Liisi Piits
Eesti Keele Instituudi vanemteadur

Meelis Mihkla
Eesti Keele Instituudi vanemteadur

Fotod: Jake Farra

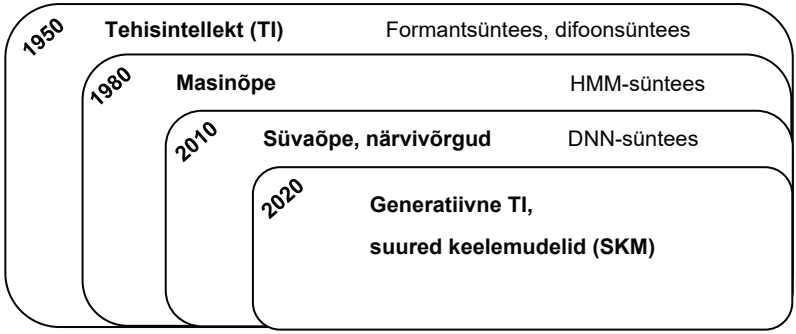
Kõnesünteesi hetkeseisust ja väljakutsetest tehisintellekti ajastul

Kujutage ette, et teie arvuti räägib teiega nagu vana sõber, sujuvalt ja loomulikult. Tehisintellekt on teinud arenguhüppe, muutes selle unistuse reaalsuseks. Kõnesüntesaatorid suudavad jäljendada inimkõnet nii hästi, et kohati on raske vahet teha, kas räägib inimene või masin. Sõnu „tehisintellekt", „kõnesüntees" ja „suured keelemudelid" kuuleb aina enam, aga kuidas need suhestuvad ja mis on eestikeelse kõnesünteesi lahendamata ülesanded, sellest nüüd kirjutamegi.

Kõnesüntees kui tehisintellekti osa

Arusaam, mida tehisintellekti (TI) all mõelda, on muutunud koos TI arenguga. Kõige laiemas mõttes mõistetakse selle all intelligentseid masinaid. 70 aasta jooksul on muutunud see, mida me neilt masinatelt ootame ja milleks nad suutelised on (vt joonist 1).

Kõnesüntees on tehnoloogia, mis suudab muuta digitaalse teksti kõneks ehk teksti ette lugeda. Kõnesünteesi süsteemid on samuti igas uues TI arenguetapis muutunud. Kui esimesed eelmise sajandi lõpus loodud eestikeelsed kõnesünteesisüsteemid formantsüntees ja difoonsüntees olid



Joonis 1. Tehisintellekti kihid Strykeri ja Kavlakoglu (2024) järgi koos vastavale kihile omaste sünteesitehnikatega

rangelt reeglipõhised, siis Markovi peitmudelitel põhinev (HMM, ingl *hidden Markov models*) süntees kasutab masinõpet. Siiani kõige paremad eestikeelsed sünteeshääled on aga loodud sügavaid närvivõrke (DNN, ingl *deep neural network*) kasutades. Generatiivne ehk tootev TI siiani veel kõnet ei tooda. Praegu veel muudetakse juturobotite väljund kõneks klassikalist kõnesünteesi kasutades, aga käib arendustöö, et generatiivne mudel suudaks ka ilma vahepealse tekstietapita kõnet toota. Sellele vaatamata saab juba praegu nautida võimalust lobiseda ChatGPT-ga ka eesti keeles. Teksti tootvad mitmekeelsed suured keelemudelid (SKM) on kaua puudu olnud lüliks kõnesünteesi ja kõnetuvastuse vahel, mis võimaldavad kasutada kõnetehnoloogiat loomulikuks vestluseks.

Millest sünteeshääli luuakse?

Kõigi tänapäevaste kõnesüntesaatorite loomiseks kasutatakse inimhäält, st kõnesünteesi treenimiseks on vaja kõnekorpusi: kõnet ja sellele vastavat teksti. Sajandi alguses valminud esimese inimkõnel põhineva eestikeelse süntesaatori jaoks loeti sisse umbes 1700 üksiksõna. Sõnad olid valitud nii, et kataksid kõiki eesti keeles esinevaid häälikuüleminekuid ehk difoone. Sellist difoone liitvat ahelsünteesi kasutades sündis difoonsüntesaator.

Masinõppelised HMM-meetodid vajasid juba aga seotud lauseid ja järgmises etapis paigutati need 1700 sõna loomulikku lausekonteksti ehk konstrueeriti 400 lauset, mis neid sõnu sisaldasid. Lisati ka sagedamaid kõnes esinevaid sõnaühendeid, fraase ning arvsõnu. Välditi lausete omavahelist sidusust, et teksti lugejal ei tekiks kiusatust emotsioone väljendada,

kuna eesmärk oli võimalikult neutraalne kõnekorpus. Esimesed HMM-meetodil tuntumad sünteeshääled olid Tõnu ja Eva[1], mis lugesid nt ajalehti ja Opiqu õpikeskkonnas õppetekste ning helindasid subtiitreid ja on siiani kasutusel pimedate ekraanilugejates.

Oluline edusamm kõnesünteesis tekkis 2020. aasta paiku transformer-mudelite[2] kasutuselevõtuga. Algselt tekstitöötlemise ülesannete jaoks loodud transformerid tõid kaasa olulise muutuse ka kõnetöötluses, mis parandas märgatavalt sünteeskõne meloodiat ja üldist kõnekvaliteeti. Selliste mudelitega tekkis aga ka vajadus suuremate kõnemahtude järele. Hakati koguma juba varem salvestatud esitusi ja tekkisid suuremad ilukirjanduskorpused[3]. Nende pealt on treenitud nii mitmed Eesti Keele Instituudis loodud sünteeshääled kui ka Tartu Ülikooli Neurokõne

1 EKI kõnesünteesi näiteid saab kuulata https://eki.ee/heli/.

2 Transformer on spetsiaalne närvivõrk, süvaõppe mudel, mis kasutab sisemise tähelepanu mehhanismi, et kaaluda sisendi osade suhtelist tähtsust sõltuvalt nende olulisusest antud kontekstis.

3 Eesti Keele Instituudi kõnekorpused https://koneveeb.ee/korpused/.

lehel[4] olevad hääled. Mida väljendusrikkamaks muutub sünteeskõne, seda enam saadakse aru, et eri liiki tekstide lugemiseks vajatakse eri stiilis rääkivaid hääli. Neutraalse kõne ja audioraamatute korpuse kõrvale hakkasime salvestama ja koguma spontaanse kõne korpusi, mille pealt treenitud kõne võiks sobida kõige paremini inimese ja arvuti vestluseks või keeleõppedialoogide tootmiseks. Neid oleme saanud kasutada spontaanses stiilis kõnesünteesi treenimiseks.

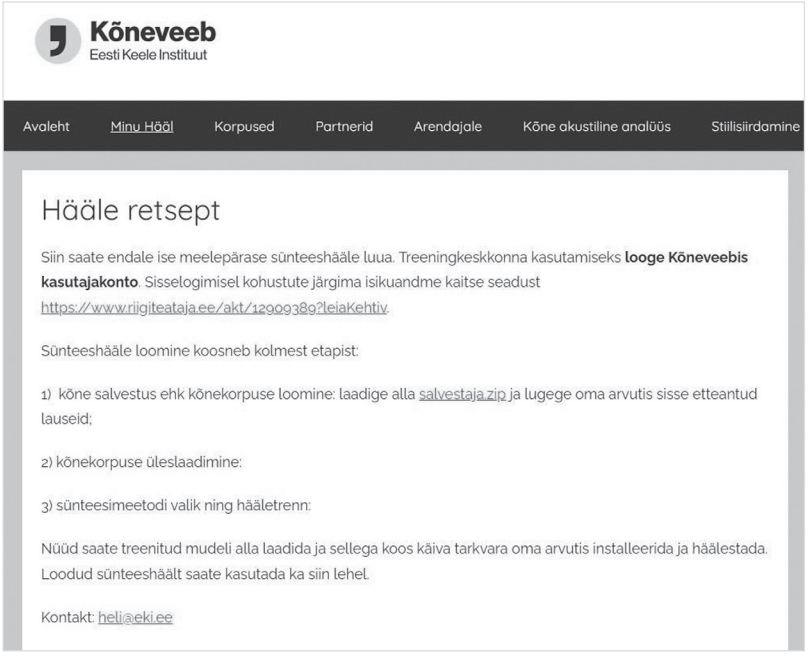
TÜ Neurokõne kõnetehnoloogid on hakanud treenima mitmehäälseid ja mitmekeelseid mudeleid, mis lubab näiteks võrukeelse Sulevi kõnelema panemiseks kasutada kõiki eestikeelseid korpusi. Sama teed on juba varem läinud ka n-ö suured tegijad, kes on loonud mitmekeelsete mudelite pealt eestikeelset sünteesi. Google'i loodud sünteeshäält on saanud igaüks kuulata tõlke- ja kaardirakendustes. Microsofti naishäälel Anul on väga loomulik kõnemeloodia, aga probleeme on numbrite käänamise ja lühendite väljalugemisega. Nende sünteeshäälte treenimiseks on juba kasutatud väga suurt hulka kõnet. Näiteks OpenAI juturoboti kõnelema panemiseks kasutatakse mitmekeelseid mudeleid mahuga 680 000 tundi kõnet.

Kõnesünteesiteenus Minu Hääl

Eestis on kõnesünteesi treenimiseks oma hääle andnud 23 inimest. Seetõttu saame valida meeste ja naiste, laste ja täiskasvanute, tavaliste tudengite ja tuntud näitlejate häälest tehtud sünteeskõne vahel. Kui kasutajale sellest ei piisa, siis aitab teenus Minu Hääl, mis võimaldab igaühel luua endale sobiva sünteeshääle ilma mingite tehniliste kõnesünteesi eelteadmisteta (Mihkla jt 2023). Sünteeshääle loomine ja kasutuselevõtmine Minu Hääles koosneb kolmest etapist (vt joonis 2).

Treeningkorpuse loomiseks on tööriist, mille abil saab oma arvutis lauseid lugeda ja sobivas vormis salvestada. Mida rohkem lauseid lugeda, seda kvaliteetsem mudel saavutatakse. Ehkki treenitud sünteeshääl pole inimhääle üks ühele kloon, kajastuvad seal doonorhääle tämber ja kõla. Järgmise sammuna treenitakse kasutaja korpuse põhjal sünteeshääled kahel erineval, HMM-il ja DNN-il põhineval meetodil. Sõltuvalt kõnekorpuse suurusest kestab iga sünteeshääle treenimine 3–8 tundi. Loodud sünteeshääli saab Minu Hääles kuulata, testida ning kasutada tekstide

[4] TÜ Neurokõne näiteid saab kuulata https://neurokone.ee/.



Joonis 2. Kõnesünteesiteenuses Minu Hääl toimub sünteeshääle loomine hääle retsepti alusel

helindamiseks. Sobivaima sünteeshääle saab alla laadida ja kasutada oma arvutis või mõnes häälrakenduses.

Ilmselt vajaks Minu Hääles kasutatavad sünteesimeetodid varsti värskendust, aga oma hääle salvestamisega ja sellest sünteeshääle treenimisega võib ka kohe alustada.

Kuhu edasi?

TI on mitmes valdkonnas (suured keelemudelid, masintõlge, kõnetuvastus ja kõnesüntees) juba olulisi edusamme teinud, aga seisame veel silmitsi mitme probleemiga, mis mõjutavad TI laialdasemat kasutuselevõttu. Kõnesünteesis vajavad lahendamist prosoodia, kõne emotsionaalsuse ning kõnestiilide ja häälduse seotud probleemid ning mõned tehnilised küsimused.

Prosoodia, mis hõlmab kõne rütmi, rõhku ja intonatsiooni, on loomuliku kõlaga sünteeskõne puhul kriitiline aspekt. Kuigi süvaõppe edusammud on sünteeshäälte prosoodiat tunduvalt parandanud, on

inimkõnele vastava taseme saavutamine endiselt katsumus. Sageli kõlab sünteetiline kõne kas liiga monotoonselt või ebaloomuliku prosoodiaga. Monotoonset kõnet võib olla tüütu kuulata, aga ebaloomulik intonatsioon ja valede sõnade rõhutamine võivad takistada kõnest arusaamist.

Inimsuhtlus on rikas emotsionaalsete nüansside poolest ja selle jäljendamine sünteeskõnes vajab lahendamist. Juturoboti emotsionaalne väljendusoskus on kaasahaarava ja lähedase suhtluse loomisel väga oluline. Selline suhtlus eeldab sünteeskõnes laia hulga inimlike emotsioonide väljendamist, aga ka vestluskaaslase emotsionaalse seisundi tuvastamist, et inimene ja virtuaalne assistent oleksid emotsionaalselt samal lainel. On lootust, et emotsioonide väljendamine paraneb oluliselt, kui suured keelemudelid hakkavad kõnet tootma ilma tekstilise vaheetapita.

Konteksti arvestamine on kõnesünteesisüsteemides oluline sõnade õigeks hääldamiseks. Näiteks sõna *palk* tuleb hääldada palataliseeritult ehk peenendatult, kui tekstis on juttu ümarpuidust, ning hääldada palataliseerimata, kui fookuses on töötasu. Selliste nüanssidega eestikeelne kõnesüntees veel hakkama ei saa ja homograafide ehk sama kirjapildiga sõnade hääldus on juhuslik. Kontekstiteadliku andmetöötluse täiustamine on kõnesünteesisüsteemide jaoks kasutaja sisendite täpseks tõlgendamiseks oluline.

Kõnesünteesi kvaliteet sõltub suuresti mudeli treenimiseks kasutatud andmete kogusest ja mitmekesisusest. Erinevaid aktsente, dialekte ja kõnestiile sisaldava suure ja mitmekesise andmestiku kogumine on aeganõudev, kuid hädavajalik. Siiani on Eestis kõnesünteesi treenitud peamiselt selleks otstarbeks kogutud kvaliteetsete ja käsitsi kontrollitud kõnekorpuste pealt, mille tüüpe eespool tutvustasime, aga kvaliteedihüppeks on vaja sadu kordi suuremaid korpusi. Selliste mahtude kogumine ei pruugi olla lihtne ja on mõnel juhul seotud ka juriidiliste küsimustega, aga ühe võimalusena on hakatud koguma vestlussaadete kõnet, sest heal tasemel kõnetuvastus suudab pakkuda ka vajalikku teksti kõnelaine kõrvale.

Sünteeshäälte treenimine on ressursimahukas ja aeganõudev ning eeldab võimsat riistvara ja küllalt keerukaid algoritme. Ka uuemate meetoditega sünteeshääled ise võivad nõuda seadmelt palju arvutusressurssi. Kõnesüntesaatorite oluliseks tehniliseks väljakutseks on kõnesünteesi kiiruse ja kvaliteedi tasakaalustamine, eriti närvivõrkudega mudelites, et need oleks kasutatavad ka vähem võimekates seadmetes. (Wang 2024)

Kokkuvõtteks

Oleme tunnistajateks keeletehnoloogia revolutsioonilisele ajastule. Suurte keelemudelite, masintõlke, kõnetuvastuse ja kõnesünteesi arengust sõltub olulisel määral üldine informatsiooni kättesaadavus nüüdisaegses digimaailmas. Tekst-kõne-süsteemidel on tähtis roll info vahendamisel suulises vormis. On tore, et eesti keel on keeletehnoloogia vallas maailma saja arenenuma keele hulgas, millel on omakeelne kompetents ja kogemus kõnesünteesi, kõnetuvastuse ja masintõlke vallas.

Viidatud kirjandus

Mihkla, Meelis, Indrek Hein, Indrek Kiissel, Jaan Pajupuu, Liisi Piits, Heete Sahkai, Hille Pajupuu, Rene Altrov, Elgar Kudritski, Liis Ermus, Egert Männisalu, Kristjan Suluste 2023. Kõneveeb ja Minu Hääl: uus kõnesünteesikeskkond ja -teenus. – Eesti Rakenduslingvistika Ühingu aastaraamat 19, 111–124.

Stryker, Cole, Eda Kavlakoglu 2024. What is artificial intelligence (AI)? – IBM, 16. august. https://www.ibm.com/topics/artificial-intelligence.

Wang, Zian (Andy) 2024. Challenging LLMs: An in-depth look at Text-to-Speech AI. – Deepgram, 10. jaanuar. https://deepgram.com/learn/text-to-speech-ai.