Pri tem so ugotovili, da lahko s spreminjanjem relativne glasnosti na posameznih področjih spektra proizvedejo zvok, ki je podoben človeškemu govoru. S tem so postavili osnove za izdelavo sintetizatorja, ki bi generiral zvok na električni osnovi. Tako je na začetku dvajsetega stoletja J.Q. Stewart na osnovi teh spoznanj skonstruiral napravo, ki je imela dve dvojni resonančni komori, vzbujani s periodičnimi električnimi impulzi. Z uglasitvijo teh resonatorjev pri različnih frekvencah, je naprava generirala glasove, podobne samoglasnikom [6, 27].

Leta 1930 so H. Dudley, R. Reisz in S. Watkins konstruirali električno izvedbo



Slika 2: Prva slika Kempelenove govorne naprave. Naprava je bila zmožna proizvesti številne soglasnike in samoglasnike s pomočjo meha in komore, ki naj bi poskušala oponašati govorni trakt (Vir: [20]).

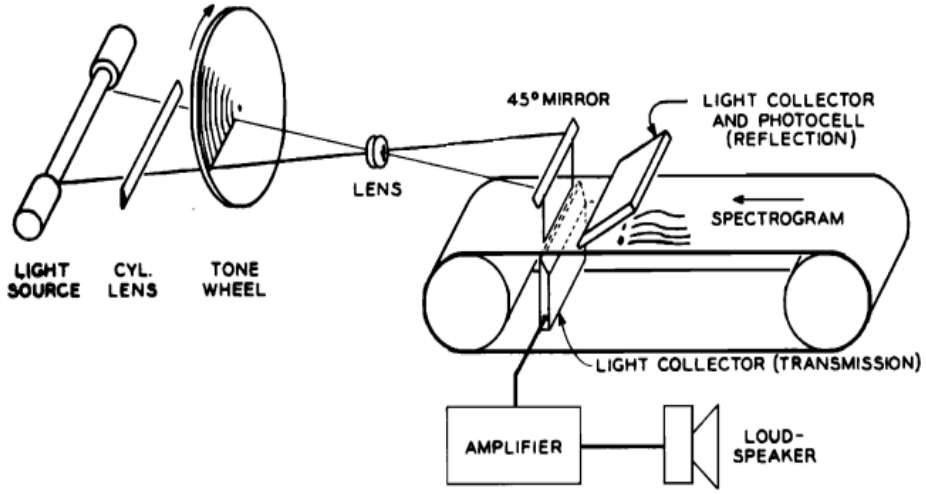
Kempelen-ovega stroja. Imenovali so jo "Voder", bila pa je razstavljena na svetovni razstavi leta 1939. Podobno, kot pri mehanskem predhodniku, je bila naprava ročno vodena preko tipkovnice, le da se je namesto oblike umetnega govornega trakta kontrolirala relativna glasnost posameznih področij spektra generiranega zvoka. Na sliki 3 lahko vidimo demonstracijo uporabe Voder [6].

Leta 1951 so se F. S. Cooper, A. M. Liberman in J.M. Borst lotili izdelave sintetizatorja na precej neobičajen način. Izdelali so napravo ("Pattern Playback"), ki je omogočila pretvarjanje vzorcev, vidnih na širokopasovnih zvočnih spektrogramih, nazaj v zvok [4].

Od 50. let 20. stoletja so raziskovalci razvijali modele sintetizatorjev, katerih je šel električni signal vira skozi filter. Izvorni signal je bil harmonični ton ali aperiodični šum [13].

Leta 1958 pa so G. Peterson, W. Wang in E. Sivertsen začeli eksperimentirati z difoni. Difoni so enote, ki zajemajo del enega in del sosednjega fonema. Po teoriji, ki so jo razvili, naj bi bili fonemi dosti bolj stabilni na sredini kot na robovih. Zato dobimo difon tako, da izrežemo polovico prvega fonema od sredine naprej in drugo polovico sosednjega fonema od začetka do polovice. Zgoraj omenjeni raziskovalci niso

Slika 3: Na sliki je gospa Harper, ki demonstrira uporabo naprave Voder. Naprava je ročno vodena preko tipkovnice (Vir: [6]).



Slika 4: Pattern Playback. Na sliki vidimo pomični filmski trak, na katerem so zapisani spektrogrami posnetega govora. Premikajoči zapisi spektrogramov na filmskem traku s pomočjo zrcal modulirajo svetlobo, ki prihaja iz tonskega generatorja v obliki kolesa in generira okoli 50 harmonično povezanih frekvenc. Te frekvence predstavljajo približek spektograma. Filtrirana svetloba se nato pretvori v zvočni signal (Vir: [13]).

izdelali popoln sintetizator govora, ampak so z nekaj besedami poskušali dokazati svoje trditve. Kljub temu, da jim je to uspelo, so naleteli na kar nekaj problemov. Nezveznost glastnosti, višine tona in spektruma povezanih segmentov je povzročala zelo nezaželene zvočne pojave, največkrat klike ali podobne zvoke. Zato so za svoj poizkus uporabili

samo difone, ki so imeli podobne akustične karakteristike [13].

Da bi lahko izdelali popolni sintetizator govora, ki bi vseboval vse difone, bi morali poskrbeti za glajenje povezav med difoni. To pa lahko storimo le takrat, ko imamo parametriziran govor. Prva naprava, ki je imela parametrizirane formante, je bila predstavljena leta 1967 [13].

Leta 1970 pa so raziskovalci naredili ogromen korak v govorni sintezi, predvsem zaradi večjih možnosti, ki so jih nudili močnejši računalniki. Najbolj znan sintetizator govora je bila digitalna izvedba sintetizatorja govora s pravili, ki ga je razvil D. H. Klatt. Le-ta je opazoval gibanje formantov, kar mu je omogočilo izdelati sintetizator visoke kakovosti. Na osnovi Klatt-ovega sintetizatorja je podjetje Digital izdelalo napravo DECTalk [13].

V Sloveniji so se raziskave na področju sinteze govora razvile relativno pozno. Prvi, ki je razvil postopek za samodejno sintezo govora slovenskega jezika, je bil Hribar. Nadaljni razvoj je potekal na Institutu Jožefa Štefana, kjer so Weilguny, Dobnikar in Štef nadaljevali raziskave. V okviru različnih projektov je bilo razvitih več sintetizatorjev govora v slovenskem jeziku:

- Govorec 1 in Govorec 2, sta sintetizatorja govora, ki uporabljata metodo sinteze govora s pomočjo lepljenja difonov. Za združevanje uporabljata postopek TD–PSOLA (Time Domain Pitch Synchronous Overlap-Add synthesis), ki temelji na deljenju signala v zaporedje kratkih prekrivajočih signalov. Govorec 2 ima bolj obširno in izpopolnjeno bazo slovarjev izgovorjav od Govorca 1 [15].

- Govorec 3 je sintetizator govora podjetja Amebis, d.o.o., ki deluje na osnovi metode HNM (angl. harmonic plus noise modulation). Bistvo te metode je dekompozicija govornega signala na harmonski del, ki se ga modelira s harmonsko sinusno sintezo, in šumni del, ki se ga modelira s filtriranjem šuma. Govorno zbirko Govorca 3, ki obsega 20 ur posnetkov, sta posnela profesionalna napovedovalca z radia [22, 23].

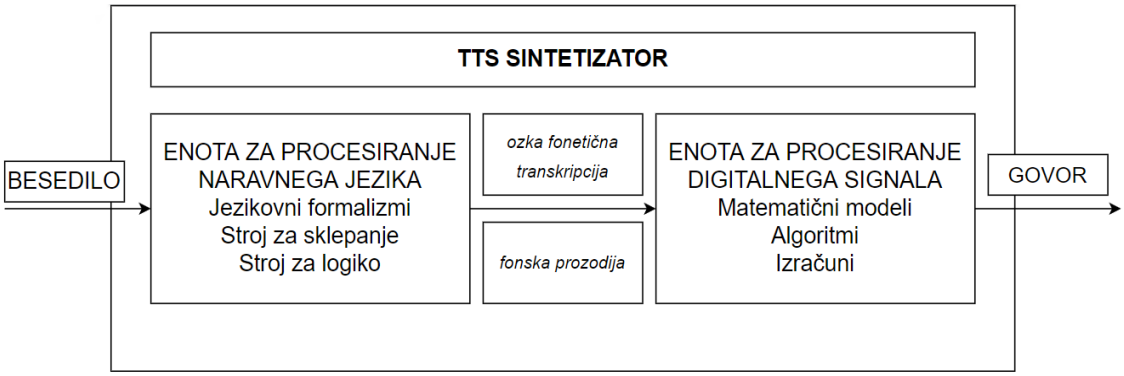
- Sintetizator govora za slovenski jezik eBralec je bil razvit v okviru projekta Knjižnica slepih in slabovidnih in je bil prvenstveno namenjen slepim in slabovidnim uporabnikom ter osebam z motnjami branja. Pri izdelavi eBralca so govorne zbirke posneli v studio RTV z desetimi profesionalnimi govorci. Govorci so imeli pri branju besedila nameščene elektrode laringografa, da so lahko spremljali nihanje glasilk za lažje označevanje osnovnih period govornega signala. Za modeliranje prozodije in tvorjenje govora so uporabili prikrite Markove modele (PMM). Postopek sinteze govora z uporabo prikritih Markovih modelov vključuje fazo učenja in fazo sinteze. Učenje so izvedli s postopkom Bauma in Welcha. [35]

Trenutno lahko brezplačno dostopajo do sintetizatorja govora eBralec na osebnih računalnikih slepi in slabovidni, osebe z motnjami branja in ustanove javnega sektorja. Za ostale pa je program plačljiv[1]. Na mobilnih telefonih z operacijskim sistemom Android pa je na voljo osnovna aplikacija eBralec, ki omogoča brezplačno pretvorbo pisanega besedila v govor za vse uporabnike. Mobilna aplikacija za razliko od plačljivega programa eBralec ne omogoča izbire različnih glasov, kakovost sintetiziranega govora pa je precej slabša.

[1]Vir: https://www.kss-ess.si/en/ebralec-sintetizator-govora-slovenskega-jezika/

4 Struktura sintetizatorja govora

Na sliki 5 vidimo funkcionalni diagram najosnovnejšega sintetizatorja TTS. Sestavljen je iz enote za procesiranje naravnega jezika (Natural Language procesing – NLP), ki skrbi za pretvorbo iz prebranega teksta v fonetični zapis z vso potrebno intonacijo [1] in ritmom [2] (prozodiko) ter enote za procesiranje digitalnega signala (Digital Signal Procesing – DSP), ki pretvori dobljen simbolični zapis v govor (Vir: [7, 29]).



Slika 5: Enostavni diagram sistema TTS. Kot vidimo iz diagrama, je sintetizator TTS sestavljen iz enote NLP in enote DSP (Vir: [7]).

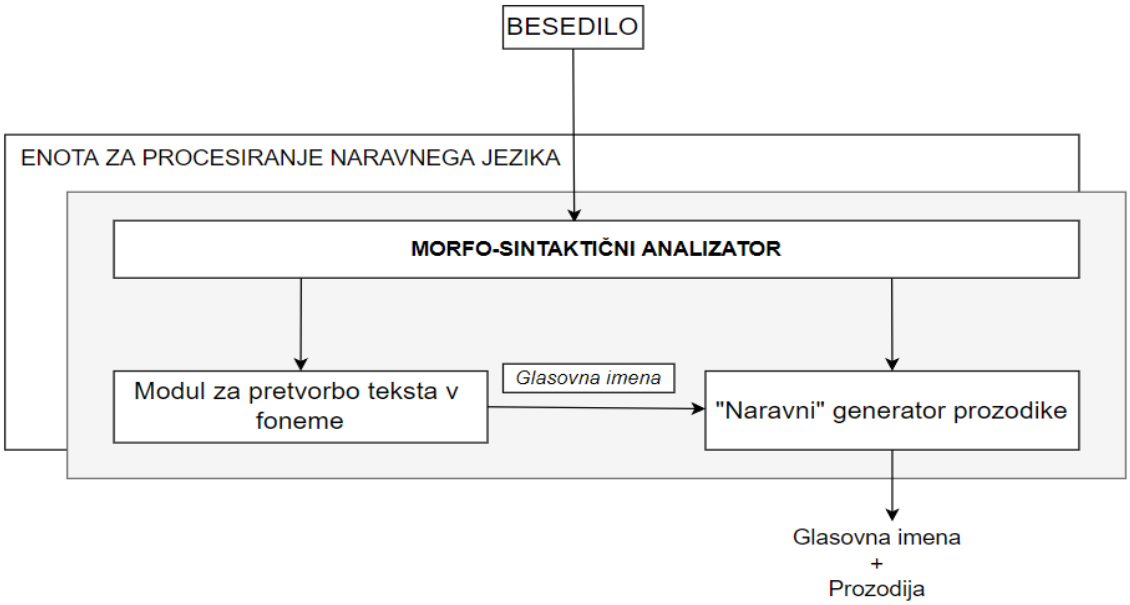
4.1 Enota za procesiranje naravnega jezika (NLP)

Kot pove že samo ime, ta enota skrbi za pretvorbo vhodnega besedila v simbolični zapis izgovorjav, ki je potreben za krmiljene DSP enote [8].

Pri izgovorjavi kakovost govora ni odvisna samo od pravilne izgovorjave besed in njenega trajanja, ampak na kvaliteto vpliva tudi pravilen poudarek, intonatika in ritem. To pa zahteva popolno razumevanje besedila in poznavanje besed. Ker tega z računalnikom seveda ne moremo doseči, je potrebno zgraditi logiko, ki bi nadomeščala potrebo po razumevanju vhodnega besedila.

[1] "jezikosl. potek osnovnega tona v posameznih segmentih povedi: intonacija povedi" [32].

[2] "lit. urejeno pojavljanje poudarjenih in nepoudarjenih ali dolgih in kratkih zlogov glede na pomensko vrednost besed v stavku: verzni ritem; ritem pesmi; ritem in metrum / jambski ritem; rastoči ritem pri katerem so poudarki na koncu govorne enote; svobodni ritem pri katerem ni stalnih poudarkov; vezani ritem pri katerem so poudarki stalni" [32].



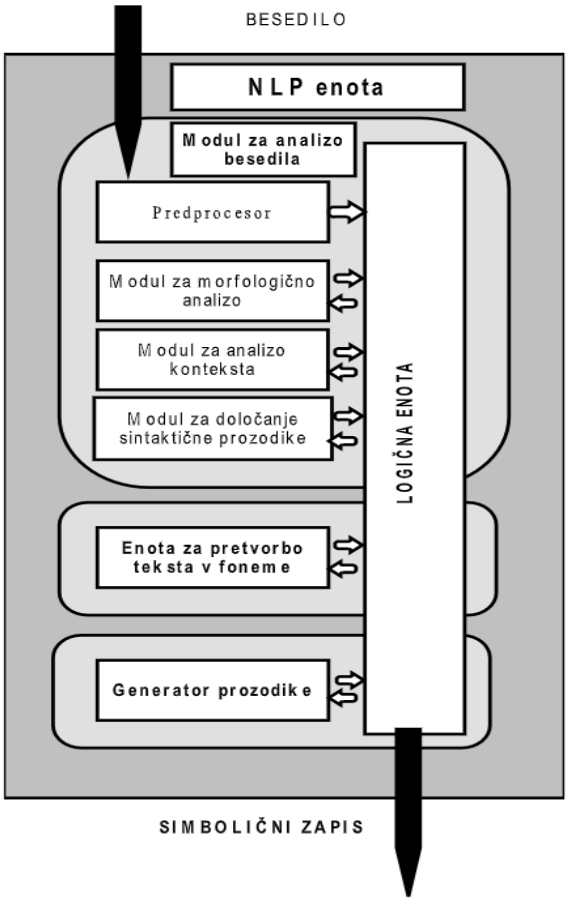
Slika 6: Slika predstavlja funkcionalni skelet NLP modula. Kot vidimo na sliki, vsebuje modul za analizo besedila, enoto za pretvorbo teksta v foneme in generator prozodike (Vir: [7]).

Pristopi k procesiranju naravnega jezika so različni, od uporabe množice lingvističnih pravil, slovničnih pravil, slovarjev, statističnih metod, do bolj eksotičnih kot npr. uporaba nevronskih mrež. Največkrat pa se uporablja kombinacija le-teh, saj se s tem lahko doseže visoka kakovost govora. Na sliki 6 vidimo funkcionalni skelet enote za procesiranje naravnega jezika (NLP) [7].

4.2 Enota za analizo besedila

Glede na to, da lahko na vhod sistema TTS dobimo poljubno besedilo, npr. v ASCII (ang. American Standard Code for Information Interchange) obliki, ga moramo najprej analizirati. Takšno besedilo vsebuje tudi številke, datume, kratice, akronime in podobno, kar moramo predhodno pretvoriti v besedilo. S postopkom segmentacije razdelimo besedilo na stavke, ki jim določimo začetek in konec, nato pa vsak stavek tokeniziramo. S postopkom tokenizacije stavke razdelimo na manjše enote, zato da se besedilo lažje obdela. Ponavadi se tokenizira glede na prazen prostor med besedami ali glede na ločila. Tokenizacija glede ločil včasih predstavlja problem, saj se lahko pike, ki označujejo konec stavka, nahajajo tudi v stavku (npr., itd., Marjeta T. Novak, 100.000,00). Velikokrat lahko posamezni znaki pripeljejo do dvoumnosti, ki jo je potrebno reševati s pravili gramatike [5, 28, 29].

Enota za analizo besedila je sestavljena iz naslednjih modulov:



Slika 7: Na sliki vidimo, da modul za analizo besedila vseboje predprocesor, modul za morfološko analizo, modul za analizo konteksta in modul za določanje sintaktične prozodike (Vir: [29]).

- Predprocesor normalizira besedilo, tako da pretvori vhodne stavke v listo besed in pri tem odstrani vse znake, ki nimajo vpliva na izgovorjavo besedila. Pri tem analizira vhodne znake in jih po potrebi spremeni v besede. To so lahko različne številke, ure, naslovi, datumi, kratice, matematični operatorji +, ', *, / ali pa posebni znaki kot sta $, & in decimalna vejica. Na prvi pogled se takšna pretvorba zdi zelo enostavna naloga, vendar kaj hitro naletimo na primere, kjer je potrebno analizirati celoten stavek, da dobimo ustrezno besedilo. Kot primer lahko navedemo simbol %, kjer se najprej preveri ali je postavljen v obliko samostalnika za njim, in nato pretvori v pravilni spol, število in sklon. Številka pred simbolom se mora pojaviti v nominativni obliki [26, 29].

- Modul za morfološko analizo poišče korene besed, ki se v govoru pojavljajo v različnih oblikah (zaradi sklanjatev, spola, množine, itd.). Ta postopek je potreben, ker se v slovarjih nahajajo samo koreni besed [29].

- Modul za analizo konteksta preučuje okolje, v katerem se nahaja beseda, ki jo analiziramo. Glede na to, da se lahko besede, ki imajo povsem identični zapis, v različnih stavkih izgovorijo drugače, je potrebno tudi ugotoviti kontekst stavka. S tem lahko določimo pravilno izgovorjavo in naglas besede. Kot primer lahko navedemo stavka »Konj je travo« in »Konj je hiter«. Na spodnji sliki je prikaz pravil izgovorjave besede "je", glede na kontekst [29].

The Slovenian word *je* has two meanings: *jè* (verb *to be*) or *jé* (verb *to eat*).

Rules	**Pronunciation**	**Example**
...	...	...
to + *je*	⇒ *jè* (verb *to be*)	To je stol. (This is a chair.)
koliko + *je*	⇒ *jè* (verb *to be*)	Koliko je ura? (What's the time?)
...	...	...
je + noun [fem, nom]	⇒ *jè* (verb *to be*)	Ona je lepa. (She is beautiful.)
je + noun [fem, accu]	⇒ *jé* (verb *to eat*)	On je čokolado. (He is eating a chocolate.)
on + *je* + noun [mas or neu, (meaning ≠ food)]	⇒ *jè* (verb *to be*)	On je pilot. (He is a pilot.)
<name> + *je* + noun [mas or neu, (meaning = food)]	⇒ *jé* (verb *to eat*)	Janez je sendvič. (John is eating a sandwich.)
...	...	...

Slika 8: Delna pravila za uporabo besede "je", ker ima beseda "je" v slovenščini dva pomena (Vir: [26]).

- Modul za določanje sintaktične prozodike izbira mikroprozodične [3] parametre posameznih besed, ki določajo trajanje in višino fonemov v posameznih besedah. Tako je važen naglas besede, način poudarka pri zlogih in dolžina izgovorjene besede [29, 34].

4.3 Modul za pretvorbo teksta v foneme (LTS)

Modul za pretvorbo teksta v foneme (Letter To Sound–LTS) skrbi za avtomatsko pretvorbo vhodnega teksta v fonetični zapis. Tudi to se na prvi pogled zdi lahko, če imamo slovar izgovorjav, vendar tudi pri tej operaciji hitro naletimo na kopico težav:

[3] Prozodija preučuje vse elemente jezika, ki prispevajo k akustičnim in ritmičnim učinkom. Mikroprozodika je izraz, ki se uporablja za označevanje nekaterih najnižjih, podrobnejših vidikov prozodije [10, 28].

- Slovarji izgovorjav se nanašajo samo na korene besed. Le-te sicer s pomočjo modula za morfologično analizo izluščimo iz besed, vendar se lahko zgodi, da se izgovorjava pri različnih morfoloških variantah spreminja. Zato je potrebno poleg slovarja izgovorjav zagotoviti še pravila, ki se nanašajo na modificirane korene. Posebna veda fonologije, ki se ukvarja s tem problemom, se imenuje morfonologija.

- Nekatere besede se kljub identičnemu zapisu izgovarjajo drugače. Za pravilno pretvorbo je običajno potrebno analizirati kontekst. Za to nalogo se uporablja kar nekaj različnih rešitev, od določevanja kontekstnih pravil s pomočjo lingvističnih ekspertov ali avtomatičnega določevanja pravil s pomočjo klasifikacijskega in regresivnega drevesa, nastalega iz učne množice, do bolj tehnično zapletenih orodij, kot so prikriti markovi razredi ali pa s pomočjo nevronskih mrež.

- Besede, izgovorjene v stavku, ne zvenijo tako, kot če bi jih izgovorili izolirano.

- Slovarji izgovorjav zagotavljajo bolj fonemičen kot fonetičen zapis. Zato se takšen govor tudi v idealnem primeru malenkostno razlikuje od naravnega.

- V slovarju izgovorjav ne najdemo vseh besed, kot npr. razna imena, nove besede, tujke, itd... [29].

Najenostavnejši in največkrat uporabljen pristop za izvedbo LTS modula je uporaba slovarja izgovorjav. Glede na to, da ima lahko posamezna beseda zaradi sklanjatev, spola itd. kar nekaj izpeljank, se običajno v slovarji shranijo samo koreni, z dodatnimi pravili pa se določi izgovorjava celotne izpeljanke. Prav tako moramo določiti pravila za izgovorjavo besed, ki jih sistem ne najde v slovarju. Po pretvorbi besede v foneme pa se običajno izvede še po procesiranje, kjer sistem na osnovi pravil še " zgladi" fonemski zapis. Lahko imamo slovar, ki ima poleg korenov še vse možne izpeljanke posameznih besed [7].

Drugi pristop pa se za svoje delovanje bolj zanaša na množico pravil, s pomočjo katerih se opravi pretvorba v fonetični zapis. Ker pa vedno obstajajo izjeme, ki jih težko zajamemo s pravili, zgradimo še slovar izgovorjav izjem, ki je dosti manjši od celotnega slovarja [7].

4.4 Generator prozodike

Pri sintetizatorju govora moramo biti pozorni tudi na to, da zveni čim bolj naravno. Izkazalo se je, da smo ljudje občutljivi na način govora, saj nas hitro zmoti, če je glas preveč mehanski in nečloveški. Robotski glas je resda kovinski, vendar nas najbolj pri

njem moti monotono brezosebno govorjene. Kako torej dosežemo čustveno govorjenje? Kako dosežemo dramatičnost besedila? Najboljše odgovore na ta vprašanja seveda poznajo poklicni igralci, ki za to uporabljajo cel kup trikov, kot npr. različno poudarjanje besed, spreminjanje glasnosti in višine tona, spreminjanje hitrosti izgovarjanja besed, uporaba različnih premorov, itd. Zato lahko igralec nek stavek pove na skoraj tisoč in en način, odvisno od tega, kaj hoče poudariti. Strokovno to poimenujemo prozodika. Da bi ločili prozodiko pri izgovorjavi posameznih besed, le-to imenujemo mikroprozodika, prozodiko pri izgovorjavi celotnega stavka pa makroprozodika. Glede na to, da sintetizator ne razume besedila, ki ga hoče ustvariti, se postavlja vprašanje, kako razbiti monotonost [28]. Najsplošnejši pristop je opazovanje spreminjanja višine osnovnega tona skozi celoten stavek (F0). Opazimo, da se F(0) na začetku rahlo zviša in počasi pada do konca stavka. Seveda se razlikuje glede na to, ali je stavek izjavni, imperativni ali vprašalni. Upoštevati je potrebno tudi poudarke na posamezne besede, ki so odvisni od zlogov v posamezni besedi. V poudarjenemu zlogu se navadno zviša ton, kateremu sledi padec. Osnovno frekvenco govorca predstavimo z matematičnim približkom, ki ga opisuje enačba 4.1 definirana v [34]:

$$F0 = G(t) + \sum_i L_i(t) \tag{4.1}$$

kjer je G(t) definirana kot:

$$G(t) = F_k e^{A_z \alpha (t+0.5) e^{-\alpha (t+0.5)}} \tag{4.2}$$

in $L_i(t)$ definirana kot:

$$L_i(t) = G(T_{pi}) A_{pi} (1 + cos(\pi \frac{T_{pi} - t}{d_i})) \tag{4.3}$$

kjer mora biti $(T_{pi} - t)$ v območju med $(-d_i, d_i)$ in veljajo parametri:

F_k: asimptotična končna vrednost F0 v intonacijski enoti.

A_z: parameter za začetno vrednost F0 v intonacijski enoti.

α: parameter za nadzor oblike F0

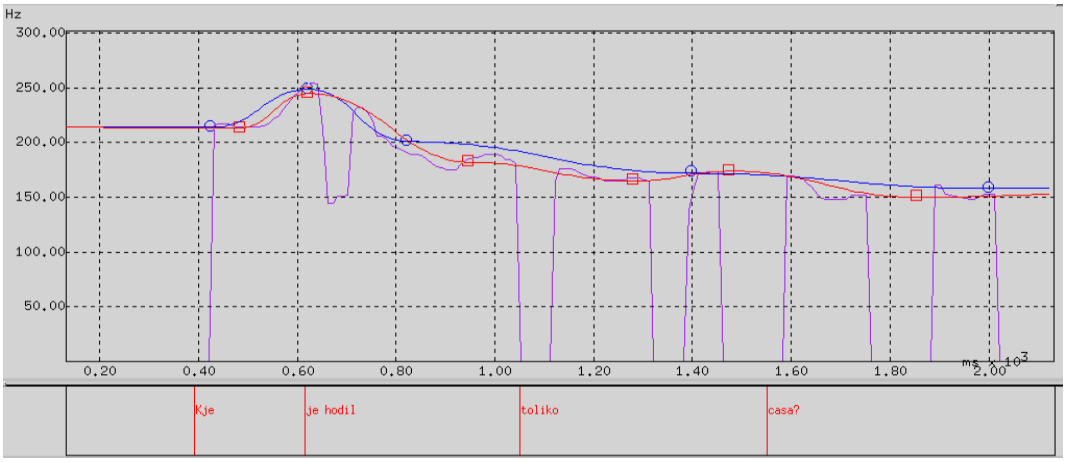
T_{pi}: čas i-tega naglasa

A_{pi}: velikost i-tega naglasa

d_i: trajanje oblike i-tega naglasa.

Parametri F_k, A_z, α in A_{pi} se spreminjajo med procesom sinteze glede na rezultate F0 konture. Na sliki 9 je prikazan primer skladnosti izračunanih vrednosti osnovne

frekvence F0, ki jih dobimo s pomočjo enačbe z vrednostmi frekvenc realnega govora.



Slika 9: Primer rezultata modeliranja F0 kontura za vprašanje: "Kje je hodil toliko časa?" Iz grafa lahko opazimo, da se F(0) na začetku malo zviša, in proti koncu vprašanja pada (Vir: [34]).

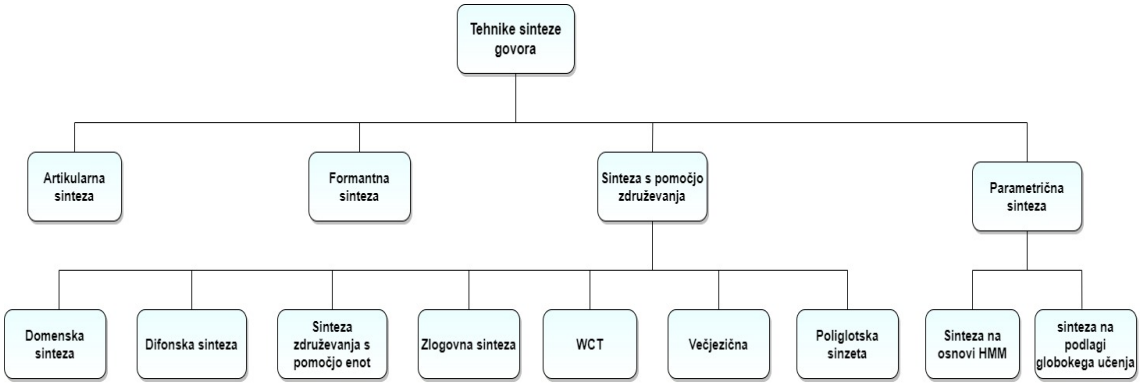
Seveda pa je za ustrezno poudarjanje potrebno stavek popolnoma razumeti. Ker je to z današnjo tehnologijo skoraj nemogoče [11], se poskuša ustvariti logika, ki bi nekako nadomestila potrebo po razumevanju. Tudi tu so raziskovalci uporabili kar nekaj povsem različnih pristopov, ki so bolj ali manj uspešni [15, 35].

4.5 Enota za procesiranje digitalnega signala (DSP)

DSP komponenta oziroma enota za procesiranje digitalnega signala skrbi za izgovorjavo besed, fraz in stavkov analogno artikulaciji človeškega govora (v audio obliko) na podlagi simbolnih informacij, ki jih prejme od NLP. Odvisno od uporabljene tehnike lahko razdelimo v skupine: artikulacijska, formantna, sinteza s pomočjo združevanja in statistično parametrična sinteza [17, 18].

4.6 Sintetizator govora na osnovi pravil

Kot pove že samo ime, delujejo takšni sintetizatorji govora na osnovi pravil. Sama izbira pravil in parametrov je zelo pomembna pri izdelavi takšnega sintetizatorja. Zato poznamo kar nekaj različnih pristopov k proučevanju te tematike. Artikulatorna sinteza je metoda sinteze govora, ki temelji na modeliranju človeškega govornega trakta. Obravnavani parametri so lahko položaj jezika, obliko ustnic pri govoru itd. Kot metoda sinteze je težje izvedljiva kot ostale metode, vendar je ena izmed najboljših metod



Slika 10: Na sliki so prikazane tehnike sinteze govora. Tehnike sinteze govora lahko v grobem razdelimo na arttikularno sintezo, formantno sintezo, sintezo s pomočjo združevanja in parametrično sintezo (Vir: [18]).

za preučevanje procesa govorne produkcije. Za boljše razumevanje govora pa so opazovali delovanje človeških organov pri govoru tudi s pomočjo rentgenskih žarkov. Kljub temu, da artikulatorna sinteza ustvarja visoko razumljiv govor, le-ta ni preveč naravno zveneč. Prednost te metode je, da ne potrebuje baze govora.

Formantna sinteza govora posnema človeški glas z generiranjem umetnih signalov na podlagi pravil povzetih iz lastnosti naravnega govora [18, 21].

4.7 Parametrična sinteza

Ta sinteza prav tako uporablja vnaprej pripravljene govorne segmente, ki jih nato modificira s pomočjo parametrov pridobljenih iz statističnih modelov. Prednost te sinteze je, da se shranijo le parametri namesto podatkov. To občutno zmanjša porabo spomina. Parametrično sintezo sestavljata dve fazi, faza treniranja in faza sinteze. V prvi fazi se iz baze govornih segmentov pridobi njihovo parametrično reprezentacijo in nato modelira s pomočjo statističnih modelov. Ker se s parametrično obdelavo izgubi kar nekaj informacij, je kakovost sinteze govora slabša kot sinteza z združevanjem enot. Prednost pa je, da s spremembo parametrov, lahko dosežemo večjo prilagodljivost spreminjanja sintetiziranega govora, ne da bi potrebovali dodatne govorne posnetke [18, 33].

Primera parametrične sinteze sta prikriti markovski model (PMM) in sinteza z globokim učenjem [33].

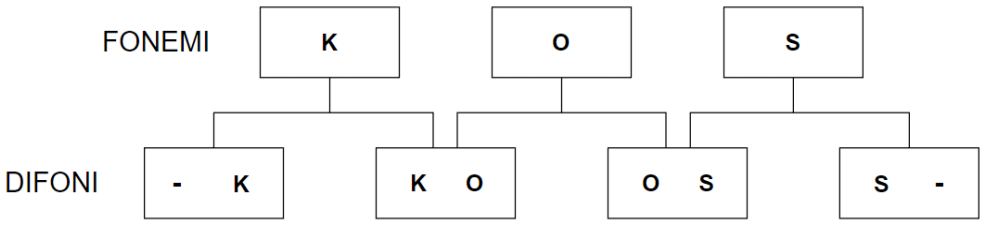
4.8 LPC Sintetizator govora

LPC (Linear Predictive coding) sintetizator govora je v bistvu nekakšna mešanica omenjenih sintetizatorjev, saj deluje podobno kot sintetizator na osnovi pravil, le-da namesto uporabe parametrov, ki opisujejo lastnosti govora (formanti), uporablja LPC parametre, ki opisujejo parametre vokalnega trakta. Ker pa LPC parametre posnamemo in predvajamo podobno kot pri sintetizatorju s pomočjo združevanja govornih segmentov, ga lahko tudi štejemo med te sintetizatorje. LPC parametri zavzamejo veliko manj prostora kot segmenti govora, shranjeni v digitalni PCM (Pulse Code Modulation) obliki, zato je njihova implementacija manj zahtevna [12].

Pri sintetizatorju govora s pomočjo združevanja ne moremo vplivati na frekvenčni spekter shranjenih segmentov govora, kar pomeni, da moramo za različne glasove uporabiti različne govorce. Pri LPC parametrih pa lahko zelo enostavno spreminjamo frekvenčni spekter [12].

4.9 Sintetizator govora s pomočjo združevanja

Sintetizatorji govora s pomočjo združevanja ne potrebujejo veliko znanja o človeškem govoru. Le-to je zapisano v segmentih govora, ki jih predhodno posname. Za produciranje govora, se segmente združene v besedo enostavno predvaja. Za segmente se lahko uporabijo kar celotne besede (npr. napovedi vlakov), zloge, posamezne foneme ali difone. Celoten postopek se sliši zelo enostavno, v praksi pa se hitro naleti na velike probleme. Če poskušamo spojiti že prej posnete foneme, ki so potrebni za neko besedo, opazimo, da je predvajani posnetek zaradi spajanja povsem nezvezen in skoraj nerazpoznaven. Takšen rezultat dobimo zato, ker so si fonemi v posamezni besedi zelo soodvisni. Boljše rezultate dobimo s uporabo difonov, ki so sestavljeni iz kombinacije dveh fonemov. Tako se difon začne na polovici prvega fonema, kjer je stabilnost največja, in konča na polovici drugega fonema. Primer difonov besede KOS si lahko ogledamo na sliki 11 [7].



Slika 11: Iz slike lahko razberemo, da se besedo "kos", katero sestavljajo trije fonemi, lahko razdeli na štiri difone.

Difonov je sicer mnogo več kot fonemov (če fonem označimo s P, jih je okoli $P^2 = 29^2 < 961$ [1]), vendar lahko z različnimi tehnikami kodiranja in komprimiranja zmanjšamo velikost potrebnih baz difonov (MBROLA algoritem [8]). Nekateri sintetizatorji (eBralec) pa uporabljajo trifone[4] ali večfone za dosego boljše kakovosti govora [35].

Kakovost takšnih sintetizatorjev je precej odvisna od izbire segmentov govora in govorca, vendar lahko hitro dosežemo visoko kvaliteto govora nasproti sintetizatorjem na osnovi pravil, kjer je za pravilno izbiro vseh potrebnih parametrov potrebno dolgo proučevati nek jezik, da dosežemo dobro kvaliteto.

4.9.1 Priprava segmentov za združevanje

V tej fazi je pomembna izbira primernega korpusa. V njem se mora pojaviti vsak segment vsaj enkrat, še bolje, večkrat. Iz korpusa je potrebno pred uporabo odstraniti vse neustrezne dele, ki bi lahko negativno vplivali na kakovost sinteze. Tak korpus je primeren za digitalni posnetek (branje).

Segmente se lahko izreže ven ročno s pomočjo signalno vizualnih orodij ali samodejno s segmentacijskimi algoritmi. Kot primer lahko podamo postopek razčlenjevanja govora, ki temelji na dveh značilkah govornega signala in sicer na jakostnem nivoju izseka ter na številu prehodov signala skozi nič [19].

Segmenti govornega signala se nato shranijo v govorno bazo podatkov, informacije o segmentih (imena segmentov, trajanje itd.) pa se shranijo v datoteko. Če podamo primer: pri difonih je potrebno shraniti mesto meje med foni, da se jim lahko modificira trajanje, ne da bi ogrozili dolžino drugega fona. Segmentom je potem pogosto dana parametrična oblika v obliki časovnega zaporedja vektorskih parametrov, ki so zbrani na izhodu analizatorja govora. Shranjeni so v podatkovni zbirki parametričnih segmentov. Prozodija je prilagojena glede na shranjene vrednosti (intonacija, trajanje) v bazi podatkov o parametričnih segmentih. Modul za usklajevanje prozodije in modul za združevanje segmentov imata bistveno olajšano delo, če so vhodni segmenti v obliki, ki omogoča enostavno spreminjanje njihove višine, trajanje in v izrednih primerih tudi ovojnico spektra.

Segmenti, ki jih je treba združiti, so pogosto iz različnih besed, torej iz različnih fonetičnih kontekstov, zato imajo neskladja v barvi in amplitudi, kar povzroča slišne prekinitve. Neskladja v amplitudi se lahko rešuje že med oblikovanjem pri generiranju podatkovne zbirke segmentov, tako da se izenačita zaključka povezanih segmentov, pri čemer se razlika porazdeli po njuni okolici. V praksi je ta operacija omejena na amplitudne parametre. Amplitudna neskladja se odpravi z nastavitvijo energije vseh

[4]Trifon se začne z drugo polovico prvega fona in konča s prvo polovica tretjega fona, vmes je en cel fon.

fonov danega fonema na njihovo srednjo vrednost. Nasprotno pa se je z barvnimi konflikti bolje spoprijeti v fazi povezovanja segmentov z glajenjem posamičnih parov, namesto da se izenačijo vsi naenkrat [7].

4.9.2 Sinteza segmentov

V sintezni fazi se najprej izpelje iz fonemskega vnosa sintetizatorja blok, imenovan generiranje seznama segmentov, ki povezuje NLP in DSP modula. Ta v podatkovni zbirki govornih segmentov poizveduje po globalnih informacijah o enotah, ki jih vsebuje. Segmentom se priredi trajanje glede na želene dolžine vgrajenih fonemov.

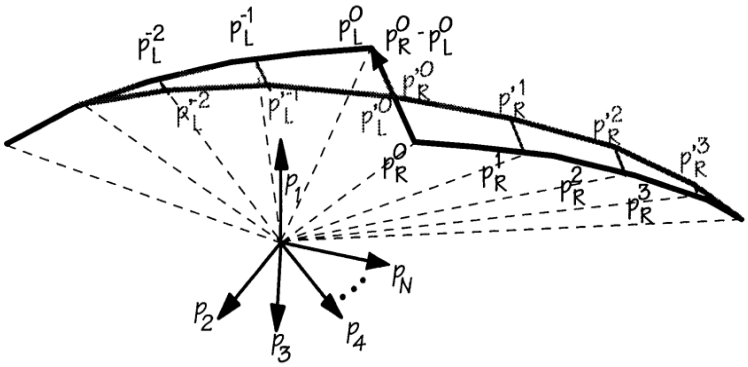
Ko so prozodični dogodki dodeljeni posameznim segmentom, govorni dekoder poizvede bazo segmentov po parametrih osnovnih zvokov, ki bodo uporabljeni in jih nato dekodira. Nato jih modul za ujemanje prozodije enega za drugim prilagodi zahtevani prozodiji.

Blok za združevanje segmentov z glajenjem prekinitev skrbi, da se segmenti ujemajo. Glede na naravno akustično bližino zvokov potrebnih veriženja, je pogosto uporabljeno linearno glajenje, kot je prikazano na sliki 12. Levi in desni segmenti so povezani z L in R. Če upoštevamo množico p parametrov $\{p_1, p_2, ..., p_n\}$, katerih vrednosti so p_L^0 na koncu L in p_R^0 na začetku R, je linearno glajenje sestavljeno iz porazdelitve razlike (p_R^0 - p_L^0) med več M_L vektorji $\{p_L^{-(M_L-1)}, p_L^{-1}, p_L^0\}$ pred vključitvijo p_L^0 ter število M_R vektorjev $\{p_R^0, p_R^1, ..., p_R^{(M_R-1)}\}$ pred in vključno s p_R^0. Če parametre po glajenju označimo s p' so interpolacijski zakoni podani z enačbo 4.4, predstavljeno v [7]

$$p_L'^{-i} = p_L^{-i} + (p_R^0 - p_L^0)\frac{(M_L - i)}{2M_L} \tag{4.4}$$

$$p_R'^{-j} = p_R^{-j} + (p_L^0 - p_R^0)\frac{(M_R - i)}{2M_R} \tag{4.5}$$

za i=0 ... M_L-1 in j=0 ... $M_R - 1$.



Slika 12: Parametrično linearno glajenje na meji zaporednih segmentov (Vir: [7]).

Tako dobljeni tok parametrov je na koncu predstavljen na vhodu bloka za sintezo, natančen ekvivalent analize. Njegova naloga je ustvarjanje govora [7].

4.9.3 Kakovost segmentov

Kakovost sintetiziranega govora s pomočjo konkatenacije segmentov je odvisna od večih dejavnikov, ki jih je pri izdelavi sintetizatorja govora treba upoštevati [7]:

- **Tip izbranih segmentov**: Za segmente lahko izberemo fone, difone, večfone, zloge itd. Daljša kot je enota izbrana za konkatenacijo, boljša je kakovost govora, ker se zmanjša število verižnih točk (točk, kjer se segmenti zlepijo). A problem je, da pri daljših segmentih ne moremo imeti obsežnega slovarja, ker je računalniški spomin omejen. Kombinacij vseh dolgih segmentov je enostavno preveč. Zaradi omejenega spomina, je bila precej popularna skozi zgodovino tehnika sinteze s pomočjo lepljenja difonov. Teh ni toliko, da bi računalniškemu spominu predstavljalo problem. Slaba stran uporabljene metode sinteze govora z lepljenjem difonov je, da je veliko verižnih točk, kar za razumljivo sintezo govora poveča potrebo po učinkovitem algoritmu. Trenutno bi bila najbolj učinkovita metoda kombinirane sinteze s pomočjo lepljenja zlogov in difonov.

- **Korpus iz katerega so bili segmenti izrezani**: Zapis in oblika korpusa morata biti prilagojena tako, da lahko iz korpusa izrežemo vse segmente, idealno dvakrat ali večkrat v različnih fonemskih okoljih. Besedilo je potrebno brati čim bolj naravno in tekoče. Snemanje posameznih besed izven korpusa ni priporočljivo, saj imamo v izoliranih besedah naglas drugačen kot v vezanem besedilu.

- **Kakovost segmentacije**: Segmentacija lahko poteka ročno ali somodejno preko algoritmov.

- **Model govornega signala, na katerega se nanašajo algoritmi analize in sinteze**: Modele lahko grobo razdelimo v dva razreda. Govorno fiziološki modeli, ki so temelj artikulacijske sinteze in fenomenološki modeli, ki obdelujejo signale digitalno.

- **Količina degradacije, ki jo povzroča faza kodiranja govora**: V govornih signalih je velika količina redundance, kompresiranje pa prinaša popačenje. Z optimizacijo danega kodirnika govora v ustrezni nastavitvi vokoderja, je potrebno nastaviti največje kompresijsko razmerje, pri katerem popačenja ostanejo neslišna.

- **Učinkovitost ujemanja prozodije**: Pogojena je z znanjem o jeziku, saj je jezik zelo kompleksna stvar z veliko pravili in izjemami. Poleg tega pa je potrebno za izdelavo kvalitetnega sintetizatorja, ne glede na izbiro načina sinteze, veliko izkušenj. Izkušnje pa se lahko pridobi samo iz praktičnega dela na sintetizatorih govora.

- **Zmogljivost algoritma za lepljenje segmentov**: V preteklosti so se razvijalci TTS soočali tako s pomanjkanjem računalniškega spomina (trenutnega in stalnega) kot tudi s pomanjkanjem procesorske moči. Ker pa računalniki postajajo vse zmoglivejši, postaja ta problem vse manj pomemben. Res pa je, da sodobni TTS težijo k "razumevanju" besedila s pomočjo umetne inteligence in ogromnem številu podatkov, ki jih podjetja pridobivajo s pomočjo internetnih programov (npr. google translator). Ker trenutno sodobni računalniki, kljub izjemni zmogljivosti ne premorejo teh nalog, se vse bolj postavlja model TTS odjemalec strežnik, kjer se besedilo preko interneta pošlje zmogljivim strežnikom, ki to besedilo obdelajo, sintetizirajo stavke in jih nato pošljejo nazaj uporabniku.

5 Implementacija

Če želimo izboljšati kakovost sintetizatorjev v slovenščini, je pomembno, da preučimo katera tehnika sinteze je najbolj primerna za sintezo v slovenskem jeziku. S časom so se tehnike sinteze govora v slovenskem jeziku spreminjale. Govorec 1 in Govorec 2 sta temeljila na lepljenju difonov, Govorec 3 uporablja metodo HNM (harmonic plus noise modulation), eBralec pa prikrite markove modele (PMM) [33].

V tem diplomskem delu želimo preveriti ali je lepljenje zlogov boljša izbira od lepljenja difonov. Želeli smo ugotoviti ali se govor izboljša z uporabo lepljenja zlogov in če zveni bolj naravno, kot če bi lepili difone. Včasih je bilo spomina malo, zato je bila najustreznejša rešitev sinteza govora z metodo lepljenja difonov. Ker spomin ne predstavlja takšnega problema kot včasih, imamo možnost, da zdaj namesto metode sinteze govora z lepljenjem difonov uporabimo sintezo govora s pomočjo lepljenja zlogov. V članku [30] je navedeno, da se ponavadi z uporabo večjih segmentov za lepljenje izboljša kakovost sinteze govora, v članku [25] pa da ima sintetizator s pomočjo lepljenja zlogov manj prekinitev in manjše število spojev v primerjavi z difonsko konkatinacijo. V čanku [24], pa je navedena ugotovitev, da se je z njihovo implementacijo sintetizatorja govora z metodo lepljenja zlogov ohranila dobra kakovost sintetiziranega govora in da se sintetiziran govor, ki ga njihov sintetizator producira, sliši naravno. Odločili smo se preveriti, ali je sinteza govora z metodo lepljenja zlogov ustrezna tudi za slovenski jezik in boljša kot sinteza govora s pomočjo lepljenja difonov.

Ker pa je skoraj nemogoče uporabiti samo zloge (slovarji zlogov bi bili preobsežni), je potrebno uporabiti kombinirano metodo lepljenja zlogov in difonov. V tem primeru se difone uporabi, ko ni vseh potrebnih zlogov v slovarju. Naša predpostavka je, da se bo z uporabo metode lepljenja zlogov namesto lepljenja difonov sinteza govora izboljšala. Z uporabo metode lepljenja zlogov pričakujemo bolj razumljiv in naraven govor, saj se zlogi ne navezujejo med sabo, tako kot fonemi in jih je posledično lažje odrezati.

Da bi preverili našo hipotezo, smo implementirali poenostavljen program sintetizator govora z omejenim slovarjem v jeziku Java. Sintetizator uporabnikom omogoča možnost izbire sinteze teksta s pomočjo združevanja difonov ali s pomočjo združevanja zlogov.

Testirali smo, kateri sintetizator se sliši bolj »naravno«, kateri sintetizator je razumljivejši in kateri se splošneje sliši bolj kvalitetno. Zbrali smo fokusno skupino ljudi, ki

so program preizkusili in ocene podali na priloženo anketo. Program je vseboval dva gumba, poimenovana »vzorec 1« in »vzorec 2«. Poleg programa so prejeli še anketo, na kateri so za vsako besedo označili kateri vzorec se jim je slišal bolje (kriteriji: naravno zveneče, razumljivost govornega besedila in primerjava kakovosti med vzorcem 1 in vzorcem 2).

5.1 Opis programa

Program je napisan v programskem jeziku Java. Uporabili smo knjižnico Sonic[1]. Program je sestavljen iz dveh delov: analiza besedila in simulator sinteze govora. Zlogi in difoni so razrezani s pomočjo programa Audacity in shranjeni v obliko .wav. Vzeti so iz 30 minut posnetega prebranega besedila.

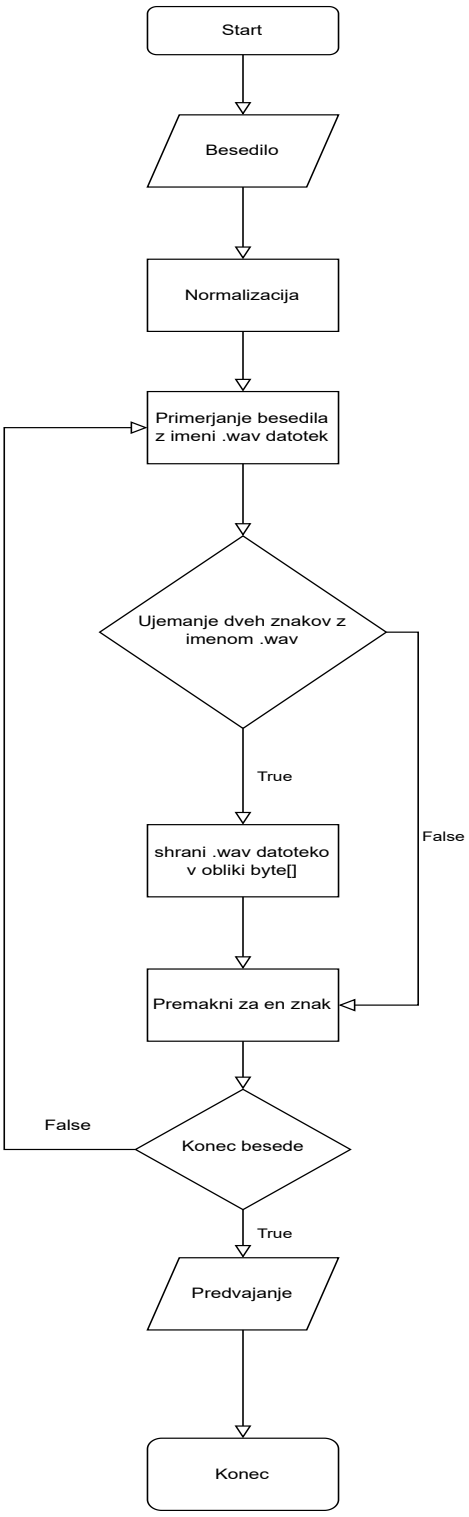
Določitev izrezanih difonov in zlogov je bila izvedena na podlagi rezultata analize besedila. Namen programa je, da simulira oba načina sinteze govora. Za ugotovitev katera tehnika sinteze je primernejša in katera proizvede bolj naraven govor, potrebujemo testni stavek, ki ga vnesemo v program.

Program ima dve verziji: prva verzija je namenjena pretvorbi teksta, ki ga zapiše uporabnik v govor. Ta je zmožen predvajati stavek iz zlogov, ki so v slovarju, če uporabnik po vpisu teksta pritisne gumb "oddaj". Druga verzija je namenjena le testiranju metode lepljenja. V tej verziji je onemogočen gumb oddaj in dodana gumba "Vzorec 1"in "Vzorec 2". Ob pritisku na gumb "Vzorec 1", se sintetizira testni stavek s pomočjo lepljenja difonov, ob pritisku na gumb "Vzorec 2", pa se testni stavek predvaja z metodo lepljenja zlogov.

Prva verzija programa uporabniku omogoča, da vnese želeno besedilo v program. Žal ima naš program omejen slovar, tako da se bo predvajalo le tisto besedilo, katerega zlogi so v slovarju. Ko uporabnik vpiše želeno besedilo, mora izbrati ali hoče sintezo govora s pomočjo zlogov ali difonov. Ko uporabnik izbere sintezo govora s pomočjo difonov, se vpisano besedilo normalizira in shrani v listo. V listi primerja dva znaka z znaki imen .wav datotek in se premakne za znak naprej, ne glede na to, ali je bil zadetek ali ne. Če je bil zadetek, shrani .wav datoteko, v oblike byte[] v rezultat. Ko gre čez vso listo, se rezultat predvaja. Na sliki 13 je prikazan algoritem v primeru, da uporabnik izbere sintezo govora s pomočjo lepljenja difonov.

Podoben algoritem je v primeru, če uporabnik izbere sintezo govora s pomočjo zlogov. Ko se vnešeno besedilo normalizira in shrani v listo, se primerja po tri znake z znaki .wav datotek. Če je zadetek, se premakne za tri znake naprej, če zadetka ni, primerja po dva znaka z .wav datotekami. Ta algoritem je narejen, zgolj za demonstrativno uporabo za preverjanje koncepta sinteze govora. Ker imamo le slovar zlogov

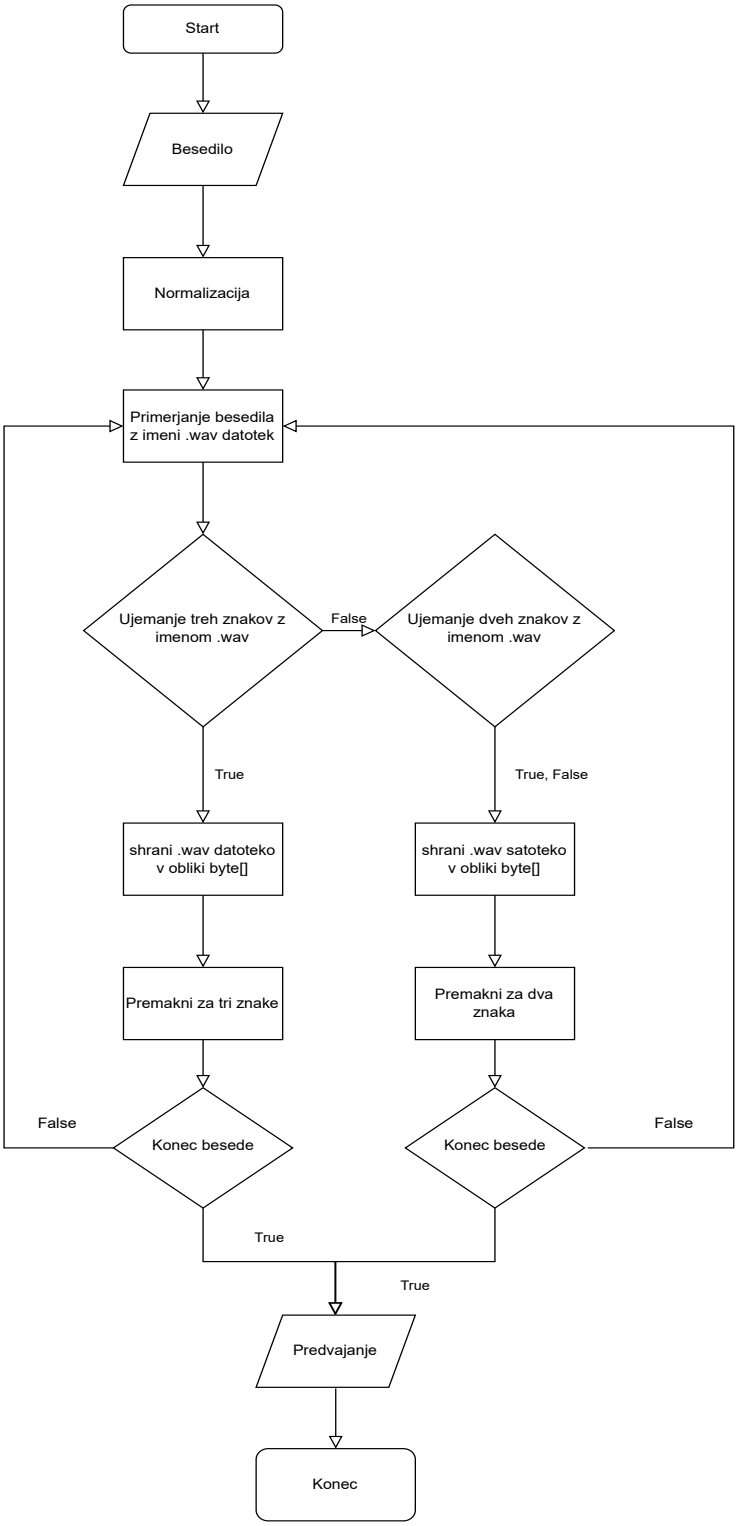
[1]Vir: https://github.com/waywardgeek/sonic



Slika 13: Na sliki je predstavljen algoritem sinteze govora s pomočjo difonov, ki smo ga uporabili pri naši implementaciji poenostavljenega sintetizatorja govora.

s tremi in dvema črkama, smo dodali " zloge", ki ne vsebujejo samoglasnikov (npr. "str"). Boljša rešitev bi bila dodati slovar zlogov s štirimi črkami ali kombiniranje

metode sinteze govora z lepljenjem difonov in zlogov. Na sliki 14 je prikazan algoritem v primeru izbire sinteze govora s pomočjo lepljenja zlogov.



Slika 14: Na sliki je predstavljen algoritem sinteze govora s pomočjo zlogov, ki smo ga uporabili pri naši implementaciji poenostavljenega sintetizatorja govora.

5.2 Izbira korpusa za analizo in izbira branega besedila

Določitev izbranih zlogov in difonov je potekala tako, da je program analiziral deset raznolikih člankov, ki so bili vzeti iz revije Delo in dom. Članki so govorili o različnih temah, da bi bila izračunana gostota pojavitve zlogov čim bolj podobna splošnemu modelu slovenskega jezika. Z raznolikostjo smo preprečili, da bi se med pogostejšimi zlogi pojavljali zlogi določene besede, ki je značilna za posamezno področje (npr. če bi bili vsi članki na temo mačka, bi bila med pogostejšimi zlogi, zloga "ma" in "čka", kar pa ne bi prikazalo pravilne gostote pojavitve zlogov).

Za brano besedilo smo izbrali naključno knjigo.

5.3 Statistična obdelava in analiza besedila

S tem delom programa smo analizirali deset različnih člankov. Cilj analize je bil najti najpogostejše zloge, ki so sestavljeni iz treh črk in vsebujejo en samoglasnik.

Poleg zlogov iz treh črk v analizi program poišče vse kombinacije dveh črk in jih razvrsti po pogostosti.

Najprej program opravi normalizacijo vhodnega besedila in vstavi besede v seznam. V naslednjem korak program sestavi sezname bigramov in trigramov iz izhoda prvega koraka. Program združi po tri oziroma dve črki (npr. Če je beseda "zjutraj", naredim "zju", "jut", "utr", "tra" in "raj"). Predvidevali smo, da bodo najpogostejše besede s tremi oziroma dvema črkama ravno zlogi.

Nato izpiše v konzoli par v obliki (ključ, vrednost) po vrsti od največje vrednosti do najmanjše. V paru je ključ zlog in vrednost število pojavitev zloga v besedilu. V tabeli 4 in tabeli 5 je prikazanih deset najpogostejših zlogov iz treh ter iz dveh črk, ki smo jih dobili z analizo desetih člankov. Zraven zlogov je zapisano kolikokrat (nenormalizirana frekvenca) se je posamezen zlog pojavil v teh desetih člankih.

5.4 Priprava zlogov in difonov

Za dobro sintezo govora je pomembno, da so zlogi in difoni čim bolj natančno izrezani. Besedilo smo posneli preko pametnega telefona, kjer je bralka besedila morala prebrati besedilo v knjižnem jeziku, brez naglasa. Poskušala je govoriti čim bolj naravno in razločno. Ko je bilo besedilo posneto, smo se lotili rezanja na zloge in difone. Odločili smo se, da za primerjalni eksperiment izrežemo 50 najpogostejših zlogov, ter poskusimo iz njih sestaviti testni stavek. Da smo iz tega lahko sestavili daljši smiselen stavek, smo

Tabela 4: 10 najpogostejših zlogov iz treh črk, kot rezultat analize 10 člankov

Zlog	Število pojavitev
pre	204
pri	195
anj	168
ost	165
rav	144
nje	133
ali	119
sti	115
ajo	104
sta	100

Tabela 5: 10 najpogostejših zlogov iz dveh črk, kot rezultat analize 10 člankov

Zlog	Število pojavitev
je	600
ko	587
ra	578
na	564
po	499
ne	486
re	443
ni	442
te	433
in	406

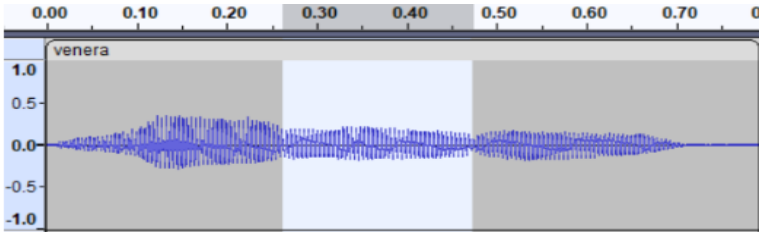
uporabili tudi zloge, ki se niso uvrstili med petdeset najpogostejših zlogov. Kot zloge smo šteli tudi najpogostejše besede s tremi črkami, četudi te ne vsebujejo zloga, kot je na primer beseda "str". Za to smo se odločili, ker nekatere besede vsebujejo zloge iz štirih črk, naš slovar pa temelji na zlogih treh ali dveh črk. Torej, če podamo primer, v naši implementaciji sintetizatorja govora delimo besedo " stroka", na "str"in "oka". Ker se takšni "*zlogi*" brez samoglasnika pojavljajo bolj redko, to ne vpliva na rezultat.

Difone smo razrezali glede na izbor zlogov ter testnega stavka.

5.4.1 Rezanje zlogov in difonov

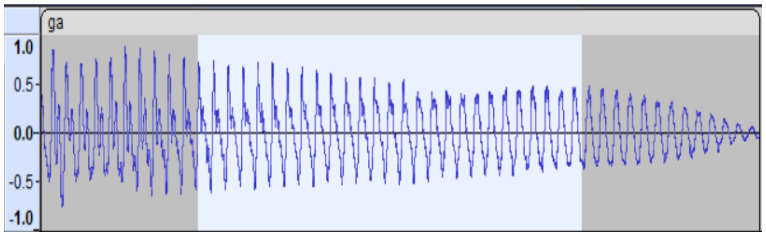
Difone in zloge smo izrezali ročno. Pred razrezom smo posneto besedilo normalizirali.

Razrez zlogov in difonov je predstavljal kar velik izziv. Problem je, da je pri rezanju govornega signala izredno pomembna natančnost, predvsem pa je potrebno paziti, da odrežemo zlog točno ob koncu določenega glasa. V programu Audacity imamo možnost razreza govornega signala. Na primeru razreza zvočnega signala “venera” (slika 15) lahko opazimo, da se lepo vidi razdelitev, kje se zlog začne in kje konča.



Slika 15: Primer zvočnega signala "venera" razdeljenega na tri zloge, kjer se jasno vidijo razmejitve med njimi. Jakost vsakega zloga na začetku začne naraščati, doseže svoj maksimum in očitno pada proti koncu zloga. Meje med zlogi se tako enostavno določijo pri lokalnih minimumih.

Razrez difona je vseeno težje delo, saj zahteva veliko časa in natančnosti. Predvsem pa je treba paziti, da odrežemo na sredini glasu, saj ne želimo medsebojnega prekrivanja med kasnejšo sintezo. Pri difonu odrežemo drugi del prvega glasu in prvi del drugega glasu. Do problema razreza pa prihaja, ker so fonemi odvisni med sabo. Pri razrezu lahko opazimo, da se velikokrat opazovani fonem že na začetku oblikuje drugače, glede na naslednji fonem. Na sliki 16 vidimo primer razreza besede “ga”, na drugo polovico fonema “g” in prvo polovico fonema “a” (difon je v svetlejšem okviru).



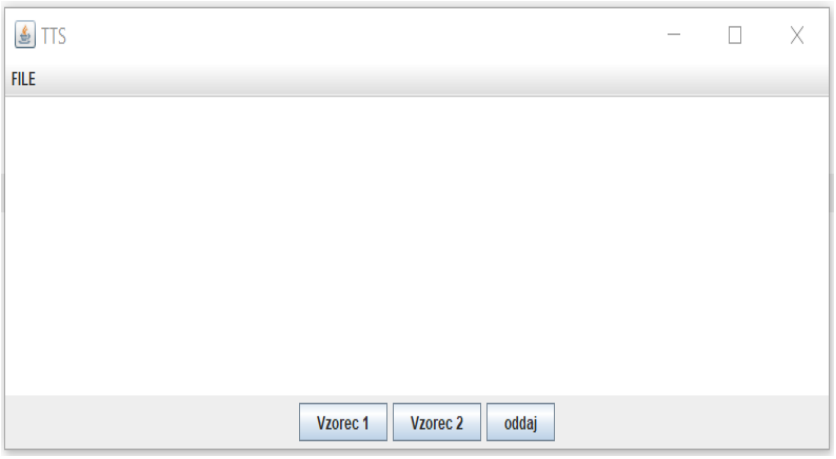
Slika 16: Primer difona "-ga-" v besedi "ga", kjer je določitev meje, ki razpolavlja fon, težje oceniti. Na sliki je razvidno, da se je meja začetka difona (osvetljeno) arbitrarno določila za maksimalno jakostjo fona g na podlagi slišane kakovosti lepljenih difonov.

5.5 Sinteza govora

Ta del programa sestavlja uporabniški vmesnik GUI, ki uporabniku prikaže okvir za besedilo, ki ga uporabnik želi predvajati (žal imamo na voljo omejen besednjak) in dva

gumba. Prvi gumb sproži sintetizator govora na podlagi zlogov, drugi gumb pa na podlagi difonov. Zaradi eksperimenta, kjer uporabniki preverijo, kateri sintetizator se jim zdi boljši, na gumbih ne piše kateri sintetizator se sproži. Na gumbih je zapisano zgolj “vzorec 1” in “vzorec 2”. Za ta del imajo uporabniki onemogočen vnos besedila.

Ko uporabnik stisne gumb, kjer piše “vzorec 1”, se v ozadju sproži analiza besedila, kjer besedilo primerja z imeni datotek .wav, ki so shranjeni v mapi. Vsaka .wav datoteka, ki se ujema z imenom, se potem predvaja. Zaporedje predvajanja datotek .wav tvori testni stavek. Ker smo uporabili knjižnico Sonic, lahko zvoku spremenimo glasnost in frekvenco.



Slika 17: Uporabniški vmesnik našega programa za testiranje sinteze govora. Uporabniki lahko izbirajo tehniko sinteze govora, glede na pritisk gumba "Vzorec 1" ali gumba "Vzorec 2". S pritiskom na gumb "oddaj" tekst analizira in izpiše gostoto posameznih zlogov iz teksta, ki ga uporabnik vnese v temu namenjen okvir.

5.6 Raziskava

Z željo, da bi preverili ali je sinteza s pomočjo lepljenja zlogov boljša od sinteze s pomočjo lepljenja zlogov, smo izvedli anketo, ki se je nanašala na naš priilagojeni poenostavljeni sintetizator. Sintetizator smo za anketo preoblikovali tako, da je imel samo dva funkcionalna gumba. Na levem gumbu je pisalo "vzorec 1", na desnem pa "vzorec 2". Gumb z napisom "vzorec 1" je sintetiziral vnaprej določeni testni stavek z lepljenjem difonov, gumb z napisom "vzorec 2" pa je sintetiziral testni stavek s pomočjo lepljenja zlogov. Anketiranci niso vedeli, kateri vzorec uporablja metodo sinteze lepljenja difonov in kateri metodo sinteze lepljenja zlogov.

Testni stavek je bil v naprej določen na podlagi najpogostejših zlogov iz analizne faze programa. Da bi preverili, ali testiranci sploh razumejo sintetizirani govor našega

programa, smo se odločili, da določimo testni stavek, ki ga anketiranci ne bodo vedeli pred pritiskom na gumb.

Zbrali smo 35 ljudi različnih starosti, ki so dobili program in v prilogi anketo. Testiranci so predvajali “vzorec 1” in “vzorec 2” ter rešili vprašanja ankete. Anketa je priložena v prilogi.

Vsak testiranec je preizkusil program in anketo reševal posamično. V anketi so testiranci poslušali oba vzorca. Pri tem so morali označiti katera verzija vzorca (vzorec 1 ali vzorec 2) se jim je zdela bolje zveneča in za obe kategoriji (naravnost govora in razločnost) označiti na lestvici od 1-5 (1 najslabše in 5 najboljše).

Poleg tega je vsak testiranec moral zapisati stavek, ki ga je slišal, da smo preverili ali je sinteza govora razumljiva za uporabnike.

5.7 Rezultati

Prvo vprašanje je bilo namenjeno testiranju, ali so anketiranci razumeli testni stavek. Testni stavek se je glasil: “Strokovnjaki so prepričani, da je porast bolnikov posledica nezdravega načina življenja.”. Testni stavek je vseboval 11 besed. Izmed petintrideset anketirancev je za "vzorec 1", ki je bil sintetiziran s pomočjo lepljenja difonov v povprečju pravilno razumelo 5,7 besed. Kar pomeni, da so anketiranci v povprečju razumeli 51,8 procentov besed.

Za "vzorec 2", ki je bil sintetiziran s pomočjo lepljenja zlogov, so anketiranci v povprečju razumeli 6,6 besed od 11. To pomeni, da so za "vzorec 2", razumeli 60 odstotkov besed. V povprečju so šteti, tudi tisti, ki niso zapisali nobene besede. Teh je bilo 6 od 35.

Na vpršanje, naj na lestvici od ena do pet (1 – nerazumljiv, 2 – slabo razumljiv, 3 – dokaj razumljiv, 4 – dobro razumljiv in 5 – odlično razumljiv) označijo razumljivost govora vzorca 1, so ljudje podali povprečno oceno 2.3.

V spodnji tabeli 6, je prikazan rezultat razumljivosti sintetiziranega govora, za vzorec 1.

Tabela 6: Razumljivost sinteze govora s pomočjo lepljenja difonov (vzorec 1

Ocena (1 – 5)	Frekvenca
1 – nerazumljiv	7
2 – slabo razumljiv	13
3 – delno razumljiv	11
4 – dobro razumljiv	3
5 – odlično razumljiv	0

Na vprašanje, naj označijo na lestvici od ena do pet (1 – zelo nenaraven(čisto robotski), 2 – nenaraven, 3 – med robotskim in naravnim človeškim govorom, 4 – dober približek človeškemu govoru in 5 – ne bi ločili sintetiziranega govora od človeka), kako naraven je govor so ljudje podali povprečno oceno 1,9.

V spodnji tabeli 7, je prikazan rezultat na vprašanje kako naraven se zdi sintetizerani govor, za vzorec 1.

Tabela 7: kako naraven se sliši testni stavek sinteze govora, s pomočjo lepljenja difonov (vzorec 1)

Ocena (1 – 5)	Frekvenca
1 – čisto robotski	13
2 – moteče nenaraven	14
3 – med robotskim in naravnim človeškim govorom	6
4 – dober približek človeškemu govoru	2
5 – ne bi ločili sintetiziranega govora od človeka	0

Na vpršanje, naj na lestvici od ena do pet (1 – nerazumljiv, 2 – slabo razumljiv, 3 – dokaj razumljiv, 4 – dobro razumljiv in 5 odlično razumljiv) označijo razumljivost govora vzorca 2, so ljudje podali povprečno oceno 2.8.

V spodnji tabeli 8, je prikazan rezultat razumljivosti sintetiziranega govora, za vzorec 2.

Tabela 8: Razumljivost sinteze govora s pomočjo lepljenja zlogov (vzorec 2)

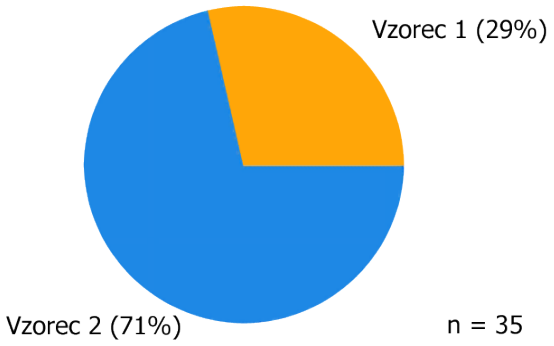
Ocena (1 – 5)	Frekvenca
1 – nerazumljiv	3
2 – slabo razumljiv	12
3 – delno razumljiv	10
4 – dobro razumljiv	10
5 – odlično razumljiv	0

Na vprašanje, naj označijo na lestvici od ena do pet (1 – zelo nenaraven(čisto robotski), 2 – nenaraven, 3 – med robotskim in naravnim človeškim govorom, 4 – dober približek človeškemu govoru in 5 – ne bi ločili sintetiziranega govora od človeka), kako naraven je govor, so ljudje podali povprečno oceno 2,2. V spodnji tabeli 8 je prikazan rezultat na vprašanje, kako naraven se zdi sintetizerani govor, za vzorec 2.
V zadnjem vprašanju smo preverjali, kateri vzorec testnega stavka se je anketirancem zdel na splošno kvalitetnejši. 71 procentov anketirancev se je odločilo za vzorec 2

Tabela 9: kako naraven se sliši testni stavek sinteze govora s pomočjo lepljenja zlogov (vzorec 2)

Ocena (1 – 5)	Frekvenca
1 – čisto robotski	8
2 – moteče nenaraven	15
3 – med robotskim in naravnim človeškim govorom	8
4 – dober približek človeškemu govoru	4
5 – ne bi ločili sintetiziranega govora od človeka	0

(sinteza govora s pomočjo lepljenja zlogov), 29 procentov se je odločilo za vzorec 1 (sinteza govora s pomočjo lepljenja difonov).



Slika 18: Na tortnem diagramu je prikazano, koliko procentov anketirancev je izbralo kateri vzorec, glede na vprašanje kateri vzorec se jim je slišal na splošno bolje.

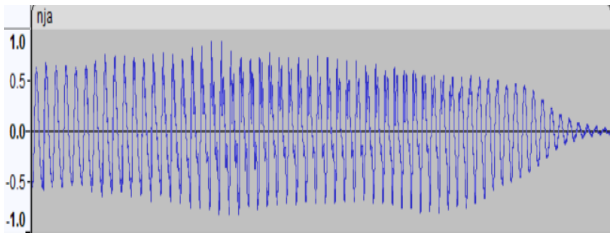
5.8 Sklep

Iz rezultatov lahko razberemo, da so bile povprečne ocene precej slabe. To lahko pripišemu temu, da je bil naš program poenostavljen sintetizator govora. To pomeni, da je vseboval zgolj slovar difonov in slovar zlogov. Če bi želeli, da je umetno ustvarjen govor bolj točen, bi morali dodati še bazo parametrov, ki vsebuje informacije o segmentih. Naš program prav tako nima implementirane logike in slovarja za razlikovanje med identično zapisanimi a različno naglašenimi zlogi in difoni.

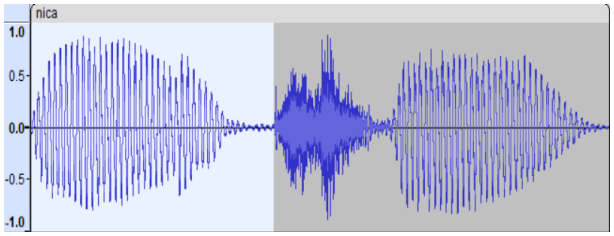
Pri prvem vprašanju ankete, kjer smo anketirance prosili, da ponovijo slišani stavek, smo ugotovili, da so pri sintezi govora z metodo lepljenja zlogov pravilno razumeli za 8,2

procentov več besed kot pri sintezi govora z metodo lepljenja difonov. Pri vprašanju, kjer so anketiranci označili razumljivost govora na lestvici od 1 do 5, so anketiranci podali za 10 procentov boljšo oceno pri vzorcu 2 (sinteza govora z metodo lepljenja zlogov), kot pri vzorcu 1 (sinteza govora z metodo lepljenja difonov), kar se ujema z rezultati prvega vprašanja. Malo odstopanja opazimo, ker 7 anketirancev ni ponovilo testnega stavka, vseeno pa so odgovorili na ostala vprašanja. To tudi zniža povprečje razumljenih besed pri obeh vzorcih.

Utemeljitev, da je sintetiziran govor z metodo lepljenja razumljivejši, bi lahko bila, da je lažje odrezati zloge kot difone. Pri rezanju difonov, se lahko hitro odreže preveč ali premalo črke, ker ni jasne meje med fonemi. Če pokažemo primer 19, lahko vidimo, da je zelo težko ločiti j od a, še težje pa poiskati, kje je sredina glasu j in glasu a. Kot je predstavljeno v poglavju 1, se fonemi različno izgovarjajo (alofoni) glede na to, kateri fonem je njihov sosed. To pomeni, da bo že znotraj enega fonema (alofona), indikator kateri fonem mu sledi, kar kaže na povezanost fonemov. Problem se pokaže, ker se zaradi te povezanosti težko določi sredina fonema, kar onemogoči točno rezanje fonemov na difone. Problem pa ni zgolj v rezanju, pač pa tudi v lepljenju, kjer energija difonov ni minimalna, kar se pozna v kvaliteti sintetiziranega govora.



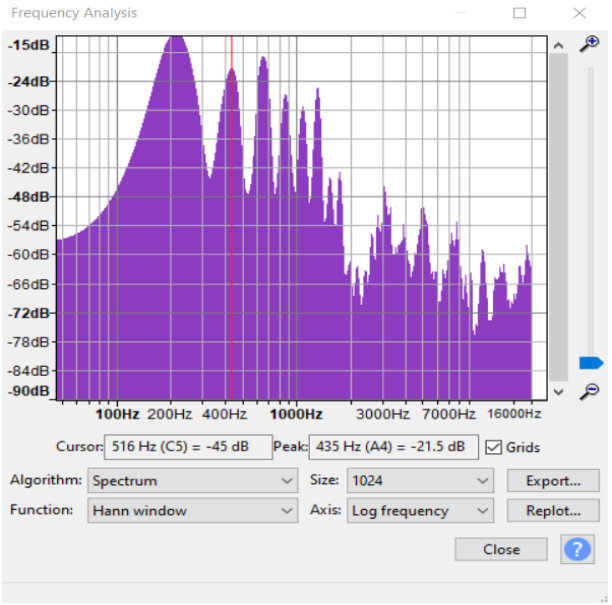
Slika 19: Primer povezanosti fonemov "n","j", "a" na zlogu "nja", kjer se meje med foni težko razločujejo, saj ni jasnih lokalnih minimumov ali maksimumov med fonemi.



Slika 20: Primer meje med zlogoma "ni" in "ca", kjer je jasno opazi minimalni signal med zlogoma (zlog "ni" je osvetljen). Kljub temu, da je opazen minimum med fonoma c in a, je ta minimum krajši od minimuma med zlogoma.

Nasprotno lahko ugotovimo za rezanje zlogov, saj se pri njih točno vidi, kje se končajo in kje začnejo. Ker so zlogi običajno povezani med seboj s samoglasniki, se točno vidi, kje se končajo in kje začnejo. To omogoča natančnejše rezanje zlogov, kar bistveno

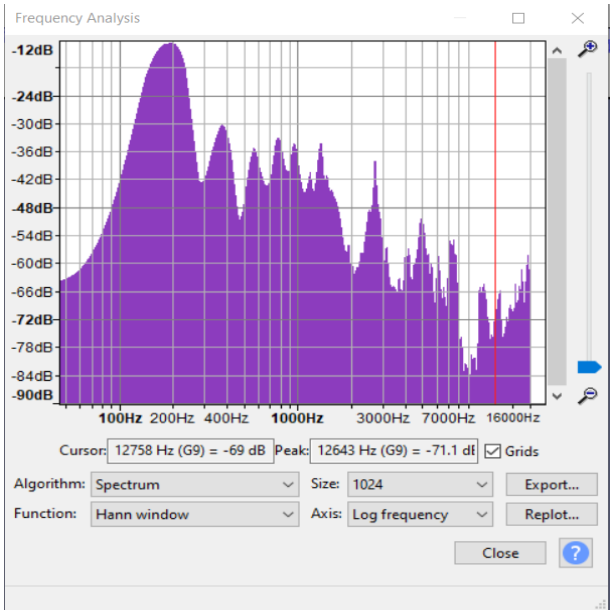
pripomore h kvaliteti zvoka. Poleg tega pa je lažje lepiti zloge, ker imajo zlogi na začetku in na koncu minimalno energijo, kar omogoča bolj povezano sintezo govora. Ugotavljamo še, da se opazi razlika med enakim zlogom, če je ta odrezan iz sredine besede ali če je odrezan iz konca besede. Opazimo lahko, da energija hitreje pada pri zlogu, ki je na koncu, kot energija zloga, ki je nekje znotraj besede. Energija je odvisna od amplitude in frekvence. Na primeru slike 21 in slike 22 lahko opazimo razliko frekvence pri črki a, ki je v začetnem zlogu, in frekvence pri črki a, ki je v končnem zlogu v besedi "mapa".



Slika 21: Primer frekvenčnega spektra izgovorjave črke a v začetnem zlogu besede "mapa". Iz slike je razviden maksimum frekvenčnega spektra, ki predstavlja osnovno frekvenco govora F(0) pri fonemu a znotraj začetnega zloga, katere vrednost je nad 200Hz.

To se seveda pozna tudi pri sintezi govora. Če bi sintezo govora s pomočjo lepljenja zlogov želeli izboljšati, bi morali imeti vsak zlog v podatkovni bazi vsaj dvakrat. Ena varianta zloga bi bila odrezana iz sredine besede, druga pa iz konca besede.

Pri vprašanju, kateri vzorec se jim je na splošno slišal kvalitetnejši, je 71 procentov anketirancev označilo vzorec 2 (sinteza govora z metodo lepljenja zlogov). Iz tega lahko sklepamo, da je bila sinteza govora z metodo lepljenja zlogov boljša izbira pri sintetizatorju govora v slovenščini, kot uporaba metode lepljenja difonov. V prihodnje predlagamo nadaljno raziskavo na področju metode sinteze govora z lepljenjem zlogov. Testirati pa bi morali tudi vpliv izgovorjave začetnega zloga.



Slika 22: Primer frekvenčnega spektra izgovorjave črke a v končnem zlogu besede "mapa". Na tej sliki je osnovna frekvenca govora F(0) pri fonemu a znotraj končnega zloga, pod 200Hz, kar jasno nakazuje razliko osnovne frekvence govora F(0) fonema a znotraj začetnega in končnega zloga.

5.9 Možne izboljšave in predlogi

Ker je sintetizator govora kompleksna stvar, smo se odločili narediti poenostavljeno različico, saj je bil namen diplomske naloge preveriti koncept sintetizatorja govora. Zato naši implementaciji manjka kar nekaj ključnih delov sintetizatorja govora.

Če bi želeli izboljšati naš sintetizator govora, bi morali implementirati ustrezno logiko za "razumevanje" konteksta. Prav tako bi morali povečati slovar zlogov in difonov. Dodati bi morali tudi različne verzije zlogov in difonov glede na naglaševanje. Kakovost bi se izboljšala tudi, če bi bilo besedilo brano in snemano v profesionalnem okolju.

Preverili smo razliko v kakovosti sinteze govora s pomočjo lepljenja zlogov ter s pomočjo lepljenja difonov in ugotovili, da je testna skupina ljudi v povprečju preferirala sintezo govora s pomočjo zlogov. Ker bi za bolj točen rezultat morali narediti obsežnejšo raziskavo, priporočamo nadalnje raziskave na področju sinteze govora z uporabo metode lepljenja s pomočjo zlogov. Predlagamo tudi, da se v podatkovno bazo shranita vsaj dve različici vsakega zloga. Ena varianta zloga naj bo odrezana na sredini besede, druga varianta pa na koncu. Predlagamo še nadalnje raziskave glede ločevanja začetnih, sredinskih ter končnih zlogov.

6 Zaključek

V diplomskem delu smo preverili koncept sintetizatorja govora v slovenskem jeziku. Implementirali smo osnovni sintetizator govora z omejenim slovarjem. Namen programa je bilo preverjanje, katera metoda sinteze govora je boljša, metoda sinteze govora s pomočjo lepljenja zlogov, ali metoda sinteze govora s pomočjo lepljenja zlogov.

Za implementacijo sintetizatorja govora smo uporabili najpogostejše zloge in difone, ki jih je program zapisal v analizni fazi iz analize desetih člankov. Članki so imeli različne vsebine. Zloge in difone smo izrezali s pomočjo programa Audacity, iz trideset minutnega posnetka branega besedila. Iz petdeset najpogostejših zlogov smo sestavili testni stavek (" Strokovnjaki so prepričani, da je porast bolnikov posledica nezdravega načina življenja."), z namenom, da bi preverili ali se testirancem sliši bolje sintetiziran govor s pomočjo zlogov ali sintetiziran govor s pomočjo difonov.

Nato smo zbrali testno skupino ljudi, ki je prejela program ter anketo. V anketi so najprej morali poslušati vzorec 1 in odgovoriti na vprašanja, ki so se nanašala na vzorec 1, nato poslušati vzorec 2 in odgovoriti na vprašanja, ki so se nanašala na vzorec 2. V prvem vprašanju pri obeh vzorcih (vzorec 1 in vzorec 2) smo najprej preverjali, če so anketiranci razumeli sintetiziran testni stavek (tako, da so ga morali zapisati), nato pa je bila njihova naloga na anketi označiti razumljivost govora na lestvici od 1 do 5 (1 je najmanj, 5 je največ), za oba vzorca. Vzorec 1 je bil sintetiziran s pomočjo lepljenja difonov, vzorec 2 pa s pomočjo lepljenja zlogov. Poleg razumljivosti testnega stavka, so anketiranci morali označiti še, kako naraven se jim je zdel sintetizirani govor, na lestvici od 1 do 5 (1 je najmanj, 5 največ). Na koncu so morali še označiti, kateri vzorec se jim je na splošno zdel boljši.

Iz rezultatov smo ugotovili, da je 71% ljudi preferiralo sintezo govorora z metodo lepljenja zlogov. Prav tako se je testirancem za 10% zdel bolj razumljiv sintetiziran govor s pomočjo zlogov, ter v povprečju za 4% bolj naraven.

Sklepali smo, da je bil rezultat takšen, ker je lažje odrezati zlog kot difon. Difone je potrebno odrezati na sredini prvega fonema, problem pa je, ker se v fonemu pozna vpliv sosednjega fonema. Ugotovili smo, da je zloge razrezati lažje, ker je njihova energija na koncu zloga majhna. Če sintetiziramo zloge, se poved sliši bolj povezano in naravno kot z difoni. Ker smo ugotovili, da ima zlog različno energijo (frekvenca in amplituda) glede na položaj v besedi, smo podali predlog, da se v bodoče naredi

sintetizator govora, ki ima v slovarju dve variaciji zloga. Ena variacija zloga mora biti odrezana iz sredine ali začetka besede, druga variacija pa iz konca besede. Predlagali smo nadalnjo raziskavo na tem področju.

7 Literatura

[1] K. A. L. Alan W Black. Building synthetic voices, 2014. Dostopano: 11.8.2022. *(Citirano na strani 22.)*

[2] F. Brackhane, R. Sproat, and J. Trouvain. Editing kempelen's"mechanismus der menschlichen sprache": Experiences and findings. In *HSCR 2017 Proceedings of the Second International Workshop on the History of Speech Communication Research Helsinki, August 18-19, 2017*, pages 16–24. TUDpress, 2017. *(Citirano na strani 8.)*

[3] S. Bunc. *Pregled slovnice slovenskega knjižnega jezika*. Jugoslovanska knjigarna, 1940. *(Citirano na strani 4.)*

[4] F. S. Cooper, A. M. Liberman, and J. M. Borst. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences*, 37(5):318–325, 1951. *(Citirano na strani 9.)*

[5] A. Cyrill, S. Felix, and L. Gladence. Text reader for blind: Text-to-speech. *International Journal of Pure and Applied Mathematics*, 117(21):119–125, 2017. *(Citirano na strani 14.)*

[6] H. Dudley, R. R. Riesz, and S. S. Watkins. A synthetic speaker. *Journal of the Franklin Institute*, 227(6):739–764, 1939. *(Citirano na straneh VIII, 8, 9 in 10.)*

[7] T. Dutoi. *An Introduction to Text-to-Speech Synthesis, vol. 3*. Springer book (Kluwer Academic Publishers), The Netherlands, 1997. *(Citirano na straneh VIII, IX, 4, 13, 14, 17, 21, 23 in 24.)*

[8] T. Dutoit and H. Leich. Mbr-psola: Text-to-speech synthesis based on an mbe re-synthesis of the segments database. *Speech Communication*, 13(3):435–440, 1993. *(Citirano na strani 22.)*

[9] J. Gros, A. Mihelic, N. Pavesic, M. Zganec, and S. Gruden. Slovenian text-to-speech synthesis for speech user interfaces. In *WEC (5)*, pages 216–220, 2005. *(Citirano na strani 1.)*

[10] H. S. Gross. prosody, 2022. Dostopano: 11.8.2022. *(Citirano na strani 16.)*

[11] K. Hao. Ai still doesn't have the common sense to understand human language, 2020. Dostopano: 11.8.2022. *(Citirano na strani 19.)*

[12] O. Kamil. Speech sound coding using linear predictive coding. 07 2017. *(Citirano na strani 21.)*

[13] D. Klatt. Text-to-speech conversion. *J Acoust Soc Am*, 82(3):737–793, 1987. *(Citirano na straneh VIII, 9, 10 in 11.)*

[14] V. Kolar. *Fonološki razvoj in govorna razumljivost predšolskih otrok: magistrsko delo*. [V. Kolar], 2017. *(Citirano na strani 3.)*

[15] J. Leskovec. Govorec — sistem za slovensko govorjenje računalniških besedil. Dostopano: 11.8.2022. *(Citirano na straneh 1, 11 in 19.)*

[16] J. J. Ohala. Christian gottlieb kratzenstein: Pioneer in speech synthesis. In *ICPhS*, pages 156–159, 2011. *(Citirano na strani 8.)*

[17] J. Onaolapo, F. Idachaba, J. Badejo, T. Odu, and O. Adu. A simplified overview of text-to-speech synthesis. *Lecture Notes in Engineering and Computer Science*, 1:582–584, 07 2014. *(Citirano na strani 19.)*

[18] S. P. Panda, A. K. Nayak, and S. C. Rai. A survey on speech synthesis techniques in indian languages. *Multim. Syst.*, 26(4):453–478, 2020. *(Citirano na straneh VIII, 19 in 20.)*

[19] N. Pavešić. *Razpoznavanje vzorcev: uvod v analizo in razumevanje vidnih in slušnih signalov*. Založba FE in FRI, 3., popravljena in dopolnjena izd. edition, 2012. 200 izv. Bibliografija na koncu poglavij Kazali. *(Citirano na straneh 3, 4 in 22.)*

[20] B. Pompino-Marschall. Von kempelen et al.: remarks on the history of articulatory-acoustic modelling. *ZAS Papers in Linguistics*, 40:145–159, 2005. *(Citirano na straneh VIII in 9.)*

[21] M. Z. Rashad, H. M. El-Bakry, I. R. Isma'il, and N. Mastorakis. An overview of text-to-speech synthesis techniques. In *Proceedings of the 4th International Conference on Communications and Information Technology*, CIT'10, page 84–89, Stevens Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS). *(Citirano na strani 20.)*

[22] N. Robida. *Sinteza govora in Govorec 3: diplomsko delo*. [N. Robida], Sep 2013. Viri in literatura: f. 47-48 Izvleček; Abstract. *(Citirano na strani 11.)*

[23] N. Robida. Sinteza govora in govorec 3. In H. Tivadar, editor, *Prihodnost v slovenskem jeziku, literaturi in kulturi: zbornik predavanj*, Prihodnost v slovenskem jeziku, literaturi in kulturi: zbornik predavanj, page 117–124. Znanstvena založba Filozofske fakultete, 2014. Izvlečka v slov. in angl. *(Citirano na strani 11.)*

[24] J. Sangeetha, S. Jothilakshmi, S. Sindhuja, and V. Ramalingam. Text to speech synthesis system for tamil. *Int J Emerging Tech Adv En*, 3:170–5, 2013. *(Citirano na strani 26.)*

[25] P. Sarma and S. Sarma. Syllable based approach for text to speech synthesis of assamese language: A review. In *Journal of Physics: Conference Series*, volume 1706, page 012168. IOP Publishing, 2020. *(Citirano na strani 26.)*

[26] T. Šef and M. Gams. Speaker (govorec): a complete slovenian text-to speech system. *International Journal of Speech Technology*, 6(3):277–287, 2003. *(Citirano na straneh VIII, 15 in 16.)*

[27] J. Q. Stewart. An electrical analogue of the vocal organs. *Nature*, 110(2757):311–312, 1922. *(Citirano na strani 8.)*

[28] P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009. *(Citirano na straneh 14, 16 in 18.)*

[29] B. K. Thakur, B. Chettri, and K. B. Shah. Current trends, frameworks and techniques used in speech synthesis–a survey. *International Journal of Soft Computing and Engineering*, 2(2):2231–2307, 2012. *(Citirano na straneh VIII, 13, 14, 15, 16 in 17.)*

[30] A. Thida and C. Su. A comparison between syllable, di-phone, and phoneme-based myanmar speech synthesis. *International Journal of Information Technology and Computer Science*, 10:58–66, 11 2018. *(Citirano na strani 26.)*

[31] H. Tivadar. Slovenska fonetika za tuje študentke in študente. *Drugačnost v slovenskem jeziku, literaturi in kulturi*, pages 111–116, 2016. *(Citirano na straneh VII, 4 in 5.)*

[32] I. za slovenski jezik Frana Ramovša ZRC SAZU. slovenski-pravopis, 2017. Dostopano: 2.8.2020. *(Citirano na straneh VII, 5, 6 in 13.)*

[33] T. Šabanov. *Generiranje slovenskega govora na podlagi učnih množic več govorcev: diplomsko delo: univerzitetni študijski program prve stopnje Računalništvo in informatika*. [T. Šabanov], 2021. Bibliografija: str. 35-40 Povzetek; Abstract:

Generating Slovene speech with multi-speaker datasets. *(Citirano na straneh 20 in 26.)*

[34] T. Šef, A. Dobnikar, and M. Gams. Text-to-speech synthesis in slovenian language. In *9th European Signal Processing Conference (EUSIPCO 1998)*, pages 1–4, 1998. *(Citirano na straneh VIII, 16, 18 in 19.)*

[35] J. Žganec Gros, B. Vesnicer, S. Rozman, P. Holozan, and T. Šef. Sintetizator govora za slovenščino ebralec = the ebralec speech synthesis system for slovenian. In T. Erjavec and D. Fišer, editors, *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september - 1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija*, Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september - 1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija, page 180–185. Znanstvena založba Filozofske fakultete = Ljubljana University Press, Faculty of Arts, 2016. Nasl. z nasl. zaslona Opis vira z dne 12. 1. 2017 Bibliografija: str. 185. *(Citirano na straneh 1, 11, 19 in 22.)*

Priloge

A Anketa

Spoštovani! sem študentka računalništva in za diplomsko nalogo sem izdelala sintetizator govora. Sintetizator govora tekst pretvori v govorno besedo. Uporablja se ga lahko predvsem kot pripomoček za slepe in slabovidne. Sintetizator govora se lahko naredi na več načinov. Da bi ugotovili kateri način je najustreznejši za slovenski jezik, pa je potrebno izvesti anketo, kjer ljudje poslušajo različne verzije sintetizatorja govora in označijo kvaliteto zvoka.

Ker me v moji diplomski nalogi zanima, ali je sintetizator govora bolje izdelati z metodo lepljenja zlogov ali z metodo lepljenja difonov(drugi del prve črke in prvi del druge črke), sem naredila program, ki je zmožen na oba načina glasovno prebrati vstavljeno besedilo. Prosila bi Vas za sodelovanje.

Navodila: Ko odprete program, se vam prikažeta spodaj dva gumba. Na levem spodnjem gumbu piše “Vzorec 1” in na desnem “Vzorec 2”. Kliknite najprej “Vzorec 1” in odgovorite vprašanja pod rubriko “Vzorec 1”. Nato kliknite “Vzorec 2” in odgovorite vprašanja pod rubriko “Vzorec 2”. Nato označite kateri vzorec se vam je slišal bolje.

Vzorec 1 -
Zapišite stavek, ki ste ga slišali ob kliku gumba “Vzorec 1”.
Obkrožite številko razumljivosti govorjenega stavka vzorca 1 na lestvici od 1 – 5:

1. 1 - nerazumljiv
2. 2 - slabo razumljiv
3. 3 - delno razumlji
4. 4 - dobro razumljiv
5. 5 - odlično razumljiv

Kako naraven se vam je zdel govor vzorca 1? Obkrožite številko na lestvici od 1 -5:

1. 1 - čisto robotski
2. 2 - moteče nenaraven
3. 3 - med robotskim in naravnim človeškim govorom
4. 4 - dober približek človeškemu govoru
5. 5 - ne bi ločili sintetiziranega govora od človeka

Vzorec 2 -
Zapišite stavek, ki ste ga slišali ob kliku gumba “Vzorec 2”.
Obkrožite številko razumljivosti govorjenega stavka vzorca 2 na lestvici od 1 – 5:

1. 1 - nerazumljiv
2. 2 - slabo razumljiv
3. 3 - delno razumlji
4. 4 - dobro razumljiv
5. 5 - odlično razumljiv

Kako naraven se vam je zdel govor vzorca 2? Obkrožite številko na lestvici od 1 -5:

1. 1 - čisto robotski
2. 2 - moteče nenaraven
3. 3 - med robotskim in naravnim človeškim govorom
4. 4 - dober približek človeškemu govoru
5. 5 - ne bi ločili sintetiziranega govora od človeka

Obkrožite kateri vzorec se vam sliši bolje

1. 1 - vzorec 1
2. 2 - vzorec 2