Kõnesüntees? … See on imelihtne

Meelis Mihkla
eesti keele instituudi vanemteadur

Kõnesüntees ja rääkivad masinad

Milline on meie esimene ettekujutus kõnesünteesist ja rääkivatest masinatest? Tõenäoliselt on see seotud ulmefilmidest nähtu ja kuulduga – alates George Lucase tähesõjafilmide robotitest ja populaarse sarja „Knight rider" rääkivast imeautost Kit kuni tänapäeva fantaasiafilmide „masinlike" kangelasteni. Paraku ei esitata neis filmides tavaliselt kõnesüntesaatori tekitatud sünteeskõnet, vaid moonutatud inimhäält. Aga mingi ettekujutuse kõnesünteesi kasutusvõimalustest see muidugi annab ja vaevalt oleksid meie ootused-lootused rääkivate masinate võimetele vähem fantaasiarikkad kui ulmefilmides. Filmidest nähtu ja kuuldu põhjal peaks sünteeskõne olema võimalikult sarnane inimhäälega, kõnesüntesaator ise mõnus vestluskaaslane, kes räägiks meiega ilmekalt meeldiva häälega, rõõmustaks ja kurvastaks koos meiega ning viskaks koos meiega nalja.

Kõnesünteesiga on maailmas tänapäeva tasemel tegeletud pea pool sajandit. Vaadates tagasi kõnesünteesi suhteliselt pikale ajaloole ja sellele, kuhu kõnetehnoloogia areng tänaseks on jõudnud, tekib kiusatus küsida: "Miks ikkagi on näiliselt nii selge ja konkreetse ülesande nagu kõnesünteesi lahendamine võtnud nii palju aega ja miks ei ole veel loodud täiuslikku tekst-kõne-sünteesi? Või miks fantaasiafilmides kujutatu ei ole veel teoks saanud, kui mitte imeautodena, siis vähemalt intelligentsete kodumasinatena?" Miks ei ole kõnesüntees nii imelihtne kui pealkirjas?

Selle kirjutise pealkiri on ajendatud eelmise sajandi seitsmekümnendatel-kaheksakümnendatel aastatel eesti keeles ilmunud prantsuse elektroonikainseneri Eugene Aisbergi kirjutatud raamatutest „Raadio? … See on imelihtne", „Televisioon? … See on imelihtne", „Transistor? … See on imelihtne" ja „Värvustelevisioon? … See on peaaegu lihtne". Dialoogivormis esitatud raamatutes püüdis E. Aisberg tol ajal suhteliselt uusi ja hiljuti laiatarbesse jõudnud teaduse- ja tehnikasaavutusi

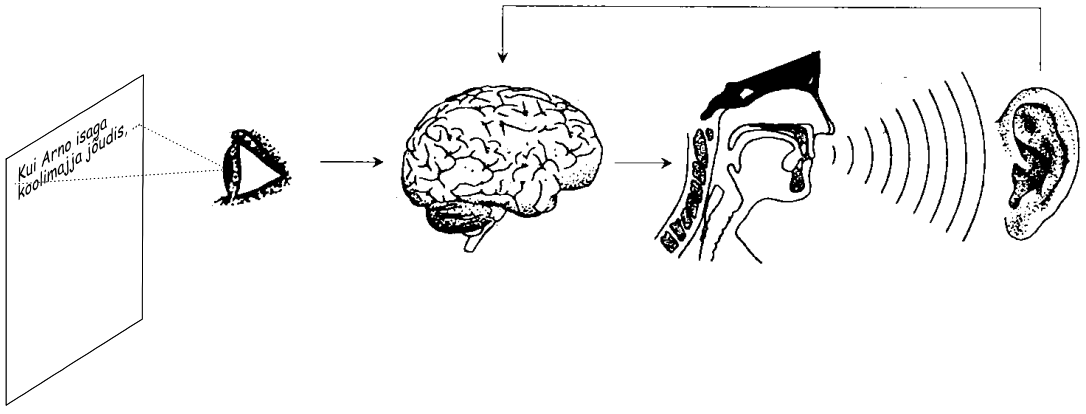
inimestele populaarses vormis esitada. Autor võttis lugeja käekõrvale, et üheskoos nende nö „mustade kastide" sisemusse piiluda.

Praegu oleks aeg küps samasuguste raamatute ilmumiseks keeletehnoloogia vallas, *à la* „Kõnesüntees – see on imelihtne" ning „Kõnetuvastus ja masintõlge – need on peaaegu lihtsad". Siis muutuks pealtnäha lihtsad ja teisalt ka keerulised asjad meile omasemaks ja arusaadavamaks. Kuni sellised raamatud veel ilmumist ootavad, püüab see kirjutis lugejas kõnesünteesist kaasamõtlemist ärgitada nii lihtsuse ja vahel ka keerukuse aspektist.

Kuidas me loeme?

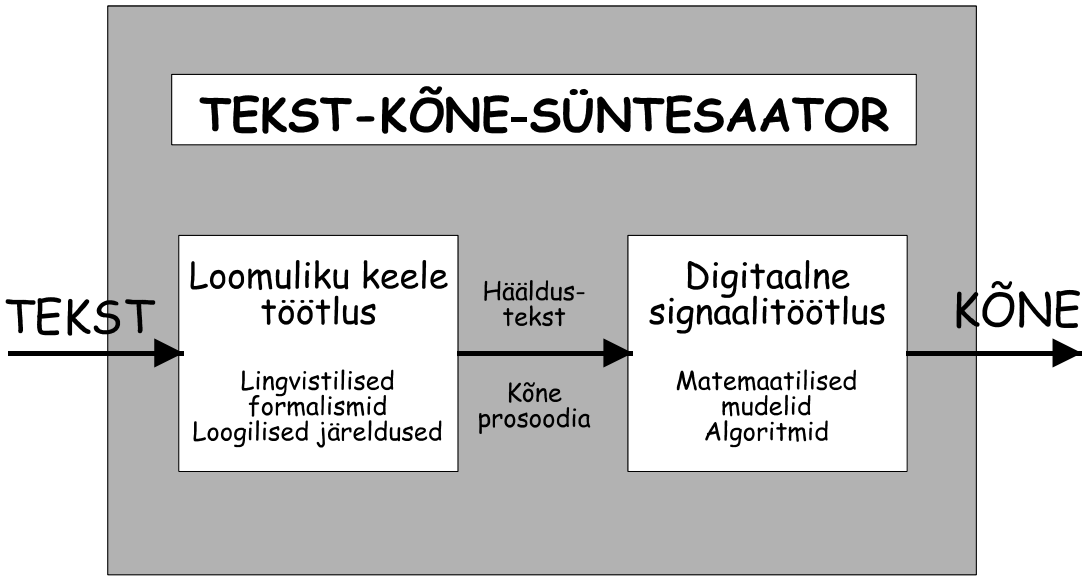
Tekst-kõne-süntesaatorite eeskujuks on olnud inimlugemine. Inimene omandab lugemisvõime esimesel elukümnendil, edasises elus lugemisoskus areneb ja täieneb. Olles lugemisvõime omandanud, muutub see automaatseks tegevuseks. Vaadeldes lugemist füsioloogia tasandil, näeme, et tegu on väga keeruka protsessiga. Joonisel 1 on esitatud teksti häälega ettelugemise lihtsustatud skeem ja kujutatud inimese lugemisprotsessi kaasatud füsioloogilised organid.

Tähemärkide kujutise haaravad silmade sensorneuronid ja kannavad kujutise elektriliste stiimulite vormis inimese ajju, kus see informatsioon töödeldakse ja formeeritakse motoorsete neuronite käsklusteks, mis kannavad hoolt kopsude, häälekurdude ja artikulatsioonilihaste aktiveerimise eest[1]. See viib kõne tekitamisele, kusjuures artikulatsiooniprotsessi jälgitakse ja juhitakse pidevalt põhiliselt kuulmisorganitest saabuva informatsiooni põhjal.



Joonis 1. *Andmevoo skemaatiline diagramm illustreerimaks lugemisprotsessi.*

Kõnesüntees on võime konkreetses keeles teisendada suvalist teksti kõneks ilma inimese osaluseta. Väljundkõne peab olema arusaadav ja loomulik. Arvutil imiteeritav tekst-kõne- süsteem on lihtsustatud mudel füsioloogilisest lugemisprotsessist (joonis 2). Tekst-kõne-süsteem eeldab oma sisendis teksti, mis on eelnevalt juba arvutisse viidud. Tähemärkide optiline tuvastussüsteem või ekraanilugeja jääb tavaliselt tekst-kõne-süsteemi käsitlusest välja. Süsteem ei sisalda ka tagasisideahelat, mis väljundi analüüsi põhjal võimaldaks automaatselt kõne valjust ja häälekõrgust kohandada vastavalt keskkonna tingimustele.



***Joonis 2.** Üldistatud tekst-kõne-sünteesi mudel.*

Nii nagu inimlugemine, sisaldab tekst-kõne-süntesaator loomuliku keele töötlusmoodulit, mis teisendab sisendteksti hääldustekstiks koos soovitud intonatsiooni ja kõnerütmiga. Loomuliku keele töötlusmoodul annab teksti foneetilise kirjelduse ja paneb paika kõne prosoodia. Üldjuhul sisaldab tekstitöötlus keele erinevaid kirjeldustasandeid: foneetikat, fonoloogiat, morfoloogiat, süntaksit ja semantikat. Kohe pärast seda, kui tekst-kõne-teisenduse loomuliku keele töötlusmoodul on lõpetanud teksti töötluse ja genereerinud väljundi, on kõnesüntesaator umbes samal tasemel kui inimene, kes teab, mida öelda, kuid ikka veel kahtleb, kuidas ennast täpselt väljendada. Signaalitöötluse moodulis sisalduvad operatsioonid on artikulatsioonilihaste ja häälekurdude võnkesageduse dünaamilise juhtimise arvutianaloogiks, nii et väljundsignaal vastaks sisendi nõudmistele. Digitaalne signaalitöötlusmoodul teisendab sisendis oleva sümbolinformatsiooni loomuliku kõlaga kõneks.

Kõnesünteesi meetodid

Kõnesünteesi meetodid võib jagada kolmeks:

- artikulatoorne süntees,
- reegelsüntees (formantsüntees),
- ahelsüntees.

Artikulatoorne süntees baseerub kõneproduktsiooni füsioloogilisel mudelil ja kõnetraktis hääle tekitamise füüsikalisel kirjeldusel. Meetod on arvutuslikult väga töömahukas, reaalajas ei toimi ja kasutatakse seni põhiliselt uurimisotstarbel. Vastupidiselt artikulatoorsele lähenemisele ei püüagi reegel- ja ahelsüntees seletada kõneorganite kinemaatikat, vaid lihtsalt kirjeldab vastavaid akustilisi lainekujusid.

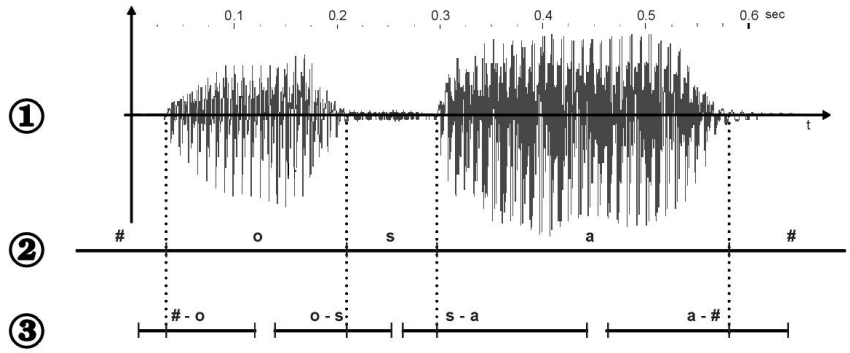
Kõneteaduses on ammu teada, et foneetilised siirded pole kõne arusaadavuse seisukohalt vähem olulised kui häälikutestatsionaarsed osad. Arusaadava ja loomuliku väljundkõne saamiseks on raskuspunkt nende kahe meetodi puhul häälikult häälikule üleminekute ja koartikulatsiooni modelleerimisel. Reegelsünteesil kirjeldatakse foneetilisi siirdeid otseselt reeglite jada vormis, mis formaalselt kirjeldab häälikute mõju üksteisele. Ajalooliselt on reegelsüntees üles ehitatud põhiliselt formantsüntesaatori kujul. Kõnet kirjeldatakse kuni 60 parameetri dünaamilise muutumisena, mis on esmajoones seotud formantide (ja antiformantide) sageduste ja ribalaiustega ja kõri lainekuju kirjeldusega.

Erinevalt reegelsünteesist on kõneüksuste ühendamisel põhinevatel süntesaatoritel väga vähe informatsiooni käsitletavate andmete kohta. Enamik infost sisaldub segmentides, mida jadas ühendatakse. Ahelsünteesil salvestatakse foneetilised siirded ja seega koartikulatsioonilised mõjustused kõnesegmentide andmebaasi ja neid kasutatakse sünteesil akustiliste üksustena häälikute asemel.

Ahelsüntees eeldab, et artikuleeritud kõnevoog ei ole lihtsalt ritta seatud häälikute jada. Pigem koosneb kõne pidevalt kattuvatest üleminekutest ühelt häälikult teisele. Difoonid* (vt joonis 3) on ahelsünteesil enim kasutatud kõneühikud, kuna kõne genereerimiseks suvalise teksti alusel on vaja suhteliselt väikest arvu difoone. Joonisel 3 toodud näite põhjal on üksiksõna *osa* kõnelaine genereeritav nelja difooni abil: pausilt üleminek *o*-le, *o*-lt üleminek *s*-ile, *s*-ilt üleminek *a*-le ja *a*-lt üleminek pausile. Eesti keele difoonide andmebaas sisaldab ligikaudu 1900

* Difoonid algavad mingi hääliku stabiilse osa keskelt ja lõpevad järgmise hääliku stabiilses osas.

difooni[2]. Korpuspõhisel ahelsünteesil kasutatakse lisaks difoonidele pikemaid kõneüksusi (fraas, sõna, kõnetakt) ja kõnelõike.



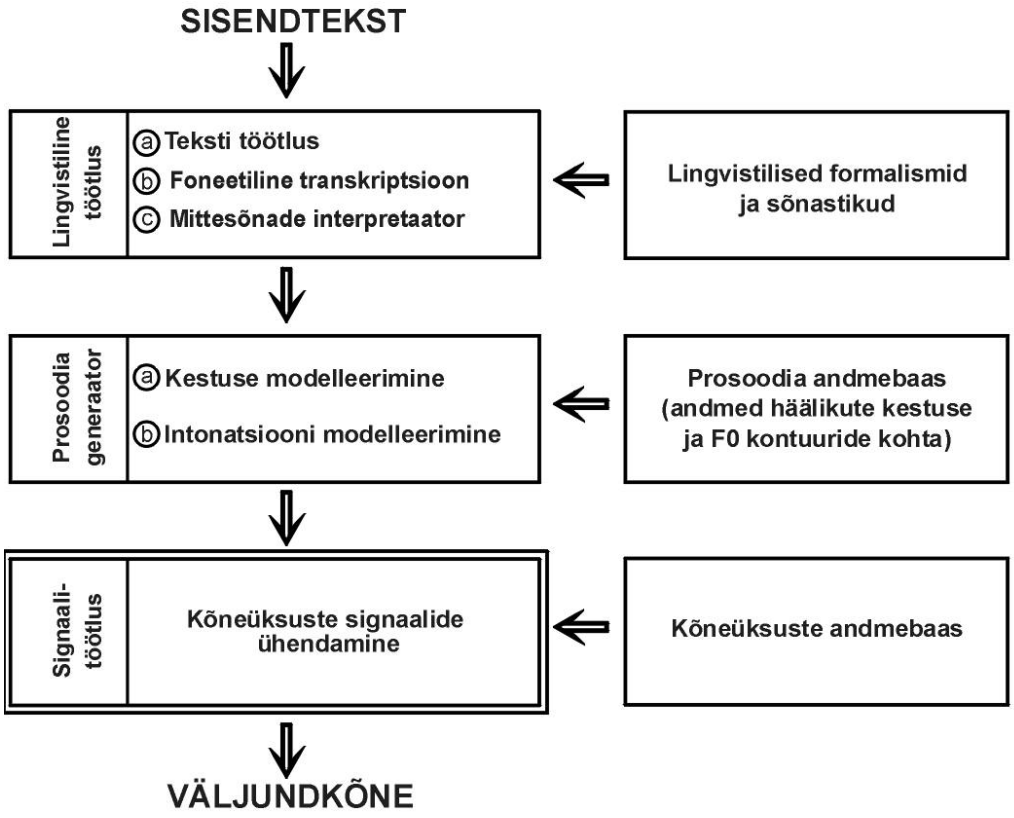
Joonis 3. *Difoonide märgendus kõnelaine (sõna '***osa***') põhjal:*
1 – kõnelaine, 2 – häälikute piirid (märgitud vertikaaljoontega),
3 – difoonidele vastavad signaalilõigud.

Ahelsüntees

Joonisel 4 on toodud tüüpiline ahelsünteesi mudel. Kuna samasugusel plokkskeemil põhineb ka eestikeelne difoonidel põhinev tekst-kõnesüntesaator, siis vaatame lähemalt, millised ülesanded on vaja lahendada kõne sünteesimiseks eestikeelse kirjaliku teksti alusel.

Lingvistilise keeletöötluse tulemusena teisendatakse ortograafiline tekst hääldustekstiks. See ei ole sugugi kerge ülesanne, sest eesti ortograafia ei ole foneetiline. Kirjapilt II ja III väldet üldjuhul ei erista (nt *lapsed mängivad kooli juures; lapsed lähevad kooli*), eristamata on palataliseeritud konsonandid palataliseerimata konsonantidest (nt *Eesti keskmine palk on ligi 14000 krooni kuus; see palk on kuus meetrit pikk*), kirjas ei ilmne pika *üü* diftongeerumine rõhutu silbi lühikese vokaali ees (nt *müüa* hääldame *müija*) ja palju muud. Sõnade õiged hääldused ja välted leitakse sõnastikest. Lisaks leitakse välte ja palatalisatsiooni märkimisel lingvistilise töötluse käigus ka liitsõnapiirid, sõnarõhud ja silbipiirid, mis on vajalikud prosoodiageneraatori tööks. Kirjalikes tekstides on ka suur hulk lühendeid, numbreid ja erimärke. Neid töötleb nn mittesõnade interpretaator, teisendades süntesaatorile tundmatud märgid loetavaks tekstiks.

Prosoodiageneraatori ülesandeks on kõne prosoodilise struktuuri modelleerimine, so häälikute kestuse ja lausetüübile vastava meloodiakontuuri (põhitooni kontuuri) genereerimine. Häälikute kestusmallid ja intonatsioonikontuurid on prosoodia andmebaasis esitatud tabelite



Joonis 4. *Ahelsünteesil põhinev tekst-kõne süntesaator.*

või loogiliste seostena, sõltuvalt hääliku ja sõna asukohast, rõhust, vältest jm. Niisugune on reeglipõhine prosoodiamudel, mille puuduseks on see, et sõltumatult tuletatud reeglite samaaegne rakendamine võib põhjustada vigu. Kui suured kõnekorpused muutusid kättesaadavaks, hakati prosoodiamudelites rakendama statistilisi meetodeid, mis võimaldasid avastada ka varjatud, kuid kõne meloodiat mõjutavaid olulisi tunnuseid. Statistiliselt rakendatakse teatud masinõppe meetodeid (regressiooni, närvivõrke või otsustuspuid) ja arvuti ise genereerib kõnekorpuste baasil optimaalseid prosoodiamudeleid.

Signaalitöötluseks rakendatakse erinevaid signaalitöötlusmeetodeid, mis ühendavad kõneüksustele vastavad signaalid sujuvaks väljundkõneks. Eestikeelsel difoonsünteesil oleme kasutanud Mon'si Ülikoolis Belgias välja töötatud MBROLA sünteesimootorit[3], mis ühendab häälikutele vastavad difoonid prosoodiageneraatorist saadud info põhjal sünteeskõneks. Näiteks lausefragmendi *"Kui Arno isaga koolimajja…"* sünteesimiseks vajalik sisendinfo on toodud tekstiboksis joonisel 5 (esimeses veerus on hääliku tähis, teises hääliku kestus, järgnevates informatsioon põhitooni kohta):

```
k0      78
u       70 0 125 100 125
i       60 0 125 100 117
a       71 0 117 100 117
r       52 0 117 100 112
n       53 0 112 100 107
o       59 0 107 100 109
i       68 0 109 100 108
s       68
a       74 0 108 100 108
g       53
a       73 0 109 100 109
k       54
o       63 0 111 100 111
o       62 0 111 100 111
l'      59 0 111 100 107
i       61 0 107 100 106
m       73 0 106 100 112
a       74 0 112 100 114
j:      111 0 114  70 105 100 102
a       66 0 102 100 103
```

***Joonis 5**. Prosoodiline info ahelsünteesil.*

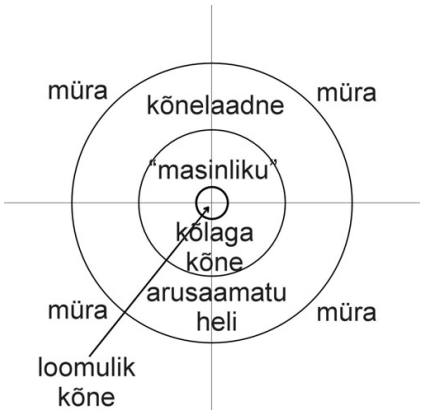
Kõnest arusaadavus, kõne loomulikkus ja ekspressiivne kõnesüntees

Kõnesünteesi tähtsaks osaks on kõnest arusaadavus. Võib ju tunduda, et selles vallas on jõutud vajalikule tasemele, pea kõigi süntesaatorite kõne on arusaadav. Aga tegelikult kätkeb kõne endas väga rikkalikult infot. Kui monomodaalse suhtluse korral, näiteks telefoniliini pidi, on oluline vaid kõnesignaali kvaliteet, siis multimodaalses suhtluses annab olulist infot ka kõneleja silmavaade, miimika, käteliigutused ja kehakeel. Selles vallas on üheks arengusuunaks sünteesi esitus nn „rääkiva pea" (*talking head*) vormis, mis loob multimodaalse suhtluskeskkonna.

Pea kõigi tänapäeva kõnesüntesaatorite pudelikaelaks on kõneprosoodia ja sellest tulenev väljundkõne monotoonsus ja ebaloomulik kõla. Ehkki erinevaid inimhääli võib konkreetses keeles olla miljoneid või isegi sadu miljoneid, suudame pea ilmeksimatult eristada inimhäält

sünteeskõnest. Ikka leidub nüansse, kus süntesaator ei suuda jääda loomulikkuse piiridesse ja kaldub loomulikkuse alast välja. Seega on kõnetehnoloogia üheks oluliseks märksõnaks kõne variatiivsus. Kui kõnetuvastuses põhjustab kõnelaine variatiivsus sageli probleeme, siis kõnesünteesis viib vähene variatiivsus sünteeskõne monotoonsusele ja masinlikule kõlale[4].

Kõneprosoodia tekitab põhisageduse ja intensiivsuse kaudu kõnes teatud korrastuse või meloodia füüsikaliste suuruste kestuse. Kõnesünteesis on kõneprosoodial oluline roll. Mida paremini oskame neid füüsikalisi parameetreid (häälikute kestusi, häälekõrgust ja signaali intensiivsust) sünteeskõnes edasi anda, seda loomulikumana me kõnet tajume. See ei ole lihtne ülesanne. Kui püüda näitlikult kujutada inimhäälele lähedasi helisid mingis kahemõõtmelises tunnusruumis, siis võtab loomulik kõne sellest vaid väga väikese osa (joonis 6).



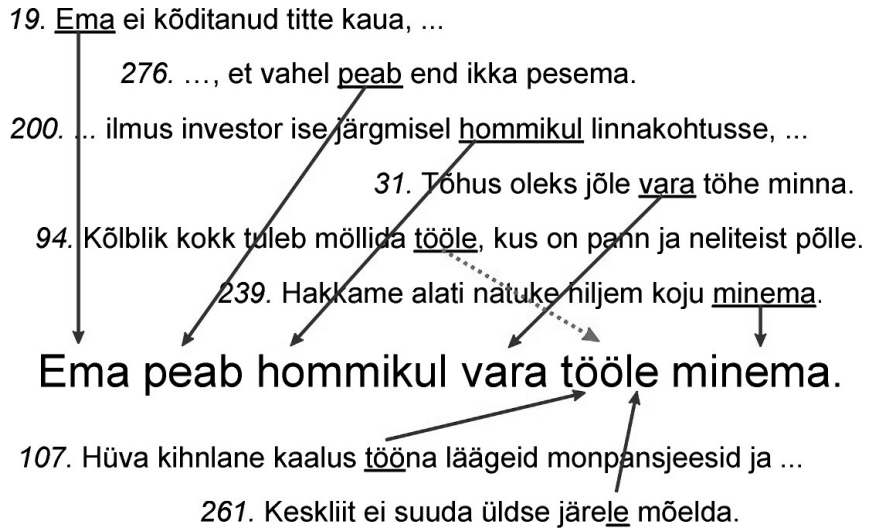
Joonis 6. *Loomulik kõne inimhäälele lähedaste helide kahemõõtmelises tunnusruumis.*

Kõne loomulikkuse probleemi püütakse lahendada kas otseselt – paremate prosoodiamudelite väljatöötamisega (siia kuuluvad ka eespool mainitud statistilised mudelid) – või kaudselt, kasutades sünteesil võimalikult pikki kõnelõike, kus kõne loomulik prosoodia on neis lõikudes juba loomulikul kujul olemas.

Üheks kõnesünteesi väljakutseks on emotsioonid kõnes ja kõne ekspressiivsus. See ei tähenda, et süntesaator peaks meile emotsioone puistama või armastust avaldama, pigem seda, et inimese väga suure rumaluse korral peaks ta häält tõstma ja võib-olla tahtmatu äparduse korral kaasa tundma. Kvaliteetse ekspressiivse kõne kommertsrakendusteni veel jõutud ei ole, aga tööd selles vallas käivad. Kes esimesena turule jõuab, saab eelise.

Korpuspõhine kõnesüntees

Korpuspõhine kõnesüntees põhineb suurtel kõnekorpustel (0,5–2 tundi salvestatud kõnet ühe diktori esituses). Kui difoonidel põhineval ahelsünteesil on andmebaasis iga häälikult häälikule ülemineku jaoks vaid üks difoon, siis korpussünteesil on kogu kõnekorpus sünteesi akustiliseks baasiks. Kõneüksuseid hakatakse korpusest otsima kõrgematest hierarhilistest tasanditest (fraas, sõna, kõnetakt) ja eelistatakse võimalikult pikki kõnelõike või -stringe. Sobiv kõneüksus valitakse vastavalt kontekstile nii lingvistiliste kui ka füüsikaliste sobivuskriteeriumide alusel. Sageli ei pruugi lingvistiliste tunnuste alusel leitud sõna või fraas sünteesitavasse lausesse sobida ehkki täheline ja ka vormiline vastavus on olemas. Seda situatsiooni illustreerib joonisel 7 toodud näide. Kui meil on vaja sünteesida näiteks lauset *Ema peab hommikul vara tööle minema*, siis oleksid kõik kõneüksused korpusest leitavad sõnatasandil. Ainus kontekstist johtuv mittevastavus tekib sellest, et sõna *tööle* on korpuses fraasi lõpus, sünteesitavas lauses aga fraasi keskel, mistõttu väljundkõnes on tajutav ebakõla. Kui me aga sõna *tööle* kompileerime sõnade *tööna* ja *järele* vastavatest silpidest, on tulemus oluliselt parem.



***Joonis 7.** Korpuspõhise sünteesi näide: lause Ema peab hommikul vara tööle minema kompileerimine kõnekorpuse erinevatest üksustest.*

Eestikeelse korpuspõhise kõnesünteesi kõnekorpus sisaldab u 50 minutit salvestatud kõnet professionaalse raadiodiktori esituses ja tekstikorpus koosneb 400 lausest ja ligi 3000 sõnast[5]. Kõnekorpus sisaldab

kõikvõimalikke difoone, kuna kõiki lauseid ei õnnestu kompileerida sõna tasandil, enamik harva esinevaid sõnu ja ka võõrsõnu tuleb suures osas kokku panna väikseimatest ühikutest difoonidest. Korpus sisaldab ka eesti keele 200 enamlevinud sõna, palju numbreid ja aastaarve ning enamlevinud kohanimesid. Võib muidugi tekkida mõte koostada kõnekorpus, mis sisaldaks kõiki eesti keele sõnu, siis saaks kõnet sünteesides jääda sõnatasandile. Paraku pole see võimalik, kuna ainuüksi õigekeelsussõnastikus on üle 50000 märksõna, sõnu aga on vähemalt kaks korda nii palju, sest regulaarsed tuletised ja liitsõnad antakse sõnaartikli sees ühe märksõna all. Lisaks võib enamik sõnu olla kümnetes eri vormides, kuna eesti keele morfoloogia on väga rikas. Selline kõnematerjal muutuks hiigelsuureks ja halvasti juhitavaks. Samuti on ühel inimesel pea võimatu niisugust pikka salvestussessiooni sooritada, sest kõik sõnad peavad olema sidusas kõnes ja mõtestatud lausetes. Kõikide sõnade salvestamise teed ei ole mindud ka flekteerivates keeltes (inglise, hiina, vietnami), kus morfoloogial on tähtsusetu roll.

Eestikeelne korpussüntees on veel väljatöötluse faasis, aga selle ressursid ja töö käigus loodavad moodulid on Internetis kättesaadavad http://www.eki.ee/keeletehnoloogia/projektid/syntees/tks.html.

Kõneüksuste valik on korpussünteesi kõige olulisemaks osaks. Algoritme ja meetodeid, kuidas ja millise kaaluga arvestada erinevaid faktoreid ning kuidas minimeerida lingvistilise ja füüsikalise sobivuse saavutamise hindu, on siin palju. Rahvusvaheline kõnekommunikatsiooni ja -tehnoloogia assotsiatsioon korraldab ideede arengu ergutamiseks iga-aastaseid võistlusi *Blizzard Challence*, kus osavõtjatele antakse kasutada mõõduka suurusega märgendatud kõnekorpus, mille põhjal püüavad osalejad valida kõneüksusi ja luua võimalikult head kõnesünteesi. Põhimõtteliselt võiks ka Eestis lähiajal huvilistel lasta kõnekorpuse baasil kõnesünteesi arendada.

Kokkuvõtteks

Eeltoodu on vaid põgus ülevaade kõnesünteesi hetkeseisust, probleemidest ja arenguperspektiividest. Loodan, et ma teile lihtsaid asju liiga lihtsustatult ei seletanud ega keerulisi liiga igavalt. Kui teil tekkis lugedes väikegi huvi kõnesünteesi vastu, siis soovitan teil kohe kõnesünteesi katsetada ja proovida, sest praktiline kogemus ja teooria peavad käima käsikäes. Alustada võiks muidugi eestikeelse kõnesünteesiga, mis on vabavarana Internetis kättesaadav (vt http://

www.eki.ee/keeletehnoloogia/projektid/syntees/tks.html). Kui tekib huvi tutvuda ka teiste eri keelte sünteesisüsteemide ja sünteesinäidetega, siis võiks külastada järgmisi kodulehekülgi: http://www.cstr.ed.ac.uk/projects/festival/, http://tcts.fpms.ac.be/synthesis/mbrola.html, http://hts.sp.nitech.ac.jp/?Voice%20Demos.

---

1 **J. N. Holmes,** Speech Synthesis and Recognition. Van Nostrand Reinhold. London, 1988.

2 **M. Mihkla, E. Meister,** Eesti keele tekst-kõne-süntees. – Keel ja Kirjandus, 2002, nr 2, lk 88–97; nr 3, lk 173–182.

3 **T. Dutoit,** An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers. Dordrecht, 1997.

4 **M. Tatham, K. Morton,** Developments in Speech Synthesis. John Wiley & Sons Ltd. Chichester, 2005.

5 **L. Piits, M. Mihkla, T. Nurk, I. Kiissel,** Designing a Speech Corpus for Estonian Unit Selection Synthesis. – Nodalida 2007 Proceedings: The 16th Nordic Conference of Computational Linguistics. Tartu, 2007, lk 367–371.