DOKTORIVÄITEKIRI EMOTSIOONIDEST EESTIKEELSES KÕNES

Kairi Tamuri. Basic emotions in read Estonian speech: acoustic analysis and modelling. (Dissertationes philologiae Estonicae Universitatis Tartuensis 39.) Tartu: University of Tartu Press, 2017. 238 lk.

6. oktoobril 2017 kaitses Kairi Tamuri Tartu Ülikooli eesti ja üldkeeleteaduse instituudis edukalt oma doktoriväitekirja „Basic emotions in read Estonian speech: acoustic analysis and modelling" („Põhiemotsioonid eestikeelses etteloetud kõnes: akustiline analüüs ja modelleerimine"). Väitekirja sisuks on emotsioonide akustika tekst–kõne situatsioonis, täpsemalt emotsioonide akustilise moodustamise ja taju analüüs ning selle tulemuste praktiline rakendamine eestikeelse kõne sünteesis. Tegemist on kuuest publikatsioonist [P1–P6] koosneva nn artikliväitekirjaga, nendest neli on avaldatud inglise ning kaks eesti keeles. Publikatsioonidele eelnev üle-

vaade on esitatud esmalt inglise keeles ning seejärel ligikaudu samas mahus ka eesti keeles. Kõik see lubab väitekirja esimeseks keeleks pidada inglise keelt. Kuuest artiklist kaks on avaldatud kaasautorluses Meelis Mihklaga. Dissertandi roll nende artiklite valmimisel on tekstis selgelt piiritletud.

Uuritud on järgmiste akustiliste parameetrite seost emotsioonidega: pausid (arv, asukoht, iseloom, st seos hingamisega, ning kestus), esimese ja teise formandi sagedus ning vokaalide redutseerumine, kõnetempo, häälikute kestus ning esimese ja teise silbi vokaalide suhe, intensiivsuse tase ja selle muutumispiirkond, põhitooni sagedus ja selle muutumispiirkond.

Publikatsioonis [P1] näidati, et pausid üksi ei ole eestikeelses kõnes emotsioone eristavaks tunnuseks. Siiski võimaldavad artiklis kirjeldatud tulemused anda soovitusi pauside kasutamiseks tekst–kõne-sünteesis.[1]

[P2]-s kirjeldatud tulemused osutavad, et viha, rõõm ja kurbus avaldavad mõju esimese formandi sagedusele vokaalidest [a], [i] ja [u] kahe esimese puhul. Artikulatsiooni täpsus langeb teatud määral kurbuse-lausungites.[2]

[P3]-s oli vaatluse all emotsioonide mõju kõne temporaalsele struktuurile, täpsemalt kõnetempole ja sõnaprosoodiale. Rõõm ja viha muudavad kõnetempo kiiremaks, kurbus aeglasemaks. Emotsioonid avaldavad mõju kõnetakti struktuurile teise- ja kolmandavältelistes vokaalipõhise kestuserinevusega sõnades ning kurbus kaotab nendes sõnades temporaalsed erinevused teise ja kolmanda välte vahel.[3]

[P4]-s on uuritud emotsioonide mõju intensiivsuse tasemele ning järeldatud, et viimane on kõrgeim neutraalsetes lausungites, madalaim aga kurbades. Uuritud parameetri muutumispiirkonnas emotsiooniti statistiliselt olulisi erinevusi ei leidunud. Lause alguses ja lõpus oli intensiivsuse tase kõrgeim neutraalsetes lausungites, madalaim aga kurbades.[4]

[P5]-s kirjeldatakse emotsioonide mõju põhitooni kõrgusele, muutumispiirkonnale ning asukohale lausete alguses ja lõpus. Põhitoon on kõrgeim rõõmsates ning madalaim vihastes lausetes. Põhitooni muutumispiirkond on suurim vihastes, väiksem aga kurbades lausetes. Põhitooni kõrguse erinevused lause alguses ja lõpus polnud emotsioonipaariti ja neutraalse kõnega võrreldes oluliselt erinevad.[5]

[P6]-s kirjeldatakse emotsioonide mõju uurimise tulemusi kõnes rakendatuna parameetrilisele eestikeelsele kõnesünteesile. Tajukatsetes testiti kolme erinevat mudelit. Uuritud emotsioonidest tuvastati sünteeskõnes kõige paremini kurbust ning kõige halvemini rõõmu. Testimise käigus selgus, et mees- ja naishäälte jaoks sobisid kõnesünteesiks kõige paremini erinevad mudelid.[6]

Kairi Tamuri väitekirja konteksti asetamiseks on vaja esile tuua kaht asjaolu. Esiteks, väitekiri tugineb Eesti Keele Instituudis loodud eesti keele emotsionaalse kõne korpusele, millele on alus pandud varem, kui Tamuri oma väitekirjaga tööd alustas. See tähendab, et väitekirja kohta küsimuste kerkimisel tuleb eristada tegureid, mida autor

[1] K. Tamuri, Kas pausid kannavad emotsiooni? – Eesti Rakenduslingvistika Ühingu aastaraamat 2010, nr 6, lk 297–306.

[2] K. Tamuri, Kas formandid peegeldavad emotsioone? – Eesti Rakenduslingvistika Ühingu aastaraamat 2012, nr 8, lk 231–243.

[3] K. Tamuri, M. Mihkla, Emotions and speech temporal structure. – Linguistica Uralica 2012, nr 3, lk 209–217.

[4] K. Tamuri, Intensity of Estonian emotional speech. – Human Language Technologies: The Baltic Perspective. Proceedings of the Fifth International Conference Baltic HLT 2012. IOS Press, 2012, lk 238–246.

[5] K. Tamuri, Fundamental frequency in Estonian emotional read-out speech. – ESUKA / JEFUL 2015, kd 6, nr 1, lk 9–21.

[6] K. Tamuri, M. Mihkla, Expression of basic emotions in Estonian parametric text-to-speech synthesis. – ESUKA / JEFUL 2015, kd 6, nr 3, lk 145–168.

on saanud oma valikutega mõjustada, ja teisi tegureid, mis tulenevad kõne all oleva korpuse parameetritest ning on seega määratletud autorist sõltumatult. Psühholoogias on tavaks nn põhiemotsioonidena vaadelda nelja (rõõm, kurbus, viha ja hirm) või kuut (lisanduvad üllatus ja vastikus). Tamuri väitekirja aluseks on emotsionaalse kõne korpus, milles eristatakse kolme emotsiooni: rõõmu, kurbust ja viha. See valik on napp, kuid autori tahtest sõltumatu.

Teiseks, väitekirjas põimuvad teineteisega tihedalt teoreetiline ja rakenduslik aspekt. Siit tuleneb, et kui ka mõned töös saadud tulemused ei osutu teoreetiliselt piisavalt kandvateks, võib nende rakendamisest ühel või teisel moel siiski kasu olla eestikeelse tekst–kõne-sünteesi kvaliteedi parandamiseks.

Järgnevalt toon välja publikatsioonide kaupa väitekirja lugemisel tekkinud mõned küsimused.

Publikatsioonis [P1] kirjeldatakse 4. alajaotuses kuulamistesti tulemusi. Sealjuures öeldakse katseisikute kohta vaid seda, et neid oli kümme. Samas on teada, et indiviidid võivad oma vanuse jt omaduste tõttu anda sellistes testides lahknevaid vastuseid. Nt Ryan, Murray ja Ruffman on näidanud, kuidas vanemad katseisikud kogevad noorematega võrreldes raskusi põhiemotsioonide, nagu rõõm, kurbus, viha ja hirm, tuvastamisel.[7] Kas oleks võimalik iseloomustada, kui homogeense grupi moodustasid katseisikud artiklis [P1] kirjeldatavas kuulamistestis? (Võrdluseks: [P6] kasutatud katseisikud olid vanuses 30 kuni 73 aastat.)

[P2] lk 232 leidub kaks lauset, mille vahel on teatud vastuolu: heliliste häälikute „foneetilist kvaliteeti tajutakse harilikult esimesest kolmest formandist" ning „[p]õhjus, miks kuuldakse foneetilisi erinevusi kahe vokaali vahel, peitub selles, et vokaalidel on erinev asukoht kahemõõtmelises vokaaliruumis". Mitut formanti siis ikkagi läheb tarvis vokaalidevaheliste erinevuste kirjeldamisel?

Mulle tundub, et artikulatsiooni täpsuse hindamisel emotsiooniti saanuks toimida lihtsamalt, ilma neutraalse hääliku asukohta formantruumis välja arvutamata. Joonisel 1 kujutatud formandikolmnurki võinuks konstrueerida ka tabelis 2 esitatud väärtuste põhjal.

[P3] loeme (lk 209), et eesti keel on sõnakeskne ning välted avalduvad sõna prosoodias („Estonian is a word-central language. It is in word prosody that quantity degrees are manifested"). Kas pole siiski õigem väita, et eesti keeles on välte domeeniks kõnetakt, mitte terve sõna?

[P3] tabelis 4 on esitatud kõneüksuste kestuste ja vokaalide V1:V2 suhte väärtused erinevates väldetes ning erinevate emotsioonide ja neutraalse kõne korral. Seejuures pole mõõdetud konsonandi C1 kestust CV[:]CV struktuuriga sõnades. Olen nõus, et C1 temporaalsetes opositsioonides rolli ei mängi, ent tahtnuks seda väitekirjast selgelt lugeda. Teine küsimus samas puudutab teise- ja kolmandavälteliste sõnade tähistust. Kui teisevältelisi sõnu tähistada CV:CV, siis kas kolmandavältelisi sõnu ei peaks tähistama CV::CV?

[P3] lk 210 loeme, et välted on määratavad ka järjestikuste häälikute kestussuhete põhjal ning lisatud viide Eek, Meister 2003 („[---] it has also been suggested that quantity degrees could be determined from the durational ratios of adjacent speech sounds (Eek, Meister 2003)"). Siia sobinuks täiendav viide Traunmüllerile ja Krullile, kes on oma töös arendanud samalaadseid mõtteid.[8]

[7] M. Ryan, J. Murray, T. Ruffman, Aging and the perception of emotion: Processing vocal expressions alone and with faces. – Experimental Aging Research 2009, kd 36, nr 1, lk 1–22.

[8] A. Eek, E. Meister, Foneetilisi katseid ja arutlusi kvantiteedi alalt. – Keel ja Kirjandus 2003, nr 11, lk 815–837; nr 12, lk 904–918; H. Traunmüller, D. Krull, The effect of local speaking rate on the perception of quantity in Estonian. – Phonetica 2003, kd 60, nr 3, lk 187–207.

[P3] lk 215 loeme, et kurbuse väljendamisel oli rõhulise ja rõhuta silbi kestussuhe V1:V2 teises (Q2) ja kolmandas (Q3) vältes peaaegu sarnane, seega kolmikvastandus on asendunud duaalsega („[---] in sad speech the duration ratio V1:V2 of the stressed and unstressed syllables was almost similar for the second and third quantity degrees, due to which the three-way opposition gave way to a dual one"). See on üsna julge üldistus. Kas pole võimalik, et Q2 ja Q3 vahelised erinevused on kõnes väljendatud mõne teise parameetri, näiteks F0 kontuuri kaudu?

Peamine küsimus [P4] kohta puudutab meetodit. Lk 240 loeme, et korpuslaused liigitati lisaks selle järgi, kas emotsiooni äratundmist sisu ei mõjutanud (lugemistesti tulemused erinevad kuulamistestist) või oli mõju võimalik (lugemistesti ja kuulamistesti tulemused langevad kokku). Uurimistöö materjaliks valiti esimest tüüpi laused, kus lugemis- ja kuulamistesti tulemused olid erinevad. Küsimus seisneb selles, kuidas neid erinevusi tõlgendada. Kahtlen, kas kahes nimetatud testis saadud erinevate tulemuste põhjal saab järeldada, et lause sisu ei avaldanud emotsiooni tuvastamisele mõju. Võib-olla järeldub tulemuste erinevusest hoopis see, et ettelugeja kuulamistestis on lause emotsionaalset sisu tajunud teistmoodi kui lugeja lugemistestis. Vrd ka eelmises lõigus: ettelugeja emotsiooni määrab passuse semantiline sisu („[---] it is the semantic content of the passage that elicits the reader's emotion").

Pisut ettevaatlikuks teeb eraldi salvestatud lausete helirõhu taseme võrdlemine üksteisega [P4]. Olgugi et emotsionaalse kõne korpuse loomisel oli suu ja mikrofoni vaheline kaugus väidetavalt alati sama (50 cm), polnud pea asend mikrofoni suhtes nähtavasti fikseeritud, mis tähendab, et see kaugus võis tegelikkuses mõnevõrra muutuda. Autor möönab seda probleemi ka ise, kui ta kirjutab, et intensiivsus on tundlik lindistamistingimuste suhtes, mistõttu on oluline arvestada mikrofoni kaugust lugejast, kas lindistamisruum on piisavalt vaikne ja taustamüra puudub, kas salvestusseade on kalibreeritud jne (lk 241).

[P4] lk 239 altpoolt 3. lõigus loeme, et millegi eriti ebameeldiva tajumine põhjustab tihti neelu ja kõri pingulolekut, selle tagajärjeks on kõnetrakti pinge ja kõnesignaali kõrgenenud helikõrgus („[---] a sensation of something extremely unpleasant often causes a tightness felt in the pharynx and larynx, resulting in tension in the vocal tract and a higher pitch of the outcoming"). Põhitooni kõrgus oleneb häälekurdude võnkumise sagedusest. Kas pinge kõnetraktis saab mõju avaldada põhitooni kõrgusele?

[P4] lk 240 võib lugeda: kuna emotsioon ei ole näideldud, vaid esile kutsutud, siis selle väljendus kõnes on tagasihoidlik, sageli vaevu tajutav; samuti pole meil täismahulisi emotsioone, vaid pigem emotsiooniga seotud seisundid („As the emotion is not acted but *elicited* the expression in speech is moderate, often hardly perceptible; neither have we got full-blown emotions, but rather *emotion-related states*"). Ei ole päris selge, kuidas selle väiteni on jõutud. Osutatud lause ei sisalda viidet ning sellele ei eelne/järgne ka väitekirja enese materjalile tuginevat ratsionaalset arutlust.

Mõnikord käiakse terminitega väitekirjas ringi üsna vabalt. Nt [P4] lk 242 kasutatakse sünonüümidena sõnu *mean* ('keskväärtus') ja *median* ('mediaan'). Merriam-Websteri sõnaraamatust loeme: „To find one type of average, called the mean, you'd simply add up the total value of money and property of everyone in the group and divide it by the number of people. To find the other type, called the median, you'd identify the net worth of the person who is richer than half the people and poorer than the other half."[9] Vea parandamiseks piisab, kui sõna *mean* asendada sõnaga *average* ('kesk-

[9] https://www.merriam-webster.com/dictionary/median (13. II 2018).

mine'). See kategooria jaguneb omakorda kaheks: *mean* ja *median*.

[P4] järelduste osas lk 245 on paar viga. Kokkuvõtte eelviimasest lõigust loeme: „The differences of intensity for emotion pairs as well as in comparison with neutral speech were not statistically significant" ('Erinevused emotsioonipaaride intensiivsuses, ka võrdluses neutraalse kõnega, ei osutunud statistiliselt oluliseks'). See on vastuolus lausega eelmisest lõigust: „The differences between the mean intensities were also significant statistically" ('Erinevused intensiivsuste keskväärtuste vahel on statistiliselt olulised'). Ilmselt tuleks esimeses lauses „differences of intensity" ('erinevused intensiivsuses') asendada väljendiga „differences of intensity ranges" ('erinevused intensiivsuse ulatustes'). Viimane lause samas: statistiliselt olulised saavad olla ikka ainult intensiivsuste vahelised erinevused, mitte intensiivsused ise.

[P6]-s jääb mõneti ebaselgeks nn referentsmudelite teke. Lk 150 loeme: esmalt koostasime iga emotsiooni jaoks parameetrilise akustilise katsemudeli, mida valideeris väike ekspertide grupp (isikud, kes osalesid kõnesünteesi arendamises ja emotsionaalse kõne uurimuses). Nad hindasid parameetrite valikut ning soovitasid ka mõne muutuse parameetrites. Ekspertide keskmise hinnangu ja nende parameetrite valiku kohta tehtud soovituste põhjal konstrueerisime emotsioonide nn etalonmudelid („First, we constructed a parametric acoustic test model for each emotion. These models were then validated by a small group of experts (persons participated in the development of speech synthesis and the study of emotional speech). They evaluated the choice of the parameters and also suggested some changes in the parameters. On the basis of the experts' mean evaluation and their suggestions concerning the choice of parameters, we constructed the so-called reference models of the emotions"). See seletus on üsna napp. Mida täpsemalt eksperdid hindasid? Mis laadi muutusi parameetrites nad soovitasid?

[P6] on mul probleeme sõnadest *to lower* ('alandama') ja *to raise* ('tõstma') arusaamisega. Loeme (lk 151): „Model 1 (M1) is a model with decreased values, where the values of M2 have been lowered by approximately 15% in the direction characterizing the emotion (towards neutrality). Model 3 (M3) is a model with increased values, in which the values of M2 have been raised by 15% in the emotion's characteristic direction (away from neutrality)" ('Mudel 1 (M1) on vähendatud väärtustega mudel, kus M2 väärtused on u 15% alandatud emotsioonile iseloomulikus suunas (neutraalsuse poole). Mudel 3 (M3) on suurendatud väärtustega mudel, kus M2 väärtused on u 15% tõstetud emotsioonile iseloomulikus suunas (neutraalsusest eemale)'). Kui vaadata joonist 1, siis rõõmu (J) ja viha (A) puhul on tõepoolest kõnetempo väärtused J1 < J2 < J3 ja A1 < A2 < A3. Ent kurbuse (S) puhul on kõnetempo väärtused S1 > S2 > S3, seega M1 tõstab ja M3 kahandab nimetatud väärtusi.

Kokkuvõttes võib öelda, et eespool välja toodud pisematele puudustele vaatamata täidab Kairi Tamuri doktoriväitekiri „Basic emotions in read Estonian speech: acoustic analysis and modelling" praeguses vormis väitekirjale esitatavaid nõudeid (või isegi ületab neid). Töö osad moodustavad väitekirjale esitatavate nõuete seisukohast küllaldase, piisavalt ühtse ja mahuka terviku.

JAAN ROSS