Izdelava govorne zbirke za sintezo slovenskega govora

Development of a Speech Corpus for Slovenian Text-to-Speech Synthesis

Tomaž Šef
Institut "Jožef Stefan"
Jamova cesta 39
1000 Ljubljana
+386 1 477 34 19
tomaz.sef@ijs.si

Miro Romih
Amebis d.o.o.
Bakovnik 3
1000 Ljubljana
+386 1 831 10 35
miro.romih@amebis.si

Jerneja Žganec Gros
Alpineon d.o.o.
Ulica Iga Grudna 15
1000 Ljubljana
+386 1 423 94 40
jerneja.gros@alpineon.si

POVZETEK

V članku opisujemo potek izgradnje govorne zbirke za potrebe sinteze slovenskega govora. Zbirka bo primerna tako za HMM-sintezo, kot tudi sodobnejšo WaveNet-sintezo, preprosto pa jo bo prilagoditi tudi na korpusno sintezo. Opisujemo postopke za izbiro povedi, izbiro govorcev, način snemanja govorne zbirke in njenega označevanja.

ABSTRACT

In this paper, we describe the process of building a speech corpus for Slovenian text-to-speech synthesis. The corpus will be suitable for both HMM synthesis and WaveNet synthesis, but it will also be easy to adapt to corpus-based synthesis. We describe the procedures for selecting text, speakers, and the process of recording and annotation a speech collection.

Ključne besede

Govorna zbirka, sinteza slovenskega govora.

Keywords

Speech Corpus, Text-to-Speech Synthesis for the Slovenian language.

1. UVOD

Za potrebe sinteze slovenskega govora v okviru projekta CityVOICE izdelujemo novo govorno zbirko z branim govorom. Takšen govor ustreza najpogostejšim oblikam rabe sintetizatorjev govora. Poleg tega je lažje izdelati njegovo transkripcijo, snemanje je bolj nadzorovano in predvidljivo. Pri spontanem govoru je govorno zbirko težko fonetično in prozodično uravnotežiti.

Najpomembnejši preostali dejavniki, ki smo jih upoštevali pri snovanju nove govorne zbirke za sintezo govora, so: izbira vsebine posnetkov, izbira govorcev, postopek snemanja in označevanje posnetkov.

Izbira velikosti govorne zbirke je posledica kompromisa med želenim številom variacij glasov oz. njihovim pokritjem na eni strani ter časom in stroški, vezanimi na razvoj, na drugi strani. Upoštevali smo tudi čas za kasnejše preiskovanje govorne zbirke in potreben prostor za njeno hranjenje.

2. IZBIRA VSEBINE POSNETKOV

Umetno generirani govor mora zveneti naravno in biti prijeten za poslušanje. Pri izgradnji nove govorne zbirke smo na podlagi preteklih izkušenj [1,2] več pozornosti namenili:

- večji prozodični pestrosti posnetega besedila, ki pokriva najrazličnejše situacije rabe sintetizatorja govora, besedilo vsebuje tudi zelo kratke in zelo dolge povedi,
- poleg najpogostejših besed smo »pokrili« tudi različne prozodične kontekste, v katerih se te besede običajno pojavljajo,
- čim bolj smo se skušali izogibati besedam, ki niso vsebovane v slovarjih izgovarjav, s katerimi razpolaga projektna skupina, saj je zanje pravilen fonetični prepis potrebno izvesti ročno,
- v besedilo smo vključiti pogoste leksikalne termine oz. pogoste besede, kot so denimo telefonske številke, ekonomska terminologija, različne valute, terminologija s področja računalništva in interneta, medicine, pogosta lastna imena, nekatera tuja imena in izrazi, glavni in vrstilni števniki, črkovanje ipd.,
- v zadnjem času je postalo pomembno, da govorna zbirka pokriva še različne situacije, ki nastopajo v dialogu (aplikacije dialoga in simultanega prevajanja; npr. raba v virtualnih asistentih),
- večji zastopanosti raznovrstnih povedi, predvsem vprašalnih in velelnih (pogostost teh povedi je večja, kot je v samem besedilnem korpusu, iz katerega smo zajemali besedilo za branje),
- besedilo zajema različne zvrsti novic, razne napovedi (npr. vremenske napovedi) in podajanje informacij (npr. stanje na cestah, borzne informacije) ter navodil (npr. napotki za vožnjo),
- izbiri ustreznega ženskega glasu – ta je nekoliko nižji in bolj aspiriran (povprečna osnovna frekvenca je nižja kot pri aktualnem ženskem glasu),
- zagotavljanju enakih snemalnih pogojev med posameznimi sejami snemanja,
- obsegu govornega korpusa, ki je precej večji od obstoječega [1].

Izbor vsebine posnetkov oz. branih besedil govorne zbirke za sintezo govora:

- ustvarila se je obsežna tekstovna zbirka besedil, ki je pokrivala različne zvrsti (dnevni časopisi, revije, leposlovje ipd.); uporabili smo tudi besedilni korpus Gigafida , ki vsebuje 1,2 milijarde besed v slovenskem jeziku,
- tokenizacija – iz zbirke besedil smo odstranili vse oznake, vezane na oblikovno podobo (glava besedila, tabele ipd.),
- okrajšave, števila ipd. smo pretvorili v polno besedno obliko (normalizacija besedil),
- besedila smo pretvorili v predvideni fonetični prepis (grafemsko-fonemska pretvorba); izvedli smo ga z modulom za grafemsko-fonemsko pretvorbo, s katerim razpolagamo projektni partnerji.

- obseg zbirke smo optimirali glede na vnaprej pripravljene kriterije (metoda požrešnega iskanja); pri tem smo si prizadevali zagotoviti statistično ustrezno vzorčenje izbranega področja govorjenega jezika.

Izbira povedi ni potekala naključno, pač pa je bila skrbno načrtovana. Postopek izbire povedi je potekal v več korakih:

1. *Statistična obdelava besedil:*
   - Statistično obdelamo celoten besedni korpus in določimo pogostost pojavljanja posameznih glasov in glasovnih nizov v besedilu.
   - Vključimo vse stavke (povedne, velelne, vprašalne itd.) in izdelamo statistiko posameznih vrst povedi oz. stavkov.
2. *Izdelava spiska glasovnih nizov z oceno zaželenosti posameznega niza:*
   - V spisek vključimo nabor vseh teoretično možnih kombinacij difonov.
   - V spisek vključimo vse trifone, štirifone in (po potrebi) ostale zaželene (najpogostejše) polifone, na katere smo naleteli pri statistični obdelavi besedil.
   - Utež oz. ocena zaželenosti niza je odvisna od pogostosti njegovega pojavljanja v besedilu.
3. *Postopek izbire povedi:*
   - Ocenimo doprinos glasovnih nizov za vsako poved iz tekstovnega korpusa.
   - Doprinos povedi je enak vsoti vseh ocen zaželenosti nizov (iz spiska), ki se v povedi pojavijo.
   - Doprinos posamezne povedi normiramo z dolžino povedi (št. besed v povedi ali št. fonemov v povedi).
   - Določimo takšno utež, da bodo dolžine izbranih stavkov čim bolj ustrezale statistični porazdelitvi dolžin stavkov iz korpusa.
   - Izberemo poved z najvišjim normiranim doprinosom.
   - Iz spiska odstranimo vse glasovne nize, ki jih izbrana poved vsebuje.
   - Ponovno ocenimo vsako poved in izberemo najboljšo (glede na novi spisek, v katerem so izločeni tisti glasovni nizi, ki smo jih že pokrili) ter popravimo spisek.
   - Postopek ponavljamo, dokler ne izberemo želenega števila povedi.
4. *Ovrednotenje rezultatov:*
   - Vsakih 1000 povedi izdelamo statistiko difonov, trifonov, štirifonov in drugih polifonov, ki jih že pokrivamo (gre za glasovne nize, ki smo jih do takrat že izločili iz zgoraj omenjenega spiska).
5. *Dodatne izboljšave algoritma:*
   - Ker mora zbirka vsebovati vse možne kombinacije difonov, algoritem popravimo tako, da difone dodatno utežimo glede na ostale polifone. Na takšen način algoritem na začetku daje prednost povedim, ki pokrijejo čim več novih difonov. Predvidoma se vsi difoni pokrijejo že po ca. 100 stavkih.
   - Pri trifonih in štirifonih upoštevamo pri robnih glasovih tudi podatek o glasovni skupini, ki ji pripadajo (npr. štirifon "krak" ne bo doprinesel prav dosti novega v našo zbirko, če ta že vsebuje štirifon "krat"; zato oceno koristnosti takega štirifona popravimo navzdol). To lahko naredimo preprosto tako, da v spisek vnesemo dodatne nize skupaj z njihovimi frekvencami pojavljanja v korpusu (primer takega štirifona: "k"+"r"+"a"+"pripornik").
   - Algoritem z različnim uteževanjem izboljšamo tako, da končni nabor vsebuje različne povedi (povedne, vprašalne, velelne, enostavne, sestavljene, naštevanje itd.). Tako lahko isti korpus učinkovito uporabimo tudi za generiranje prozodičnih parametrov pri sintezi govora.

3. IZBIRA GOVORCEV

Pridobili smo posnetke preko 10 različnih govorcev. Te krajše posnetke (nekaj deset stavkov z dobrim pokritjem difonov) smo nato strojno označi in preizkusili na aktualnem sintetizatorju govora. Posnetke je poslušalo več poslušalcev, ki so podali svojo oceno glede naravnosti in razumljivosti govora, pa tudi glede subjektivne ocene, kateri glas se jim je zdel najprijetnejši za poslušanje. V praksi se izkaže, da so nekateri glasovi preprosto bolj primerni za izdelavo sintetizatorjev govora kot drugi. Pri tem je težko vnaprej napovedati, ali je nek glas primeren ali ne, pri tem ni splošno veljavnih pravil.

Pri izbiri govorca smo upoštevali tudi njegovo sposobnost sledenja napotkom, potrebne ponovitve med snemanjem, čas snemanja ipd. Posneli smo en moški in en ženski glas. Smotrno je, da sintetizator govora razpolaga s po dvema glasovoma za vsak spol.

Kandidate smo vnaprej seznani z namenom snemanja in možne uporabe tako pridobljenih glasov. Pred snemanjem so morali izbrani govorci podpisati pogodbo oz. privolilo, da dovolijo rabo posnetkov za potrebe sinteze govora.

4. POSTOPEK SNEMANJA

Snemanje govornega gradiva je potekalo ob prisotnosti izkušenega snemalnega operaterja z namenom, da se je preprečilo neustrezne izgovarjave besed in napake pri snemanju govora. Govorcu smo pred snemalnimi sejami podali ustrezna navodila in ga zaprositi, da povedi prebira razločno in enakomerno hitro. Med branjem besedila so imeli govorci nameščene elektrode laringografa, s katerimi smo spremljali nihanje njihovih glasilk zaradi lažjega kasnejšega označevanja osnovnih period govornega signala.

Govor smo snemali preko kvalitetnega mikrofona v digitalni obliki in sicer na namenski računalnik v studiu. Pred mikrofonom se je nahajal ustrezen filter (angl. »anti-pop filter«), ki je zadušil razne poke, tleske ipd. Potrebna je bila še ustrezna mešalna miza, zaslon in slušalke (preko katerih govorec prejema navodila in posluša povratni govor). Posnetke govora smo shranili v digitalni obliki na trajne računalniške pomnilniške medije. Frekvenca vzorčenja je 44,1 kHz, 24-bitno.

Samo snemanje celotne govorne zbirke je zaradi obsežnosti besedila, ki ga je bilo potrebno prebrati, trajalo več mesecev. Pri tem so morale nastavitve snemalne opreme ves čas ostati nespremenjene. Oseba, ki je snemanje nadzorovala, je preverila položaj govorca pred vsako snemalno sejo in jo primerjala s položaji v predhodnih sejah. Pred vsakim snemanjem je govorec poslušal svoje predhodne posnetke, s čimer se je skušal zagotoviti čim bolj enoten način govora med posameznimi snemalnimi sejami. Na začetku snemanja posamezne seje je govorec prebral nekaj vnaprej določenih fiksnih stavkov, ki so omogočali primerjavo glasnosti in višine govora med posameznimi snemalnimi sejami. Govorec ne sme biti preveč utrujen, zato je 10-minutno snemanje potekalo znotraj pol urnega intervala. Posamezna seja je trajala dve uri; znotraj tega časa smo lahko posneli 40 minut govornega materiala. Vsak govorec je opravil le eno dvourno sejo na dan.

5. OZNAČEVANJE POSNETKOV

Uporabljamo tri nivoje anotacij oz. prepisov govorjenega besedila: grafemski prepis, fonetični prepis in prozodijske oznake. Ker je ročna segmentacija govora na fonetičnem nivoju naporna in

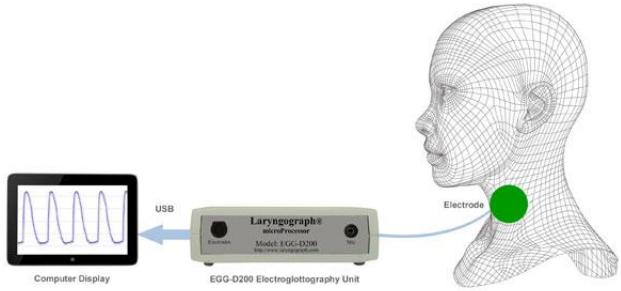
dolgotrajna, smo pri tem uporabili vsaj delno avtomatizirane postopke, ki so bolj učinkoviti, če vnaprej poznamo grafemski prepis govorjenega gradiva. Avtomatskim metodam in postopkom je sledilo »ročno« popravljanje oznak, kar je ne glede na hiter razvoj tehnologije še vedno zelo zamudno. Osnovne periode govornega signala smo označili s posebnim algoritmom temelječim na izhodnem signalu larigografa.

OPIS LARINGOGRAFA

Laringograf (slika 1) je elektronska naprava za nadzor in analizo vibriranja glasilk. Obsega strojno opremo (elektrode, mikrofon, laringograf mikroprocesor priključen na računalnik) in programsko opremo (Speech Studio). Akustični signal (govor) je posnet preko mikrofona, signal elektrolaringografa pa preko dveh elektrod nameščenih preko oz. na obeh straneh glasilk. Mikrofon in elektrode so priključene na laringograf mikroprocesor, slednji pa je priključen na računalnik (slika 2).

Deluje tako, da se preko dveh elektrod, nameščenih na obeh straneh govorčevega vratu na nivoju ščitnice, spusti šibak električni tok. Z napravo nato merimo električno impedanco vratu. Ker se impedanca spreminja z nihanjem glasilk, dobimo signal, ki je dobro koreliran z osnovnim tonom izgovorjenih glasov. Časovni potek impedance je označen kot signal Lx. Lx signal se lahko nadalje obdela z namenom pridobitve informacije o času trajanju period nihanja vokalnega trakta. Dobra lastnost signala laringografa je, da je odporen na akustični šum prisoten v signalu mikrofona [3, 4].

Slika 1: Laringograf z elektrodama [5].



Slika 2: Uporaba laringografa [6].

ZAJEM PODATKOV LARINGOGRAFA

Zajem podatkov laringografa obega pravilno izbiro elektrod glede na fiziološke značilnosti govorca, njihovo pravilno namestitev in shranjevanje signala laringografa Lx.

Pravilna izbira velikosti in namestitev elektrod laringografa je zelo pomembna. Za vsakega uporabnika laringografa si je bilo potrebno vzeti dovolj časa in prilagoditi nameščene elektrode tako, da je bil signal čim močnejši. Signal laringografa in njegovo jakost pri tem opazujemo na zaslonu računalnika. Signal je najmočnejši takrat, ko sta elektrodi na nivoju glasilk. Takrat je namreč električno polje med elektrodama najbolj podvrženo vplivu samega stika med glasilkama. Pri osebah z velikim izrazitim grlom je bila namestitev elektrod preprostejša, saj smo lahko enostavno locirali obe strani ščitničinega hrustanca, dovolj velik kot elektrod pa je omogočal koncentracijo električnega polja okoli samih glasilk. Kadar grlo ni bilo tako izrazito, še posebej ko je bilo obloženo z večjimi količinami maščobnih blazinic, ga je bilo težje natančno locirati (položaj se med govorjenjem znatno spreminja). Posledično je bilo težje pridobiti »dober« in ustrezno močan signal laringografa Lx. Vendar pa za ugotavljanje osnovnih period k sreči zadostuje že relativno šibak signal. Detajli Lx signala pa so slabše razvidni oz. določljivi.

Pri ženski govorki (sliki 3) smo preizkusili večje število elektrod. Kot primerne so se izkazale elektrode s premerom med 30 mm in 22 mm, medtem ko so se manjše elektrode (s premerom pod 16 mm) izkazale za neustrezne. Na koncu smo kot optimalni izbrali elektrodi s premerom 22 mm.

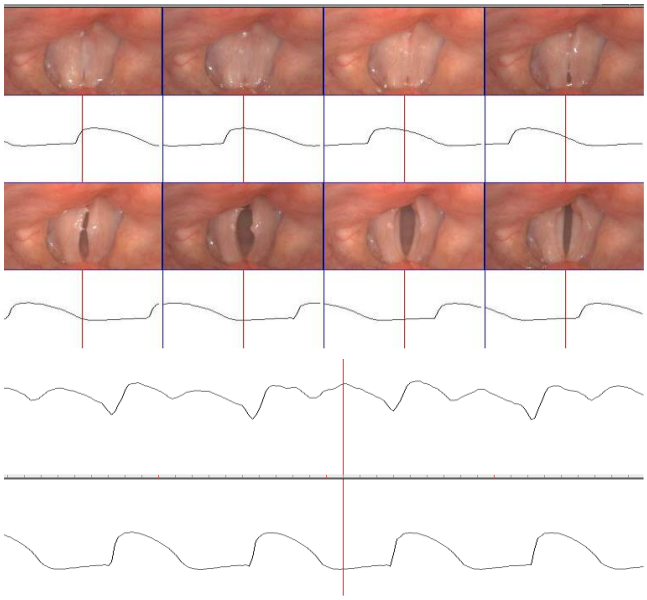
Slika 3: Govorka med snemanjem oz. branjem pripravljenega gradiva v tonskem studiu. V ozadju tonski tehnik, ki med snemanjem na zaslonu spremlja tako signal laringografa Lx kot tudi mikrofonski signal Sp.

OBDELAVA PODATKOV LARINGOGRAFA

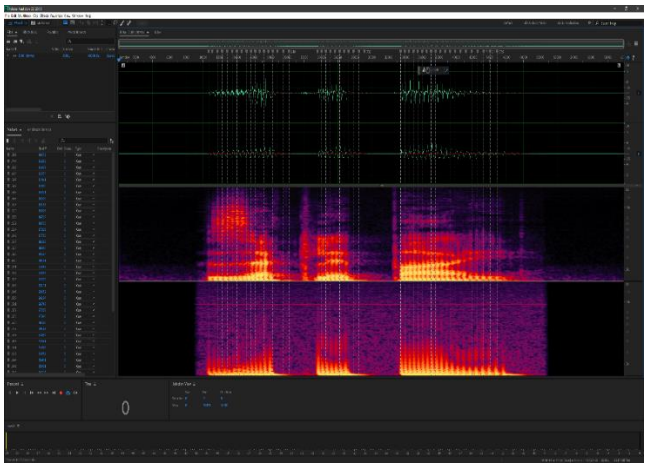
Obdelava podatkov laringografa obega označevanje osnovnih period govornega signala s pomočjo zajetega signala.

Na sliki 4 je prikazanih osem zaporednih položajev glasilk med tvorjenjem zvenečega glasu moškega govorca. Začetno zapiranje glasilk se odraža v hitro naraščajoči krivulji Lx signala. To zapiranje povzroči akustični odziv Sp (mikrofon) signala. Pri Sp signalu tipično opazimo krajši zamik glede na Lx signal. Najvišja točka signala Lx predstavlja trenutek, ko preko elektrod laringografa steče maksimalen tok. To ustreza trenutku, ko sta glasilki v največjem kontaktu oz. najbolj skupaj. Ko sta glasilki najbolj narazen, je amplituda signala Lx najnižja. Širina konice signala pove, kako dolgo sta glasilki zaprti znotraj posameznega cikla vibriranja. Čas potreben za prehod iz ene točke signala Lx do enake točke na naslednji konici imenujemo perioda. Na takšen način lahko dobimo informacijo o osnovni frekvenci vibriranja glasilk (Fx). Informacijo o načinu odpiranja in zapiranja glasilk pa

lahko dobimo iz naklona dviganja in spuščanja signala Lx. Prav tako lahko razberemo, kako dolgo sta bili glasilki zaprti.



Slika 4: Gibanje glasilk, moški glas, 120 Hz. Prikazana je serija osmih slik položaja glasilk, pridobljenih s stroboskopom, s pripadajočo oznako mesta na Lx signalu pod vsako sliko. Spodaj sta Sp (mikrofon) in Lx (laringograf) signala za šesto sliko. [5]



Slika 5: Primer govornega signala z označenimi osnovnimi periodami. Zgoraj je govorni signal posnet z mikrofonom, sledi signal laringografa Lx in spektralna prikaza obeh signalov. Navpične črte predstavljajo oznake period govornega signala.

Najprej smo preizkusili preprost algoritem za zaznavanje pulza in izračun časa trajanja med dvema impulzoma [7]. Začetek impulza predstavlja prvi vzorec, katerega amplituda je manjša od nič in hkrati manjša ali enaka amplitudi naslednjega vzorca. Konec impulza pa predstavlja zadnji vzorec, katerega amplituda je manjša kot nič in manjša ali enaka amplitudi predhodnega vzorca. Širina pulza je definirana kot časovna razlika med začetkom in koncem impulza. Vrh pulza predstavlja vzorec z največjo amplitudo med začetkom in koncem impulza (vrednost je vedno večja od nič). Trenutek pulza je določen kot prvi vzorec z amplitudo vrha pulza. Širina pulza je morala biti širša od od štirih vzorcev, vrh pulza pa je moral presegati neko arbitrarno določeno vrednost. Izračunal se je čas trajanja med dvema pulzoma, vrednost pa se je pretvorila v Hertze.

Potrebno je bilo izvesti še razločevanje med zvenečimi in nezvenečimi deli signalov. V nadaljevanju pa smo uvedli še omejitve signala Lx in sicer tako, da je Lx > 50 Hz za moški glas in Lx > 120 Hz za ženski glas. Znotraj posameznega zvenečega predela pa so morali biti vsaj trije pulzi. Takšen algoritem je že uporabna referenca za določevanje osnovne frekvence govornega signala.

Zanesljivost prvotno zasnovanega algoritma smo izboljšali tako, da smo ga dopolnili z avtokorelacijsko metodo začetne ocene osnovne frekvence. Prav tako smo predlagali dodatne dopolnitve pri prepoznavanju osnovnih period zahtevnejših fonemov, kot je to npr. fonem »r«. Z analizo koeficientov HNM smo povečali natančnost določevanja osnovne frekvence ter sredine okna. Ob tem je bilo potrebno detektirati situacije, ko je osnovni algoritem »odpovedal« in takšne situacije še posebej obravnavati in razrešiti. Z obsežnim testiranjem smo prišli do končnega algoritma, ki se je izkazal kot dovolj robusten in natančen za potrebe izgradnje govorne zbirke za sintezo govora v okviru projekta CityVOICE.

6. ZAKLJUČEK

Pri izdelavi govorne zbirke CityVOICE smo posebno pozornost namenili določanju optimalnih pogojev za snemanje govornih zbirk, določanju optimalnih fonetično in drugače uravnoteženih vsebin za snemanje govornih zbirk ter rešitvam za iskanje optimalnih govorcev.
Pri označevanju osnovni govornih period smo si pomagali s signalom laringografa. Algoritem za označevanje osnovnih period prebranega besedila, skupaj s dodatno spremljevalno kodo, je bil implementiran v programskem orodju Matlab. Prvotno zasnovani algoritem je bilo tekom preizkušanja potrebno dopolniti z obravnavo bolj zahtevnih fonemov, kot je to denimo fonem »r«, poleg tega je bilo potrebno razviti še detekcijo in obravnavo posebnih situacij, ko je osnovni algoritem »zgrešil« pravilno namestitev oznak period.

7. ZAHVALA

Operacijo CityVOICE sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj, in sicer v okviru »Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020«.

8. LITERATURA IN VIRI

[1] Žganec Gros, J., Vesnicer, B., Rozman, S., Holozan, P., Šef, T. 2016. Sintetizator govora za slovenščino eBralec. *Konferenca Jezikovne tehnologije in digitalna humanistika*, Ljubljana.

[2] Šef, T., Romih, M. 2011. Zasnova govorne zbirke za sintetizator slovenskega govora Amebis Govorec, *Informacijska družba IS 2011*.

[3] Wikipedia (https://en.wikipedia.org/wiki/Electroglottograph), 2019.

[4] Laryngograph and Nasality processor, tehnična dokumentacija podjetja Laryngograph Ltd., 2019.

[5] http://www.laryngograph.com, 2019.

[6] http://www.rose-medical.com/electroglottography.html, 2019.

[7] Bagshaw, P., C., Hiller, S., M., Jack, M., A., 1993. *Enhanced Pitch Tracking and the Processing of F0 contours for computer aided intonation teaching*. Center for Speech Technology Research, University of Edinburgh, UK.