

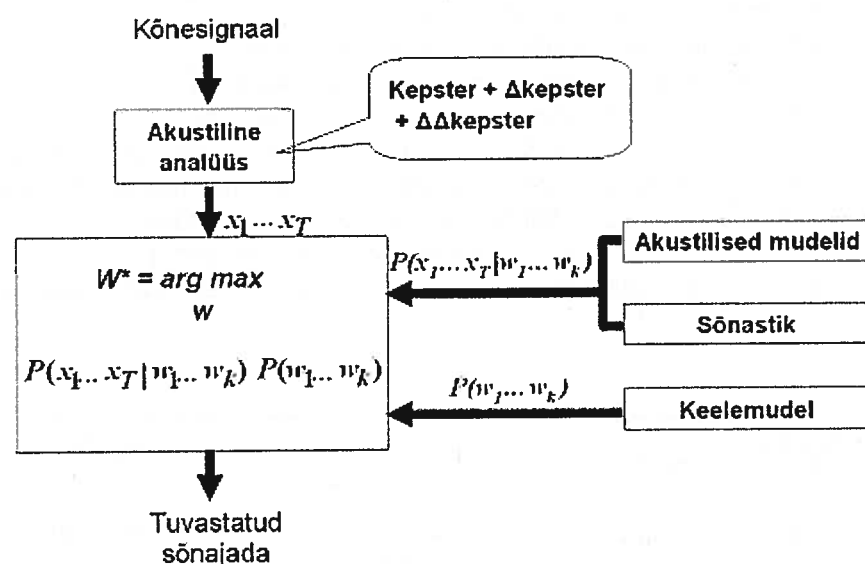


## STATE-OF-ART KÕNETUVASTUSES

Kõnetuvastuse ülesandeks on leida akustilisele signaalile vastav sõnajada. Inimesele on see ülesanne üldjuhul lihtne, samas on selle realiseerimine arvutis üsnagi keeruline. Kõne on oma olemuselt pidev, mitte diskreetsetest üksustest koosnev akustiline tekst. Ülesande teeb eriliselt raskeks kõnesignaalil esinev suur variatiivsus – sama sõna identseid hääldusi praktiliselt ei eksisteeri ja iga hääldus realiseerub erineva akustilise muustrina.

Variatiivsuse põhjusteks on:

- kõnelejate vanus ja sugu
- kõnestiil (vestlus sõbraga vs vestlus võõraga; avalik esinemine vs teksti dikteerimine; jne)
- keeletaust (emakeel vs võõrkeel),
- emotsionaalne ja tervislik seisund
- jpm.



Joonis 2. Kõnetuvastussüsteemi struktuur [3]

Joonisel 2 esitatud skeemilt näeme, et kõnetuvastus koosneb kahest põhiastmest – (1) akustilisest analüüsist ja (2) muustrituvastusest.

**Akustilise analüüsi** eesmärgiks on leida sisendsignaalist tuvastamiseks olulist informatsiooni sisaldavad tunnused ja maha suruda ebaolulised variatsioonid. Kõige sagedamini kasutatakse kõnetuvastuses mel-kepstri kordajaid (ingl k. mel-frequency cepstral coefficient – MFCC), mis saadakse inimkõrvale sarnase mitterilineaarse signaalitöötlemise tulemusena [4]. Sisendsignaal esitatakse kindla intervalli (tavaliselt 25 ms) järel arvutatavate tunnusvektorite jadana:

$X = x_1 x_2 \dots x_T$ .

**Mustrituvastuse** ülesandeks on leida sõnajada  $W^* = w_1 w_2 \dots w_n$ , mis kõige tõenäolisemalt vastab sisendsignaali  $X$ :

$$W^* = \underset{w}{\operatorname{argmax}} P(W|X) = \underset{w}{\operatorname{argmax}} \frac{P(W)P(X|W)}{P(X)}$$

Loobudes komponendist  $P(X)$  (sest meid huvitab sõnajada, mille tinglik tõenäosus on kõige suurem, mitte tõenäosuse  $P(W|X)$  täpne väärtus), saame:

$$W^* = \underset{w}{\operatorname{argmax}} P(W)P(X|W)$$

Seega, kõige tõenäolisem sõnajada  $W^*$  sõltub:

- sõnajada *a priori* tõenäosusest  $P(W)$  – see leitakse keelemudelist
- tõenäosusest  $P(X|W)$ , mis leitakse akustiliste mudelite põhjal.

Nii keelemudel kui ka akustilised mudelid on realiseeritud Markovi peitmudelitena, mille treenimiseks vajatakse suuremahulisi korpusi – keelemudeli puhul tekstikorpusi, akustiliste mudelite puhul kõnekorpusi.

Markovi mudelite kasutamisest kõnetuvastuses on kirjutanud A&As Tanel Alumäe [5], teemaga põhjalikumaks tutvumiseks peaks lugeja pöörduma mõne erialase allika poole (näiteks [4, 6] jpt).

Valdav osa kommertsrakendusi ja arendatavaid prototüüpe maailmas kasutab eelkirjeldatud statistilise modelleerimise meetodit – see on tänane kõnetuvastuse *state-of-art*.

Ka eestikeelse kõnetuvastuse arenduses on järgitud maailmatrende ja kohandatud üldlevinud meetodeid eesti keele spetsiifikale. TTÜ Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris on loodud selleks vajalikud kõneressursid [7, 8] ning infrastruktuur. Labori teaduri ja TTÜ doktorandi Tanel Alumäe töö tulemusena on loodud mitmeid spetsiaalselt eestikeelse kõne tuvastuseks vajalikke mudeleid ning esimesed piiratud sõnavaraga prototüübid, uuringud jätkuvad piiramatult sõnavaraga kõnetuvastuse loomiseks [9, 10].

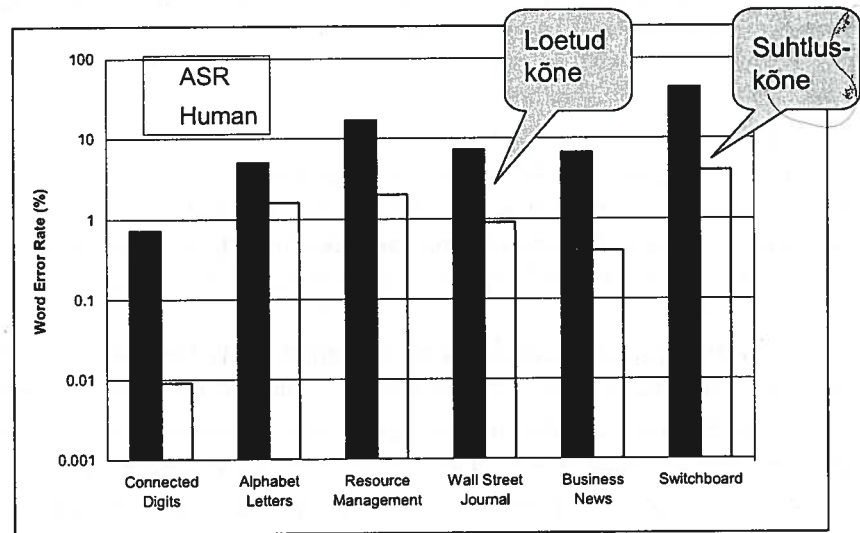
## VÖRDLEME INIMEST JA MASINAT

Kuigi uute tehnoloogiliste lahenduste, rakenduste ja tuvastatavate keelte hulk kasvab aasta-aastalt, on automaatne kõnetuvastus veel üsna kaugel inimese võimekusest. Joonisel 3 esitatud inimese ja automaatse kõnetuvastuse võrdlus [11] erinevat tüüpi kõne puhul näitab, et inimese kõnetuvastusvõime on kõigi kõnetüüpide puhul masinast parem. Eriti suur vahe on numbrijada tuvastuses, mille puhul masin teeb peaaegu 100 korda rohkem vigu kui inimene, ometigi on see ülesanne näiliselt nii lihtne – tuleb ära tunda vaid kümne sõna erinevad kombinatsioonid.

Samuti näeme, et nii inimene kui masin teevad spontaanse suhtluskõne (Switchboard korpus) tuvastamisel rohkem vigu kui loetud ajaleheteksti (Wall Street Journal korpus) tuvastamisel. Selle põhjuseks on asjaolu, et kõnetuvastussüsteemide akustilised mudelid on valdavalt treenitud laboratoorse kõne (etteantud tekstide lugemine müravabas akustilises keskkonnas) baasil ja keelemudelid on treenitud kirjalike tekstide alusel. Uuemad uuringud on näidanud, et kasutades akustiliste ja keelemudelite treenimiseks spontaanse kõne andmebaasi, on spontaanse kõne tuvastusvigade protsent umbes kaks korda väiksem, võrreldes laboratoorse kõne baasil treenitud mudelitega [12].

Kõige lähemal inimvõimetele on masin üksikult hääldatud tähestiku tähtede tuvastamisel. Selle tulemuse hindamisel on vajalik arvestada keelespetsiifikat – joonisel 3 esitatud tulemused on saadud inglise keele kohta ja näiteks eesti keele puhul võime foneetiliste iseärasuste tõttu saada oluliselt erinevad tulemused. Nii on eesti keeles ka inimesel (masinast rääkimata!) raske eristada

tähepaaride p – b ja t – d isoleeritud hääldust, sest sõna alguses need klusiidid foneetiliselt ei eristu (p ja b – mõlemad hääldatakse /pee/, t ja d hääldatakse /tee/).



Joonis 3. Inimese ja masina kõnetuvastuse võrdlus erinevat tüüpi kõne puhul [11]. Horisontaalteljel kõnetüübid, vertikaalteljel sõnatuvastuse viga protsentides; mustad tulbad on kõnetuvastussüsteemi tulemused, hallid tulbad inimese tulemused

Kas automaatse kõnetuvastuse kvaliteet saab kunagi võrreldavaks inimvõimetega? Selle ja veel palju muid kõnetehnoloogia arengut puudutavaid küsimusi esitas Roger Moore (mitte see, kes kunagi James Bondi mängis) 1997. ja 2003. aastal paljudele kõnetehnoloogia ekspertidele maailmas [13]. Küsitluse tulemused näitasid, et pessimistide osakaal on kahe küsitluse vahel oluliselt kasvanud (vt joonis 4).

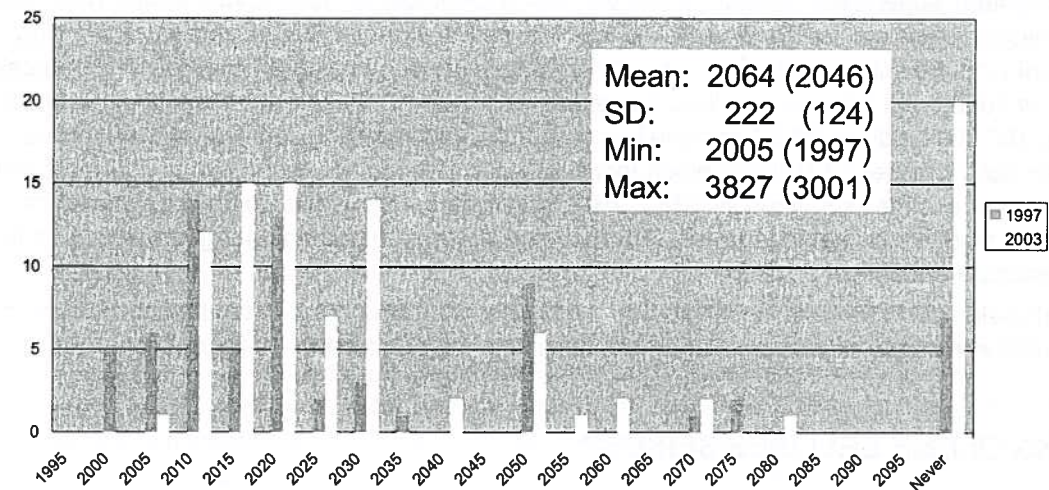
Analüüsidest vastuste jaotust joonisel 4, näeme, et 1997. aastal langeb märkimisväärne osa vastustest vahemikku 2010 – 2025 (üldkeskmine 2046), aastal 2003 aga vahemikku 2010 – 2035 (üldkeskmine 2064). Samas on 2003. aastal oluliselt suurem nende ekspertide hulk, kelle arvates ei saa masin kõnetuvastuses inimesega kunagi võrdseks.

## PROBLEEMID KÕNETUVASTUSES

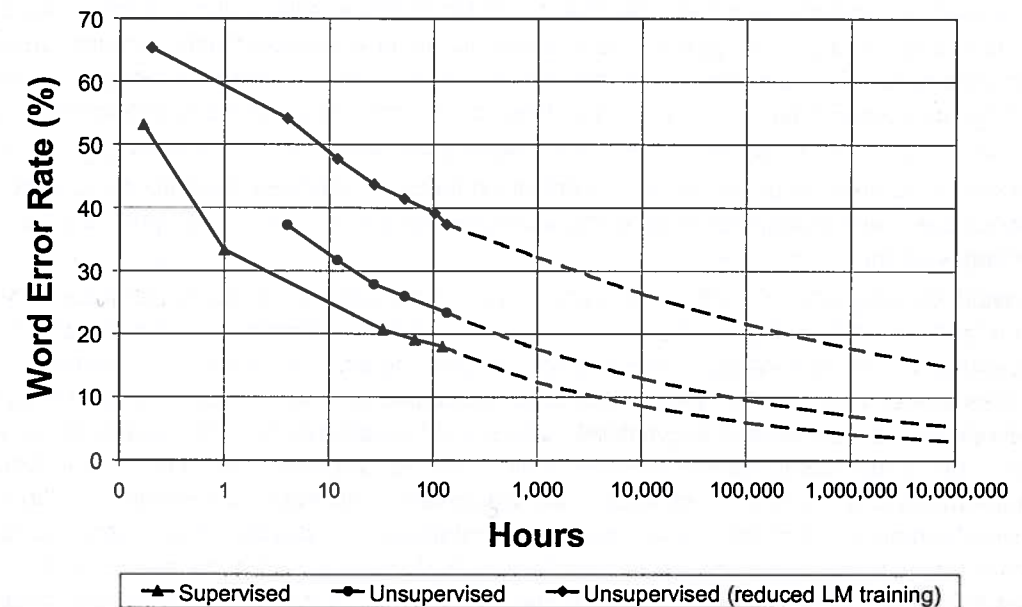
Miks siis ikkagi pärast 50 aastat progressi on suur hulk tippeksperthe nii pessimistlikul seisukohal? Aga seepärast, et progress kõnetuvastuses ei ole aset leidnud mitte tänu olulistele avastustele inimaju kõnetöötlusprotsesside olemusest (sellest teatakse endiselt väga vähe!), vaid eelkõige tänu arvutusvõimsuste kiirele kasvule, suurte kõneandmebaaside loomisele ja statistiliste meetodite laialdasele kasutamisele [14, 15].

Praegust, nn jõumeetodil toimuvat arengut iseloomustakse raskestitõlgitava ingliskeelse fraseologismiga *There's no data like more data!*, mida võiks lahti seletada järgmiselt: paremate tuvastustulemuste saamiseks vajame süsteemide treenimiseks üha suuremaid andmebaase. Kui palju treeningumaterjali on siis vaja, et jõuda inimesele lähedaste võimeteeni? Vastuse sellele küsimusele leiame R. Moore'i artiklist [14] (joonis 5).

12. Speech recognition accuracy equals that of the average (individual) human transcriber.



Joonis 4. Vastuste jaotus küsimusele "Mis aastal saab automaatse kõnetuvastuse kvaliteet võrdseks inimese omaga?" aastatel 1997 (heledad tulbad) ja 2003 (tumedad tulbad) [13]. Horisontaalteljel on aastaarvud, vertikaalteljel vastuste hulk. Keskväärtuse (Mean), standardhälbe (SD), minimaalse (Min) ja maksimaalse (Max) hinnangu väärtused 1997. aasta kohta on esitatud sulgudes



Joonis 5. Tuvastuskorrektuse sõltuvus treeninguks kasutatud kõnematerjali hulgest kolme erineva treenimismeetodi puhul [14]

Jooniselt näeme, kuidas spontaanse kõne tuvastusvigade protsent väheneb sõltuvalt treeningumaterjali hulgest. Reaalsed tulemused on saadud ca 100 tunni treeningumaterjali abil (pidev joon), katkendlik joon on saadud ekstrapoleerimisel. Valides parimaid tulemusi andva juhendatud (supervised) treeningu, on võimalik saavutada ca 12% tuvastusviga 1000 tunni treeningumaterjaliga, ca 8% viga 10 000 tunniga, ca 6% viga 100 000 tunniga ja ca 3% viga 10 miljoni tunni treeningumaterjaliga. Et paremini ette kujutada nimetatud kõnematerjali mahtusid, siis ca 1000 tundi kõnet on kuulnud 2-aastane laps ja ca 10 000 tundi on kuulnud 10-aastane laps; 100 000 tundi vastab 80-aastase inimese kuulnud kõnemahule ja 10 miljoni tundi kõnet on enam kui 70 inimese keskmise eluea jooksul kuuldu (toodud võrdlused esitas R. Moore oma suulises ettekandes konverentsil SPECOM'2005, viidates USA uurijate andmetele). On ilmne, et selliste gigantsete mahtudega kõnekorpusi koguda on ebareaalne ja "jõumeetodil" kõnetuvastuse arendamisel on piirid.

Võrdluseks: eestikeelse SpeechDat-tüüpi andmebaasi maht on ca 240 tundi kõnet salvestatud ca 1300 kõnelejal; salvestuste kogumine, kontroll ja märgendamine kestis umbes 2 aastat.

## MIKS OLEME SELLES SEISUS?

Tsiteerides veel kord R. Moore'i [14]: "Kõnetöötlus on universumis teadaoleva kõige keerukama elusorganismi kõige keerulisem talitus." (EM tõlge.) Teisisõnu, kõnetöötlemisest inimajus ja kõnekommunikatsiooni protsessidest teame tänasel päeval veel liialt vähe, et seda keeleteaduslike ning matemaatiliste meetodite abil edukalt modelleerida.

Maailmas on olemas suur hulk killustatud teadmisi paljudest kõnekommunikatsiooniga seotud valdkondadest ja palju insener-tehnilisi meetodeid ning keeleteaduslikke mudeleid, kuid üldist kõnetuvastuse teooriat kui sellist pole olemas! Eksisteerivad ka teatud vastuolud insenerliku ja keeleteadusliku lähenemise vahel, mis on omandanud isegi folkloorse väljenduse, näiteks: 'Kõnetuvastussüsteemi tuvastuskorrektsus on pöördvõrdeline selle väljatöötamises osalenud keeleteadlaste arvuga' või 'Iga kord, kui ma vallandasin ühe keeleteadlase, paranes süsteemi tuvastuskorrektsus' – see lause on omistatud IBMis kõnetuvastuse loomisega tegelenud uurimisgrupi juhile Frederick Jelinek'ile, kes hiljuti on püüdnud seda ümber lükata artiklis "Some of my best friends are linguists" (mõned minu parematest sõpradest on lingvistid).

Kõnekommunikatsiooni eri aspektide uurimisel on lähtunud põhiliselt laboratoorsest kõnest ja tekstipõhisest keelematerjalist ning nendel uurimistulemustel põhineb ka enamik tehnoloogias rakendatavaid mudeleid.

Erinevaid kõneaspekte on uuritud teineteisest (peaaegu) sõltumatult: kõnetuvastuse puhul on oluline eelkõige kõne lingvistiline sisu, s.o mida öeldi, ja kõnelejalast tingitud variatiivsust käsitletakse kui müra; kõnelejatuvastuse puhul on põhiline ekstralingvistiline informatsioon, s.o kes ütles ja see, mida öeldi, on sageli ebaoluline; dialoogide kirjeldamisel pööratakse tähelepanu kõnevoorude vahetumisele ja lingvistilisele sisule, kuid paralingvistiline informatsioon – kuidas ütles – on olnud teisejärguline. Mitmed uurimused on näidanud [16, 17], et suhtluskõne (*talk-in-interaction*) erineb oluliselt laboratoorsest kõnest, sisaldades hulgaliselt akustilis-foneetilisi tunnuseid, mis dialoogi kontekstis edastavad olulist paralingvistilist informatsiooni. Nende tunnuste – kõnerütm, tempo, kestus, valjus, põhitoon, häälekvaliteet – roll kõnesuhtluses on jäänud praktiliselt ilma tähelepanuta. See, et suhtluskõne akustilis-foneetilisi tunnuseid on väga vähe uuritud ja neid ei rakendata kõnetuvastussüsteemides, on osaliselt tingitud ka adekvaatse kõnematerjali kogumise raskustest.

## KUIDAS EDASI?

Üha rohkem uurijaid on mõistnud, et selline teadmusignorantne (ingl k *knowledge-ignorant*) tehnoloogiaarendus ei saa lõpmatult jätkuda, edasiminekuks saab toimuda ainult teadmusrikka (ingl k *knowledge-rich*) tehnoloogia arenduse teel. Sellise mõtteviisi üks aktiviste maailmas Chin-Hui Lee on välja pakkunud kõnetuvastuse arendamiseks järgmisi ideid [15]:

- häälikuspetsiifilised tunnused – lisaks kepsrile on vaja kasutada mitmeid akustilis-foneetilisi tunnuseid: kestus, valjus, põhitoon jm
- võtmesõnade tuvastus ja lause verifitseerimine – inimene ei pea kuulma kõiki sõnu, lausest arusaamiseks piisab võtmesõnade tuvastusest
- teadmuspõhised tunnused – saadakse neuronvõrkude abil, kasutatakse HMM treenimiseks
- inimese kõnetöötluse mudelid – inimene ei teisenda kõnesignaali sõnajadaks otse, vaid tuvastab signaalist akustilisi ja auditiivseid sündmusi, mille põhjal formuleeritakse kognitiivsed hüpoteesid, neid verifitseerides jõutakse konteksti sobiva tulemuseni.

Tõepoolest, inimestevahelises suhtluses on võrdset tähtsust nii akustiline ja lingvistiline kui ka para- ja ekstralingvistiline komponent ning need peaksid olema adekvaatselt modelleeritud ka inimese-masina suhtluse mudelis. Sellise integreeritud suhtluse mudeli loomine eeldab:

suhtlusolukorrale tüüpilise andmestiku kogumist ja mitmekülgset analüüsi

ühtse kõnekommunikatsiooniteooria väljaarendamist.

Inimese-inimese ja inimese-masina kommunikatsiooni uurimiseks vajaliku multimodaalse andmestiku kogumiseks on tarvis realiseerida intelligentse ruumi prototüüp, mis võimaldaks tulemuslikult modelleerida reaalseid suhtlusolukordi. Ühtse kõnekommunikatsiooniteooria loomine ja väljaarendamine on uue interdistsiplinaarse teadusvaldkonna – kognitiivse informaatika – üks olulisemaid väljakutseid [14, 18].

Kuid jäägu need mõisted – intelligentne ruum, kognitiivne informaatika – siinses kirjatükis avamata, see vajaks terve A&A numbrijagu trükiruumi; asjast huvitatud lugeja leiab vastavat infot ka Internetist otsides.

## KEELETEHNOLOOGIA SEIS EESTIS

Kuigi keeletehnoloogias kasutatavad matemaatilised meetodid on keelest sõltumatud, on paljud komponendid (näiteks akustilised mudelid, sõnastik ja keelemudel kõnetuvastuses) rangelt keelespetsiifilised ja nende puhul ei tule kõne alla mingi muu keele jaoks loodud mudelite kohandamine (mida mõned nn eksperdid on meile soovitanud). Seetõttu pole erilist vahet, kas luuakse keeletehnoloogiat eesti või suahiili keele jaoks – see on iga keele puhul ühtemoodi teadmiste- ja töömahukas ning kallid. See, et meil täna pole keeletehnoloogilisi rakendusi kuigi palju ja et meie arengutase on mõnevõrra tagasihoidlikum, võrreldes mitmete suuremate keelte tehnoloogiliste võimalustega (inglise, prantsuse, saksa, hispaania, jaapani, hiina jpt), on otseselt tingitud inim- ja finantsressursside vähesusest. Suure kõnelejakonnaga keelte puhul on tehnoloogiaarengu stimulaatoriks eelkõige turunõudlus, mis tagab ka keeletehnoloogilise uurimis- ja arendustöö kasumlikkuse. Väikeste keelte korral (alla 10 miljoni kõneleja) ei tasu arendustööks vajalikud investeeringud end niipea ja loodetav kasumimarginaal võib osutada olematuks. Seetõttu ei leia me täna firmat (ei Eestis ega välismaal), kes oleks huvitatud investeerimisest näiteks eestikeelse kõnetuvastuse arendusse.

On ilmne, et keeletehnoloogia arendus Eestis ei saa toimuda turumajanduse reeglite kohaselt. Eesti keele tehnoloogilise arengu tagamiseks on mõõdapäasmatu, et keeletehnoloogilist uurimis- ja arendustööd finantseeritaks riigieelarvest. Õnneks on keeletehnoloogia leidnud äramärkimist mitmes olulises riikliku tähtsusega dokumendis, näiteks Eesti teadus- ja arendustegevuse strateegia "Teadmistepõhine Eesti" võtmevaldkonna "Kasutajasõbralikud infotehnoloogiad ja infoühiskonna areng" osana ja ka "Eesti keele arendamise strateegia (2004–2010)" sisaldab vastavat peatükki. Viimasest lähtuvalt on koostatud riiklik programm "Eesti keele keeletehnoloogiline tugi (2006–2010)", mille vastuvõtmine annaks olulise tõuke (nii moraalse kui ka finantsilise) valdkonna kiiremaks arenguks.

Keeletehnoloogia on ka üks Euroopa Liidu prioriteete, mida toetatakse mitme programmi kaudu. Kuid vaatamata kõigi liikmesriikide keelte võrdse staatuse deklareerimisele, domineerivad EL asjaajamises 3–4 keelt. ELi teadus-arendustöö programmidest suunatakse igal aastal kümneid (kui mitte sadu) miljoneid eurosid keeletehnoloogia arendamiseks, kuid see läheb eelkõige nendesamade 3–4 majanduslikult domineeriva keele tehnoloogiliseks arendamiseks [19]. Kuidas ELi kultuuride ja keelte Paabelis oma keele ja kultuuri arengut tagada ning milline roll on selles keeletehnoloogial – ka sellega saab hea lugeja tutvuda artiklit [19] lugeses.

## LÕPETUSEKS

Kuigi loo pealkirjas lubatakse juttu teha ka kõnelevatest masinatest, ei mahtunud see valdkond käesoleva kirjatüki raamidesse. Kuid lugejat võib lohutada – olukord kõnesünteesi valdkonnas on üsna sarnane kõnetuvastusega. Nii on ka seal suurte kõnekorpusete kasutuselevõtuga saadud väga häid tulemusi ja kui näiteks ühelt poliitikut salvestada 5–10 tundi kõnet, siis võib teda virtuaalselt rääkima panna mida tahes. Muidugi oleks see hea võimalus poliitiliste intriigide tekitamiseks, eriti valimiseelsel perioodil, kuid tõenäoliselt lõpeksid need katsed kohtuliku menetlusega, kus asjatundlik ekspert pettuse suure tõenäosusega avastab.

Kuid jällegi on see progress eelkõige jõumeetodite tulemus ja kõneloome protsesside olemusest ei tea me täna veel kuigi palju. Seetõttu ei oska me kuigi hästi sünteesida ei emotsionaalset ega dialoogipartneriga kohanduvat kõnet.

## KIRJANDUS

1. Meister, E. *Kõnetehnoloogia olemusest*. //A&A (2002) 5, 20-33
2. Furui, S. *50 years of progress in speech and speaker recognition*. // SPECOM'2005 – 10th International Conference Speech and Computer, 17-19 October, 2005, Patras, Greece, Proceedings vol. 1: 1-7
3. Furui, S. *Toward Robust Speech Recognition*. //Second Baltic Conference on Human Language Technologies, Tallinn 2005. Tutorials Day, April 6, 2005: 1-41
4. Huang, X., Acero, A., Hon, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall. (2001)
5. Alumäe, T. *Varjatud Markovi mudelid*. //A&A (2002) 4, 27-36
6. Cole, R. A. et al (eds). *Survey of the State of the Art in Human Language Technology*. <http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>

7. Meister, E., Eek, A. *Estonian Phonetic Database*. EU Copernicus Programme, Project No. 1304 "BABEL – A Multi-Language Database". Tallinn 1999
8. Meister, E., Lasn, J., Meister, L. *SpeechDat-Like Estonian Database*. // Proceedings of 6th International Conference TSD 2003, Lecture Notes in Artificial Intelligence 2807, Springer, 2003: 412-417
9. Alumäe, T. *Large vocabulary continuous speech recognition for Estonian using morpheme classes*. // Proceedings of ICSLP 2004 – Interspeech, Jeju, Korea, 389-392
10. Alumäe, T. *Phonological and morphological modeling in large vocabulary continuous Estonian speech recognition system*. // Proceedings of the Second Baltic Conference on Human Language Technologies, April 4-5, 2005, Tallinn, Estonia, 89-94
11. Lippmann, R. *Speech Recognition by Machines and Humans*. //Speech Communication, vol. 22, pp 1-15, 1997
12. Furui, S. *Spontaneous speech recognition and summarization*. // Proceedings of the Second Baltic Conference on Human Language Technologies, April 4-5, 2005, Tallinn, Estonia, 39-50
13. Moore, R. K. *Speculating on the Future for Automatic Speech Recognition*. //IEEE workshop on Automatic Speech Recognition and Understanding (ASRU), St. Thomas, US Virgin Islands, (2003)
14. Moore, R. K. *Cognitive Informatics: The Future of Spoken Language Processing? // SPECOM'2005 – 10th International Conference Speech and Computer, 17-19 October, 2005, Patras, Greece, Proceedings vol. 1: 11-15*
15. Lee, C.-H. *From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition*. // ICSLP 2004– Interspeech, Jeju, Korea, 2004
16. Campbell, N. *Getting to the heart of the matter: Speech is more than just the expression of text of language*. Keynote speech in Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal, 2004
17. Local, J. *Phonetics and talk-in-interaction*. // Proceedings of ICPHS 2003
18. Moore, R. K. *Towards a unified theory of spoken language processing*. Proceedings of 4th IEEE International Conference on Cognitive Informatics, Irvine, CA, USA, 8-10 August 2005
19. Krauwer, S. *How to survive in a multilingual EU? // Proceedings of the Second Baltic Conference on Human Language Technologies, April 4-5, 2005, Tallinn, Estonia, 61-66*

Einar Meister  
TTÜ Küberneetika Instituut  
Foneetika ja kõnetehnoloogia labor