

INIMKÕNE ON ARVUTILE VÕORKEEL

TANEL ALUMÄE, TOOMAS KIRT



Kõnetuvastus on osa automaatselt kõnest arusaamisest, mille eesmärgiks on inimkõne teisendamine mingile abstraktsele kujule, mis esindab inimkõnes olnud mõtet teatud formalismile tuginedes ning mis on arvutil töödeldav. Selline tehnoloogia võimaldab luua automaatseid süsteeme, kus inimene saab arvutiga suhelda kõne abil – mõte sellest on inimesi alati võlunud. Hoolimata näilisest lihtsusest, on automaatne kõnetuvastus väga keerukas.

Igapäevaelus oleme harjunud, et meid ümbritsevad kõikvõimalikud helid ja me suudame suuremate raskusteta nende hulgast eraldada kõne, tänavamüra, linnulaulu. Järgnevalt püüame heita pilku sellele, kuidas toimub arvutikeskkonnas kõne tuvastamine. Eluslooduses toimival evolutsiooniprotsessil kulus miljoneid aastaid keelt kasutava inimese loomisele. Nüüd püüab inimene seda sama keelest arusaamist õpetada arvutile. Käsil on pikema protsessi esimene etapp, mille käigus püütakse jõuda selleni, et arvuti suudaks eristada kuuldu kõnet. Selle mõistmiseni arvuti poolt kulub veel kahtlemata aastaid.

INIMENE ÕPIB KOGU ELU

Loodus on inimese aju programmeerinud nii, et pärast sündi hakkab imik kohe õppima teda ümbritsevate inimeste keelt. Inimese puhul on õppimisprotsess kaunis pikk ning kuna keel areneb pidevalt edasi, siis võib öelda, et see kestab kuni surmani.

Keele õppimisel on siiski olulised esimesed eluaastad ja selleks on teatav kriitiline aeg. Pärast kuuendat eluaastat muutub keele omandamine inimestest eraldatud keskkonnas elanud lapsele praktiliselt võimatuks. Sünnijärgselt on laps valmis õppima mis tahes keelt, ta suudab eristada kõiki talle esitatud häälikuid, ükskõik, mis keelega tegu. Ent vanuses 6–10 kuud kohandub laps teda ümbritsevate inimeste poolt kasutatavate häälikutega. Steven Pinkeri raamatus "The Language Instinct" toodud näite kohaselt suudavad kuuekuused hispa-

nia lapsed eristada veel inglise keele foneeme "pa" ja "ba", kuid alates kümnenast elukuust see enam ei õnnestu. Kuna hispaania keeles neid ei kasutata, tajub laps neid ühe foneemina. Kuuldud häälikut muudetakse ajus alateadlikult sellele kõige lähedasema hääliku suunas ning inimene ei taju seda sellisena, nagu see tema kõrva kostub. Pärast kümnead elukuud on laps võimeline eraldama oma vanemate keele foneeme ning see loob võimaluse foneemidest sõnade ja nende seostest omakorda grammatiliste struktuuride kokkupanekuks.

Esimese sünnipäeva paiku hakkab laps eristama sõnu ja neid ka ise moodustama. Ühesõnaliste lausete periood kestab kahest kuust kuni aastani. Selles vahemikus eristab laps lausest ainult talle tuttavaid sõnu. Kaheksateistkuuselt saab alguse keel – kuigi laused on kahe sõnalised, on need juba korrektses järjestuses. Teise eluaasta lõpuks või kolmanda eluaasta keskpaigaks on laps omandanud ka grammatika ning on praktiliselt valmis vabaks keelekasutuseks.

Niisiis kulub inimesel keele omandamiseks aastaid. Tänapäeva inimene püüab õpetada ka arvutit keelest aru saama ning järgnevalt vaatamegi, millised probleemid meil tegelikult ees seisavad.

HÄÄLIKUD NAGU SÕRMEJÄLJED

Esmapilgul tundub, et inimese jaoks on kõne mõistmine loomulik ja lihtne. Tõepoolest, me suudame peaaegu alati eksimatult aru saada, millised sõnad on

jutus, mida me kuuleme, ja mida nendega öelda tahetakse. Igapäevase kõne mõistmisega tulevad suurepäraselt toime nii päris väikesed lapsed kui ka vanurid. Inimesel pole suuri probleeme kõne mõistmisega ka olukorras, kus rääkija hääli on moonutatud (näiteks läbi telefoni kostev hääli), samuti küllaltki suure mürataustaga keskkonnas. Seda imelikum võib tunduda fakt, et aastakümnetepikkuse uurimistöö järel on arvuti samaväärse kõne mõistmisel küllalt saamatu. Kui aga meenutada mõnda korda, mil oleme kuulnud meie senikuulmata ja väga kaugelt keelt, võib arvuti olukorras natukenegi aru saada: võõra keele sõnad tunduvad omavahel väga sarnased ning võimatu on eraldada sõnadevahelisi piire. Seesugused raskused esinevad ka automaatses kõnetuvastuses. Arvuti teeb tavaliselt väga palju vigu juba häälikute äratundmisel, mis omakorda tingib valede sõnade tuvastamise. Selle põhjuseks on erinevate sõnaosade (häälikute ja silpide) äärmiselt suur kvalitatiivne ja kvantitatiivne varieeruvus.

Foneetika üks põhitõdesid on, et kõnes ei esine kahte identselt häälikut, nagu ei esine kahte ühesugust puulehte või sõrmejälge. Ei erine ainult erinevad häälikuklassid, vaid ka sama hääliku erinevad realiseeringud. Varieeruvuse põhjusi on mitmeid. Juba hääle sisestamisel arvutisse on palju selliseid keskkonnaparameetreid, mis töödeldavat signaali mõjutavad: sellisteks muutujateks on näiteks mikrofoni tüüp ja positsioon rääkija suhtes, signaali transleerimisel kasutatavad filtrid (eriti telefonikõne puhul), müratase, ruumi akustika. Kõne varieerub väga tugevalt olenevalt rääkimisviisist: valjus, emotsioonid, tempo, kooperatiivsus, intonatsioon. Kõik need asjaolud mõjutavad kõnet, mille tõttu samad häälikud ja sõnad räägituna sama inimese poolt erinevas olukorras on oma pikkuselt ja spektraalsetelt tunnustelt väga erinevad. Häälikute ja sõnade akustikat mõjutab tugevasti nende naabruskontekst – nii näiteks on häälik /k/ sõnas 'koli' veidi teistsugune kui sõnas 'kilo'. Lõpuks, kõne varieerub ka sõltuvalt rääkija füüsilisest ja sotsiaalsest eripärast. Häälikute akustikat mõjutab rääkija vanus, sugu, kõnetrakti ja muude

kõnelemisprotsessis osalevate organite suurus ja kuju, tervislik seisund (näiteks on nohusel kasutajal alati suuri probleeme tänapäeva kõnetuvastussüsteemidega suhtlemisel), samuti sotsiaalne ja regionaalne päritolu.

KOGEMUS JA MITMETÄHENDUSLIKKUS

Teadusfilosoofiast on teada, et eksperimentaatori vaatlusotsuseid mõjutavad tugevalt varasem kogemus ja olemasolevad hüpoteesid. Sarnane on inimõistuse tegevus kõne mõistmisel, ilma selleta langeks inimese kõnetuvastusoskus madalamale praeguste arvutite tasemest. Nimelt on inimesel kõne dekodeerimisel abi väga paljudest asjaoludest – kõne artikulatsioonist, tempost, morfoloogiast, süntaksist ja semantikast. Inimene oleks küllalt abitu, kui ta ei suudaks kasutada igapäevaelust kogunenud teadmisi, situatsioonikonteksti ja kogu intellekti, et enese teadmata automaatselt parandada vigu, mis võivad olla tingitud taustamürast, kõneleja vigadest või lohakusest. Tundub ka, et häälikute varieeruvus, mis nende tuvastamise arvutile nii raskeks teeb, just aitab inimesel kõnet paremini mõista. Iga kõne varieeruvuse parameeter “on omal kohal nii kõneleja kui kuulaja jaoks”, ütleb Mati Hint oma raamatus “Häälikutest sõnadeni”. Ta toob järgmise näite: kui Jürile hüütakse “Mari tuleb!”, kuid Jüri ei kuulnud algushäälikut /m/ mingi äkilise taustamüra tõttu, saab Jüri hoiatusest ikkagi õigesti aru, kuna häälik /m/ tingis muutusi järgneva hääliku /a/ algusosades. See, ning antud olukorra kontekst, laseb Jüril eksimatult taastada sõna alguse.

Automaatses kõnetöötles tekkitab suuri probleeme ka keeles tihti esinev mitmetähenduslikkus. Kõnetuvastuse kui kõne-tekst teisenduse seisukohast on üks probleem homofoonia. Homofoonia puhul saab mingit kõnelõiku “legaalselt” tekstiks teisendada mitut moodi: näiteks sõnu “kas sa” ja sõna “kassa” võib hääldada täpselt sama moodi. Kõne mõistmise seisukohast on palju suuremad probleemid süntaktiline ja semantiline mitmetähenduslikkus. Süntaktilise mitmetimõistetavuse puhul võib ühest lausest mitut moodi aru saada, kuna on mitu viisi lause grammatilist struktuuri korrektset tõlgendada. Näiteks laused “Me andsime ahvidele banaane, kuna nad olid näljased” ning “Me andsime ahvidele banaane, kuna nad olid küpsed” ei erine oma struktuurilt üldse, kuid sõna “nad” võib viidata nii ahvidele kui banaanidele. Inimesel on tänu oma

teadmiste ahvide ja banaanide kohta lihtne toodud lauseid tõlgendada, kuid alati see nii ei ole. Lause “Ma nägin mäe otsas puud” puhul on vaja teada ja mõista konteksti, kus see esines, et aru saada, kas nähti puud, mis kasvab mäe otsas, või keegi nägi mäe otsas seisest ühte puud. Semantilise mitmetähenduslikkuse puhul tekitavad probleeme sõnad, millel on mitu tähendust. Fraas “ukse hinged” viitab ilmselt hingedele kui füüsilistele objektidele, kuid teatud kontekstis on võimalik, et räägitakse hingedest kui ukse vaimust. Semantilise mitmetähenduslikkuse alla võib paigutada ka metafooride tõlgendamise. Lause “Mu hobi sööb palju raha” räägib tõenäoliselt mingist palju raha nõudvast (ka siin lauses on mitmetähenduslikkus!) hobist, kuid ei saa välistada, et jutt käib söömise selle otseses mõttes.

Eelneva põhjal peaks olema mõistetav, miks on automaatne kõnetuvastus keeruline. Mõned uurijad lähevad koguni niikaugemale, et väidavad: kõnetuvastus on kogu tehisisintellekti suhtes täielik, st see on redutseeritav tehisisintellekti loomise probleemile, ning seda ei saa täielikult lahendada enne, kui tehisisintellekt on olemas. Tahtmata anda tehisisintellekti definitsiooni, võib öelda, et tehisisintellekti all mõeldakse siin sellist intellektitaset, mis on võrreldav inimintellektiga.

KÕNETUVASTUS ARVUTIS

Automaatses kõnetuvastuse teevad arvuti jaoks keeruliseks mitmed asjaolud: taustamüra, mikrofonide erinevad omadu-

sed, inimeste kõnetrakti erinevustest tingitud häälikute varieeruvus, ning sotsiaalsest ja situatsioonilisest kontekstist tulenev erinev hääldus- ja kõnelemismanner. Et saavutada inimesega vähegi võrreldavat kõnetuvastustäpsust, peaks automaatne tuvastussüsteem olema suuteline kõiki neid varieerumisallikaid piisavalt täpselt modelleerima. Paraku pole tänapäeva arvutiteadus veel kaugeltki nii arenenud, et eriti just sotsiaalse ja situatsioonikonteksti hästi ära kasutada. Seetõttu on praegused tuvastussüsteemid kõik teatud määral piiratud – näiteks dikteerimisprogrammid, mis paljude keelte jaoks juba olemas on, “saavad aru” ainult korralikult artikulatsioonist sujuvast kõnest, samuti on tavaliselt vaja, et iga üksikkõneleja süsteemi eelnevalt lühiajaliselt trenniks. Teised süsteemid aktsepteerivad küll spontaanset kõnet, kuid ainult mingi teatud kitsa ülesande piires (näiteks busside sõiduplaanide automatiseeritud infotelefon).

AKUSTILINE MUDEL

Tüüpilise kõnetuvastussüsteemi (vt joonis) põhilisteks komponentideks on akustilised häälikumudelid, millega modelleeritakse erinevate häälikute unikaalseid akustilisi ja artikulatoorseid omadusi, ning keelemudel, mis omakorda koosneb sõna- ja lausemudelitest. Kõnetuvastuse käigus sisestatakse kõnesignaal läbi mikrofoni või telefoni arvutisse, kus see digitaliseeritakse. Digitaliseeritud kõnesignaal jagatakse lühikeseks, tüüpiliselt 10 millisekundiga pik-



kusteks lõikudeks. Iga kõnelõigu spektrist arvutatakse tunnusvektor, mis paarikümne koefitsiendiga iseloomustab antud lõigus olevat informatsiooni. Tunnusvektorite arvutamisel on kaks põhilist eesmärki: vähendada infohulka ning tuua esile sellised tunnused, mis erinevate häälikute vahel võimalikult suurelt varieeruvad. Praktikas kasutatakse selleks kõige sagedamini nn mel-cepstrumkoefitsiente, millega saab kompaktselt kirjeldada antud kõnelõigu spektris esinevaid sagedusi. Spektraalanalüüsi tulemusena saadakse sõnale vastav tunnusvektorite jada. Süsteemi treenimisel töödeldakse nii väga suuri kõnekorpuseid, ning saadud tunnusvektorijadade ja korpuse kõne foneetiliste transkriptsioonide võrdlemise tulemusena saadud statistilised häälikumudelid suudavad küllalt hästi modelleerida erinevate häälikute unikaalseid omadusi ning varieeruvust. Saadud häälikumudelitest sünteesitud sõnamudelite abil saab iga suvalise sisendkõne lõigu jaoks arvutada tõenäosuse, et antud lõigus esines just sellele mudelile vastav sõna.

Seda põhimõtet on lihtne seletada optilise käekirjatuvastamise (OCR) näitel. Käekirjatuvastamise strateegia on kõnetuvastusega äärmiselt sarnane: ka siin leitakse käekirjast teatud tunnused (kriipsude, haakide, täppide asend ja arv jms). Saadud tunnuseid võrreldakse iga kandidaatsõna mudeliga. Kandidaatsõnade mudelid on treenitud paljude inimeste käekirjade näidetel. Iga mudel on võimeline "ütlemata", kui suure tõenäosusega on kirjutatud sõna just temale vastav sõna. Näiteks joonisel esitatud käekirjatuvastamisjuhtumi puhul annavad sõnade "null" ja "ühiksa" mudelid üsna väikese tõenäosuse, sõna "üks" mudeli poolt arvutatud tõenäosus on kõige suurem, ning "üks" saadetaksegi väljundisse. Alati ei pruugi tegelikult kirjutatud sõnale vastava mudeli tõenäosus kõige suurem olla, sel juhul tuvastatakse vale sõna.

KEELEMUDEL

Akustiliste mudelite kõrval on kõnetuvastussüsteemi teiseks tähtsaks komponendiks keelemudel. Keelemudelid võib laias laastus jagada kaheks: formaalsed ja statistilised. Formaalse grammatika kasutamisel on süsteemi poolt tuntavad laused mingite reeglite põhjal defineeritud. Sellist keelemudelit kasutatakse tavaliselt mitmesuguste juhtimis- ja kontrollsüsteemide juures. Keelemudelit tutvustaval joonisel toodud grammatika aktsepteerib näiteks lauseid "Helista koju", "Helista isale", "Helista üks üks kaheksa kaheksa", kuid mitte "Helista Toomasele".

KEELEMUDEL

<lause> = helista <sihtkoht>

<sihtkoht> = koju | tööle | emale | isale | <telefoninumber>

<telefoninumber> = <number> <number>

<telefoninumber> = <telefoninumber> <number>

<number> = üks | kaks | kolm | neli | viis | kuus | seitse | kaheksa | üheksa | null

Statistilist keelemudelit rakendatakse suurema ja üldisema sisendkeele rakendustes, näiteks dikteerimissüsteemides. Statistilise keelemudeli ülesandeks on võimalikult hästi ära arvata, millise sõna inimene järgmisena ütleb. Keelemudeli abil saab tuvastussüsteem automaatselt parandada valesti või halvasti kuulnud sõnu. Inimesel on see võime väga hästi arenenud: näiteks minu arhitektist sõber kaebab selle üle, et küsides ajalehekioskist ajakirja Maja, antakse talle tavaliselt hoopis ajakiri Maaja. Ilmselt juhtub see selle tõttu, et ajakiri Maaja on palju populaarsem ning lausel "Üks Maaja palun" on müüja kogemuste põhjal tunduvalt suurem tõenäosus kui lausel "Üks Maja palun". Kui lauset korrata, hääldades sõna 'maja' väga püüdlukult, saab müüja ostjast lõpuks siiski aru, sest lause akustikast saadud informatsioon kaalub üle keelemudelis oleva.

Statistilist keelemudelit treenitakse väga suurte keelekorpuste põhjal (vähe-

malt paar miljonit sõna, soovivatult palju rohkem). Korpuste alusel on võimalik teatud määral hinnata, millised sõnad ja sõnakombinatsioonid keeles sagedamini esinevad, ning millised kombinatsioonid on väga ebatõenäolised. Statistiline keelemudel püüab nõnda simuleerida inimese igapäevateadmisi ja kogemusi ning situatsioonikonteksti kasutamise oskust. Loomulikult on selline statistiline keelemudel siiski küllalt primitiivne.

Tallinna Tehnikaülikooli Küberneetika Instituudi foneetika- ja kõnetehnoloogialaboris arendatav üldise sõnavara eesti keele tuvastussüsteemi prototüüp tuvastab praeguse seisuga korrekt-

selt ligikaudu 74 protsenti sõnadest. Kuna tuvastustäpsus sõltub olulisel määral sellest, kui hästi sisendkeel vastab süsteemi kee-

lemudelile, ei saa seda protsenti siiski väga tõsiselt võtta. Antud prototüübi statistiline keelemudel on treenitud põhiliselt ajalehetekstide põhjal, seega võib näiteks teadusajakirja artikli dikteerimisel oodata märgatavalt viletsamat tulemust.

UNIVERSAALNE LAHENDUS PUUDUB

Kõnetuvastuse eesmärk on tõlkida inimkõne arvuti abil automaatselt sellele vastavaks tekstiks. See protsess on osa automaatselt inimkõne mõistmise tehnoloogiast. Kõne on inimese jaoks loomulik suhtlusvahend ning annab tihti suuri eeliseid teiste suhtlemisviiside ees ka arvuti ja teiste automaatsete süsteemide kasutamisel. Kõnetuvastus, hoolimata oma keerukusest, on teinud viimaste aastate jooksul suuri edusamme, kuid universaalselt toimiv lahendus puudub siiani. On olemas palju reaalselt kasutatavaid rakendusi, kus kõnetuvastus toimib väga hästi. Inglise ja teiste suuremate keelte jaoks on olemas väga hea kvaliteediga dikteerimissüsteemid, st rakendused, kus inimese poolt soravalt dikteeritud kõne teisendatakse sellele vastavaks tekstiks. Nende kvaliteet on hüppeliselt paranenud just viimase viie aasta jooksul. Tihti ütlevad selle ala spetsialistid, et kõnetuvastus on tuleviku tehnoloogia. Peab lootma, et ta ei jää selleks alatiseks. ■

TOOMAS KIRT (1971) on TTÜ doktorant. Uurimisteenaks andmeanalüüsi meetodid ja neurovõrgud.
TANEL ALUMÄE (1976) on TTÜ doktorant. Tegeleb TTÜ Küberneetika Instituudi foneetika- ja kõnetehnoloogialaboris eestikeelse kõnetuvastuse uurimisega.

