TEOORIA JA PRAKTIKA VAHEKORRAST KÕNETEHNOLOOGIAS: TÄIUSLIKUMA TEHNOLOOGIA SUUNAS

Einar Meister
TTÜ Küberneetika Instituut

Kõnetehnoloogia valdkonnas on viimase paarikümne aasta jooksul kogu maailmas toimunud oluline edasiminek – turul on mitmeid kõnetuvastus- ja sünteesiprogramme ning edu on saavutatud inimene-masin dialoogsüsteemide rakendamisel. Kuid siiski on selle valdkonna areng olnud pigem evolutsiooniline – kauaoodatud revolutsioonilist läbimurret (selleks võiks olla klaviatuurita arvuti turuletulek, mille puhul põhiliseks kasutajaliideseks oleks kõneliides; Microsoft prognoosis seda aastaks 2003) ei näe me ilmselt ka veel lähiaastatel.

Et mõista kõnetehnoloogia (põhiliselt on jutt siiski kõnetuvastusest) arengu eripära ja selle tänast seisu, esitan alljärgnevalt põgusa ülevaate olulisematest arengusuundadest läbi aastakümnete.

„50 aastat progressi kõne- ja kõnelejatuvastuses"

Sellise pealkirjaga ettekande esitas Tokyo ülikooli professor Sadaoki Furui 2005. aasta oktoobris rahvusvahelisel konverentsil SPECOM' 2005 – 10th International Conference on Speech and Computer (17 –19. oktoober, Kreeka, Patras) (Furui 2005a).

Kõnetuvastuse alased uuringud said alguse 1950ndatel aastatel USAs (Bell Laboratories, RCA Laboratories, MIT Lincoln Laboratories), Jaapanis (Radio Research Lab, Kyoto Ülikool) ja 1960ndatel tollases Nõukogude Liidus (Kiievi Küberneetika Instituut). Olulist rolli valdkonna arengus (paljude teiste uurimisgruppide hulgas) on etendanud IBM Labs, AT&T Bell Labs ja Carnegie Mellon Ülikool USAs ja JSRU (Joint Speech Research Unit) Suurbritannias. Paljusid uurimis- ja arendusprojekte on finantseerinud USA Kaitseministeerium läbi

DARPA (Defence Advanced Research Projects Agency) programmi, mille tulemusena loodi mitmeid maailmas tunnustust võitnud kõnetuvastus- ja dialoogsüsteeme (näiteks Hearsay I, Hearsay II, Harpy, SPHINX, BYBLOS, DECIPHER jt).

Olulisemateks saavutusteks kõnetuvastuses peab professor Furui järgmisi arenguid:

- mustrituvastuse asemel kasutatakse korpustel baseeruvat statistilist modelleerimist (HMM[1] ja N-gram[2] mudelid),
- akustiliste tunnustena kasutatakse spektri resonantside asemel kepstri-tunnuseid[3] (kepster + Δkepster + ΔΔkepster),
- heuristiline ajanormeerimine on asendunud dünaamilise normeerimisega,
- tunnuste klassifitseerimisel rakendatakse tõenäosuslikke meetodeid,
- isoleeritud sõnade tuvastus on asendunud sidusa kõne tuvastusega,
- väike sõnavara on kasvanud suureks või peaaegu piiramatuks,
- kontekstist sõltumatute tuvastusüksuste asemel kasutatakse kontekstist sõltuvaid üksusi,
- müravaba kvaliteetse kõne tuvastus on arenenud müritaustaga kõne tuvastuseks,
- ühele kõnelejale treenitud süsteemide asemel luuakse kõnelejast sõltumatuid süsteeme,
- monoloog-kõne tuvastus on asendunud dialoog-kõne tuvastusega,
- loetud kõne kõrval tuvastatakse ka spontaanset kõnet,
- tuvastusele on lisandunud kõne mõistmine,
- arendatakse ka audio-visuaalse kõnetuvastuse meetodeid,
- riistvaraliste süsteemide asemele on tulnud tarkvarasüsteemid,
- prototüüpidest on välja kasvanud mitmed kommertsrakendused.

Progressi kõnetuvastuses illustreerib ka joonisel 1 esitatud graafik (Furui 2005b), mille kohaselt on kõnetuvastus viimase 30 aasta jooksul arenenud isoleeritult hääldatud piiratud arvu häälkäskluste tuvastusest (enne 1980. aastat) suure sõnavaraga teksti dikteerimiseni (1990ndad aastad); sel sajandil on kõnetuvastuse areng liikunud piiramatu sõnasti-

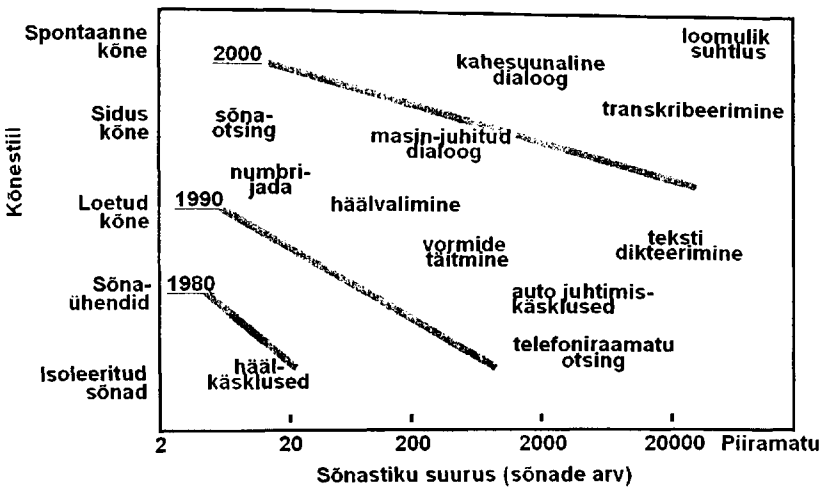
---

[1] HMM (ingl *Hidden Markov Model*) – Markovi peitmudel.

[2] Statistiline keelemudel, mis leiab mingi sõna esinemise tõenäosuse sellele eelnenud N-1 sõna põhjal.

[3] Inimese kuulmistaju omadusi modelleerivad tunnused, mis leitakse kõnesignaali logaritmilise spektri teisendamisel mittelineaarsesse sagedusskaalasse ja sellele Fourier' pöördteisenduse rakendamise teel.

kuga loomulikku suhtlust võimaldavate süsteemide suunas, mida tänased tuvastussüsteemid siiski veel ei toeta.



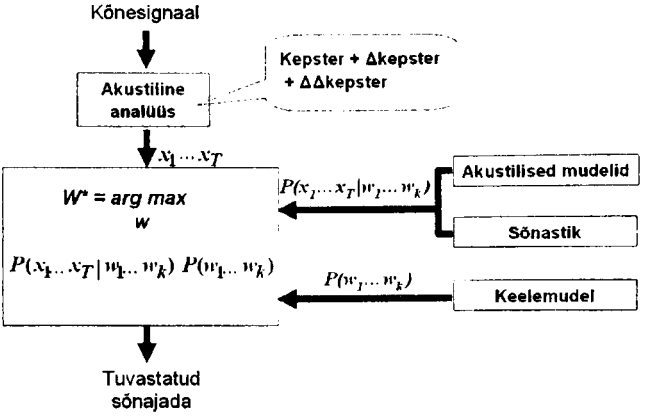
Joonis 1. Kõnetuvastuse areng viimase 25 aasta jooksul (Furui 2005b).

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 kõnesignaalis esinev suur variatiivsus – sama sõna identseid häälduisi praktiliselt ei eksisteeri ja iga hääldus realiseerub erineva akustilise mustrina.

Variatiivsuse põhjusteks on:

- kõnelejate vanus ja sugu,
- kõnestiil,
- keeletaust (emakeel *vs* võõrkeel),
- emotsionaalne ja tervislik seisund,
- jpm.



Joonis 2. Kõnetuvastussüsteemi struktuur (Furui 2005b).

Joonisel 2 esitatud skeemilt näeme, et kõnetuvastus koosneb kahest põhietapist: (1) akustilisest analüüsist ja (2) mustrituvastusest.

Akustilise analüüsi eesmärgiks on leida sisendsignaalist tuvastamiseks olulist informatsiooni sisaldavad tunnused ja maha suruda ebaolulised variatsioonid. Kõige sagedamini kasutatakse kõnetuvastuses mel-kepstri kordajaid (ingl *mel-frequency cepstral coefficient* – MFCC), mis saadakse inimkõrvale sarnase mittelineaarse signaalitöötluse tulemusena (Huang *et al* 2001: 304).

Sisendsignaal **X** esitatakse kindla intervalli (tavaliselt 25 ms) järel arvutatavate tunnusvektorite jadana:

$$\mathbf{X} = \mathbf{x}_1\, \mathbf{x}_2 \dots \mathbf{x}_T.$$

Mustrituvastuse ülesandeks on leida sõnajada $\mathbf{W^*} = \mathbf{w}_1\ \mathbf{w}_2 \dots \mathbf{w}_n$, mis kõige tõenäosemalt vastab sisendsignaalile **X**:

$$W^* = \arg\max_W P(W \mid X) = \arg\max_W \frac{P(W)\, P(X \mid W)}{P(X)}$$

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

$$W^* = \arg\max_W P(W)\, P(X \mid W).$$

Seega, kõige tõenäosem sõnajada **W*** sõltub:
1. sõnajada *a priori* tõenäosusest **P(W)** – see leitakse keelemudelist,
2. 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.

Tutvumaks Markovi peitmudelite kasutamisega kõnetuvastuses võib soovitada eestikeelset artiklit (Alumäe 2002), põhjalikumad käsitlused leiab lugeja erialastest allikatest (nt Huang *et al* 2001; Cole *et al* 1995; jt).

Valdav osa kommertsrakendustest ja arendatavatest prototüüpidest 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 (Meister, Eek 1999; Meister jt 2003) ning infrastruktuur. Labori teaduri, TTÜ doktorandi Tanel Alumäe töö tulemusena on loodud mitmeid spetsiaalselt eestikeelse kõne tuvastuseks vajalikke mudeleid ning esimesed piiratud sõnavaraga prototüübid; uuringud jätkuvad piiramatu sõnavaraga kõnetuvastuse loomiseks (Alumäe 2004, 2005a ja 2005b).

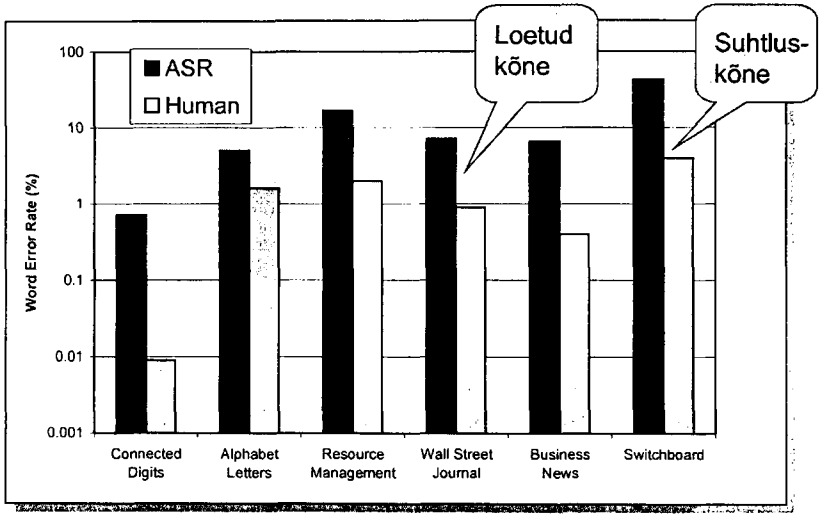
Võrdleme inimest ja masinat

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 (Lippmann 1997) 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 erinevaid kombinatsioone.

Samuti näeme, et nii inimene kui ka masin teevad spontaanse suhtluskõne (Switchboardi korpus) tuvastamisel rohkem vigu kui loetud ajaleheteksti (Wall Street Journali korpus) tuvastamisel. Masinate puhul on

selle põhjuseks 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 spontaankõne tuvastusvigade protsent umbes kaks korda väiksem võrreldes laboratoorse kõne baasil treenitud mudelitega (Furui 2005c).

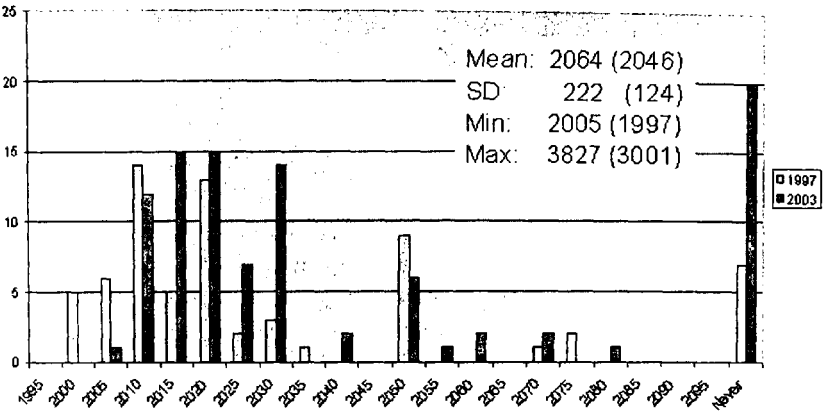
Kõige lähemal inimvõimetele on masin üksikult hääldatud tähestiku tähtede tuvastamisel. Selle tulemuse hindamisel on vajalik arvestada keelespetsiifikat – joonisel 3 esitatud tulemused on saadud inglise keele kohta ja näiteks eesti keele puhul võime saada oluliselt erinevad tulemused eelkõige foneetiliste iseärasuste tõttu. Nii on eesti keeles ka inimesel (masinast rääkimata!) raske eristada tähepaaride p – b ja t – d isoleeritud hääldust, sest sõna alguses need klusiilid foneetiliselt ei eristu (p ja b mõlemaid häälda takse /pee/, t ja d häälda takse /tee/) (Eek, Meister 1996).



Joonis 3. Inimese ja masina kõnetuvastuse võrdlus erinevat tüüpi kõne puhul (Lippmann 1997). Horisontaalteljel on esitatud erinevad kõnestiilid, 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 1997. ja 2003. aastal mitmetele kõnetehnoloogia ekspertidele maailmas (Moore 2003). Küsitluse tulemused näitasid, et pessimistide osakaal on kahe küsitluse vahel oluliselt kasvanud (vt joonis 4).



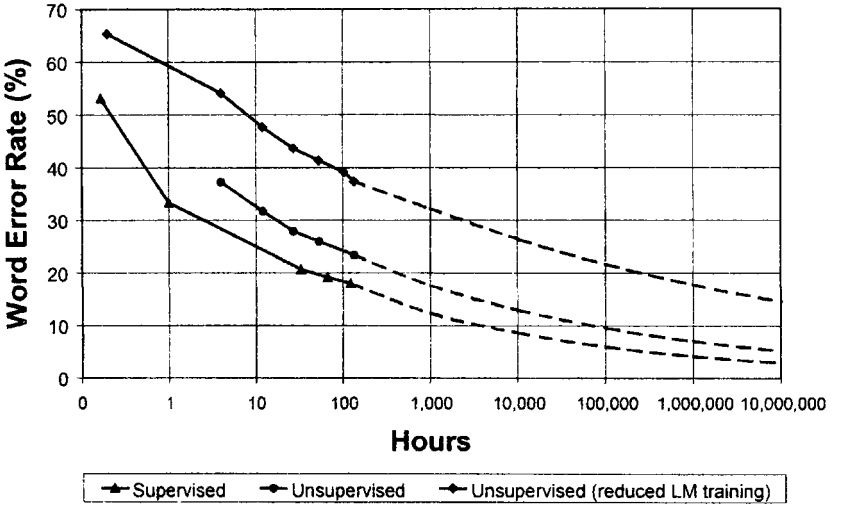
Joonis 4. Vastuste jaotus küsimusele „Mis aastal on automaatse kõnetuvastuse kvaliteet võrdne inimese kõnetuvastusega?" aastatel 1997 (heledad tulbad) ja 2003 (tumedad tulbad) (Moore 2003). Horisontaalteljel on aastaarvud, vertikaalteljel vastuste hulk. Keskväärtuse (Mean), standardhälbe (SD), minimaalse (Min) ja maksimaalse (Max) hinnangu väärtused 1997. a kohta on esitatud sulgudes.

Analüüsides 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.

Miks siis ikkagi peale 50 aastat progressi on suur hulk tippeksperte 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 kõneandme-

baaside loomisele ja statistiliste meetodite laialdasele kasutamisele (Moore 2005; Lee 2004, vt ka Mare Koidu artiklit käesolevas kogumikus).

Praegust nn jõumeetodil toimuvat arengut iseloomustatakse 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 treeningmaterjali on siis vaja, et jõuda inimesele lähedaste võimeteni? Vastuse sellele küsimusele leiame R. Moore'i artiklist (2005) (joonis 5).



Joonis 5. Tuvastuskorrektsuse sõltuvus treeninguks kasutatud kõnematerjali hulgast kolme erineva treenimismeetodi puhul (Moore 2005). Horisontaalteljel treeningmaterjali hulk tundides, vertikaalteljel sõnatuvastuse viga protsentides.

Jooniselt 5 näeme, kuidas spontaanse kõne tuvastusvigade protsent väheneb sõltuvalt treeningmaterjali hulgast. Reaalsed tulemused on saadud eksperimentidest kuni ca 100 tunni treeningmaterjaliga (pidev joon), katkendlik joon on saadud eksperimendiandmete ekstrapoleerimisel. Valides parimaid tulemusi andva juhendatud (*supervised*) treeningu, on võimalik saavutada *ca* 12% tuvastusviga 1000 tunni treeningmaterjaliga, *ca* 8% viga 10 000 tunniga, *ca* 6% viga 100 000 tunniga ja *ca* 3% viga

10 miljoni tunni treeningmaterjaliga. 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 kuuldud kõnemahule ja 10 miljonit 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 (Meister jt 2003) maht on *ca* 240 tundi kõnet, salvestatud *ca* 1300 kõnelejalt; salvestuste kogumine, kontroll ja märgendamine kestis umbes kaks aastat.

Miks oleme sellises seisus?

Tsiteerides veelkord R. Moore'i (2005): „kõnetöötlus on universumis teadaoleva kõige keerukama elusorganismi kõige keerulisem talitlus" (EM tõlge). Teisisõnu, kõnetöötlusest inimajus ja kõnekommunikatsiooni protsessidest teame tänapäeval 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.
>
> Iga kord, kui ma vallandasin ühe keeleteadlase, paranes süsteemi tuvastuskorrektsus.

Viimane ütlus on omistatud IBMis kõnetuvastuse loomisega tegelenud uurimisgrupi juhile Frederick Jelinekile, kes on hiljuti püüdnud seda ümber lükata (Jelinek 2005).

Kõnekommunikatsiooni eri aspektide uurimisel on lähtutud põhiliselt laboratoorsest kõnest ja tekstipõhisest keelematerjalist ning nendel

uurimistulemustel põhineb ka enamik tehnoloogias rakendatavaid mudeleid (Campbell 2005; Furui 2005c).

Erinevaid kõneaspekte on uuritud teineteisest (peaaegu) sõltumatult: kõnetuvastuse puhul on oluline eelkõige kõne lingvistiline sisu, s.o **mida ütles**, ja kõnelejast 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 (Campbell 2004; Local 2003), 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 **teadmus-ignorantne** (ingl *knowledge-ignorant*) tehnoloogiaarendus ei saa lõpmatult jätkuda; edasiminek saab toimuda ainult **teadmusrikka** (ingl *knowledge-rich*) tehnoloogia arenduse teel. Üks sellise mõtteviisi propageerija maailmas Chin-Hui Lee on välja pakkunud kõnetuvastuse arendamiseks järgmisi ideid (Lee 2004):

- hääliku-spetsiifilised tunnused – lisaks kepstrile on vajalik 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 tehisneuronvõrkude[4] abil, kasutatakse statistiliste mudelite treenimiseks;
- inimese kõnetöötluse mudelid – inimene ei teisenda kõnesignaali sõnajadaks otse, vaid tuvastab signaalist akustilisi ja auditiivseid

[4] Kogum vastastikku ühendatud arvutuselemente, mis modelleerivad bioloogiliste närvirakkude käitumist.

sündmusi, mille põhjal formuleeritakse kognitiivsed hüpoteesid, neid verifitseerides jõutakse konteksti sobiva tulemuseni.

Tõepoolest, inimestevahelises suhtluses on võrdselt tähtsad nii akustiline ja lingvistiline kui ka para- ja ekstralingvistiline komponent ning need peaksid olema adekvaatselt modelleeritud ka inimene-masin suhtlusmudelis. Sellise integreeritud suhtlusmudeli loomine eeldab:

- suhtlusolukorrale tüüpilise andmestiku kogumist ja mitmekülgset analüüsi,
- ühtse kõnekommunikatsiooniteooria väljaarendamist.

Inimene-inimene ja inimene-masin kommunikatsiooni uurimiseks vajaliku multimodaalse andmestiku kogumiseks on tarvilik realiseerida **intelligentse ruumi** prototüüp, mis võimaldaks efektiivselt modelleerida reaalseid suhtlusolukordi. Ühtse kõnekommunikatsiooniteooria loomine ja väljaarendamine on uue interdistsiplinaarse teadusvaldkonna – **kognitiivse informaatika** – üks olulisemaid väljakutseid (Moore 2005, vt ka Mare Koidu artiklit käesolevas kogumikus).

Kuid jäägu need mõisted – intelligentne ruum, kognitiivne informaatika – siinses kirjatükis avamata, asjast huvitatud lugeja leiab vastavat infot ka Internetist otsides.

Lõpetuseks

Vaatamata üldise kõnetuvastusteooria puudumisele, on tänu arvutusvõimsuste kiirele kasvule ja statistiliste ning insener-tehniliste meetodite oskuslikule rakendamisele loodud mitmeid küllalt hästi töötavaid kõnetuvastus-süsteemide lahendusi. Kuigi erinevate inimkommunikatsiooni käsitlevate teooriate rolli suurenemine kõnetehnoloogia edasises arengus näib olevat möödapääsmatu, võime siiski küsida, kas inimese kõnetöötluse matkimine teoreetilistes mudelites (sisuliselt looduse pealt spikerdamine) on ainuõige tee? Näiteks teame ju ajaloost, et liikuvate tiibadega lennumasinate arendamine (linnud ju lendavad tiibu liigutades!) oli üsna lootusetu ettevõtmine ja lennunduse areng sai hoo sisse alles peale aerodünaamika seaduspärasuste avastamist. Tõenäoliselt on inimkommunikatsiooni olemuse täieliku mõistmiseni veel üsna pikk tee ning selleni jõudmine ei sõltu sugugi ainult keeleteadlaste ja -tehnoloogide pingutustest.

Kirjandus

Alumäe, Tanel 2002. Varjatud Markovi mudelid. – Arvutitehnika ja Andmetöötlus 4, 27–36.

Alumäe, Tanel 2004. Large vocabulary continuous speech recognition for Estonian using morpheme classes. – Proceedings of ICSLP 2004 – Interspeech. Jeju, Korea, 389–392.

Alumäe, Tanel 2005a. Phonological and morphological modeling in large vocabulary continuous Estonian speech recognition system. – Proceedings of the Second Baltic Conference on Human Language Technologies. Tallinn, Estonia, 89–94.

Alumäe, Tanel 2005b. Using Adaptive Stochastic Morphosyntactic Language Model for Two-pass Large Vocabulary Estonian Speech Recognition. – Proceedings of SPECOM'2005 – 10th International Conference on Speech and Computer. Patras, Greece. Vol. 2, 515–518.

Campbell, Nick 2004. 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.

Campbell, Nick 2005. Expressive speech synthesis: what is the goal? – Proceedings of the Second Baltic Conference on Human Language Technologies. Tallinn, Estonia, 15–26.

Cole, Ronald A. et al (eds) 1995. Survey of the State of the Art in Human Language Technology. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

Eek, Arvo, Einar Meister 1996. Eesti sõnaalguliste sulghäälikute akustika ja tajumine. – Keel ja Kirjandus 3–5, 164–170, 241–253, 314–321.

Furui, Sadaoki 2005a. 50 years of progress in speech and speaker recognition. – Proceedings of SPECOM'2005 – 10th International Conference on Speech and Computer. Patras, Greece. Vol. 1, 1–7.

Furui, Sadaoki 2005b. Toward Robust Speech Recognition. – Proceedings of the Second Baltic Conference on Human Language Technologies. Tallinn, Estonia. Tutorials Day. 1–41.

Furui, Sadaoki 2005c. Spontaneous speech recognition and summarization. – Proceedings of the Second Baltic Conference on Human Language Technologies. Tallinn, Estonia, 39–50.

Huang, Xuedong, Alejandro Acero, Hsiao-Wuen Hon 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall.

Jelinek, Frederic 2005. Some of my best friends are linguists. – Language Resources and Evaluation. Vol. 39, Nr 1, 25–34.

Koit, Mare 2006. Ratsionalism ja empirism keeletöötluses: vastasseis või koostöö? Käesolevas kogumikus.

Lee, Chin-Hui 2004. From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition. – Proceedings of ICSLP 2004 – Interspeech. Jeju, Korea.

Lippmann, Richard 1997. Speech Recognition by Machines and Humans. – Speech Communication. Vol. 22, 1–15.

Local, John 2003. Phonetics and talk-in-interaction. – Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona.

Meister, Einar, Arvo Eek 1999. Estonian Phonetic Database. EU Copernicus Programme, Project No. 1304 „BABEL – A Multi-Language Database". Tallinn.

Meister, Einar; Jürgen Lasn, Lya Meister 2003. SpeechDat-Like Estonian Database. – Proceedings of the 6th International Conference TSD 2003, Lecture Notes in Artificial Intelligence 2807. Springer. 412–417.

Moore, Roger K. 2003. Speculating on the Future for Automatic Speech Recognition. – IEEE workshop on Automatic Speech Recognition and Understanding (ASRU). St. Thomas, US Virgin Islands.

Moore, Roger K. 2005. Cognitive Informatics: The Future of Spoken Language Processing? – Proceedings of SPECOM'2005 – 10th International Conference on Speech and Computer. Patras, Greece. Vol. 1, 11–15.