https://doi.org/10.7592/MT2017.68.meister

Automaatse segmentimise hindamine

Einar Meister
Tallinna Tehnikaülikooli vanemteadur
einar.meister@ttu.ee
Lya Meister
Tallinna Tehnikaülikooli teadur
lya.meister@ttu.ee

Teesid: Töös hinnatakse kahe eestikeelse kõne automaatsegmentimise programmiga leitud hääliku- ja sõnapiiride erinevust võrreldes käsitsi määratud piiridega ning uuritakse, kui palju erinevad käsitsi ja automaatselt saadud märgendustest mõõdetud segmendikestused. Uuringus kasutatakse 14 keelejuhi kõnenäiteid (kokku 208 lauset), neist neli eesti emakeelega (L1) täiskasvanut, kuus L1 last ning neli eesti keelt võõrkeelena (L2) kõnelevat täiskasvanut. Tulemused näitavad, et mõlemad automaatsüsteemid annavad paremaid tulemusi L1 täiskasvanute kõne puhul võrreldes lastekõne ja L2 kõnega. Automaatsegmenditud materjalist mõõdetud segmendikestused on lähedased käsitsi segmenditud kõnest leitud kestustele.

Märksõnad: automaatne märgendus, eesti keel, häälikupiirid, kõnekorpused, segmendikestused, sõnapiirid

Sissejuhatus

Loomuliku keele automaattöötluse arengus eristatakse kahte peamist perioodi: (1) 1950. aastate lõpust kuni 1980. aastate keskpaigani domineeris ratsionalistlik (reeglipõhine) keelekäsitlus, mis lähtus Noam Chomsky generatiivse lingvistika alastest töödest (Chomsky 1956, 1957 jm), (2) alates 1980. aastate keskelt said üha populaarsemaks empiirilised ehk andmepõhised (statistilised) meetodid, mis on tänapäeval levinuimaks lähenemiseks praktiliselt kõigis loomuliku keele arvutitöötluse valdkondades (teksti analüüs, masintõlge, kõnesüntees ja -tuvastus jm) (vt diskussiooni reeglipõhise ja statistilise keeletöötluse kohta nt Koit 2006; Jelinek 2005; Hajič & Hajičová 2007). Andmepõhiste mudelite treenimiseks vajatakse märgendatud teksti- ja kõnekorpusi, statistiliste

meetodite laialdane kasutuselevõtt andis nende loomisele olulise tõuke. Suuremahuliste korpuste kogumine ja märgendamine on töö- ja ajamahukas ning seetõttu kallis, eriti kui selleks kasutatakse inimtööjõudu. Korpuste loomise, eelkõige märgendamise kiirendamiseks kasutatakse ka automaatseid vahendeid (nt teksti morfoloogiline ja süntaktiline märgendus, kõne automaatne segmentimine jm), kuid üldjuhul on inimese loodud märgendus täpsem.

Kõnekorpusi kasutatakse laialdaselt automaatse kõnetuvastuse ja statistilise kõnesünteesi akustiliste mudelite treenimiseks, korpus-põhise kõnesünteesi akustilise baasina ja kõne eksperimentaal-foneetilistes uuringutes. Üldiselt mõistetakse termini "kõnekorpus" all digitaliseeritud kõnesignaalide kogumit, mis on varustatud märgenduse, meta-andmestiku ja dokumentatsiooniga. Korpuse märgendus on füüsilise signaaliga seotud diskreetne kirjeldus, see koosneb piiratud hulgast sümbolitest, mis on lingitud mingite ajahetkede või segmentidega; metaandmestik sisaldab informatsiooni salvestusprotseduuride ja keelejuhtide kohta (Schiel & Draxler *et al*. 2004).

Kõnekorpuste kasutamine erinevates uuringutes sõltub suurel määral sellest, kui rikkalik on kõnematerjali märgendus ja kui täpselt on määratud eri kõnesegmentide piirid. Näiteks eesti laste kõnekorpuse (Meister & Meister 2017) ja aktsendikorpuse (Meister & Meister 2012a) salvestused on märgendatud sõna- ja häälikutasemel, eesti keele spontaanse kõne foneetilises korpuses[1] kasutatakse märksa põhjalikumat märgendusskeemi, mis lisaks sõna- ja häälikutasemele sisaldab ka häälikustruktuuri, silpide, kõnetaktide, eri kõnelejate kõnevoorude, häälelaadi (kärin, kähin) ja paralingvistiliste nähtuste (sisse- ja väljahingamine, köhimine, nuuskamine, naer jms) märgenduskihte. Eelnimetatud korpuste märgendamisel leitakse sõna- ja häälikupiirid esmalt automaatselt, seejärel korrigeeritakse segmentide piirid käsitsi.

Erinevalt tekstist, on kõnesignaal olemuslikult pidev ja sidusas kõnes puuduvad selged sõnade vahelised piirid, rääkimata silpide või häälikute vahelistest piiridest. Siiski on akustilise helilaine segmentimine sõnadeks ja häälikuteks võimalik kokkuleppeliste reeglite alusel[2,3], mis lähtuvad eri häälikute artikulatoorsetest ja akustilistest tunnustest. Kõnesignaali käsitsi segmentimisel-märgendamisel lähtutakse üldjuhul helilainest ja selle spektrogrammist (vt joonis 1), mis võimaldab visuaalselt leida üleminekud ühelt häälikult teisele. Samas on käsitsi märgendamisel probleemiks eri märgendajate variatiivsus segmendipiiride asukoha määramisel. On leitud, et eri märgendajate segmendipiiride keskmine hälve on 10 ms (Wesenick & Kipp 1996, saksakeelne kõne) kuni 16 ms (Pitt & Johnson *et al*. 2005, ingliskeelne kõne); erinevate märgendajate käsitsi määratud häälikupiiridest mahuvad 87%, 96% ja 99% vastavalt 10 ms, 20 ms ja 32 ms suuruse ajaakna sisse (Wesenick & Kipp 1996). Ingliskeelse

TIMIT-korpuse käsitsi leitud häälikupiiridest mahub 20 ms sisse 93,5%, mitmete muude korpuste keskmine on vastavalt 93,8% (Hosom 2009).

Mitmete keelte jaoks on olemas ka kõnetuvastustehnoloogial põhinevad kõne automaatse segmentimise programmid, sealhulgas ka kaks eestikeelse kõne jaoks loodud veebirakendust: (1) Tallinna Tehnikaülikooli Küberneetika Instituudis loodud automaatne segmentija[4] (edaspidi TTÜ süsteem), ja (2) Müncheni ülikoolis loodud WebMAUS[5] (edaspidi WebMAUS) (Kisler & Reichel *et al*. 2016). Mõlemad programmid genereerivad esmalt ortograafilise teksti põhjal foneetilise transkriptsiooni ja seejärel joondavad selle akustilise signaaliga, st leiavad foneetilisele transkriptsioonile vastavate kõnesegmentide piirid kõnesignaalis, kasutades automaatse kõnetuvastuse akustilisi GMM-mudeleid (ingl *GMM – Gaussian Mixture Models*) ja sundjoondamise (ingl *forced alignment*) algoritmi.

TTÜ süsteem aktsepteerib sisendina erinevas kodeeringus tekstifaile (nt UTF-8, Windows-1257 või muu) ja erinevas formaadis helifaile (nt WAV, MP3, OGG, jm), kusjuures tekstifail peab sisaldama helifailis salvestatud kõne ortograafilist transkriptsiooni. Süsteemi väljundiks on Praati (Boersma & Weenink 2017) TextGrid-fail sõna- ja häälikupiiridega. Sõnatasemel kasutatakse ortograafilist kirjaviisi, häälikutasemel aga kohaldatud foneetilist transkriptsiooni. Programm leiab automaatselt ka eri liiki täidetud pausid (nt köhatused, kõhklused, mürad jms) ja tähistab need erinevate märgenditega. Automaatseks töötluseks saab üles laadida korraga mitmeid faile andes ette kausta arvutikettal, milles asuvad teksti- ja helifaili paarid. TTÜ süsteem sobib ainult eestikeelse kõne jaoks; see kasutab Java Web Start tehnoloogiat (eeldab Java installeerimist arvutisse) ja on sõltumatu veebibrauserist.

WebMAUSi veebiliides võimaldab valida baasversiooni (WebMAUS Basic) ja suuremate valikuvõimalustega versiooni (WebMAUS General) vahel. WebMAUSi keeltevalikus on 17 keelt (lisaks veel mitmed baski, inglise ja saksa keele piirkondlikud variandid), sealhulgas eesti keel. Baasversioon võimaldab töötlemiseks üles laadida UTF-8 kodeeringus tekstifaile (peab sisaldama kõne ortograafilist transkriptsiooni) koos vastava helifailiga WAV või NIST/SPHERE formaadis. Väljundiks on sõna- ja häälikupiiridega märgendfail Praati TextGrid formaadis. Laiendatud versioon WebMAUS General võimaldab valida tekstifailis kasutatavat transkriptsiooni (SAMPA[6] või IPA[7]), väljundfaili formaati (lisaks TextGridile veel mitmeid muid formaate) ja transkriptsioonisüsteemi (SAMPA või IPA), lisada segmenteerimiskihte ja määrata muid parameetreid. WebMAUSi on soovitav kasutada Chrome'i veebisirvijas.

Automaatsed segmentimis-märgendusprogrammid võimaldavad väga kiiresti töödelda suuremahulisi salvestusi, kuid nende täpsus pole kunagi nii hea kui käsitsi segmentimise puhul. Näiteks WebMAUSi algse versiooni (MAUS)

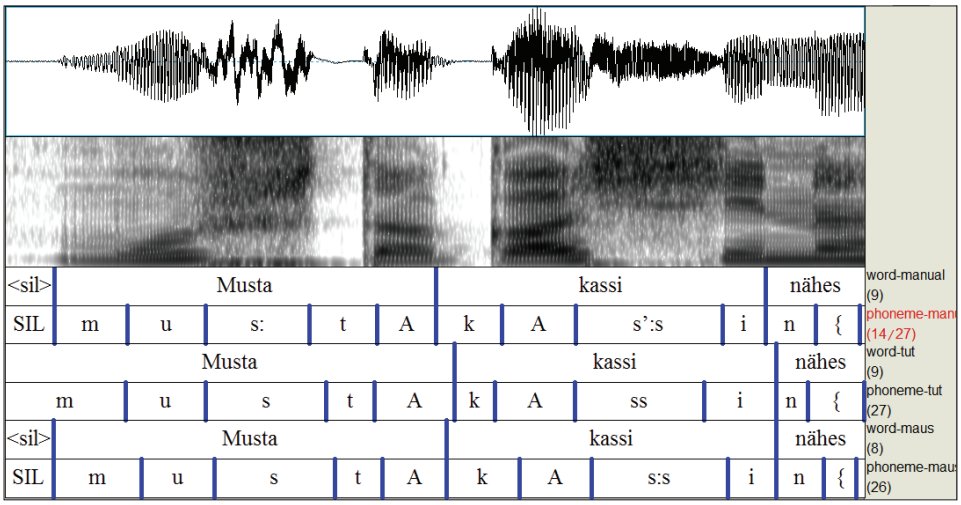
puhul mahtus 10 ms sisse 61%, 20 ms sisse 84% ja 32 ms sisse 90% saksakeelses kõnes automaatselt leitud häälikupiiridest (Wesenick & Kipp 1996; vrd käsitsi leitud piiridega eespool). Kahjuks pole hilisemaid võrdlusanalüüse WebMAUSi kohta teadaolevalt tehtud. TIMIT-korpuse automaatsegmentimisel on saadud tulemuseks 79,5% ja 92,8% häälikupiiridest vastavalt 10 ms ja 20 ms sees (Hosom 2009). Üks hilisem uuring tutvustab automaatsegmentimise meetodit, mis on andnud sama korpuse puhul tulemuseks vastavalt 84,6% ja 95,4% (Rendel & Sorin *et al*. 2012).

Vaatamata väiksemale täpsusele kasutatakse automaatseid programme märgenduste kiireks genereerimiseks, mida hiljem käsitsi korrigeeritakse. Kui hinnanguliselt kulub 1 sekundi kõne käsitsi segmentimiseks-märgendamiseks 100 kuni 1000 sekundit (sõltuvalt salvestuse kvaliteedist, märgendaja kogemustest ja märgenduse keerukusest), siis sama pika kõnesalvestuse automaatse märgenduse saamiseks TTÜ süsteemi ja WebMAUSiga kulub ainult mõni sekund (sõltuvalt võrguühenduse kiirusest ja serveri koormusest).

Töö eesmärk on hinnata kahe eelnimetatud automaatse programmiga leitud hääliku- ja sõnapiiride hälbeid käsitsi määratud piiridest ning uurida, kui palju erinevad automaatsest ja käsitsi märgendusest leitud segmendikestused.

Materjal ja meetod

Uuringuks valitud materjal sisaldab 14 keelejuhi kõnenäiteid, neist neli eesti emakeelega (L1) täiskasvanut (kaks meest, kaks naist), kuus L1 last (kolm poissi ja kolm tüdrukut vanuses 10–13 aastat) ning neli eesti keelt võõrkeelena (L2) kõnelevat täiskasvanut (kaks läti ja kaks rootsi emakeelega keelejuhti, kummagi keeletaustaga üks mees ja üks naine). Täiskasvanud L1 keelejuhtidelt valiti igaühelt 27 etteloetud lauset, teistelt keelejuhtidelt kümme lauset, kokku 208 lauset. Kõik laused segmenditi käsitsi ja automaatselt sõna ning hääliku tasandil. Uurimismaterjali märgendasid käsitsi kolm kogemustega märgendajat lähtudes ühtsetest märgendusreeglitest, kusjuures iga osa (L1 täiskasvanud, L1 lapsed, L2 täiskasvanud) märgendas erinev märgendaja. Kahe automaatse süsteemiga saadud märgendus teisendati Praati skriptidega käsitsi märgendustega samasugusele kujule (automaatselt kaheks komponendiks jagatud pikad vokaalid ja geminaadid liideti üheks segmendiks ja märgendati vastavalt topeltvokaali või -konsonandiga, järjestikused pausid liideti üheks segmendiks, häälikutaseme märgendus teisendati SAMPA formaati). Joonisel 1 on esitatud näide kolmel viisil saadud märgendustest.



Joonis 1. Lause "Musta kassi nähes lõi ta araks" signaalilõik (ülemine aken), spektrogramm (keskmine aken) ja märgendused (alumine aken). Kahes ülemises märgenduskihis on käsitsi leitud sõna- ja häälikupiirid, kahes keskmises kihis on TTÜ süsteemiga automaatselt leitud piirid ja kahes alumises kihis on WebMAUSi leitud piirid.

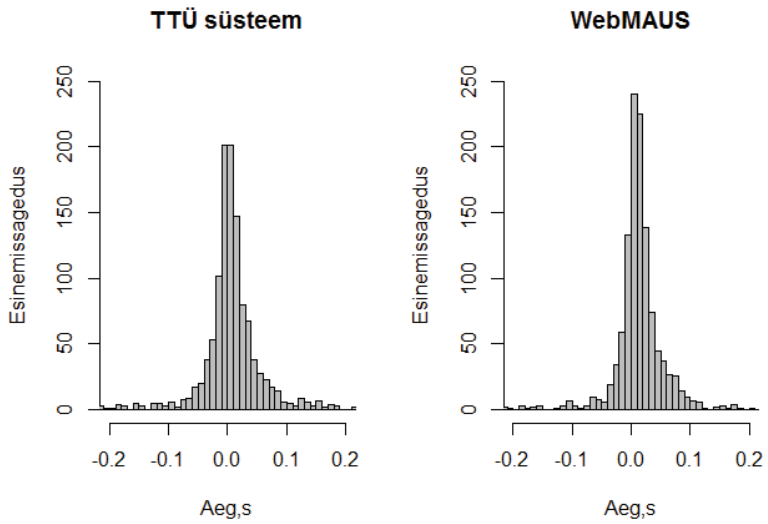
Sõna- ja häälikutaseme piiridele vastavad ajaväärtused salvestati Praati skripti abil andmefaili ja nende põhjal arvutati automaatselt leitud piiride hälbed käsitsi määratud piiride suhtes. Andmestikust filtreeriti välja automaatse märgenduse ilmsed vead, s.t juhtumid, mille puhul automaatses märgenduses piirimärgend puudus või automaatsele piirile ei leitud vastet käsitsi märgenduses (ca 23% sõnapiiridest ja 11,5% häälikupiiridest). Kokku võrreldi 1179 sõnapiiri ja 5050 häälikupiiri. Tulemuste statistiliseks analüüsiks kasutati programmi RStudio (RStudio Team 2015).

Tulemused

Sõnapiiride võrdlus

Sõnapiiride võrdlus näitas, et mõlemad automaatsed süsteemid paigutavad sõnapiirid valdavalt hilisemaks võrreldes käsitsi leitud piiridega (vt ka joonis 2). Kogu uurimismaterjalist automaatselt leitud piirid hälbivad käsitsi leitud piiridest TTÜ süsteemi puhul keskmiselt 9,2 ms (sh = 76,5 ms; mediaan = 5 ms) ja WebMAUSi puhul keskmiselt 18,2 ms (sh = 76,8 ms; mediaan = 12 ms).

Ootuspäraselt saadi parimad tulemused L1 täiskasvanute kõne puhul, kus keskmine sõnapiiride erinevus on 4,5 ms (sh = 38,9 ms; mediaan = 3 ms) ja 8,9 ms (sh = 38,2 ms; mediaan = 12 ms) vastavalt TTÜ süsteemi ja WebMAUSi puhul. Võrreldes L1 täiskasvanutega on laste kõnes käsitsi ja automaatselt leitud piiride erinevused suuremad: keskmine hälve TTÜ süsteemi puhul on 9,5 ms (sh = 87,4 ms; mediaan = 11 ms) ja WebMAUSi puhul 24,8 ms (sh = 57,4 ms; mediaan = 12 ms) . L2 kõnes on automaatselt leitud piiride hälbed veelgi suuremad: TTÜ süsteemi keskmine = 23,6 ms (sh = 126,5 ms; mediaan = 8 ms) ja WebMAUSi keskmine = 35,6 ms (sh = 154,6 ms; mediaan = 14 ms).



Joonis 2. Automaatselt leitud sõnapiiride hälvete histogrammid.

Sõnapiiride jaotuse (joonis 2) asümmeetriakordaja (ingl *skewness*) on mõlema juhul suurem nullist: TTÜ süsteemi puhul 1,7 ja WebMAUSi puhul 4,6, s.t mõlema süsteemi piiride jaotused on parempoolse ebasümmeetriaga, kuid TTÜ süsteemi jaotus on sümmeetrilisem. Parempoolse ebasümmeetriaga jaotuse korral on sõnapiirid määratud sagedamini hilisemaks võrreldes käsitsi määratud sõnapiiridega, seda näitavad ka hälvete positiivsed mediaanväärtused (vt eespool). Sõnapiiride jaotus on järsem WebMAUSi puhul – järsakuskordaja e ekstsess (ingl *kurtosis*) on 60,6; TTÜ süsteemi puhul vastavalt 34,7. Seega on WebMAUSi abil sama ajaintervalli sees leitud rohkem piire kui TTÜ süsteemi puhul, nt ±50 ms ajaakna sisse mahub 82,6% kõigist WebMAUSi leitud piiridest, TTÜ süsteemi puhul aga 80,6% piiridest (vt tabel 1).

Järgnevalt on kahe süsteemi segmenteerimiskvaliteedi võrdlemiseks arvutatud automaatselt leitud piiride osakaal, mille absoluutne hälve käsitsi määratud piiridega võrreldes jääb 10–50 ms vahele. Tulemused on esitatud tabelis 1. Kui aktsepteeritavaks hälbeks võtta 20 ms (seda ajaintervalli loetakse aktsepteeritavaks mitmetes analoogsetes uurimustes, nt Ljolje & Hirschberg *et al*. 1997; Toledano & Gomez *et al*. 2003; Hosom 2009), siis L1 täiskasvanute kõne puhul paikneb TTÜ süsteemiga leitud sõnapiiridest 20 ms ajaintervalli sees 72,1%, WebMAUSi määratud piiridest aga 62,3%. L1 laste ja L2 kõne puhul annab paremaid tulemusi WebMAUS – laste kõne puhul on vastav osakaal 49%, ja L2 kõnes 49,2%; TTÜ süsteemi vastavad näitajad on 35,7% ja 40,2%.

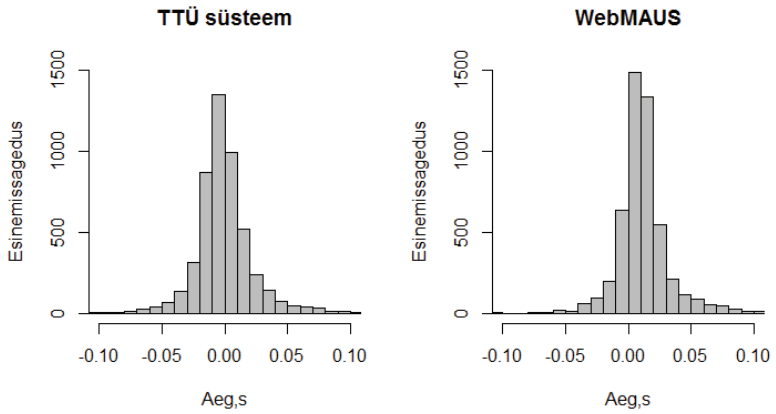
Tabel 1. Automaatselt leitud sõnapiiride osakaal (%) erinevate hälbeintervallide korral.

	Kõik keelejuhid		**L1 täiskasvanud**		**L1 lapsed**		**L2**	
Hälve	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**
<10 ms	35,7	32,6	47,5	35,1	23,1	29,8	21,6	29,6
<20 ms	55,6	56,1	72,1	62,3	35,7	49,0	40,2	49,2
<30 ms	67,2	70,7	84,9	79,5	46,5	59,3	49,2	63,3
<40 ms	75,5	78,4	90,5	87,3	57,9	67,1	60,3	70,9
<50 ms	80,6	82,6	92,6	90,8	66,3	71,9	68,8	76,4

Häälikupiiride võrdlus

Automaatselt leitud häälikupiiride hälbed osutusid võrreldes automaatsete sõnapiiridega oluliselt väiksemateks – kõigi keelejuhtide keskmine on TTÜ süsteemi puhul 0,6 ms (sh = 37,6 ms; mediaan = 0 ms) ja WebMAUSi puhul 13,8 ms (sh = 36,9 ms; mediaan = 10 ms). L1 täiskasvanute kõnes hälbisid TTÜ süsteemi piirid –0,4 ms (sh = 21,4 ms; mediaan = 0 ms), ja WebMAUSi piirid 11,9 ms (sh = 21,2 ms; mediaan = 11 ms). L1 laste kõnes on TTÜ süsteemi keskmine hälve 3 ms (sh = 46,7 ms; mediaan = 0 ms) ja WebMAUSi hälve 17,1 ms (sh = 37,5 ms; mediaan = 10 ms) ning L2 kõnes vastavalt –0,6 ms (sh = 55,5 ms; mediaan = 0 ms) ja 14,2 ms (sh = 64,3 ms; mediaan = 9,5 ms).

Häälikupiiride jaotused (joonis 3) on sarnased sõnapiiride jaotustega (joonis 2) – TTÜ süsteemi asümmeetriakordaja on 1,4 ja WebMAUSil 5,4. TTÜ süsteemi hälvete jaotus on lähedane sümmeetrilisele kujule ja hälvete mediaan = 0, s.t võrdsel määral esineb positiivseid ja negatiivseid hälbeid käsitsi määratud piiridest. WebMAUSi histogramm on selgelt parempoolse ebasümmeetriaga (mediaan = 10 ms), st rohkem esineb hilisemaks määratud häälikupiire. TTÜ süsteemi histogrammi järsakuskordaja on 98,4, WebMAUSi jaotusel aga 153,4.



Joonis 3. Automaatselt leitud häälikupiiride hälvete histogrammid.

Tabelis 2 on esitatud automaatselt leitud häälikupiiride osakaal, mille absoluutne hälve käsitsi määratud piiridega võrreldes jääb 10–50 ms vahele. Tulemused näitavad, et TTÜ süsteem töötab paremini L1 täiskasvanute kõne puhul (hälbeintervalli 20 ms sisse mahub 81,4% automaatselt leitud häälikupiiridest; WebMAUSi puhul 75%). WebMAUS aga leiab 20 ms intervalli piires 68,4% ja 72,7% häälikupiiridest vastavalt lastekõnes ja L2 kõnes, mis on veidi parem TTÜ süsteemist (vastavalt 66,2% ja 69,1%).

Tabel 2. Automaatselt leitud häälikupiiride osakaal (%) erinevate hälbeintervallide korral.

	Kõik keelejuhid		**L1 täiskasvanud**		**L1 lapsed**		**L2**	
Hälve	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**
<10 ms	48,5	42,7	53,4	42,8	41,8	41,9	44,8	43,8
<20 ms	74,8	72,7	81,4	75,0	66,2	68,4	69,1	72,7
<30 ms	85,3	85,4	91,3	88,8	77,4	79,3	80,3	84,9
<40 ms	90,6	90,6	95,0	94,3	84,7	84,0	87,0	90,4
<50 ms	93,5	93,2	97,1	96,5	88,4	87,4	90,9	92,5

Häälikupiiride võrdlus häälikuühendites

Järgnevalt analüüsime automaatselt leitud piiride hälbeid vokaal-klusiil, klusiil-vokaal, vokaal-frikatiiv, frikatiiv-vokaal, vokaal-nasaal, nasaal-vokaal ja vokaal-vokaal ühendites. Tulemused on esitatud tabelis 3.

TTÜ süsteem annab paremaid tulemusi L1 täiskasvanute kõne puhul, kus 20 ms hälbeintervalli sisse mahub 89–90,9% vokaal-klusiil, klusiil-vokaal, frikatiiv-vokaal ja nasaal-vokaal ühendite piiridest; vokaal-frikatiiv ja vokaal-nasaal piire on leitud veidi vähem, vastavalt 85,7% ja 80,8%, ning vokaalide vahelisi piire ainult 74%. L1 lastekõnes leitud piiride osakaal on enamuses häälikuühenditest ca 3–10% võrra väiksem võrreldes L1 täiskasvanud kõnega, erandiks on vokaal-klusiilühendid, kus leitud piiride osakaal on ainult 51,7% ja klusiil-vokaalühendid, kus tulemus on isegi 0,8% võrra parem kui täiskasvanute kõnes. L2 kõnes leitud piiride osakaal on L1 tulemustest väiksem kõigis häälikuühendites, halvim tulemus (50%) on saadud vokaalide vahelise piiri määramisel.

L1 täiskasvanute kõne puhul on WebMAUS TTÜ süsteemist parem klusiil-vokaal, vokaal-frikatiiv ja vokaal-nasaal ühendite piiride leidmisel vastavalt 2,4%, 4,1% ja 11,4% võrra; teiste häälikuühendite puhul annab paremaid tulemusi TTÜ süsteem, eriti vokaal-klusiilühendite (26,2% võrra) ja vokaal-vokaalühendite (28,5% võrra) puhul. Lastekõne puhul edestab WebMAUS TTÜ süsteemi enim vokaal-frikatiivühendite (9,2% võrra) puhul ja TTÜ süsteem omakorda WebMAUSi vokaalide vahelise piiri leidmisel (12,5% võrra). L2 kõnes annavad mõlemad süsteemid identseid tulemusi vokaal-klusiil ja klusiil-vokaal ühendite puhul, TTÜ süsteem leiab enam piire ainult nasaal-vokaalühendite puhul, ülejäänud häälikuühendites annab paremaid tulemusi WebMAUS, eriti vokaal-frikatiiv ja vokaal-nasaalühendites (vrd andmeid tabelis 3).

Tabel 3. Automaatselt leitud häälikupiiride osakaal (%) 20 ms hälbeintervalli korral.

Häälikuühend	**Kõik keelejuhid**		**L1 täiskasvanud**		**L1 lapsed**		**L2**	
	TTÜ	**WebMAUS**	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**	**TTÜ**	**WebMAUS**
Vokaal-klusiil	76,8	61,0	90,5	64,3	51,7	46,4	77,8	77,8
Klusiil-vokaal	89,9	91,5	90,1	92,5	90,9	91,9	87,2	87,2
Vokaal-frikatiiv	81,0	89,3	85,7	89,8	76,3	85,5	77,1	94,0
Frikatiiv-vokaal	88,3	83,5	90,9	85,1	85,5	79,0	84,4	87,5
Vokaal-nasaal	79,1	88,7	80,8	92,2	77,1	79,5	76,5	92,2
Nasaal-vokaal	87,1	88,8	89,0	80,4	85,9	89,1	83,3	79,6
Vokaal-vokaal	66,8	48,5	74,0	45,5	63,8	51,3	50,0	52,6

Segmendikestuste võrdlus

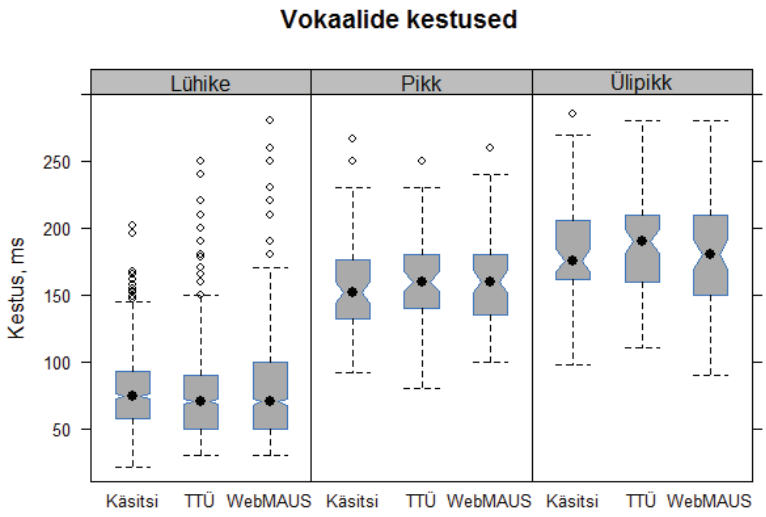
Mõõtmisusku foneetikuid (vt Hint 2016) huvitab kindlasti küsimus, kas ja kui palju hälbivad automaatselt segmenditud kõne akustilise analüüsi tulemused käsitsi segmenditud kõnematerjalil tehtud analüüsi tulemustest. Kuna automaatselt leitud häälikupiirid hälbivad käsitsi määratud piiridest, võib eeldada, et ka segmentide kestusmõõtmiste tulemused on kahe segmentimisviisi puhul erinevad. Samas nägime eelnevalt, et automaatselt leitud piiride hälbed on nii positiivsed kui ka negatiivsed (vt hälvete jaotusi joonisel 3) ja küllalt suure analüüsitava kõnematerjali korral kompenseerivad erisuunalised hälbed teineteist ning seetõttu ei pruugi mõõtmistulemused oluliselt erineda käsitsi segmenditud kõnematerjali analüüsil saadud tulemustest. Järgnevalt võrdlemegi segmendikestusi automaatselt ja käsitsi segmenteeritud kõnes.

Tabelis 4 on esitatud käsitsi ja kahe automaatse süsteemiga segmenditud kõnematerjalist mõõdetud häälikute kestusandmed. Lühikeste vokaalide kestus on arvutatud kõigi lühikeste vokaalide keskmisena sõltumata rõhust ja asukohast sõnas. Pikkade ja ülipikkade vokaalide kestused on leitud kui teise- ja kolmandavälteliste taktide esisilbi topeltvokaalide kestus. Lühikeste konsonantide kestuse puhul on kaasatud kõik üksikkonsonandid sõltumata asukohast sõnas. Pikkade ja ülipikkade konsonantide puhul on mõõdetud topeltkonsonantide kestusi vastavalt teise- ja kolmandavältelistes taktides. Joonistel 4 ja 5 on esitatud mõõdetud segmendikestuste karpdiagrammid.

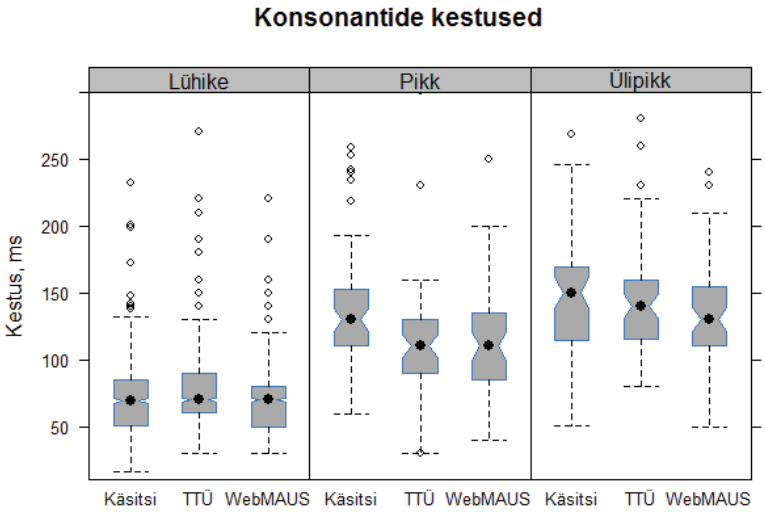
*Tabel 4. Häälikute keskmised ja mediaan- (sulgudes) kestused ning standardhälbed käsitsi ja automaatselt segmenditud kõnes.**

		Käsitsi		**TTÜ süsteem**		**WebMAUS**	
		Kestus	**Sh**	**Kestus**	**Sh**	**Kestus**	**Sh**
Vokaalid	**Lühike**	77 (74)	26,8	76 (70)	31,0	80 (70)	36,0
	Pikk	157 (152)	34,6	160 (160)	35,9	162 (160)	37,1
	Ülipikk	181 (176)	38,9	189 (190)	40,5	182 (180)	41,7
Konsonandid	**Lühike**	69 (69)	24,9	73 (70)	27,1	66 (70)	24,6
	Pikk	135 (130)	45,7	119 (110)	54,4	113 (110)	42,3
	Ülipikk	145 (150)	45,0	144 (140)	42,2	133 (130)	39,4

* Tabelis toodud kestusandmed on esitatud erinevate segmentimismeetodite võrdluseks ega ole mõeldud fonoloogiliseks tõlgenduseks ühegi välteteooria kontekstis.



Joonis 4. Käsitsi ja automaatselt segmenteeritud kõnest mõõdetud vokaalikestuste karpdiagrammid.



Joonis 5. Käsitsi ja automaatselt segmenteeritud kõnest mõõdetud konsonandikestuste karpdiagrammid.

Dispersioonanalüüs (ANOVA) näitas, et segmentimise viis on vokaalikestusi mõjutav faktor ainult lühikeste vokaalide puhul ($F[2, 3045] = 5{,}8$; $p < 0{,}01$). Tukey test (Tukey HSD – Honestly Significant Difference) näitas, et lühikeste vokaalide keskmine kestus ei erine käsitsi segmenteeritud (77 ms) ja TTÜ süsteemi autosegmenteeritud (76 ms) kõnematerjalis ega käsitsi ja WebMAUSi segmentimise (77 ms vs 80 ms) puhul ($p < 0{,}1$). Kestuserinevused (75 ms vs 80 ms) on aga olulised TTÜ süsteemi ja WebMAUSi võrdluses ($p < 0{,}05$).

Pikkade vokaalide kestuste erinevused (käsitsi 157 ms, TTÜ süsteem 160 ms, WebMAUS 162 ms) ei ole statistiliselt olulised ($F[2, 237] = 0{,}5$; $p = 0{,}6$), samuti osutusid ebaolulisteks ülipikkade vokaalide kestuserinevused (käsitsi 181 ms, TTÜ süsteem 189 ms, WebMAUS 182 ms; $F[2, 219] = 0{,}9$; $p = 0{,}4$).

Konsonantide kestuste analüüsil leidis ANOVA olulisi erinevusi lühikeste ($F[2, 2349] = 15$; $p < 0{,}001$) ja pikkade ($F[2, 189] = 3{,}4$; $p < 0{,}05$) konsonantide puhul, kuid mitte ülipikkade konsonantide ($F[2, 177] = 1{,}4$; $p = 0{,}3$) kestustes. Tukey võrdlustest näitas, et eri viisidel segmenteeritud kõnest saadud lühikeste konsonantide keskmiste kestuste (käsitsi 69 ms, TTÜ süsteemiga 73 ms, WebMAUS 66 ms) erinevused on statistiliselt olulised käsitsi ja TTÜ süsteemi võrdluses ($p < 0{,}01$) ning TTÜ ja WebMAUSi võrdluses ($p < 0{,}001$). Kestuste vahe on väikseim käsitsi ja WebMAUSi segmentimiste puhul ja see jääb olulisusnivoo ($p = 0{,}055$) piirile. Pikkade konsonantide kestuste erinevus on ainsana oluline käsitsi ja WebMAUSi võrdluses ($p < 0{,}05$).

Kokkuvõte

Kokkuvõttena tõdeme, et mõlemad automaatsüsteemid annavad paremaid tulemusi L1 täiskasvanute kõne puhul võrreldes lastekõne ja L2 kõnega. TTÜ programmiga leitud sõna- ja häälikupiirid on L1 täiskasvanute kõne puhul pisut täpsemad kui WebMAUSi abil saadud piirid (piiride keskmine hälve on TTÜ süsteemi puhul väiksem ja 20 ms ajaintervalli sees paiknevate piiride osakaal suurem kui WebMAUSil). Kuid L2 kõne ja lastekõne puhul on nii sõna- kui häälikupiiride tulemused WebMAUSi kasuks. Mõlema süsteemi treenimisel on kasutatud L1 täiskasvanute kõnet eestikeelsest BABELi korpusest (Roach *et al*. 1996), WebMAUSi puhul lisaks ka Tartu ülikooli eesti keele spontaanse kõne foneetilist korpust. Just seetõttu on mõlema süsteemi tulemused paremad L1 täiskasvanute kõne puhul. WebMAUSi tulemused lastekõne ja L2 kõne puhul on TTÜ süsteemist paremad tõenäoliselt tänu treeningmaterjalis sisaldunud spontaanse kõne suuremale variatiivsusele võrreldes etteloetud kõnega BABELi korpuses.

Mõlema automaatse süsteemi süstemaatiline viga sõnapiiride hilisemaks määramisel võib olla tingitud erinevusest automaatselt leitud erineva spektraalse koostisega kõnesegmentide optimaalsete piiritingimuste ja käsitsi segmenteerimise reeglite vahel. Kuna lause lõpus paiknevale sõnale järgneb tüüpiliselt helitu hõngatus, mille spekter kannab eelneva hääliku spektri jälge, siis klassifitseerib algoritm ka selle eelnevasse kõnesegmenti kuuluvaks ja nihutab seega sõnapiiri hilisemaks; käsitsi segmenteerimisel aga pannakse piir kokkuleppeliselt hõngatuse ette.

Häälikuühenditest hälbisid automaatsed piirid enim diftongiosiste piiri määramisel. Tõenäoliselt on ka siin põhjuseks akustiliste mudelite abil leitud optimaalsete piiritingimuste ja käsitsi segmenteerimisel kasutatavate kriteeriumide erinevus.

Vaatamata automaatselt leitud segmendipiiride hälvetele käsitsi määratud piiridest, on automaatsegmenditud materjalist mõõdetud segmendikestused lähedased käsitsi segmenditud kõnest leitud kestustele (vokaalide kestuserinevused ei ole statistiliselt olulised, konsonantide puhul kohati on). Mõlema süsteemi segmentimise täpsust võime hinnata piisavaks kõnetehnoloogiliste rakenduste seisukohalt ja osalt ka eesti emakeelega täiskasvanute kõne akustiliseks analüüsiks, kuid L2 kõne segmendikestuste mõõtmiseks (nagu nt Meister *et al*. 2015; Meister & Meister 2014, 2013, 2012b) on siiski usaldusväärsem kasutada käsitsi märgendatud kõnet.

Mõlemad automaatsed süsteemid vajavad edasiarendamist saavutamaks eestikeelse kõne segmenteerimisel lähedasi tulemusi parimate inglise keele jaoks loodud segmentimisprogrammidega. Selleks tuleb eelkõige suurendada akustiliste mudelite treeningmaterjali mahtu ja lisada sellesse erinevaid kõnevariatsioone (nt lastekõne, L2 kõne, spontaanne kõne). Alles seejärel võime kiirelt ja vähese tööjõukuluga produtseerida piisavalt täpseid autosegmenditud märgendusi ning saada sama usaldusväärseid akustilise analüüsi tulemusi kui käsitsi segmenteeritud kõnematerjali kasutades. Usaldusväärseid mõõteandmeid ei vaja mitte ainult foneetikud, vaid ka fonoloogid erinevate hüpoteeside ja teoreetiliste mudelite arendamiseks ning testimiseks. Sest mis väärtus oleks fonoloogilisel mudelil, mis ei ole kooskõlas akustiliste andmetega.

Tänusõnad

Töö valmimist on toetanud Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus) ja riikliku programmi "Eesti keeletehnoloogia 2011–2017" projekt EKT70 "Kõnekorpuste arendus".

Kommentaarid

[1] www.keel.ut.ee/et/foneetikakorpus/

[2] http://www.keel.ut.ee/sites/default/files/www_ut/ekskfk_margendamise_juhend_2-0.pdf

[3] http://www.helsinki.fi/~lennes/annotation_guide/annotation_guide.pdf

[4] https://phon.ioc.ee/dokuwiki/doku.php?id=projects:tuvastus:est-align.et

[5] https://clarin.phonetik.uni-muenchen.de/BASWebServices/

[6] http://www.phon.ucl.ac.uk/home/sampa/

[7] http://www.internationalphoneticalphabet.org/

Kõik lingid kontrollitud 31. oktoobril 2017.

Kirjandus

Boersma, Paul & Weenink, David 2017. *Praat: doing phonetics by computer* (http://www.praat.org – 20. oktoober 2017).

Chomsky, Noam 1956. Three models for the description of language. *IEEE Transactions on Information Theory* 2 (3), lk 113–124 (doi: 10.1109/TIT.1956.1056813).

Chomsky, Noam 1957. *Syntactic Structures*. The Hague/Paris: Mouton.

Hajič, Jan & Hajičová Eva 2007. Some of Our Best Friends Are Statisticians. Matoušek, Václav & Mautner, Pavel (toim). *Text, Speech and Dialogue*. 10th International Conference TSD 2007, Proceedings. Berlin: Springer-Verlag, lk 2–10.

Hint, Mati 2016. Mõõtmised ei loo teooriat. *Keel ja Kirjandus* 8–9, lk 627–637 (http://kjk.eki.ee/ee/issues/2016/8-9/825 – 20. oktoober 2017).

Hosom, John-Paul 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication* 51 (4), lk 352–368 (doi: 10.1016/j.specom.2008.11.003).

Jelinek, Frederick 2005. Some of my Best Friends are Linguists. *Language Resources and Evaluation* 39, lk 25–34 (doi: 10.1007/s10579-005-2693-4).

Kisler, Thomas & Reichel, Uwe & Schiel, Florian & Draxler, Christoph & Jackl, Bernhard & Pörner, Nina 2016. BAS Speech Science Web Services – an Update of Current Developments. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), lk 3880–3885.

Koit, Mare 2006. Ratsionalism ja empirism keeletöötluses: vastasseis või koostöö? Tragel, Ilona & Õim, Haldur (toim). *Teoreetiline keeleteadus Eestis* II. Tartu: Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 7, lk 41–54.

Ljolje, Andrej & Hirschberg, Julia & van Santen, Jan P. H. 1997. Automatic Speech Segmentation for Concatenative Inventory Selection. van Santen, Jan P. H. & Sproat, Richard W. & Olive, Joseph P. & Hirschberg, Julia (toim). *Progress in Speech Synthesis*. New York: Springer-Verlag, lk 305–311.

Meister, Lya & Meister, Einar 2012a. Aktsendikorpus ja võõrkeele aktsendi uurimine. *Keel ja Kirjandus* 55, lk 696–714.

Meister, Lya & Meister, Einar 2012b. The production and perception of Estonian quantity degrees by native and non-native speakers. *Interspeech 2012: 13th Annual Conference of the International Speech Communication Association, September 9-13, 2012, Portland, Oregon, Proceedings*, lk 886–889.

Meister, Einar & Meister, Lya 2013. Production of Estonian quantity contrasts by native speakers of Finnish. *Interspeech 2013: 14th Annual Conference of the International Speech Communication Association, 25-29 August 2013, Lyon, France, Proceedings*, lk 330–334.

Meister, Einar & Meister, Lya 2014. Estonian quantity degrees produced by Latvian subjects. *Linguistica Lettica* 22, lk 85–106.

Meister, Einar & Meister, Lya 2017. Eesti laste kõne I. Põhitooni akustiline analüüs. *Keel ja Kirjandus* 7, lk 518–533.

Meister, Einar & Nemoto, Rena & Meister, Lya 2015. Production of Estonian quantity contrasts by Japanese speakers. *Eesti ja Soome-ugri Keeleteaduse Ajakiri / Journal of Estonian and Finno-Ugric Linguistics* 6 (3), lk 79–96 (doi: 10.12697/jeful.2015.6.3.03).

Pitt, Mark A. & Johnson, Keith & Hume, Elizabeth & Kiesling, Scott 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication* 45, lk 89–95 (doi: 10.1016/j.specom.2004.09.001).

Roach, Peter & Arnfield, Simon & Barry, William & Baltova, Julia & Boldea, Marian & Fourcin, Adrian & Gonet, Wiktor & Gubrynowicz, Ryszard & Hallum, Elisabeth & Lamel, Lori & Marasek, Krzysztof & Marchal, Alain & Meister, Einar & Vicsi, Klára 1996. BABEL: An Eastern European multi-language database. *Spoken Language.* Proceedings of ICSLP '96 – Fourth International Conference on Spoken Language Processing 3, lk 1892–1893 (doi: 10.1109/ICSLP.1996.608002).

Rendel, Asaf & Sorin, Alexander & Hoory, Ron & Breen, Andrew 2012. Towards automatic phonetic segmentation for TTS. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, lk 4533–4536 (doi: 10.1109/ICASSP.2012.6288926).

RStudio Team 2015. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA (http://www.rstudio.com/ – 23. oktoober 2017).

Schiel, Florian & Draxler, Christoph & Baumann, Angela & Ellbogen, Tania & Steffen, Alexander 2004. The Production of Speech Corpora (https://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/ – 23. oktoober 2017).

Toledano, Doroteo Torre & Gomez, Luis A. Hernández & Grande, Luis Villarrubia 2003. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing* 11 (6), lk 617–625 (doi: 10.1109/TSA.2003.813579).

Wesenick, Maria-Barbara & Kipp, Andreas 1996. Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals. *Proceedings of ICSLP '96 – Fourth International Conference on Spoken Language Processing* 1, lk 129–132 (doi: 10.1109/ICSLP.1996.607054).

Summary

Evaluation of automatic speech segmentation

Einar Meister
senior researcher, Tallinn University of Technology
einar.meister@ttu.ee

Lya Meister
researcher, Tallinn University of Technology
lya.meister@ttu.ee

Keywords: automatic segmentation, Estonian, phone boundaries, segment durations, speech corpora, word boundaries

The use of large speech corpora in phonetic research depends to a great extent on the availability and quality of phonetic segmentation and transcriptions. As a rule, the best quality of segmentation is achieved by human transcribers who perform time-consuming and tedious manual work. However, tools for automatic segmentation exploiting typically HMM-based forced alignment methods have been developed for different languages. In recent years, two automatic systems as free online services have become available for Estonian: (1) the system developed at Tallinn University of Technology (https://phon.ioc.ee/dokuwiki/doku.php?id=projects:tuvastus:est-align.et), and (2) the multi-lingual tool WebMAUS (https://clarin.phonetik.uni-muenchen.de/BASWebServices/).

In this study we evaluate the performance of the two systems against human transcribers. The test set includes Estonian read speech produced by: (1) four L1 adult subjects, (2) six L1 adolescents, and (3) four L2 adult subjects. The reference segmentation data including 27 sentences from L1 subjects and 10 sentences from the other subjects were produced manually as Praat textgrid files with two tiers (word-level orthographic and phoneme-level SAMPA transcription); the automatic systems have produced similar textgrid files. In total, 1179 word boundaries and 5050 phone boundaries were compared.

The results show that both systems performed more accurately for L1 adult speech and were less accurate in the case of adolescent and L2 speech. While the TUT system outperformed WebMAUS in L1 adult speech, then in L1 adolescents and L2 speech WebMAUS produced more accurate results. Despite the deviations in phone boundaries, the durations of vowel and consonant segments measured from automatic and manual segmentations of L1 adult speech differ only marginally. This suggest that the accuracy of both automatic systems seems to be sufficient for speech technology needs and could also be used in acoustic studies of L1 adult speech. However, both systems need improvements in order to reach the accuracy of automatic segmentation tools available for English.