Väldete analüüs sünteesi teel

MEELIS MIHKLA

Oleme harjunud teadmisega, et alusuuringud on tehnoloogiate teenistuses: fundamentaalteadus annab uusi teadmisi ja rakendusteadus püüab nende teadmiste alusel uusi tehnoloogilisi vahendeid ja erinevaid rakendusi luua. Tegelikult on ammu kätte jõudnud aeg, mil võiks ka vastupidi väita: tehnoloogia abiga saame uusi teadmisi. Eriti hästi sobivad uurimisobjektideks just teaduse ebamääraseks, nn hallid alad, kus tehnoloogia abil võiks selgust saada. Välted on eesti foneetikas ja sõnaprosoodias vist kõige enam uuritud ja käsitletud teema. Kas eesti keele vältevastandust võib kõneprosoodia „halliks alaks” nimetada, on muidugi maitseküsimus. Aga näiteks suhteliselt hiljuti ilmunud suurteoses „Eesti keele hääldus”, mis võtab kokku kolmveerand sajandi pikkuse foneetikaalase uurimistöö (Asu jt 2016), on fonoloogiline kategooria *välde* jäetud defineerimata. See viitab kaudselt sellele, et eesti keele väldete osas ei ole veel kõik päris selge ja läbi uuritud.

Igapäevase keelekogemuse põhjal võime tõdeda, et eesti emakeelega inimene oskab teksti ette lugedes või kõneldes pea eksimatult sõnu õiges vältes hääldada, seda ka sama kirja-pildiga erinevas vältes sõnade puhul. See tõdemus motiveerib tehnoloogilist uurimisiideid: kuivõrd selline väldete keelevaist on tehnoloogiliste vahendite poolt õpitav ja häälduskogemus treenitav? Kui näidata arvutile piisaval hulgal eestikeelseid tekste ja anda kuulata vastavat kõnet, siis ühelt poolt saab näha ja kuulda, millisel tasemel on kõnesüntesaatorid peale masinõpet võimelised välteid iseõppijatena produtseerima, ja teiselt poolt on huvitav, milliseid akustilisi tunnuseid nende vältemudelid sealjuures kasutavad.

Inimkõne kunstlikku modelleerimist on saatnud pidev foneetikute ja inseneride vaheline dialoog. Kõnesünteesi algaastatel mõjutas foneetika kõnetehnoloogiat suuresl määral: formantsünteesi päevil oli andmeid hõredalt ja modelleerimine toimus pigem riistvara, mitte tarkvara abil (King 2014). See on täheldatav ka eestikeelse kõnesünteesi arenguloos (Mihkla jt 2012). Nüüdisaegsed sünteesimeetodid võimendavad õppimist, kasutades suurenenud ressursse, nagu mahukaid kõnesalvestusi ja nende automaatseid märgendusi. Järsult suurenenud arvutusressursid on läbimurde eelduseks, nagu ka uued meetodid kõnesünteesis: närvivokooder WaveNet kõne lainekujude loomiseks (van den Oord jt 2016) ja generatiivne tekst–kõne–süsteem Tacotron sisendteksti analüüsimiseks (Wang jt 2017). Aastakümneid väldanud järkjärguline edasimineku kõnesünteesi vallas on viimastel aastatel kulmineerunud väljundkõne kvaliteedi olulise paranemisega, mis on eelkõige seotud süvaõppega (Malisz jt 2019). Võib julgesti öelda, et foneetikute, fonoloogide ja kõnetehnoloogide dialoogis on saabunud etapp, mil kõnesüntees võib keeleteadusele pakkuda nii olulist tuge kõneeksperimentide läbiviimisel kui ka usaldusväärseid modelleerimisraamistikke fonoloogiliste mudelite testimiseks, võimaldades sünteesi abil keele-nähtusi analüüsida (ingl *analysis by synthesis*). Sellist analüüsimeetodit on juba

kasutatud näiteks põhitooni meloodiakontuuride esitamiseks (Xu, Prom-on 2014) ja tekst–kõne mudelites mandariini keele leksikaalsete toonide uurimiseks (Zhu 2020). Siinses artiklis analüüsitakse sünteesi kaudu eesti keele välteid.

Ajalooliselt on eesti keele välteid püütud selgitada ligi sada viiskümmend aastat. Mihkel Veske kirjeldas oma 1879. aastal ilmunud raamatus „Eesti keele healte õpetus ja kirjutuse viis” välteid kui hääliku omadust (Weske 1879). Väldet on ka käsitletud kui silbi omadust: lühike silp vs. pikk silp ja kerge silp vs. raske silp. Silbivälte teoreetikud on samuti rõhutanud, et eesti keeles pole tegu kolmese vastandusega, vaid pigem kahe kahese vastandusega erinevatel tasanditel (Tauli 1973; Hint 1998). Silbivältega paralleelselt arendati välja taktivälte teoreetilised alused. Kõige selgemini oli taktivälde esil Arvo Eegi ja Einar Meistri töödes, kes leidsid, et väldet ei saa tuvastada üksnes rõhulise silbi põhjal, vaid pigem kogu kõnetakti akustiliste omaduste kaasabil. Sealjuures ei vastandutud otseselt silbivälte teooriale, rõhutati vaid, et silp on allutatud kõnetaktile ja silbivälte kolmikvastandus ilmneb takti piires (Eek, Meister 2003). Välteid on püütud lahti seletada ka moorameetrikas (Ehala 1999; Eek, Meister 2004; Prillop 2018), kuid seejuures pole tegu päris uue välteteooriaga, vaid pigem sõnaprosoodilise kirjelduse interpretatsioonidega.

Käesoleval sajandil on väldetega seotud uurimistöid ilmunud küllalt palju, aga vaid vähesed autorid on väldet täpsemalt formuleerinud. Nagu juba mainitud, kõneteaduslikke uurimistöid kokkuvõtvas teoses „Eesti keele hääldus” on fonoloogiline kategooria *välde* jäänud defineerimata. Sellest on kahju, sest peale uurimistöö „keeleeõppes ja populaarses keelekirjelduseski vältest hästi mööda ei pääse ja mingit ilusat õpikudefinitsiooni oleks hädasti vaja” (Ross, Piits 2019). Mitteõpiklik ja ilmselt kõige mahukam ning suure hulga akustiliste korrelaatide kandidaatidega vältedefinitsioon on formuleeritud taktivälte teooriast lähtudes järgnevalt: „Välde on suprasegmentaalne (prosoodiline) üksus, mis põhineb rõhulise silbi riimist ja järgneva silbi tuumast koosneva, isokrooniale püüdleva taktiosa kestusmallide vastandusel, kusjuures kestusmallidele lisaks aitavad välteid lahus hoida ka põhitooni-kontuuri, intensiivsuskulu, vokaali-konsonandi liitumisviisi ja muud erinevused.” (Eek, Meister 2004: 252)

Definitsiooni järgi on välde suprasegmentaalne prosoodiline üksus, mille põhiliseks akustiliseks korrelaadiks on taktiosa rõhulise silbi riimi ja rõhuta silbi tuuma kestussuhe. Esimestena, üksteisest sõltumatult jõudsid selle suhtelise kestusliku tunnuseni Ilse Lehiste (1960) ja Georg Liiv (1961). Kolmeteistkümne senise uurimistöö (1960–2013) põhjal on rõhulise silbi riimi ja rõhuta silbi tuuma keskmine kestussuhe: I vältes 0,71, II vältes 1,74 ja III vältes 2,77 (Asu jt 2016: 136, tabel 4.1 põhjal). Nii silbi- kui ka taktivälte teoorias on esile toodud rõhulise silbi põhitooni roll II ja III välte eristamisel prosoodilise markeerituse näol. Mitmete taju-uurimuste ja -eksperimentidega on näidatud, et põhitooni kontuur, eriti selle tipu asukoht rõhulises silbis, on määrava tähtsusega II ja III välte eristamisel (Lehiste 1975; Remmel 1975; Lippus jt 2011). Ehkki intensiivsust on peetud ka põhitooni varieerumise kaasnähtuseks (Lehiste 1970: 143–144), on hiljutised intensiivsuse uurimused näidanud, et ka intensiivsuse taseme erinevusel kõnetakti rõhulises ja rõhuta silbis võib olla teatud roll väldete eristamisel (Sahka, Mihkla 2019).

Aga unustagem hetkeks konkreetseid välteteooriad, keerulised definitsioonid ja akustilised parameetrid. Vaevalt et tavaline eesti emakeelega inimene neid vältete produktseerimisel ja tuvastamisel teadlikult kasutab. Võib-olla polegi mõttekas näiteks kõnesünteesis arvutile silbiriimide kestussuhteid, erinevaid põhitooni kontuure ja intensiivsuse tasemeid ette kirjutada. Piisab sellest, kui iseõppivatele kõnesüntesaatoritele ette sööta piisaval hulgal eestikeelseid kirjalikke tekste ja anda kuulata vastavat eestikeelset kõnet. Kuidas lahendab masinõpe eesti keele välted kui fonoloogilise fenomeni?

Käesolevas töös püütakse analüüsiga sünteesi teel läheneda vältetele veidi teise nurga alt ja uue metodoloogia abil fonoloogilise kategooria mõne varjatud aspekti osas pisut selgust saada. Sünteeskõne analüüsi tulemusi kõrvutatakse inimkõne analüüsiga ja võrreldakse erinevaid kõnesünteesi meetodeid omavahel. Kõnesüntesaatorite vältete hääldustäpsust hinnatakse tajutestidega ja vältetunnuste klassifitseerimisvõimet diskriminantanalüüsiga. Artiklis otsitakse vastuseid järgmistele uurimisküsimustele:

- Millised akustilised parameetrid on süntesaatorite prosoodiamudelites olulised ja millised mitte?
- Millistes aspektides sünteeskõne ja loomuliku kõne vältemudelid erinevad?
- Kas kõnesünteesi erinevad meetodid kasutavad vältet samu akustilisi korrelaate?
- Millised akustilised parameetrid põhjustavad sünteeskõnes vältete hääldusvigu?

Kõnesünteesisüsteemid, meetodid, andmed ja parameetrid

Vältete analüüsimiseks kasutati kolme erinevat iseõppivat kõnesünteesisüsteemi: Ossiani, Merlini ja Mozilla TTS-i. Ossiani kõnesünteesisüsteem loodi rahvusvahelise Euroopa projekti Simple4All raames. Süsteemis kasutatakse Markovi peitmudeleid (ingl *hidden Markov model* – HMM), milles kõnesignaal esitatakse Markovi protsessi olekute jadana, kestuste ja põhitooni kõneprosoodia mudelid on realiseeritud otsustuspuudena. Kõnesünteesi rahvusvaheline kogemus on näidanud, et regressioonil põhinevate otsustuspuude asendamine närvivõrkudega parandab väljundkõne kvaliteeti (Watts jt 2016). Merlini tööriistakomplekt kõnesünteesisüsteemina on loonud Edinburghi ülikooli teadlased ja selles kasutatakse sünteesihäälte treenimiseks sügavaid närvivõrke¹ (ingl *deep neural network* – DNN). Viimaste aastate olulised uuendused, heli lainekujude genereerimine närvivõrkude abil, nagu WaveNet (van den Oord jt 2016), ja sisendteksti analüüsimine takotronidel (Wang jt 2017), põhinevad rekurrentsetel närvivõrkudel² (ingl *recurrent neural network* – RNN). Artiklis

¹ Sügavad närvivõrgud, mille tehisneuronid on korraldatud kahte või enamasse hierarhilisse kihti.

² Rekurrentsed närvivõrgud, mille tehisneuronitevahelised seosed moodustavad vähemalt ühe tagasisideahela.

kasutatud Mozilla TTS on rekurrentsetel närvivõrkudel põhineva kõnesünteesisüsteemi üks esindaja.

Sünteeshäälte treenimiseks anti nende kolme masinõppe süsteemi sisendisse *ca* 1000 lauset eestikeelset teksti ja teksti alusel salvestatud keelejuhtide kõnet. Tuhandelauselised korpused on üldjuhul iseõppivate kõnesüntesaatorite alumine mahupiir, mis tagab sünteesil rahuldava kvaliteediga arusaadava väljundkõne. Kõnematerjal salvestati Eesti Keele Instituudi (EKI) helistuudios kondensaatormikrofoniga Neumann TLM 102, monosalvestuste digitaalne formaat oli 48 kHz, 16 bitti. Keelejuhtideks olid kolm raadiouudiste lugejat ja toimetajat, üks näitleja ja üks koolitama hääledoonor (kaks meest (M1, M2) ja kolm naist (N1, N2, N3)). Kõik keelejuhid ei lugenud ette päris sama teksti. Küll aga sisaldas nende ettelooetav tekst häälikult häälikule ülemineku seisukohast foneetiliselt tasakaalustatud korpuse osa (400 lauset), mille lausetes kajastusid kõikvõimalikud eesti häälikute üleminekud (Piits jt 2007). Tekstikorpuse ja vastava kõnekorpuse info põhjal treeniti erinevates sünteesisüsteemides keelejuhtide häälele põhinevad kõnesüntesaatorid. Iga sünteesihääle õpi- ja treeningprotsess vältas 7 kuni 25 tundi. Kokku treeniti erinevate meetoditega kaheksa sünteesihäält: peidetud Markovi mudeli põhjal neli (M1, N1, N2, N3), sügavate närvivõrkudega kolm (M1, N1, N3) ja rekurrentsete närvivõrkudega üks (M2).

Ühtegi eesti keelega seotud moodulit või analüsaatorit iseõppivad kõnesünteesisüsteemid ei sisaldanud, nende sisendis olid ainult eestikeelsed tekstid ja vastavad kõnelainete failid.

Kui hästi kõnesüntesaatorid masinõppel vältete hääldamise omandasid, seda testiti lühikeste lausetega, mis sisaldasid minimaalsete erinevustega testsõnu, erinevas vältes kolmikuid. Testsõnadest moodustus kolm vokaalikeskset CV[V:]CV vältemalli (*lina*, *Liina*, *vii:na*; *vene*, *seene*, *vee:ne*; *valu*, *vaalu*, *vaa:lu*) ja kolm konsonandikeskset CV[C:]CV vältemalli (*nõme*, *Nõmme*, *nõm:me*; *linu*, *linnu*, *lin:nu*; *male*, *Malle*, *Mal:le*). Testsõnad olid sünteesitavates neljasõnalistes lausetes eelviimasel kohal sihitisena, nt:

Voova tõmbas vene randa.

Reeli pani seene korvi.

Luule löigub veene veriseks.

Viivi paneb linu vanni.

Meeli pani linnu puuri.

Leelo paneb linnu kaardile.

Testsõnu sisaldavaid lauseid lugesid ette kõik kaheksa kõnesüntesaatorit. Kõnesüntesaatorite väljundkõnet võrreldi ka inimkõne salvestistega ehk loomuliku kõne materjaliga. Ka kaheksa keelejuhti (neli meest ja neli naist) lugesid neid testsõnu sisaldavaid lauseid ette dialoogi vormis vastusena küsimusele „Mis toimub?“. Lauseid salvestati müravabas ruumis, EKI helistuudios sama aparatuuri ja samas digitaalses formaadis kui kõnesünteesi korpusigi.

Et aga analüüsida, millised akustilised tunnused on vältete eristamisel testsõnades olulised, valiti välja üheksa potentsiaalset akustilist parameetrit:

1. rõhulise silbiriimi ja rõhuta silbi tuuma kestussuhe;³
2. põhitooni kontuuri tipu asukoht rõhulises silbis;⁴
3. intensiivsuskontuuri tipu asukoht rõhulises silbis;⁵
4. põhitooni ulatus = põhitooni maksimum rõhulise silbi riimis – põhitooni miinimum rõhuta silbi tuumas;
5. põhitooni keskmiste väärtuste vahe rõhulises ja rõhuta silbis;
6. põhitooni maksimumide vahe rõhulises ja rõhuta silbis;
7. intensiivsuse ulatus = intensiivsuse maksimum rõhulise silbi riimis – intensiivsuse miinimum rõhuta silbi tuumas;
8. intensiivsuse keskmiste väärtuste vahe rõhulises ja rõhuta silbis;
9. intensiivsuse maksimumide vahe rõhulises ja rõhuta silbis.

Esimesed kaks parameetrit on klassikalised völdete eristamistunnused. Intensiivsuskontuuri iseloomustamiseks kaasati analüüsi parameetriks intensiivsuse maksimumi asend rõhulises silbis. Põhitooni ja intensiivsust testsõnades kirjeldatakse kolme suhtelise vöörtusega parameetriga: ulatus, keskmiste vahe ja maksimumide vahe. Ehkki intensiivsust on völdete kontekstis peetud ka põhitooni varieerumise kaasnähtuseks, näitas hiljutine intensiivsuse uurimus, et sõnatasandil sõltuvad völteastmete eristamisel intensiivsuse parameetrid vähe põhitoonist (Sahkai, Mihkla 2019), seetöttu arvati hinnatavate parameetrite hulka põhitooni ja intensiivsuse parameetrid eraldi.

Könesüntesaatorite völjundköne ja keelejuhtide loetud laused märgendati automaatselt (WebMAUS Basic), aga testsõnade häälikupiire kontrolliti käsitsi Praat-keskkonnas (Boersma, Weenink 2018). Köigi üheksa parameetri vöörtused mõödeti ja arvutati Praati skripti alusel. Statistiliseks analüüsiks kasutati statistikaprogrammi SYSTAT 13.

Analüüs sünteesi teel

Joonisel 1 on toodud sünteeskönes testsõnade üheksa parameetri karpdiagrammid vöälteti. Halli taustaga on need vöalte seisukohast olulised tunnused, mille keskmised erinesid paarikaupa köigi völdete tasemete vahel vöaltefaktori ANOVA-analüüsis. Näiteks intensiivsuse maksimumide vahe oli köiki vöalte tasemeid arvesse vöttes oluline parameeter ($F[2, 142] = 5,9$; $p = 0,003$), paraku II ja III vöalte vaheline erinevus polnud oluline ($F[1, 95] = 0,4$; $p = 0,551$), mistöttu vöalte seisukohast see tunnus ei kvalifitseerunud hallil taustal esiletöstmist vöärivaks parameetriks. Sünteesköne vöalte-

³ Köik testsõnad olid CVCV struktuuriga, häälikusegmentide kestustega C1, V1, C2 ja V2. Kestussuhe (KS) arvutati vokaalikeskse malli ja konsonandikeskse malli I vöalte testsõnades $KS = V1 : V2$, konsonandikeskse malli II-vöaltelistes testsõnades $KS = (V1 + 0,5 \times C2) : V2$ ja III-vöaltelistes testsõnades $KS = (V1 + 0,67 \times C2) : V2$.

⁴ Põhitooni tipu asukoht on esitatud protsendina rõhulise silbi kestusest, seega: (tipu asukoht silbi algusest : silbi kestus) $\times 100$.

⁵ Intensiivsuse tipu asukoht on esitatud protsendina rõhulise silbi kestusest, seega: (tipu asukoht silbi algusest : silbi kestus) $\times 100$.

mudeli kõige olulisemad tunnused olid: silbiriimide kestussuhe ($F[2, 142] = 168,1$; $p < 0,005$), põhitooni tipu asukoht rõhulises silbis ($F[2, 142] = 20,7$; $p < 0,005$), põhitooni maksimumide vahe ($F[2, 142] = 26,8$; $p < 0,005$) ja intensiivsuse ulatus ($F[2, 142] = 16,4$; $p < 0,005$). Neist kolmel parameetril on kas kasvavate erinevate keskmistega (silbiriimide kestussuhe ja põhitooni maksimumide vahe) või kahanevate keskmistega (põhitooni tipu asukoht) välteti loogiliselt tõlgendatavad jaotused. Vaid intensiivsuse ulatusel oli suurim keskmine väärtus II vältel ja seega teatud ebaloomilisus vältetevahelises järjestuses.

Joonis 1. Sünteeskõne testsõnade vältete akustiliste parameetrite jaotuste karpdiagrammid (halli taustaga on vältete seisukohast oluliste akustiliste parameetrite karpdiagrammid).

Sünteeskõne vältemudeli nelja kõige olulisema parameetri (silbiriimide kestussuhe, põhitooni tipu asukoht rõhulises silbis, põhitooni maksimumide vahe, intensiivsuse ulatus) alusel tehti klassikaline diskriminantanalüüs. Joonisel 2A on kujutatud testsõnade välteti paiknemine kahemõõtmelises faktorruumis akustiliste parameetrite väärtuste alusel. Erinevate vältete ala tähistavad ellipsid (usalduspiir

$p = 0,85$). Faktorruumis on küllaltki suur II ja III vâlde ala kattuvus. Tabeli 1 sünteeskõne veergudes on toodud eksimismatriks, mis näitab vâldete klassifitseerimisvõimet oluliste parameetrite alusel. I vâldet eristatakse korrektselt (100%), II vâldet 77,1% täpsusega ja enim hajuvat III vâldet klassifitseeriti õigesti vaid 60,4%-l juhtudest. Sünteeskõne vâldete üldine klassifitseerimistäpsus oli 79,2%.

Tabel 1. Diskriminantanalüüsil saadud vâldete klassifikatsiooni eksimismatriksid sünteeskõne ja loomuliku kõne kohta (ridades olevaid vâltekategoriaid klassifitseeriti veergudes olevate näitajate põhjal).

Vâlde	Sünteeskõne				Loomulik kõne			
	I	II	III	Õigesti klassifitseeritud vâlde	I	II	III	Õigesti klassifitseeritud vâlde
I	48	0	0	100,0%	46	2	0	95,8%
II	3	37	8	77,1%	2	38	8	79,2%
III	0	19	29	60,4%	2	6	40	83,3%
Kokku				79,2%				86,1%

Joonis 2. Diskriminantanalüüs testsõnade vâlteti paiknemisest kanoonilises faktorruumis (A – sünteeskõne, B – loomulik kõne), vâldete alad on piiritletud usalduspiiri ($p = 0,85$) ellipsitega.

Joonisel 3 on võrdlusena toodud loomuliku kõne testsõnade üheksa potentsiaalse akustilise parameetri karpdiagrammid. Halli taustaga on jällegi nende parameetrite karpdiagrammid, mille vâltefaktori tasemed erinevad paarikaupa kõigi vâldete vahel: silbiriimide kestussuhe ($F[2, 142] = 155,7$; $p < 0,005$), põhitooni ja intensiivsuse tipu asukoht rõhulises silbis ($F[2, 142] = 24,4$; $p < 0,005$ ja $F[2, 142] = 35,5$; $p < 0,005$) ning põhitooni ja intensiivsuse maksimumide vahe ($F[2, 142] = 12,7$; $p < 0,005$ ja $F[2, 142] = 19,3$; $p < 0,005$). Loomulikus kõnes oli kõigil viiel olulisel parameetril vâlteti kas kasvavate või kahanevate erinevate keskmiste loogilised järjestused.

Loomuliku kõne vâltemudeli viie olulisema parameetri (silbiriimide kestussuhe, põhitooni ja intensiivsuse tipu asukoht rõhulises silbis ning põhitooni ja intensiiv-

suse maksimumide vahe) põhjal tehti samuti diskriminantanalüüs. Joonisel 2B on kujutatud testsõnade välteti paiknemine kahemõõtmelises faktorruumis akustiliste parameetrite väärtuste alusel. Ka loomuliku kõne korral on täheldatav küllaltki suur II ja III välte alade kattuvus. Mõni I välte testsõna on faktorruumis üllatavalt II välte alale sattunud, mistõttu tabeli 1 loomuliku kõne eksimismaatriksis on näha, et I välde polegi päris korrektselt klassifitseeritud (95,8%). II välde eristati pisut kehvemini (79,2%) kui III välde (83,3%). Loomuliku kõne välde üldiseks klassifitseerimisvõimeks parameetrite alusel kujunes 86,1%.

Joonis 3. Loomuliku kõne testsõnade välde akustiliste parameetrite jaotuste karpdiagrammid (halli taustaga on välde seisukohast oluliste akustiliste parameetrite karpdiagrammid).

Ehkki sünteeskõne karpdiagrammides paistavad oluliste parameetrite jaotused kompaktsemad ja vähem hajuvad, eriti põhitooni maksimumide vahe korral, on ainuüksi jooniste 1 ja 3 põhjal visuaalselt küllalt keeruline hinnata välteti oluliste tunnuste jaotuste erinevusi sünteeskõnes ja loomulikus kõnes. Samuti pole lihtne diskriminantanalüüsi andmepilvede kujude ja usaldusellipsite järgi teha välde kohta sisukaid järeldusi (vt joonist 2). Seetõttu on tabelisse 2 võrdluseks koondatud

sünteeskõne ja loomuliku kõne vâlte aspektist oluliste ühiste (halli taustaga karp-diagrammid nii joonisel 1 kui ka joonisel 3) parameetrite keskmised väärtused ja standardhälbed. Kuna sünteeskõne ja loomulik kõne on olemuselt erinevad kõne-materjalid, siis kõigepealt võrreldi F-testiga kolme ühise parameetri dispersioone: kas sünteeskõne ja loomuliku kõne parameetrite jaotused välteti on sarnased või mitte (vt jaotuste sarnasuse veergu tabelis 2). F-testi analüüsi tulemuste põhjal rakedati sünteeskõne ja loomuliku kõne parameetrite keskväärtuste võrdlemiseks Studenti t-testi ühiste parameetrite dispersioonide sarnasusest või erinevusest lähtuvalt. Tabeli 2 viimases veerus on toodud olulisuse tõenäosuse väärtused (p) kahepoolse hüpoteesi põhjal. Sünteeskõne ja loomuliku kõne oluliselt erinevad keskmiste väärtused ja seda näitav p -väärtus on toodud poolpaksus kirjas. Silbiriimide kestussuhete keskmistes on olulised erinevused sünteeskõne ja loomuliku kõne I (0,89 vs. 0,76; $p = 0,004$) ja ka II (2,06 vs. 2,28; $p = 0,039$) vältes. Kui võrrelda nende testsõnade jaotuste variatiivsusi, siis silbiriimide kestussuhetel on sünteeskõnes standardhälbed nii I (0,12 vs. 0,28) kui ka II (0,45 vs. 0,70) vältes märgatavalt väiksemad ning jaotused kompaktsemad ja väiksema hajuvusega kui loomulikus kõnes. Sünteeskõne kompaktsemad jaotused on täheldatavad ka põhitooni maksimumide vahes, mille III välte parameetri keskmine on oluliselt erinev sünteeskõnes ja loomulikus kõnes (2,57 pt vs. 3,36 pt; $p = 0,044$). Vaid olulise erinevusega põhitooni tipu asukohal I vältes (71,1% vs. 78,5%; $p = 0,042$) on sarnase variatiivsusega jaotused sünteeskõnes ja loomulikus kõnes (standardhälbed 19,3 ja 18,0).

Tabel 2. Vâldete oluliste parameetrite keskmised väärtused ja standardhälbed ning sünteeskõne ja loomuliku kõne jaotuste sarnasused ja p -väärtused (poolpaksus kirjas on need keskväärtused, mille erinevus on valimites statistiliselt oluline ehk $p \leq 0,05$).

Parameeter	Vâlde	Sünteeskõne		Loomulik kõne		Jaotuste sarnasus	P
		keskmine väärtus	standard-hâlve	keskmine väärtus	standard-hâlve		
Silbiriimide kestussuhe	I	0,89	0,12	0,76	0,28	–	0,004
	II	2,06	0,45	2,28	0,70	–	0,039
	III	3,05	0,95	3,12	0,89	+	0,693
Põhitooni tipu asukoht (%)	I	71,1	19,3	78,5	18,0	+	0,042
	II	55,7	20,2	60,6	24,7	+	0,259
	III	46,2	20,5	46,3	24,4	+	0,973
Põhitooni maksimumide vahe (pt)	I	0,80	0,82	0,70	1,94	–	0,707
	II	1,84	1,26	1,81	2,69	–	0,931
	III	2,57	1,58	3,36	2,51	–	0,044

Vâldete hâaldustâpsuse kontrolliks korraldati tajutest, mille käigus 13 inimest (seitse naist ja kuus meest vanuses 15–66 aastat) hindasid sünteesitud ja loomuliku kõne lausete testsõnade vâlteid. Et lause kontekst ei mõjutaks vâlte määramist, esitati testsõnu veebilehel kuulamiseks eraldi, lausest välja lõigatult. Vâldete hindamisel võisid testijad veebilehel testsõnu kuulata mitu korda. Testsõna määrati sellesse vâlte kategooriasse, mida tajusid üle poole testijatest. Tabelis 3 on eksimismaatriksis sünteeskõne ja loomuliku kõne tajutestidel registreeritud vâldete õigete ja vâärade

liigitusjuhtude arvud ja õigesti määratud välte osakaalud. Üldine hääldustäpsus oli loomulikus kõnes üle kümne protsendipunkti võrra kõrgem kui sünteeskõnes (94,4% vs. 80,6%), kusjuures vokaali- ja konsonandikesksete testsõnade vältejajutajate suuri erinevusi ei olnud. Loomulikus kõnes tajuti õiges vältes veidi paremini konsonandikeskseid testsõnu (95,8% vs. 93,1%) ja sünteeskõnes vokaalikeskseid testsõnu (81,9% vs. 79,2%). Välteti oli kõige suurem erinevus II vältes testsõnade tajumisel (93,7% vs. 66,7%). Sünteeskõne 48-st I-vätelisest sõnast tajuti I vältes 46 testsõna ja II vältes kaks testsõna, mis annab hääldustäpsuseks 95,8%; loomuliku kõne I-vätelisi sõnu tajuti korrektselt (100%). Huvitaval kombel oluliste parameetrite põhjal tehtud diskriminantanalüüsil olid klassifitseerimistulemused vastupidised (vt tabelit 1): siis eristati sünteeskõne I vältes testsõnu korrektselt ja loomuliku kõne kahte sõna liigitati II välte alla. Kui sünteeskõne tajutestil tuvastati kõige kehvemini II vältet, vaid kahel kolmandikul juhtudest, siis diskriminantanalüüsil klassifitseeriti kõige halvemini III vältet (60,4%). Üldised õigesti tajutud ja õigesti klassifitseeritud vältete protsendimäärad olid sünteeskõne puhul suhteliselt sarnased (80,6% ja 79,2%). Loomuliku kõne vastavad protsendimäärad erinesid teineteisest palju rohkem: õigesti tajuti 94,4% testsõnade välteid, aga parameetrite väärtuste põhjal klassifitseeriti 86,1% testsõnu õigesse vältesse. Ilmselt taju ja diskriminantanalüüsi mittevastavus otseselt ei viita mõne akustilise korrelaadi puudumisele viie olulise parameetri hulgast. Parameetrite väärtuste suurem varieeruvus loomulikus kõnes (vt tabelit 2) osutab pigem kompensatsioonimehhanismide olemasolule loomuliku kõne vältemudelites, millega ühe parameetri kõrvalekallet kompenseeritakse teiste parameetrite väärtustega.

Tabel 3. Sünteeskõne ja loomuliku kõne tajutestidel testsõnade vältete tuvastamise eksimismatriks (ridades olevaid vältekategooriaid tajuti testil veergudes olevate vältetena).

Välde	Sünteeskõne				Loomulik kõne			
	I	II	III	Õigesti tajutud välde	I	II	III	Õigesti tajutud välde
I	46	2	0	95,8%	48	0	0	100%
II	0	32	16	66,7%	0	45	3	93,7%
III	0	10	38	79,2%	0	5	43	89,6%
Kokku				80,6%				94,4%

Peale sünteeskõne ja loomuliku kõne erinevuste pakkus töös huvi, millised on eri sünteesimeetodite lähenemisviisid vältete produtseerimisele. Ehk milliseid analüüsiks valitud üheksast parameetrist eri masinõppe algoritmid enda loodud sõna-prosoodia mudelites kasutavad ja milliseid mitte. Eri meetodite akustilisi parameetreid analüüsides võis täheldada olulistest vältetunnustes küllalt suuri erinevusi (vt tabelit 4). Kõigis meetodites oli vaid silbiriimide kestussuhe ühine olulisim vältetunnus, teisi akustilisi parameetreid kasutati välte produtseerimiseks küllaltki erinevalt. Markovi peitmudeleid rakendatavates süntesaatorites (HMM) oli välte olulise lisatunnusena kasutusel põhitooni tipu asukoht ja põhitooni maksimumide vahe, sügavatel närvivõrkudel põhinevates süntesaatorites (DNN) intensiivsuse ulatus ning RNN-süntesaatorites põhitooni ulatus ning põhitooni keskmiste ja maksi-

mumide vahe. Tabeli 4 alumises reas on toodud veldete haldustapsus erinevates konesuntee si meetodites, mille pohjal voib elda, et ehkki veldete produtseerimisel olid kullalt vaikeseid erinevused haldustapsuses (77,8–81,5%), siis oluliste akustiliste parameetrite valikus oli variatiivsus snagi suur. See naitab, et valte pohitunnusele – silbiriimide kestussuhtele – voivad masinoppealgoritmid lisatunnuseid kaasata suhteliselt vabalt sarnase haldustulemuse saavutamiseks.

Tabel 4. Valte tunnuste kasutamine erinevate sunteesimeetoditega loodud konesuntesaatorite prosoodiamudelites ja suntesaatorite veldete haldustapsus.

Valte tunnused	HMM	DNN	RNN
Silbiriimide kestussuhe	+	+	+
Pohitooni tipu asukoht	+	–	–
Intensiivsuse tipu asukoht	–	–	–
Pohitooni ulatus	–	–	+
Pohitooni keskmiste vahe	–	–	+
Pohitooni maksimumide vahe	+	–	+
Intensiivsuse ulatus	–	+	–
Intensiivsuse keskmiste vahe	–	–	–
Intensiivsuse maksimumide vahe	–	–	–
Veldete haldustapsus	80,6%	81,5%	77,8%

Sunteeskone valtevigade akustiliseks analuüsiks arvutati vaatluse all olnud uheksa parameetri keskmised vartused kahes osas: esiteks, tajutestides tuvastatud oige valte halduse korral ja teiseks, tajutud valte haldusvea korral. Tulemused on toodud tabelis 5. Tabelisse on kaasatud koik valte potentsiaalsed akustilised parameetrid, mitte ainult olulised, kuna konkreetse valte korral voivad koik parameetrid mingit spetsiifilist moju avaldada. Valtevigade hindamisel arvutati, kas vastava parameetri valteti keskmiste vartuste vahel on oluline statistiline erinevus (p-vartus), st kas vastava parameetri oige valte halduse ja valtevea valimid on erinevate keskvartustega voi mitte. Tabelis 5 on poolpaksus kirjas toodud need keskmised vartused koos p-vartusega, mis on veldete seisukohast olulised erinevused ja voivad pohjustada valtevigu. Silbiriimide kestussuhte vartus oli oluline parameeter ule koigi veldete. I valte haldusvea korral oli kestussuhte vartus 1,19, oigesti tajutud I valte korral oli silbiriimide keskmine kestussuhe 0,87. II ja III valte haldusvigade puhul olid veldete kestussuhted vahetusse lainud: II valte keskmine 2,59 ja III valtel 1,94. Rohulise ja rohuta silbi pohitooni ulatuses, keskmiste vahes voi maksimumide vahes eksimine oli oluline II valte halduses. III valte haldusvigu voib pohjustada peale kestussuhte intensiivsuskontuuri tipu suhteliselt hiline asukoht (77% rohulise silbi algusest). Aga I-valteliste sonade ebatapset haldust naib mojutavat ebaharilikult suur intensiivsuse maksimumide vahe (3,7 dB) rohulises ja rohuta silbis.

Tabel 5. Akustiliste parameetrite keskmised väärtused välteti tajutestidega tuvastatud õige välte häälduse ja vältevigade korral ning p-väärtused (poolpaksus kirjas on need keskväärtused, mille õige välte häälduse ja vältevea erinevus on valimites statistiliselt oluline ehk $p \leq 0,05$).

Parameeter	I välde			II välde			III välde		
	õige välde	välte- viga	p-väärtus	õige välde	välte- viga	p-väärtus	õige välde	välte- viga	p-väärtus
Silbiriimide kestussuhe	0,87	1,19	< 0,005	1,79	2,59	< 0,005	3,40	1,94	< 0,005
Põhitooni tipu asukoht (%)	71	78	0,634	56	56	0,999	44	53	0,144
Intensiivsuse tipu asukoht (%)	76	89	0,307	72	68	0,572	61	77	0,020
Põhitooni ulatus (pt)	3,10	3,65	0,587	3,87	5,03	0,020	4,50	4,31	0,809
Põhitooni keskmiste vahe (pt)	1,63	1,78	0,832	1,90	2,95	0,009	2,19	2,20	0,984
Põhitooni maksimumide vahe (pt)	0,77	1,55	0,189	1,60	2,29	0,043	2,62	2,05	0,278
Intensiivsuse ulatus (dB)	5,6	7,7	0,305	10,8	9,9	0,551	8,6	8,5	0,895
Intensiivsuse keskmiste vahe (dB)	1,5	2,8	0,394	3,0	3,8	0,254	2,9	2,0	0,203
Intensiivsuse maksimumide vahe (dB)	1,2	3,7	0,034	2,1	3,2	0,167	2,9	2,4	0,493

Kokkuvõte

Käesolev töö oli esimene analüüsikatse püüda sünteesi teel masinõppe meetoditega vältele kui fonoloogilisele kategooriale uue nurga alt läheneda ja erinevate aspektide osas selgust saada. Analüüsi tulemusi võrreldi loomuliku kõne materjaliga ja kõne-sünteesi eri meetodeid omavahel. Väldete hääldustäpsust kontrolliti tajutestidega ja vältetunnuste klassifitseerimisvõimet diskriminantanalüüsiga.

Sünteeskõne ja loomuliku kõne väldete analüüs kinnitas, et silbiriimide kestussuhe on olulisim akustiline korrelaat vältekategooriate tajumisel ja eristamisel. Kui võrrelda siinse artikli uurimismaterjali testsõnade kolme välte kestussuhete keskmisi väärtusi sünteeskõnes (I vältes 0,89, II vältes 2,06, III vältes 3,05) ja loomulikus kõnes (I vältes 0,76, II vältes 2,28, III vältes 3,12) varasemate uurimistööde kestussuhete koondkeskmisega (I vältes 0,71, II vältes 1,74, III vältes 2,77; Asu jt 2016: 136, tabel 4.1 põhjal), võib täheldada suurimaid suhtelisi erinevusi sünteeskõne I vältes (erinevus 25%, 0,71 vs. 0,89) ja loomuliku kõne II vältes (erinevus 31%, 1,74 vs. 2,28).

Teistel juhtudel jäävad kestussuhete suhtelised erinevused 7–17% piiridesse. Siinses materjalis olid kestussuhete väärtustes nii I vâlte (0,89 vs. 0,76) kui ka II vâlte keskmised (2,06 vs. 2,28) süntees- ja inimkõnes teineteisest oluliselt erinevad. Teistest vâlte-tunnustest olid olulised põhitooni kontuuri tipu asukoht rõhulises silbis ja põhitooni maksimumide vahe nii sünteeskõnes kui ka loomulikus kõnes. Sünteeskõne välteti jaotused olid väiksema variatiivsusega ja kompaktsemad kui loomulikus kõnes, eriti põhitooni maksimumide vahes ja silbiriimide kestussuhetes.

Väldete tajutesti üldtulemused olid sünteeskõne puhul sarnased parameetrite alusel klassifitseerimisvõimele diskriminantanalüüsil (80,6% ja 79,2%). Loomuliku kõne puhul olid taju ja parameetrite põhjal klassifitseerimise erinevused palju suuremad (94,4% vs. 86,1%). Paremad vâldete õigesti tajumise tulemused loomulikus kõnes, vaatamata oluliste parameetrite suuremale variatiivsusele, viitavad parameetritevahelisele interaktsioonile, millega ühe parameetri kõrvalekallet kompenseeritakse teiste parameetrite väärtustega. Selline kompensatsioonimehhanism näib sünteeskõne mudelites puuduvat.

Ehkki vâldete produtseerimisel eri sünteesimeetoditega olid hâäldustâpsuses küllalt väikesed erinevused (77,8–81,5%), siis oluliste akustiliste parameetrite valikus oli variatiivsus üsnagi suur. See näitab, et peale vâlte põhitunnuse (silbiriimide kestussuhe) võivad masinõppealgoritmid mudelitesse kaasata suhteliselt vabalt lisatunnuseid sarnase hâäldustulemuse saavutamiseks.

Vâlte hâäldusvigades prevaleeris sünteeskõnes esimese põhjusena eksimine silbiriimide kestussuhetes ja seda kõigi vâldete korral. Teiste parameetrite vâltevea valimid olid oluliselt erineva keskvaârtusega vaid üksikute vâldete puhul: I vâltes intensiivsuse ulatus, II vâltes kõik kolm põhitooni parameetrit ja III vâltes intensiivsuskontuuri tipu asukoht rõhulises silbis.

Ehkki vâldete pârîs uusi aspekte artiklis esile ei kerkinud, näitas see katseuurimus, et analüüs sünteesi kaudu on arvestatav ja perspektiivne meetod erinevate fonoloogiliste kategooriate testimiseks ja võib-olla ka kõne automaatseks fonoloogiliseks esitluseks. Uute aspektide ja eri sünteesimeetodite vaheliste süstemaatiliste erinevuste esilekerkimine võis ka jääda kõnekorpusete suhteliselt tagasihoidliku mahu (1000 lauset) taha. Seetõttu kindlamateks järeldusteks ja varjatud vâlteaspektide ilmsiks tulekuks analüüsil sünteesi kaudu peaks arvutitele ette sõõma oluliselt mahukamaid teksti- ja kõnekorpusi.

Artikli valmimist on toetanud Euroopa Liidu Regionaalarengu Fondi kaudu Eesti-uuringute Tippkeskus, samuti on see seotud Eesti Haridus- ja Teadusministeeriumi uurimisprojektiga IUT35-1 „Kõnestiilid, lauseprosoodia ja fonoloogiline varieerumine: kirjeldus, teooria ja modelleerimine”.

VEEBIVARAD

Boersma, Paul; Weenink, David 2018. Praat: doing phonetics by computer. Version 6.1.08.

Installitud 25. I 2020. <http://www.praat.org>

Merlin. <https://github.com/CSTR-Edinburgh/merlin>

Mozilla TTS. <https://github.com/mozilla/TTS>

Ossian. <http://simple4all.org/product/ossian>

WebMAUS Basic. <https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic>

KIRJANDUS

Asu, Eva Liina; Lippus, Pärtel; Pajusalu, Karl; Teras, Pire 2016. Eesti keele hääldus. (Eesti keele varamu II.) Tartu: Tartu Ülikooli Kirjastus.

Eek, Arvo; Meister, Einar 2003. Foneetilisi katseid ja arutlusi kvantiteedi alalt (I). Hääliku-kestusi muutvad kontekstid ja välde. – Keel ja Kirjandus, nr 12, lk 904–918.

Eek, Arvo; Meister, Einar 2004. Foneetilisi katseid ja arutlusi kvantiteedi alalt (II). Takt, silp ja välde. – Keel ja Kirjandus, nr 4, lk 251–271; nr 5, lk 336–351.

Ehala, Martin 1999. Eesti völdete probleemi üks lahendusi. – Keel ja Kirjandus, nr 6, lk 378–386; nr 7, lk 453–466.

Hint, Mati 1998. Why syllabic quantity? Why not the foot? – *Linguistica Uralica*, kd 34, nr 3, lk 172–177.

King, Simon 2014. Measuring a decade of progress in Text-to-Speech. – *Loquens*, kd 1, nr 1, e006.

Lehiste, Ilse 1960. Segmental and syllabic quantity in Estonian. – *American Studies in Uralic Linguistics. Uralic and Altaic Series 1*. Toim Thomas A. Sebeok. Bloomington: Indiana University Publications, lk 21–82.

Lehiste, Ilse 1970. Suprasegmentals. Cambridge (Mass.)–London: The M.I.T. Press.

Lehiste, Ilse 1975. Experiments with synthetic speech concerning quantity in Estonian. – *Congressus tertius internationalis fenno-ugristarum, Tallinnae habitus*, 17.–23. VIII 1970. Pars I: *Acta Linguistica*. Toim Valmen Hallap. Tallinn: Valgus, lk 254–269.

Liiv, Georg 1961. Eesti keele kolme völtusastme vokaalide kestus ja meloodiatööbid. – Keel ja Kirjandus, nr 7, lk 412–424; nr 8, lk 480–490.

Lippus, Pärtel; Pajusalu, Karl; Allik, Jüri 2011. The role of pitch cue in the perception of the Estonian long quantity. – *Prosodic Categories: Production, Perception and Comprehension. Studies in Natural Language and Linguistic Theory 82*. Toim Sónia Frota, Gorka Elordieta, Pilar Prieto. Dordrecht: Springer, lk 231–242.

Malisz, Zofia; Henter, Gustav Eje; Valentini-Botinhao, Cassia; Watts, Oliver; Beskow, Jonas; Gustafson, Joakim 2019. Modern speech synthesis for phonetic sciences: A discussion and an evaluation. – *Proceedings of the 19th International Congress of Phonetic Sciences*. Toim Sasha Calhoun, Paola Escudero, Marija Tabain, Paul Warren. Canberra: Australasian Speech Science and Technology Association Inc, lk 487–491.

Mihkla, Meelis; Hein, Indrek; Kalvik, Mari-Liis; Kiissel, Indrek; Sirts, Risto; Tamuri, Kairi 2012. Estonian speech synthesis: Applications and challenges. – *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”*. Toim A. E. Kibrik. Moskva: PFTY, lk 443–453.

van den Oord, Aaron; Dieleman, Sander; Zen, Heiga; Simonyan, Karen; Vinyals, Oriol; Graves, Alex; Kalchbrenner, Nal; Senior, Andrew; Kavukcoglu, Koray 2016. WaveNet: A generative model for raw audio. – arXiv preprint. arXiv:1609.03499.

- Piits, Liisi; Mihkla, Meelis; Nurk, Tõnis; Kiissel, Indrek 2007.** Designing a speech corpus for Estonian unit selection synthesis. – Nodalida 2007: Proceedings of the 16th Nordic Conference of Computational Linguistics. Toim Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, Mare Koit. Tartu: University of Tartu, lk 367–371.
- Prillop, Külli 2018.** Mida teeb moora eesti keeles? – Keel ja Kirjandus, nr 5, lk 345–364.
- Remmel, Mart 1975.** The phonetic scope of Estonian: Some specifications. Preprint KKI-5. Tallinn: Academy of Sciences of the Estonian S. S. R., Institute of Language and Literature.
- Ross, Kristiina; Piits, Liisi 2019.** Mõlgutusi tõest eesti keelekirjelduses. – Keel ja Kirjandus, nr 8–9, lk 686–694.
- Sahkai, Heete; Mihkla, Meelis 2019.** Intensity and spectral parameters as correlates of phrasal stress and word quantity in Estonian. – Proceedings of the 19th International Congress of Phonetic Sciences. Toim Sasha Calhoun, Paola Escudero, Marija Tabain, Paul Warren. Canberra: Australasian Speech Science and Technology Association Inc, lk 2475–2479.
- Zhu, Jian 2020.** Probing the phonetic and phonological knowledge of tones in Mandarin TTS models. – Proceedings of 10th International Conference on Speech Prosody. Tokyo, Japan, lk 930–934.
- Tauli, Valter 1973.** Standard Estonian grammar. Part I: Phonology, Morphology, Word-formation. (Acta Universitatis Upsaliensis 8.) Uppsala: Studia Uralica et Altaica Upsaliensia.
- Wang, Yuxuan; Skerry-Ryan, RJ; Stanton, Daisy; Wu, Yonghui; Weiss, Ron J.; Jaitly, Navdeep; Yang, Zongheng; Xiao, Ying; Chen, Zhifeng; Bengio, Samy; Le, Quoc; Agiomyriannakis, Yannis; Clarck, Rob; Saurous, Rif A. 2017.** Tacotron: Towards end-to-end speech synthesis. – Proceedings of Interspeech 2017, lk 4006–4010.
- Watts, Oliver; Henter, Gustav Eje; Merritt, Thomas; Wu, Zhizheng; King, Simon 2016.** From HMMs to DNNs: Where do the improvements come from? – Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, lk 5505–5509.
- Weske, Mihkel 1879.** Eesti keele healte õpetus ja kirjutuse wiis. Tartu: Schnakenburg.
- Xu, Yi; Prom-on, Santitham 2014.** Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. – Speech Communication, kd 57, nr 2, lk 181–208.

Meelis Mihkla (snd 1955), PhD, Eesti Keele Instituudi vanemteadur (Roosikrantsi 6, 10119 Tallinn), meelis.mihkla@eki.ee

Analysis of quantity degrees by synthesis

Keywords: quantity degrees, analysis by synthesis, speech synthesis, machine learning

The article is the first attempt to provide a fresh view of the quantity degree as a phonological category by using synthesis and machine learning. The results of the analysis are compared with recorded natural speech and with synthetic speech achieved by different methods of synthesis. Self-training speech synthesizers based

on hidden Markov models (HMM), on deep neural networks (DNN) and on recurrent neural networks (RNN) were used. The input for machine learning consisted of written Estonian texts (ca 1000 sentences) and audio texts recorded from different informants who used the written texts as a basis. The self-training TTS systems did not contain any Estonian-based module or analyser. To check the precision of pronunciation, the output speech of each synthesizer was subjected to a perception test. According to the results obtained, the general precision of pronunciation reached up to 80.6% for synthetic speech, vs. 94.4% for natural speech. Another evaluative mechanism used involved an analysis of nine acoustic parameters. The analytic comparison of synthetic and natural speech confirmed that the main acoustic cue to differentiate between quantity degrees is the durational ratio between syllable onsets. As for additional characteristics, it was the position of the peak of the pitch contour in the stressed syllable and the difference measured between the pitch maxima that proved to be significant parameters both in synthetic and natural speech. Synthetic speech displayed less intra-degree variability than natural speech, especially for the difference between pitch maxima and between the duration ratios of syllable onsets. While the different methods used for synthesizing quantity degrees ended up in relatively small differences in pronunciation precision (77.8–81.5%), involvement of relevant acoustic parameters brought about a considerable rise in variability. This indicates that although the duration ratio of syllable onsets will remain the main characteristic feature of a quantity degree, the algorithms of machine learning enjoy a relative freedom of choice among additional characteristics to achieve a good enough audio result. The prevalent cause of error in the pronunciation of words with different quantity degrees was the choice of a wrong duration ratio between syllable onsets, which applied to all three quantity degrees. The rest of error-causing parameters only worked for single quantity degrees. In conclusion, although the present study did not reveal any new aspects of Estonian quantity, it serves well as a pilot study indicating that analysis by synthesis is quite a considerable and promising method to test various phonological categories and phonological representation of speech.

Meelis Mihkla (b. 1955), PhD, Institute of the Estonian Language, Senior Researcher (Roosikrantsi 6, 10119 Tallinn), meelis.mihkla@eki.ee