doi:10.5128/ERYa19.07

KÕNEVEEB JA MINU HÄÄL: UUS KÕNESÜNTEESIKESKKOND JA -TEENUS

Meelis Mihkla, Indrek Hein, Indrek Kiissel, Jaan Pajupuu, Liisi Piits, Heete Sahkai, Hille Pajupuu, Rene Altrov, Elgar Kudritski, Liis Ermus, Egert Männisalu, Kristjan Suluste

Ülevaade. Tekst-kõne-süntees on tehnoloogia, mis muudab kirjaliku teksti kõneks. Kõnesünteesi kasutusvaldkond on tänapäeval väga lai. Vajadus kõnesünteesi oma seadmesse paigaldada või tootesse integreerida on muutunud üldiseks. Käesolevas artiklis tutvustame Eesti Keele Instituudis arendatavat interaktiivset kõnesünteesi veebikeskkonda Kõneveeb, mille eesmärgiks ongi ühiskonna, ettevõtjate ja arendajate vajadusi arvestades pakkuda tasuta ühest kohast ja mugavalt kõiki Eesti Keele Instituudi kõnesünteesiga seotud teenuseid ja ressursse. Põhjalikumalt tutvustame Kõneveebis saadaval olevat uut teenust Minu Hääl, mis võimaldab igaühel luua ise sünteeshääle ilma mingite tehniliste eelteadmisteta kõnesünteesist. See teenus on mõeldud üksikisikutele ja ettevõtjatele, kelle vajadusi olemasolevad sünteeshääled ei rahulda, näiteks kellel on vaja spetsiifilises stiilis või unikaalset sünteeshäält või omaenda häälest tehtud sünteeshäält.*

Võtmesõnad: kõne, kõnetehnoloogia, kõnekorpused, masinõpe, kõnesünteesiteenus, eesti keel

1. Sissejuhatus

Tekst-kõne-süntees on tehnoloogia, mis muudab kirjaliku teksti kõneks. Kaasaegne kõnesüntees kasutab salvestatud inimhäält treeningkorpusena, mis koosneb piisavast hulgast tekstilistest lausetest ja neile vastavatest etteloetud lausetest helifailidena. Salvestatud treeningkorpust nimetatakse ka kõnesünteesi doonorhääleks. Korpuse põhjal treenitakse kõnemudelid, kasutades erinevaid masinõppemeetodeid, nagu Markovi peitmudelid ja, üha enam, tehisnärvivõrkudel põhinevad meetodid.

* Kõneveebi keskkonna ja Minu Hääle teenuse valmimist toetas Eesti Keele Instituudi projekt "Kõnesünteesi veebikeskkond KÕNEVEEB" ASTRA II taotlusvoor (Euroopa Regionaalarengu Fond, projekti kood: 2014-2020.4.01.20-0291), projekt TK145 "Eesti-uuringute tippkeskus" (Euroopa Regionaalarengu Fond), projekt EKI-BAAS-2021/2 " Kõnesüntees keeleteaduse teenistuses ja vice versa I etapp (2021–2022)" (Haridus- ja Teadusministeeriumi Eesti Keele Instituudi baasfinantseerimine), Haridus- ja Teadusministeeriumi Eesti Keele Instituudi baasfinantseerimise projekt EKI-BAAS-2023/2 "Kõneuuringud kõnetehnoloogia teenistuses (2023–2024)" ja riikliku programmi "Eesti keeletehnoloogia (2018–2027)" projekt "Väljendusrikas ja mitmekesine eestikeelne kõnesüntees".

Kõnesüntesaatori esikomponendiks on tavaliselt tekstitöötlusmoodulid, mis lisavad morfoloogilise märgenduse, muudavad ortograafilise teksti häälduspäraseks, teisendavad numbrid ja lühendid sõnadeks jmt. Kõnesünteesil treenitavad kõnemudelid sisaldavad kõneüksuste akustilisi mudeleid ning kõneprosoodia kestuste ja põhitooni mudeleid. Kõnemudelite abil teisendatakse eeltöödeldud ortograafiline sisendtekst ortoeepiliseks väljundkõneks (ülevaateid kõnesünteesi kohta vt nt Klabbers 2019, Georgila 2017).

Kõnesünteesi kasutusvaldkond on tänapäeval lai. Kõnesünteesi üks oluline funktsioon on erivajadustega inimeste juurdepääsuvõimaluste avardamine. Kõnehäirega, hääle kaotanud või häälekahjustusega inimesed saavad süntesaatori abil siiski teksti vahendusel häälega suhelda. Vaegnägijatel ja vaeglugejatel (düslektikutel) on tänu kõnesünteesile juurdepääs elektroonilistele tekstidele. Ka väljaspool neid sihtrühmi on kõnesünteesi kasutajaskond üha suurem. See võimaldab ammutada tekstidest informatsiooni paralleelselt muude tegevustega nagu autojuhtimine, bussisõit, jalutamine või sportimine, mistõttu kõnesüntees on muutunud näiteks e-raamatuid pakkuvate rakenduste ja meediaväljaannete veebiportaalide standardseks komponendiks. Samuti on kõnesünteesi vaja nii kodumasinates ja isejuhtivates autodes kui ka uutes nutikates abivahendites (assistendid, juturobotid, nõuandjad, lepingute sõlmijad), mis suhtlevad tehisintellekti ajastul inimkeeles (Mihkla, Piits 2022). Kombinatsioonis kõnetuvastuse ja masintõlkega rakendatakse kõnesünteesi häältõlkes ehk kõne-kõneks tõlkes ja automaatses dubleerimises.

Eestikeelse kõnesünteesiga tegeletakse Eesti Keele Instituudis (EKI), Tartu Ülikoolis[1], Google'is ja Microsoftis. Eestikeelne kõnesüntees on muutunud igapäevaelu koostisosaks ning seda rakendatakse paljudes valdkondades. Mõned rakendusnäited on audioraamatute genereerimine näiteks kooliõpikute portaalis Opiq[2], Elisa Raamatu Iselugejas ja EKI keskkonnas Vox Populi; veebiartiklite ettelugemine Eesti Meedia ja Ekspress Meedia meediaportaalides; audioväljund raamatukogude digitaalarhiivile Digar[3]; EKI sõnaraamatute näitelausete helindamine; riigi virtuaalne abiline Bürokratt; laste suhtlusabivahend robot Pepper; kommunikatsioonabivahendid nagu Communicator 5[4]; Google'i rakendused ja Microsofti kontoritarkvara. Eesti Televisiooni kanalites on võimalik subtiitritega varustatud võõrkeelseid saateid ja filme jälgida koos sünteeshäälega etteloetavate subtiitritega. See lahendus on eelkõige mõeldud vaegnägijatele ja -lugejatele, kes sageli ei suuda või ei jõua subtiitreid lugeda. EKI sünteeshääled on üles astunud isegi teatrietendustes ja õhtujuhtidena.

Eestikeelse kõnesünteesi kasutuspotentsiaal ja -vajadus on siiski tunduvalt laiem. Võimalikud täiendavad kasutusalad on kõnelevad virtuaalassistendid, kodumasinad, isejuhtivad autod, arvutimängude jmt dubleerimine, subtiitrite helindamine kõigis telekanalites, keeleõpe, logopeedia, veel laialdasem e-raamatute helindamine, kõnetuvastuse ja masintõlkega kombineeritud lahendused jpm.

Kõnesünteesi rakendusvaldkonna laienemine tähendab seda, et üha rohkematel üksikisikutel ja ettevõtetel on vajadus see kasutusele võtta. Samuti tähendab see seda, et kõikideks praegusteks ja tulevasteks funktsioonideks ei ole võimalik pakkuda sobivaid valmis sünteeshääli. Seega võiks kõnesünteesi kasutuselevõtt ja uute sünteeshäälte loomine olla sama lihtne ja enesestmõistetav nagu raamatupidamis- või personalihaldusprogrammi või pildi- või helitöötlustarkvara soetamine.

[1] https://neurokone.ee (29.3.2023).
[2] https://www.opiq.ee (29.3.2023).
[3] https://www.digar.ee (29.3.2023).
[4] https://www.tobiidynavox.com/pages/communicator-5-ap (29.3.2023).

Lihtne kasutuselevõtt on oluline ka sellepärast, et keeletehnoloogiline tugi on vajalik keele säilimiseks (vt nt Mihkla, Piits 2022). Kõnesünteesi ja selle rakendusvõimaluste peamine areng toimub ingliskeelses maailmas. Eesti keele säilimiseks on seetõttu vaja tagada, et nii olemasolevate kui uute rakenduste eestikeelseks muutmine oleks võimalikult lihtne. Keeletehnoloogial on ka suur potentsiaal eesti keele kasutusala senisega võrreldes oluliselt laiendada, näiteks tõlkides ja helindades tooteid, mida inimressursiga ei jõutaks kunagi eestikeelseks muuta, nagu arvutimängud.

Käesolevas artiklis (ptk 2) tutvustame Eesti Keele Instituudis arendatavat interaktiivset kõnesünteesi veebikeskkonda Kõneveeb (kõneveeb.ee), mille eesmärgiks ongi ühiskonna, ettevõtjate ja arendajate vajadusi arvestades pakkuda tasuta ühest kohast ja mugavalt kõiki Eesti Keele Instituudi kõnesünteesiga seotud teenuseid ja ressursse. Kõneveebi keskkond on loodud selleks, et toetada kõnesünteesi kasutuselevõttu ja arendamist: et kõigil üksikisikutel oleks lihtne seda oma seadmetesse paigaldada, et ettevõtjad saaksid seda mugavalt oma teenustesse ja toodetesse integreerida, et arendajatel oleks võimalikult lihtne uusi sünteesimeetodeid arendada ja katsetada.

Lisaks Kõneveebi keskkonnale tutvustame põhjalikumalt Kõneveebis saadaval olevat uut teenust Minu Hääl, mis võimaldab igaühel luua ise sünteeshääle ilma mingite tehniliste eelteadmisteta kõnesünteesist (ptk 3). See teenus on mõeldud üksikisikutele ja ettevõtjatele, kelle vajadusi olemasolevad sünteeshääled ei rahulda, näiteks kellel on vaja spetsiifilises stiilis või unikaalset sünteeshäält või omaenda häälest tehtud sünteeshäält.

Artikkel lõpeb kokkuvõtte ja tulevikuplaanide tutvustusega (ptk 4).

2. Kõnesünteesikeskkond Kõneveeb

Eesti Keele Instituudis arendatav kõnesünteesi veebikeskkond Kõneveeb koondab EKI eestikeelse kõnesünteesiga seotud teenuseid ja ressursse üksikisikutest kasutajatele, ettevõtetele ja kõnesünteesi arendajatele. EKIs on arendatud eestikeelset kõnesünteesi alates 1980. aastate algusest. Viimasel aastakümnel on arendatud tänapäevast eestikeelset korpuspõhist tekst-kõne-sünteesi, treenitud mitmeid sünteeshääli ja loodud liidesed nende kasutamiseks platvormide Windows, Android ja OSX häälrakendustes. Samuti on loodud häälrakendusi ja tekstide helindamise teenuseid. Kõneveebi eesmärk on nii neile olemasolevatele kui tulevastele ressurssidele mugavat juurdepääsu pakkuda. Kõik teenused ja ressursid on tasuta, kuid mõned nõuavad kasutajaks registreerumist.

Peamise teenusena pakub Kõneveeb tekstide helindamist mitmesuguste sünteeshäältega. Teksti helindamiseks kirjutab või kleebib kasutaja veebilehele teksti, mis teisendatakse kõneks. Tulemust saab veebis kuulata või helifailina alla laadida.

Pikkade tekstide ja raamatute helindamiseks on loodud eestikeelsete heliraamatute genereerimiskeskkond Vox Populi (Mihkla jt 2017, 2018). Kasutaja laadib üles tekstifaili ja saab vastu valitud häälega tekitatud helifaili. Teine abivahend Vox Populi teenuse juures on transkriptsioonisõnastik, mis esitab võõrnimed loetaval häälduspärasel kujul.

Kuna kõnesünteesi eri kasutusfunktsioonid nõuavad erinevaid tehnilisi parameetreid, hääli ja kõnestiile, on vajalik sünteeshäälte mitmekesisus, see tähendab,

eri meetoditel ja korpustel treenitud sünteeshäälte valik (Kato jt 2020, Piits jt 2022, Shin jt 2022). Praegu on esindatud hääled, mis on treenitud mees-, nais- ja lapshäälte ning eesti-, võru- ja kihnukeelsetel korpustel. Kõnestiilidest on esindatud neutraalse loetud kõne, etteloetud ilukirjanduse ja spontaanse kõne korpused. Hääled on treenitud eri sünteesimeetodeid, arenduskeskkondi ja tööriistu kasutades.

EKIs on praegu kasutusel viis kõnesüntesaatorit.

1. Häälikupõhine HMM[5] on Markovi peitmudelitel põhinev kõnesüntees, mis baseerub eesti keele häädusreeglitel, st teksti analüüsil määratakse kindlaks sõnade välde, konsonantide palatalisatsioon jne.
2. Tähemärgipõhine HMM[6] on Markovi peitmudelitel põhinev kõnesüntees, kus ei kasutata tähemärkide häälikuteks teisendamist ja väljundkõne genereeritakse vaid tähemärke sisaldava teksti põhjal.
3. Tähemärgipõhine DNN[7] on sügavatel närvivõrkudel põhinev kõnesüntees, mis on tähemärgipõhine ega kasuta eesti keele häädusreegleid.
4. Häälikupõhine DNN[8] on sügavatel närvivõrkudel põhinev kõnesüntees, mis baseerub eesti keele häädusreeglitel.
5. Transformer TTS[9] on sügavatel närvivõrkudel põhinev tähemärgipõhine kõnesüntees.

Peale tekstide helindamise veebis saab kõnesüntesaatoreid alla laadida. Sünteeshäälte kasutuselevõtmiseks oma seadmes on Kõneveebis saadaval kõnesünteesiliidesed ja nende installimisjuhendid platvormidele Windows, Android ja OSX. Paigaldatud eestikeelseid hääli saab kasutada Windowsi, Androidi ja OSX-i häälrakendustes, nagu ekraanilugejad, e-raamatute ettelugemise rakendused, navigeerimisrakendused, sõnumite ettelugemine jms. Lisaks valmis sünteeshäältele pakub Kõneveeb oma sünteeshääle treenimise, veebis kasutamise ja alla laadimise võimalust, mida tutvustame järgmises peatükis.

3. Uus kõnesünteesiteenus Minu Hääl

Kui kõnesünteesi kasutajate hulk suureneb, ei piisa enam valmis häältest. Kasutajal võib olla vaja unikaalset või spetsiifilist sünteeshäält, näiteks ettevõtte brändina, konkurentidest eristumiseks, spetsiifilist stiili nõudvateks kasutusteks, nagu arvutimängude või multifilmide dubleerimine, kindla teemavaldkonna jaoks vms. Samuti võib kasutaja vajada enda häälel põhinevat sünteeshäält, näiteks mõne elukutse puhul, kus on vaja palju ettekirjutatud teksti salvestada (nt taskuhäälingu saatejuht), või kui hääl on mingi seisundi või haiguse tõttu kas ajutiselt või püsivalt kadumas või lihtsalt soovitakse oma häält konserveerida mingi tulevikurakenduse tarbeks (nt Judge, Hayton 2022).

Eestikeelse kõnesünteesi küllalt lai kasutajaskond ja uute sünteesimeetodite levik oli motivatsiooniks vajadusele luua interaktiivne kõnesünteesi teenus Minu Hääl, kus juba praegused ja ka tulevased häälrakendusi pakkuvad ettevõtted ja ka üksikisikud saaksid erinevatel meetoditel endale sobivate häältega kõnesüntesaatoreid luua.

Teenus Minu Hääl võimaldab kasutajal tasuta ja ilma igasuguste tehniliste eelteadmisteta kõnesünteesi kohta ise sünteeshääle treenida. Teiste keelte jaoks on loodud sarnaseid, enamasti tasulisi keskkondi (nt Microsofti Custom Neural

[5] https://github.com/ikiissel/synthts_et (29.3.2023).
[6] https://github.com/jaotus/grafeem (29.3.2023).
[7] https://github.com/jaotus/grafeem (29.3.2023).
[8] https://github.com/ikiissel/mrln_et (29.3.2023).
[9] https://github.com/TartuNLP/TransformerTTS (29.3.2023).

Voice[10], Google'i Custom Voice[11], Acapela My-own-voice[12]), kuid eesti keele jaoks on see meile teadaolevalt ainuke selline teenus.

Sünteeshääle loomine ja kasutuselevõtmine koosneb kolmest etapist:

1) kõnekorpuse salvestamine (ptk 3.1);
2) sünteeshäälte treenimine (ptk 3.2);
3) treenitud hääle veebipõhine kasutamine või installimine (ptk 3.3).

3.1. Kõnekorpuse salvestamine

Sünteeshääle treenimiseks on vaja treeningkorpust (vähemalt 1 h), mis koosneb helifailidest (wav) ja neile vastavatest tekstifailidest (txt, UTF-8). Iga fail peab vastama ühele lausele. Kui kasutajal ei ole sobivat korpust olemas, leiab ta Kõneveebist kõik vajaliku selle salvestamiseks. Korpuse salvestamiseks on loodud salvestusprogramm Salvestaja ning tekstikorpus, mis on piisavalt suur sünteeshääle treenimiseks ja katab kõik eesti keele häälikud ja häälikuühendid.

Tekstikorpus[13] koosneb 1192 lausest, mille loomisel on lähtutud mitmesugustest põhimõtetest. Juba esimeste eestikeelsete korpuspõhiste süntesaatorite loomisel hakati koostama ettelugemiseks sobivat tekstikogu. See materjal on aastate jooksul muutunud olenevalt sellest, mis meetodil sünteesi kasutati. Difoonsüntesaatori loomisel koostati sõnaloend, mis sisaldaks kõiki eesti keeles võimalikke häälikuid ja difoone ehk külgnevaid häälikupaare (vt Mihkla jt 1998). Järgmises etapis paigutati need sõnad üksuste valiku sünteesi tarbeks lausekonteksti, lisati sagedamaid sõnu, sõnavorme ja sõnaühendeid, sagedamaid nimesid ja saadi 400-lauseline korpus (Piits jt 2007). Järgnevatel aastatel lisati tekstikorpusesse peamiselt tüüpilisemaid viisakusväljendeid ja otsiti ajakirjanduskorpusest juurde lauseid, mis sisaldaksid harvem esinevaid häälikuühendeid. Lõpuks oli neutraalseid üksiklauseid sisaldava tekstikorpuse maht kasvanud 3350-lauseliseks[14].

Algselt olid laused mõeldud lugemiseks professionaalsetele lugejatele nagu näitlejad ja diktorid. Minu Hääle teenuse loomisega laienes võimalike lugejate ring ja tekstikorpuses esinevad laused tuli kohendada tavalugeja oskustele vastavaks. Lihtsustasime lauseid, paigutades keerulisema häälduse sõnu eraldi lausetesse. Vähendasime korpuse mahtu *ca* 1200 lauseni, mis osutus testimisel piisavaks hulgaks, et treenida hästi kõlavat sünteeshäält. Välja jätsime laused, mis võiks lugejas tekitada negatiivseid emotsioone (nt haiguste, õnnetuste või surmaga seotud laused), samuti asendasime kindlatele isikutele viitavad pärisnimed väljamõeldud nimedega.

Korpuse salvestamiseks on võimalik kasutada vabalt valitud salvestustarkvara (nt Audacity ®[15] või SpeechRecorder[16]), kuid need ei pruugi kasutajatele piisavalt lihtsad olla ning meil puudub võimekus nende osas tuge pakkuda, mistõttu otsustasime luua oma rakenduse Salvestaja, mis sisaldab ainult vajalikku funktsionaalsust ning on kõigile kasutajatele mugav ja arusaadav (joonis 1).

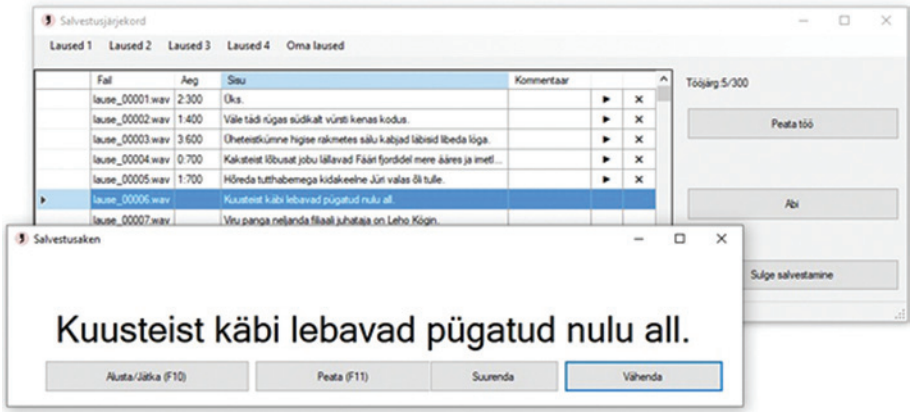
Salvestaja on lihtne ja väikesemahuline rakendus Windowsi op-süsteemile. Programm on oma tööpõhimõttelt mõeldud maksimaalselt vastama Kõneveebi vajadustele: kasutajale kuvatakse programmis sisalduvast tekstikorpusest järjest 1192 lauset ning nende ettelugemisel tekkiv signaal salvestatakse wav-failina vastavasse

[10] https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/custom-neural-voice (29.3.2023).
[11] https://cloud.google.com/text-to-speech/custom-voice (29.3.2023).
[12] https://www.acapela-group.com/solutions/my-own-voice (29.3.2023).
[13] https://koneveeb.ee/arendajale (30.3.2023).
[14] https://www.eki.ee/heli/index.php/korpused (30.3.2023).
[15] https://www.audacityteam.org (30.3.2023).
[16] https://www.bas.uni-muenchen.de/Bas/software/speechrecorder (30.3.2023).

kausta, moodustades helikorpuse. Lisavõimalustena saab tekstikorpusse lauseid lisada või sealt lauseid välja jätta, salvestusi üle kuulata ja üle salvestada ning salvestuses pause teha. Ettelugemise lõppedes saab kasutada funktsionaalsust "paki salvestus saatmiseks", mis pakib teksti- ja helikorpuse Kõneveebi üleslaadimiseks sobivasse zip-formaati.

Saadaval on ka salvestusprogrammi lähtekood, juhuks kui kasutaja soovib seda oma vajadustele kohandada.

Hääle treenimiseks tuleb korpus üles laadida Kõneveebi keskkonda. Üles laaditud andmete privaatsus on tagatud ja need kustutatakse pärast hääle treenimist automaatselt.

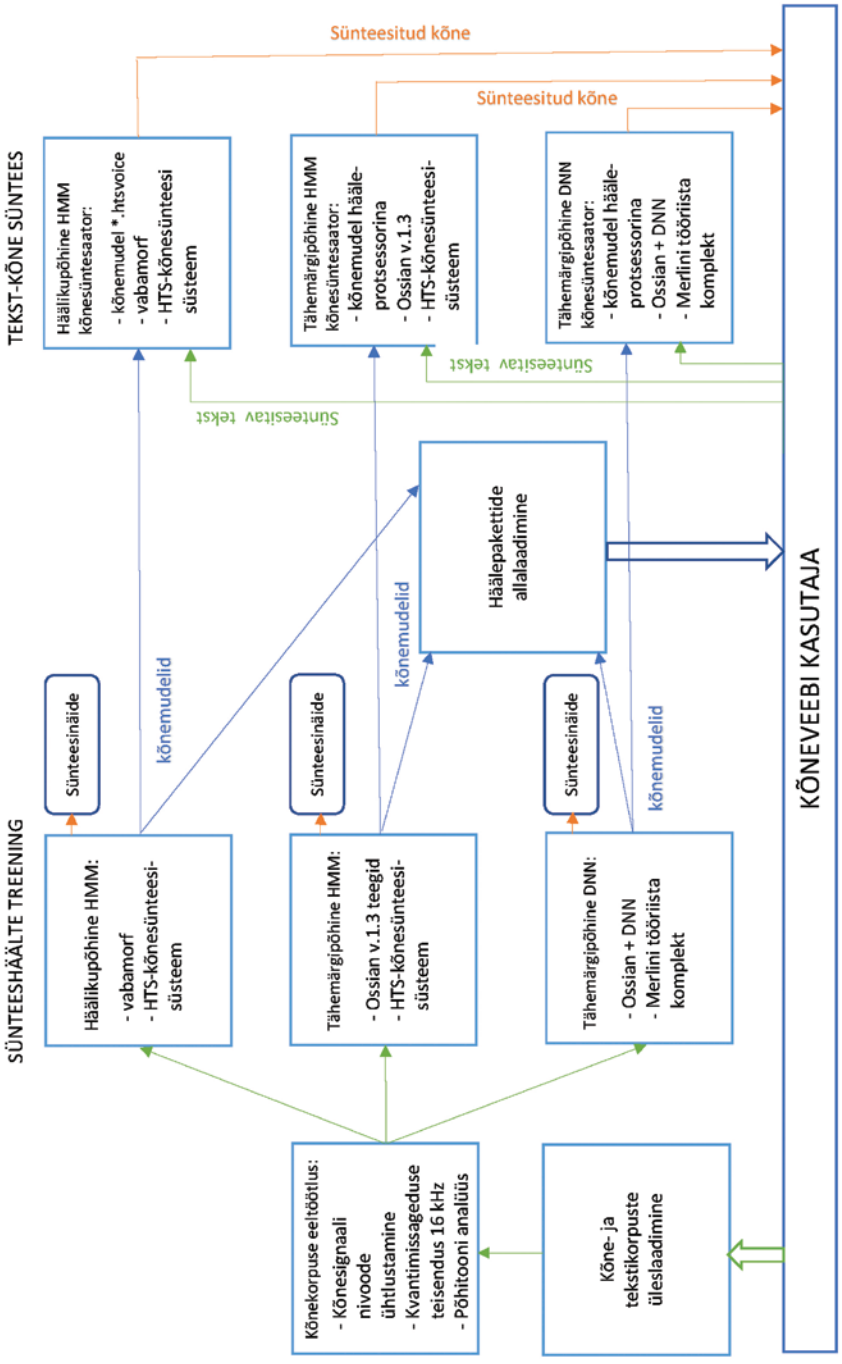


Joonis 1. Salvestaja vaated

Lisaks oma treeningkorpuse salvestamise võimalusele pakutakse Kõneveebi keskkonnas arendajatele või uurijatele juba olemasolevaid kvaliteetseid treeningkorpusi. Esindatud on meeshääled (Indrek, Peeter, Tambet) ja naishääled (Liivika, Külli, Kersti), nii etteloetud ilukirjandus (millest on eraldatud ja märgendatud vastavalt saatelause asukohale otsekõne ehk tegelaskõne) kui ka neutraalsed üksiklaused, mis sisaldavad kõiki eesti keeles võimalikke häälikuid ja häälikuüleminekuid.

3.2. Sünteeshäälte treenimine

Teenuse Minu Hääl arhitektuur on kujutatud joonisel 2. Kõneveebi kasutaja poolt üles laaditud korpustele tehakse kõigepealt eeltöötlus, mille käigus ühtlustatakse wav-failides kõnesignaali nivood ja teisendatakse signaali diskreetimissagedus 16 kHz peale. Kõnelaine 16 kHz diskreetimissageduse valik tagab kõne kvaliteetse edasiandmise dünaamilises sagedusdiapasoonis 0–8 kHz ja sellega optimeeritakse sünteeshäälte treenimiseks kuluvat aega. Eeltöötluse käigus analüüsitakse ka kõne põhitooni (f_0), et hinnata doonorhääle kõrgust ja ulatust, kuivõrd iga inimese hääl on unikaalne ja oma häälekõrguse ja ulatusega. Igale kõnekorpusele fikseeritakse analüüsi põhjal põhitooni ulatuse piirid miinimum- (f_{0_min}) ja maksimumväärtusega (f_{0_max}), millises vahemikus konkreetse doonorhääle põhitoon muutub. Treeningprotsessis võetakse arvesse ainult sellesse vahemikku jääva põhitooniga



Joonis 2. Kõneveebi kõnesünteesiteenuse Minu Hääl protsessidiagramm

üksusi. See tagab ühtlasema treeningandmestiku ja seega kvaliteetsema väljundkõne. Kõneveebi kasutaja tekstikorpus, eeltöödeldud kõnekorpus ja doonorhääle põhitooni ulatuse info on sünteeshäälte treeningprotsessi sisendiks (vt joonis 2).

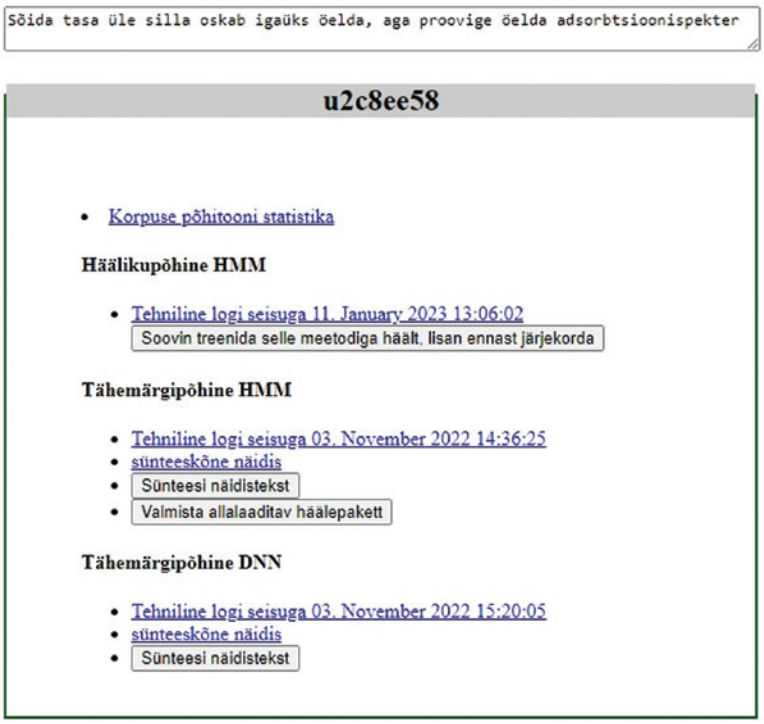
Järgmise sammuna treenitakse kasutaja korpuse põhjal automaatselt kolm sünteeshäält. Sünteeshääli treenitakse Kõneveebis kahel erineval statistilis-parameetrilisel kõnesünteesimeetodil: Markovi peitmudelitel (HMM, ingl *hidden Markov models*) põhineval ja sügavaid närvivõrke (DNN, ingl *deep neural network*) rakendaval meetodil. Markovi peitmudelitel põhineval kõnesünteesil rakendatakse HMM-sünteesisüsteemi versiooni HTS 2.0 (Zen jt 2007). Avatud koodiga sügavatel närvivõrkudel DNN-kõnesünteesisüsteemi Merlin treeningu- ja sünteesiprotsess toetub Theano arvutusteekidele (Wu jt 2016). HMM ja DNN kõnesünteesi treeningsüsteemide valikul lähtusime pragmaatilistest kaalutlustest, et mõõduka suurusega kõnekorpustest (*ca* 1000 lauset) oleks mõistliku aja jooksul (3–8 tundi) võimalik treenida tavapärase serveri arvutusressursi (ei eelda graafikakaartide kasutamist) baasil hea kvaliteediga sünteeshääli.

Mõlema kõnesünteesisüsteemi rakendamisel kerkis esile keeleüksuste valik: kas kasutada kõnesünteesi üksustena häälikuid või tähemärke? Ehkki eesti keele kirjas ja kõnes pole tähemärgi ja hääliku suhe päris üksühene, on eesti ortograafia siiski võrdlemisi foneetiline (EKG II 1993). Eestikeelses kirjalikus tekstis üldjuhul ei eristu vaid teisevältelised sõnad kolmandavältelistest sõnadest (nt lausetes *Rong jõudis jaama* ja *Türi jaama taga oli turg* hääldub ühesuguse kirjapildiga sõna *jaama* erinevalt, esimeses lauses III ja teises lauses II vältes) ning palataliseeritud konsonandid palataliseerimata konsonantidest (nt sõnas *palk* sõltub *l*-i palataliseerimine sellest, kas mõeldakse töötasu või langetatud puutüve, ja selliseid sõnu on eesti keeles küllalt palju, nt *tulp*, *hall*, *nutt*). Ka pika *üü* diftongistumine on eesti keelele omane (nt *müüa* hääldub [müija], *hüüe* [hüije]). Aga eelpool toodud näited väljendavad vaid nõrgalt lahknevat suhet tähemärgi ja hääliku vahel eesti keeles ning see ärgitas HMM-põhisel kõnesünteesil kasutama kahte erinevat lähenemisviisi:

1) häälikupõhine ehk juhendatud hääldusreeglitega, tähemärk-häälik teisendust sisaldav variant, mille korral määratakse liitsõnapiirid, tuvastatakse kolmandavältelised sõnad, eristatakse lühikesi ja pikki häälikuid ning palataliseerituid konsonante palataliseerimata konsonantidest. Hääldusreeglite tuvastamiseks oleme kaasanud HMM-süsteemi esikomponendina eesti keele morfoloogilise analüsaatori[17] ja eesti keele õigekeelsussõnaraamatu (ÕS 2018);
2) tähemärgipõhine ehk eesti keele hääldusreegliteta, juhendamata hääldusega variant, mille korral treenitakse keelemudelid vaid ortograafilise teksti põhjal ja sünteesil tehakse tekst-kõne teisendus häälikutasandit kasutamata.

Juhendamata variandi korral on kasutatud HMM-süsteemi esikomponendina keelest sõltumatuid Ossian-süsteemi tekstitöötlusteeke (Vainio jt 2014). Kõneüksuste erinevaid valikuprintsiipe on õigustanud ka see, et mõnes uurimuses on tähemärgipõhine lähenemine saanud kõrgemaid hinnanguid (vt Piits jt 2022), kuigi võiks eeldada, et häälduspärasem sisend annab kvaliteetsema tulemuse. Ka sügavatel närvivõrkudel põhineval DNN-kõnesünteesis on kõneüksusteks valitud tähemärgid. Joonisel 2 on eri treeningmeetodite ja kõnesüntesaatorite moodulites toodud viited treeningprotsessis kasutatud tarkvarale, süsteemidele ja ressurssidele.

[17] https://github.com/Filosoft/vabamorf (30.3.2023).



Joonis 3. Kõneveebi sünteeshäälte treeninguleht

Kõneveebis treenitakse kasutaja korpuse põhjal sünteeshääled järjekorras kõigil kolmel viisil: häälikupõhine HMM, tähemärgipõhine HMM ja tähemärgipõhine DNN (vt joonis 3). Treeningprotsessi käigus luuakse kõigile häälikutele või tähemärkidele akustilised mudelid ning sõnadele, fraasidele ja lausetele treenitakse Kõneveebi kasutaja doonorhääle põhjal iseloomulikud meloodiamudelid. Ehkki treenitav sünteeshääl pole inimhääle üks-ühele klooniks, võib doonorhääle meeldiv variatiivsus rikastada sünteeshääle kõla ja ilmekust. Ühe sünteeshääle treenimine vältab 3–8 tundi, sõltuvalt korpuse mahust ja konkreetsest meetodist. Iga edukalt treenitud häälega sünteesitakse automaatselt "sünteeskõne näidis" *Sõida tasa üle silla oskab igaüks öelda, aga proovige öelda adsorbtsioonispekter*, mille põhjal saab kasutaja anda sünteeskõne kvaliteedile esmase hinnangu. Juhul kui mingi meetodi sünteeskõne näidist ei ole treeningulehel väljas, siis treeningprotsess veel käib või ei ole olnud edukas. Treeningprotsessi käiku saab jälgida treeningulehel tehniliste logide põhjal (vt joonis 3). Treeningprotsessi ebaõnnestumist võivad põhjustada näiteks heli ja teksti mittevastavused või akustilised ebaühtlused treeningandmetes.

3.3. Treenitud hääle veebipõhine kasutamine ja installimine

Treeningprotsessi käigus loodud sünteeshääli saab Kõneveebi treeningulehel kuulata ja põhjalikumalt testida ning tekstide helindamiseks kasutada, sisestades Kõneveebi treeningulehe ülaservas olevasse demoaknasse oma teksti (kuni 500

tähemärki, sest brauseri ooteaeg on piiratud). Tekst söödetakse ette doonorhääle põhjal loodud kõnesüntesaatoritele, mis sünteesivad tekstist väljundkõne. Sünteesitud helifaile saab kuulata ja alla laadida.

Kui kasutaja on leidnud endale sobivaima sünteeshääle, saab Kõneveebi treeningulehelt vastava häälepaketi alla laadida. Sellist isiklikku kõnesüntesaatorit võib kasutada nii oma arvutis kui ka lülitada erinevatesse häälrakendustesse. Süntesaatori kasutamiseks oma arvutis on lisaks kõnemudelitele vajalik ka seda toetav tarkvara. Üldjuhul eeldavad Kõneveebi rakendusjuhendid, et häälepakette kasutatakse Linuxis, vaid häälikupõhine HMM on platvormivaba kõnesüntesaator, mis installeerub ka Windowsis ja OSX-is. Tähemärgipõhise HMM-i ja DNN-iga treenitud mudelid on alla laadimiseks pakendatud isoleeritud tarkvarakeskkonda (konteineritesse). Konteinerite jooksutamiseks Windowsi all peab kasutaja endale kõigepealt installima Docker Desktop[18].

Minu Hääle teenuses kasutatud süsteeme, tarkvara ja ressursse saab kasutada ka väljaspool Kõneveebi. Kõneveebis "Arendajale" saki all on spetsiaalne lehekülg, kus on olemas konteinerid treeningu ja sünteesikeskkonna (tähemärgipõhine HMM ja DNN) Linuxis kasutamiseks. Need võimaldavad ise hääli treenida, kasutades oma korpuseid, ja sünteesida. Häälikupõhise HMM-kõnesüntesaatori tarkvara on samuti huvilistele kättesaadav.

4. Kokkuvõte ja edasine tegevus

Artiklis tutvustasime Eesti Keele Instituudis arendatavat interaktiivset kõnesünteesi veebikeskkonda Kõneveeb, mille eesmärgiks on ühiskonna, ettevõtjate ja arendajate vajadusi arvestades pakkuda tasuta ühest kohast ja mugavalt kõiki EKI kõnesünteesiga seotud teenuseid ja ressursse. Põhjalikumalt tutvustasime Kõneveebis saadaval olevat uut teenust Minu Hääl, mis võimaldab igaühel luua ise sünteeshääle ilma mingite tehniliste eelteadmisteta kõnesünteesist. See teenus on mõeldud üksikisikutele ja ettevõtjatele, kelle vajadusi olemasolevad sünteeshääled ei rahulda, näiteks kellel on vaja spetsiifilises stiilis või unikaalset sünteeshäält või omaenda häälest tehtud sünteeshäält.

Kõneveebi keskkonda ja Minu Hääle teenust on kavas järjepidevalt täiendada uute ressursside ja võimalustega. Keskkonda on kavas lisada uusi sünteeshääli, sünteesimeetodeid, treeningkorpusi ja arendajatele mõeldud ressursse. Samuti lisatakse uurijatele mõeldud ressursse, nagu kõnesignaali analüsaator (Eyben jt 2016). Minu Hääle teenusesse on plaanis lisada häälikupõhine DNN, mis on katsete käigus näidanud head väljundkõne kvaliteeti. Samuti on olnud fookuses siirdeõppe meetodid (ingl *transfer learning methods*), mis võimaldavad eeltreenitud mudeli põhjal luua uusi sünteeshääli senisest väiksemal kõnematerjalil (Neekhara jt 2021, Zhang jt 2021). Kõnestiilide mitmekesistamiseks pakume välja stiilisiirdamise võimaluse – olemasolevale (süntees)häälele saab üle kanda kõnestiiliga seotud omadusi, säilitades samal ajal kõneleja identiteedi ja teksti lingvistilise sisu (Gao jt 2019), näiteks muuta melanhoolselt kõlav kõne energiliseks. Kavas on lisada ka võimalus teha oma treeningkorpus ja treenitud hääl teistele kasutajatele kättesaadavaks, et mitmekesistada eestikeelset kõnesünteesi.

[18] https://www.docker.com/products/docker-desktop (30.3.2023).

Viidatud kirjandus

EKG II 1993 = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Vare, Silvi 1993. Eesti keele grammatika II. Süntaks. Lisa: Kiri. ['The Grammar of the Estonian Language II: Syntax']. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Eyben, Florian; Scherer, Klaus R.; Schuller, Bjorn W.; Sundberg, Johan; Andre, Elisabeth; Busso, Carlos; Devillers, Laurence Y.; Epps, Julien; Laukka, Petri; Narayanan, Shrikanth S.; Truong, Khiet P. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. – IEEE Transactions on Affective Computing, 7 (2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Gao, Jian; Chakraborty, Deep; Tembine, Hamidou; Olaleye, Olaitan 2019. Nonparallel emotional speech conversion. – Proceedings of INTERSPEECH 2019, 2858–2862. https://doi.org/10.21437/Interspeech.2019-2878

Georgila, Kallirroi 2017. Speech synthesis: State of the art and challenges for the future. – Judee K. Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, Alessandro Vinciarelli (Eds.), Social Signal Processing. Cambridge: Cambridge University Press, 257–272. https://doi.org/10.1017/9781316676202.019

Judge, Simon; Hayton, Nicola 2022. Voice banking for individuals living with MND: A service review. – Technology and Disability, 34 (2), 113–122. https://doi.org/10.3233/TAD-210366

Kato, Shuhei; Yasuda, Yasuke; Wang, Xin; Cooper, Erica; Takaki, Shinji; Yamagishi, Junichi 2020. Modeling of rakugo speech and its limitations: Toward speech synthesis that entertains audiences. – IEEE Access, 8, 138149–138161. https://doi.org/10.1109/ACCESS.2020.3011975

Klabbers, Esther 2019. Text-to-speech synthesis. – Michael Filimowicz (Ed.), Foundations in Sound Design for Embedded Media: A Multidisciplinary Approach. New York: Routledge. https://doi.org/10.4324/9781315106359

Mihkla, Meelis; Eek, Arvo; Meister, Einar 1998. Creation of the Estonian diphone database for text-to-speech synthesis. – Linguistica Uralica, 34 (3), 334–340.

Mihkla, Meelis; Hein, Indrek; Hiiepuu, Andrus; Kiissel, Indrek; Ruusalepp, Raivo; Sinisalu, Urmas 2017. Raamat sünnib kuulata ['Books for listening']. – Keel ja Kirjandus, 60 (2), 114–129. https://doi.org/10.54013/kk711a3

Mihkla, Meelis; Hein, Indrek; Kiissel, Indrek 2018. Self-reading texts and books. – Frontiers in Artificial Intelligence and Applications, 307, 79–87. https://doi.org/10.3233/978-1-61499-912-6-79

Mihkla, Meelis; Hein, Indrek; Kiissel, Indrek; Räpp, Artur; Sirts, Risto; Valdna, Tanel 2013. Subtiitrite helindamine – kas, kuidas, kellele ja milleks? ['Spoken subtitles – if, how, for whom and why?'] – Keel ja Kirjandus, 56 (11), 819–828. https://doi.org/10.54013/kk672a3

Mihkla, Meelis; Hein, Indrek; Kiissel, Indrek; Räpp, Artur; Sirts, Risto; Valdna, Tanel 2014. A system of spoken subtitles for Estonian Television. – Frontiers in Artificial Intelligence and Applications, 268, 19–26. https://doi.org/10.3233/978-1-61499-442-8-19

Mihkla, Meelis; Piits, Liisi 2022. Vaimult suureks keeletehnoloogia toel. – Arenguseire Keskuse trendiraport "Pikksilm", aprill 2022. https://arenguseire.ee/wp-content/uploads/2022/04/2022_pikksilm_keeletehnoloogia_artikkel.pdf (27.11.2022).

Neekhara, Paarth; Li, Jason; Ginsburg, Boris 2021. Adapting TTS models for new speakers using transfer learning. – arXiv:2110.05798. https://doi.org/10.48550/arXiv.2110.05798

Piits, Liisi; Mihkla, Meelis; Nurk, Tõnis; Kiissel, Indrek 2007. Designing a speech corpus for Estonian unit selection synthesis. – Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007), 367–371.

Piits, Liisi; Pajupuu, Hille; Sahkai, Heete; Altrov, Rene; Ermus, Liis; Tamuri, Kairi; Hein, Indrek; Mihkla, Meelis; Kiissel, Indrek; Männisalu, Egert; Suluste, Kristjan; Pajupuu,

Jaan 2022. Audiobook dialogues as training data for conversational style synthetic voices. – Proceedings of the 13th International Conference on Language Resources and Evaluation, LREC 2022. Marseille: The European Language Resources Association (ELRA), 1047–1053.

Shin, Yookyung; Lee, Younggun; Jo, Suhee; Hwang, Yeongtae; Kim, Taesu 2022. Text-driven emotional style control and cross-speaker style transfer in neural TTS. – Proceedings of Interspeech 2022, 2313–2317. https://doi.org/10.21437/Interspeech.2022-10131

Zen, Heiga; Nose, Takashi; Yamagishi, Junichi; Sako, Shinji; Masuko, Takashi; Black, Alan W.; Tokuda, Keiichi 2007. The HMM-based speech synthesis system (HTS) version 2.0. – 6th ISCA Workshop on Speech Synthesis, SSW 2007, 294–299.

Zhang, Mingyang; Zhou, Yi; Zhao, Li; Li, Haizhou 2021. Transfer learning from speech synthesis to voice conversion with non-parallel training data. – IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1290–1302. https://doi.org/10.1109/TASLP.2021.3066047

Vainio, Martti; Grönroos, Stig-Arne; Smit, Peter; Suni, Antti; Watts, Oliver 2014. Deliverable D2.2. Description of the final version of the new front-end. https://simple4all.org/wp-content/uploads/2014/11/Simple4All_deliverable_D2.2.pdf (20.11.2022).

ÕS 2018 = Eesti õigekeelsussõnaraamat ÕS 2018. Eesti Keele Instituut. Maire Raadik (Toim.). Tiiu Erelt, Tiina Leemets, Sirje Mäearu, Maire Raadik (Koost.). Tallinn: Eesti Keele Sihtasutus, 2018.

Wu, Zhizheng; Watts, Oliver; King, Simon 2016. Merlin: An open source neural network speech synthesis system. – Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW 9), 202–207. https://doi.org/10.21437/SSW.2016-33

KÕNEVEEB AND MINU HÄÄL: AN INTERACTIVE WEB ENVIRONMENT FOR SPEECH SYNTHESIS RESOURCES AND A SERVICE FOR CUSTOM SYNTHETIC VOICE CREATION

Meelis Mihkla, Indrek Hein, Indrek Kiissel, Jaan Pajupuu, Liisi Piits, Heete Sahkai, Hille Pajupuu, Rene Altrov, Elgar Kudritski, Liis Ermus, Egert Männisalu, Kristjan Suluste
Institute of the Estonian Language

Text-to-speech synthesis – a technology that converts written text into speech – has become part of everyday applications. This means that there is a general need for individuals to install speech synthesis in their devices and for companies to integrate it into their products and services. It is therefore important to make speech synthesis available for maximally easy uptake, especially for languages that are in danger of being dominated by English. The paper describes the interactive web environment Kõneveeb that is being developed to this end at the Institute of the Estonian Language. The purpose of the environment is to make the Institute's Estonian speech synthesis resources and services easily available for individuals, companies, and developers alike. In addition to various free text voicing services, application interfaces, training corpora and training software, Kõneveeb offers a custom synthetic voice creation service, Minu Hääl. The service is intended for users who need a synthetic voice that is unique, represents a specific speaking style, or resembles their own voice. It allows the user to create a synthetic voice without any technical knowledge about speech synthesis. The service includes a dedicated program that enables an effortless recording of the donor voice and outputs a training corpus in the required format. After the user has uploaded the training corpus to Kõneveeb, three synthetic voices are automatically trained, using different synthesis methods. The resulting speech synthesisers can be used for text voicing either directly in Kõneveeb or installed in the user's computer. Both the Kõneveeb environment and the Minu Hääl service will be constantly updated with new resources and possibilities.

Keywords: speech, speech technology, speech corpora, machine learning, text voicing service, Estonian

Meelis Mihkla on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna vanemteadur.
Roosikrantsi 6, 10119 Tallinn, Estonia
meelis.mihkla@eki.ee

Indrek Hein on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna vanemtarkvaraarendaja.
Roosikrantsi 6, 10119 Tallinn, Estonia
indrek.hein@eki.ee

Indrek Kiissel on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna vanemtarkvaraarendaja.
Roosikrantsi 6, 10119 Tallinn, Estonia
Indrek.kiissel@eki.ee

Jaan Pajupuu on vabakutseline tarkvaraarendaja ja konsultant.
Roosikrantsi 6, 10119 Tallinn, Estonia
j.pajupuu@gmail.com

Liisi Piits on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna vanemteadur.
Roosikrantsi 6, 10119 Tallinn, Estonia
liisi.piits@eki.ee

Heete Sahkai on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna vanemteadur.
Roosikrantsi 6, 10119 Tallinn, Estonia
heete.sahkai@eki.ee

Hille Pajupuu on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna juhtivteadur arvutiparalingvistika alal.
Roosikrantsi 6, 10119 Tallinn, Estonia
hille.pajupuu@eki.ee

Rene Altrov on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna vanemteadur arvutiparalingvistika alal.
Roosikrantsi 6, 10119 Tallinn, Estonia
rene.altrov@eki.ee

Elgar Kudritski on tarkvaraarendaja ettevõttes Avatud Lahendused.
Pärnu mnt 105, 11312 Tallinn, Estonia
e.kudritski@estysoft.com

Liis Ermus on Eesti Keele Instituudi keeleajaloo, murrete ja soome-ugri keelte osakonna nooremteadur ja Tartu Ülikooli doktorant eesti ja soome-ugri keeleteaduse erialal.
Roosikrantsi 6, 10119 Tallinn, Estonia
liis.ermus@eki.ee

Egert Männisalu on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna tarkvaraarendaja ning Tallinna Tehnikaülikooli magistrant.
Roosikrantsi 6, 10119 Tallinn, Estonia
egert.mannisalu@eki.ee

Kristjan Suluste on Eesti Keele Instituudi kõneuurimise ja kõnetehnoloogia osakonna keeletehnoloog ning Tallinna Ülikooli magistrant digitaalsete õpimängude erialal.
Roosikrantsi 6, 10119 Tallinn, Estonia
kristjan.suluste@eki.ee